The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

# 回帰型ニューラルネットを用いた会話エージェントのアイドリング動作生成の検討

## Considering Idling Motion Generation using RNNs for Virtual Conversational Agents

黄 宏軒 *1*2 Hung-Hsuan Huang　福田匡人 *1 Masato Fukuda　西田豊明 *1*2 Toyoaki Nishida

*1国立研究開発法人理化学研究所革新知能統合研究センター
RIKEN Center for Advanced Intelligence

*2京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

This work aims to develop a model to generate fine grained and reactive non-verbal behaviors of the virtual character when the human user is talking to it. The target *micro* non-verbal idling behaviors are micro facial expression, head movements, and postures. We explored the use of recurrent neural network (RNN) to learn these behaviors in reacting to the human communication interlocutor's corresponding micro non-verbal behaviors. The models are trained on an active listening data corpus which features elderly speakers talking with young active listeners and was collected by ourselves.

## 1. はじめに

擬人化会話エージェントはコミュニケーションロボットのような物理的制約がなく，高い自由度で精緻でリアルな顔の表情や身体の仕草を表出でき，ユーザとコミュニケーションが取れるインターフェイスとして期待されている．しかし，これまでの会話エージェントの多くは事前に定義された顔表情や体の動きのアニメーションシーケンスをインタラクション中にルールベースで呼び出して再生する仕組みになっていた [Huang 11]．そのため，エージェントが同じことをする際に，例えば，笑うと必ず同じ笑いをしたり，視線をそらすと，同じ視線の動きをしたりすることになる．また，エージェントが何らかの意図を表出しようとしない際には，静止する，または，固定のアイドリングモーションを繰り返し再生するようになっている．こうしたエージェントとしばらくインタラクションをしていると，その固定パターンが見えて自然さを損なってしまう．

しかし，生身の人間は固定パターンの動作をしない。笑うたびに毎回少し違う笑いをするはずであり，微細なブレが含まれている．そのブレは物理的な要因以外に会話相手，文脈，本人の内的状態などの影響を受けると思われる．本研究はそのブレを捉えようとするものである．エージェントが何らかの意図を表出しようとしない間，いわゆるアイドリングの状態でも，リアルタイムで会話相手の振る舞いから微細なアイドリングモーションの生成を目指している．

近年，深層学習が様々な分野で著しい成果を挙げており，普及が急速に進んでいる．そのなかでは，データの経時的変化を捉えるように再帰型ニューラル (Recurrent Neural Networks, RNN) が考案されている．本研究では，エージェントの振る舞い生成モデルの検討として，エージェントの会話相手の振る舞い (頭部の回転，視線，表情，声の韻律情報) からエージェントの振る舞い (頭部の動き，表情，上半身の動き) を予測することを課題とし，我々が収集した傾聴会話データコーパスを基に，異なるネットワーク構造の予測性能について考察する．RNN の代表的な構造である Long Short-term Memory (LSTM) [Hochreiter 97] と忘却ゲートや出力ゲートを省きより簡素な構造になっている Gated Recurrent Unit (GRU)[Cho 14] と，ベースラインのマルチレイヤパーセプトロン (MLP) を比較対象とする．

連絡先: 黄 宏軒，hhhuang@acm.org

## 2. データコーパス

本研究の目的は利用者の年齢層を限定しないが，実装する会話エージェントが高齢者のメンタル支援での活用が視野に入るため，会話データコーパス収集実験は高齢者を対象とした傾聴会話課題で行った．実験参加者は高齢者 4 名 (男性 2 名，女性 2 名．年齢 69〜73 歳，平均 71 歳) と大学生 4 名 (男性 2 名，女性 2 名，平均年齢 21 歳) であり，参加者毎の組み合わせで 15〜30 分間の 4x4 計 16 会話を収録した．エージェントとの会話は画面越しになるのと高齢の参加者は遠隔地からだったため，実験は Skype でのテレビ会話であった．Skype 通信用の Web カメラの他，フル HD のビデオカメラで参加者の上半身の映像を正面から記録した (図 1)．高齢者が話し手役で若者が聴き手役の会話であったが，話題は限定せず初対面の設定で自由に会話してもらった．結果として若者が質問をして高齢者がそれに答える場面が多く見られ，話題は双方の自己紹介のほか，高齢者の過去の仕事，趣味，最近はまったことなどであった．

わずかであったがネットワーク通信の遅延があったため，分析データは実時間のデータではなく，高齢者側で収録した若者の音声を基準に若者側の映像を同期させた．その上で高齢者と若者それぞれ直接撮影したビデオカメラの映像と音声を抽出した．映像に対して，OpenFace*1 を用いて Ekman が提唱しており，顔表情を記述する Facial Action Coding System (FACS)[Ekman 02] の Action Unit (AU)44 個のうち 17 個，そして，頭部の回転 (3 次元)，視線方向 (8 次元) の 30 fps データを抽出した．さらに，OpenPose *2 を用いて肩関節の 2 次元座標からカメラまでの距離と肩の傾きを抽出した (30 fps)．音声については OpenSmile *3 を用いて Interspeech 2009 の

∗1 https://github.com/TadasBaltrusaitis/OpenFace
∗2 https://github.com/CMU-Perceptual-Computing-Lab/openpose
∗3 https://www.audeering.com/what-we-do/opensmile/

図 1: 傾聴会話コーパスの収録映像．自然な会話に近づけるように大画面のテレビとスクリーンに映像を映し，視線が合うように位置を調整した

Emotion Challenge[Schuller 09] で定められたローレベル特徴量，energy (1 次元), mfcc (12 次元), zero-crossing rate of time signal (1 次元)，voicing probability (1 次元)，F0(1 次元) の計 16 次元 100 fps のデータを抽出した．聴き手の振る舞いは発話状態によって変わると思われるため，Elan[?] のツールを用いて聴き手の発話区間を割り出した．

話し手の振る舞い (OpenFace で抽出した映像情報 28 次元と OpenSmile で抽出した韻律情報 16 次元) を説明変数とし，聴き手の振る舞い (OpenFace の 28 次元と OpenSmile の 16 次元と OpenPose の 2 次元) を目的変数と設定した．計算が重くてリアルタイムでの入力が困難であったため，OpenPose の情報を説明変数にしていなかった．会話セッション長にばらつきがあったため，セッションの最初から最長 20 分間の部分を切り出した．入力データに 1 秒間と 2 秒間のウィンドウ幅を持たせ，出力は 1 フレーム分になるようにデータを整形した．整形した約 4.5 時間分のデータセットの概要を表 1 に示す．このデータから，男性聴き手の発話時間が短く，話し手の話す時間が長くなっており，女性聴き手自身の発話が長く，話し手の話す時間が相対的に短ったことが分かる．実験時に，男性聴き手は一方的に質問を聞いており相手の回答から話題を掘り下げない傾向に対して，女性聴き手の方は聞き上手で全体的に会話が弾んでいる観察と一致している．

表 1: データセットの概要．データ量は 2 秒時間幅のもの

| | 発話状態 | フレーム | 時間 (s) | データ量 |
|---|---|---|---|---|
| 男性 | On | 69,818 | 2,334 | 1.4GB |
| | Off | 169,387 | 5,663 | 13.6GB |
| 女性 | On | 82,261 | 2,750 | 1.7GB |
| | Off | 147,165 | 4,920 | 11.8GB |

## 3. 学習モデルの比較

### 3.1 実験手順

前節のデータセットの全体，男性聴き手，女性聴き手，そして聴き手個人のサブデータセットに対して，MLP, GRU, LSM の 3 つのモデルと時間幅 1 秒，2 秒の条件で機械学習実験を行った．ネットワーク構造については入力モダリティの fps 数が違うため，各モダリティの時系列をそれぞれに対応する LSTM/GRU 層 (入力の 8 倍、それぞれ 28 x 8, 16 x 8 個のノード) に入力して、そのあと全モダリティを結合する全結合層 (2 x 128 個のノード)，そして，さらに 1 層の全結合層 (512 個のノード) の出力層のシンプルな 4 層構成で統一している (図 2)．入出力データはすべて 0 から 1 までの区間に正規化した．なお，MLP については，時系列表現ができないため，同等のデータ入力を横方向 1 次元に展開して扱う．

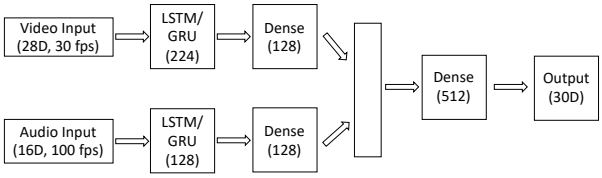


図 2: 評価実験に用いたネットワークの構造

学習モデルの評価には平均二乗誤差の平方根 (Root Mean Squared Error, RMSE) を用いる．各条件で一人の話し手をテストデータとして抽出して残り 3 人のデータを用いて学習し，これを 4 回繰り返して平均値を求める leave-one-subject out 法で交差検証を行った．

### 3.2 結果の考察

表 2 から，どのデータセットを用いても GRU が最も高い精度 (回帰誤差を示す RMSE 値が低い) を得ており，LSTM がそれに次いで，MLP の精度が最も低い結果を示す．このことから，同等の入力データに対してやはり時系列に特化したモデルの方は話し手と聴き手のインタラクションの特徴を捉えられている．そのなかでも今回のデータセットに対してよりシンプルな構造の GRU の方は性能が高い点，2 秒間の入力時間幅よりも 1 秒間幅の方は性能が高いことから一瞬一瞬の顔表情は反射的な要因が強いことを示唆される．また，LSTM モデルにおいても GRU モデルにおいても，男性聴き手の両名とも長い時間幅の方はより精度が高く，女性聴き手の両名とも短い時間幅の方は精度が高いことも興味深い．

## 4. 結論

本研究は仮想会話エージェントがアイドリング状態における意図の含まないモーションを生成するために，傾聴会話コーパスを基に話し手の振る舞いの映像と音声情報から聴き手の振る舞いを回帰的ニューラルネットの LSTM，GRU とベースラインの MLP で予測する試みをした．0〜1 の値域のデータに対して，どのモデルも 0.15 程度以下の精度を出力しており，会話パートナー同士の心理状態や話の内容を問わず，単純に相手の顔・声から顔表情と上半身のポスチャーをある程度予測できている結果を示した．しかし，この結果はあくまで数値データの観点での精度であり，生成されたエージェントの動きに対して，人間らしさを感じるかどうかの主観評価も今後実施する必要がある．

## 参考文献

[Cho 14] Cho, K., Merrienboer, van B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, *CoRR*, Vol. abs/1406.1078, (2014)

[Ekman 02] Ekman, P., Friesen, W. V., and Hager, J. C.: Facial Action Coding System (FACS), Website (2002)

表 2: 機械学習の結果

| | 時間幅 | 全員 | 男性 | 女性 | 全員平均 | 男性平均 | 女性平均 |
|---|---|---|---|---|---|---|---|
| LSTM | 1 sec | 0.1335 | 0.1439 | 0.1219 | 0.1219 | 0.1356 | 0.1081 |
| LSTM | 2 sec | 0.1416 | 0.1415 | 0.1343 | 0.1271 | 0.1357 | 0.1185 |
| GRU | 1 sec | 0.1299 | 0.1369 | 0.1187 | 0.1174 | 0.1282 | 0.1065 |
| GRU | 2 sec | 0.1323 | 0.1343 | 0.1278 | 0.1207 | 0.1264 | 0.1150 |
| MLP | 1 sec | 0.1377 | 0.1513 | 0.1263 | 0.1278 | 0.1393 | 0.1162 |
| MLP | 2 sec | 0.1514 | 0.1525 | 0.1442 | 0.1354 | 0.1406 | 0.1303 |

[Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long Short-term Memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (1997)

[Huang 11] Huang, L., Morency, L.-P., and Gratch, J.: Virtual Rapport 2.0, in *11th International Conference on Intelligent Virtual Agents (IVA 2011)*, pp. 68–79 (2011)

[Schuller 09] Schuller, B., Steidl, S., and Batliner, A.: The INTERSPEECH 2009 Emotion Challenge, in *10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, United Kingdom (2009)