RNN を利用した傾聴会話中の聞き手視線生成モデルの試作

Prototyping an RNN-based Listener's Gaze Generation Model in Attentive Conversations

石原 理央　　黄 宏軒 *
Riou Ishihara　　Hung-HsuanHuang

福知山公立大学情報学部
Faculty of Informatics, The University of Fukuchiyama

Abstract: 会話エージェントは，心理的負担が少なく，いつでも利用可能な対話相手として，孤独感の軽減に貢献する可能性がある．しかし，既存の会話エージェントは不自然な視線挙動を示すことが多く，親しみやすさや対話の没入感が損なわれている．本研究では，実際の対話データを活用することで，会話エージェントの視線生成手法を改良し，会話相手の視線に応じた視線変化を実現することを目指す．RNN を用い，LSTM や GRU を活用してモデルを構築し，対話中の聞き手の視線方向を予測する．本手法が既存の視線モデルと比較してより自然な視線挙動を実現し，会話エージェントの親しみやすさや没入感の向上に寄与することを目指す．

1　はじめに

現代の日本は，単身世帯の増加や少子高齢化に加えコロナウイルスの蔓延による外出規制などによって人々の繋がりが弱くなり，孤独を感じる人が増加している．加えて，人と積極的に話す機会が少なくなったため人と話すことが苦手という人も増加している．

2021 年 12 月から 2022 年 1 月にかけて政府によって初めて行われた孤独・孤立に関する全国調査「人々のつながりに関する基礎調査」[1] によると,「孤独の状況（直接質問）」という項目では孤独感が「しばしばある・常にある」と回答した人の割合は 4.5 %,「時々ある」が 14.5 %,「たまにある」が 17.4 %となっている．また,「孤独の状況（間接質問）」という項目では,「UCLA 孤独感尺度」に基づく孤独感スコアでは,「10〜12 点」（常にある）という人の割合は 6.3 %,「7〜9 点」（時々ある）が 37.1 %となっている．

孤独は誰かと会話することで緩和，解消することができる．しかし，孤独を感じる瞬間はその時々であり人によって異なるため孤独を感じるときに話せる相手は限られる．また，話すさいに相手に気を遣う状況などは逆にストレスになってしまう．そのため話す相手は関係の薄い人や気まずい人ではなく，友人の様な気軽に話せる人が好ましい．このようないつでも，気軽に話せる環境が普及すれば孤独を感じる人は少なくなる．しかし，このような環境は現代では珍しい．

*連絡先：福知山公立大学情報学部情報学科
〒 620-0886 京都府福知山市字堀 3370
E-mail:hhhuang@acm.org

いつでも，気軽に話せる環境を実現するためには話し相手が人では幅広いタイプの人をカバーすることは難しい．そこで，会話を行なえるシステムとして会話エージェントがある．会話エージェントとは音声認識技術や自然言語処理を組み合わせ，音声等により自動的な会話を行うプログラムである．会話エージェントであれば時間の制限などなくいつでも会話を行なえる．また，会話エージェントと話すさいに気を遣う必要がないため，関係の薄い人や気まずい人に比べ会話エージェントと会話しやすい．

このような背景から日常的に会話ができる環境を普及させることを目的とした会話エージェントの開発が重要である．一方で現在の会話エージェントは人と比べ会話するときの動きや表情が不自然なものが多く，会話をしていても自分の話を聞いているのか分からず積極的に話がしたい相手とは言いづらい．

そこで本研究では，会話エージェントにおいて視線を通じて「話を聞いてくれている」と感じられることを目指し，自然で動的な視線生成手法を提案する．

会話相手が会話エージェントに，話を聞いてくれていると感じられる，会話に興味がある，威圧感などがない，これらを満たせることを目標としている．対話相手が話を聞いてくれていると感じられることは，会話への積極さなどにもつながるため重要である．

相手が話を聞いていると感じるポイントの一つに視線がある．Adam Kendon(1967) は視線には 3 つの機能があることを指摘している [2]．その機能の中に，自分の態度や感情を話している相手に伝達する感情表出機能と発言の交代を促すといった会話の調整機能があ

る．また，それ以外にも他者に視線を向けることは相手への好意や関心を知らせるサインにもなる．相手に好意をもつとアイ・コンタクトが活発になる，自分や話題にしているものに視線が向いていると話を聞いているように感じる．反対にどれだけ会話内容や動作を話しに合わせている場合でも，視線が話に関係ないものに向いていると話に興味ないように感じる．現実で感じるこれらは，全て視線が会話において重要な役割を果たしている 1 例である．これらのことから，自然と会話したくなるという条件には視線を上手く利用する必要がある．そこで，相手の話す内容や視線などに対応して自然な視線を生成できることを目指す．

本研究の独自性は，対話中の視線生成手法において，話し手と聞き手の動的な相互作用をデータ駆動型のアプローチで学習し，自然な視線を生成するという点にある．従来の研究では，キャラクターの視線動作が単独でトレーニングされていることが多く，会話中の相互作用や文脈を反映した視線生成はほとんど行われていない．本研究では，対話中のデータセットを活用し，相手の話し手の視線と頭部動作に基づいて聞き手の視線を生成するモデルを構築する．さらに，将来的には会話内容そのものを反映した視線生成を目指しており，エージェントが発話内容に基づいて関心を示す視線動作を行えるようにすることを目標としている．これにより，単に動きが自然なだけでなく，会話の意味に寄り添った視線生成が可能になる．

2　関連研究

人と人以外が会話することに関する研究として，内田ら [3] が，人，ロボット，アンドロイドへの第一印象で被験者がそれぞれに対してどの程度自己開示するのか検証している．この実験から，人はロボットやアンドロイドに対しても被験者は一定の自己開示を行うことが示唆された．また，人に対してはポジティブな事柄に関して被験者自身に関する情報を開示される割合が高く，ロボット・アンドロイドに対してはニュートラル，ネガティブな事柄になるほど割合が大きくなっていることが分かった．このことから，会話エージェントにおいても人が一定の自己開示を行うことを推測でき，対話相手として適切であると考える．

人間同士の対話において，石井ら [4] が複数の視線情報と頭部動作の情報から会話参加態度を推定する手法について検証している．この研究から注視パターン，注視時間，注視位置の移動距離，瞳孔径に関する複数の視線情報，および眼球検出によらないものとしてヘッドトラッキングによる頭部姿勢の情報と会話参加態度との間に相関があることを示された．そのうえで，複数の視線情報と頭部動作の情報の両方を用いた推定モデルが最も性能が高く，頑健な推定モデルを構築可能であることを示した．今回の実験でも視線情報だけでなくいくつかの頭部動作の情報も収集し生成に利用する．

視線動作のアニメーション生成の研究として，Alex Klein ら [5] が， RNN(Recurrent Neural Network)[6] を使用したデータ駆動型の視線アニメーション手法を紹介している．近年データ駆動型アニメーションと機械学習のアプローチにより、アニメーションの品質向上に有望な結果が示されており，複雑な環境での移動などのアニメーション タスクにディープ ニューラル ネットワークを適用することにより，多くの有望な結果が提示されている．この研究から，データ駆動型のアプローチを使用し RNN を使用して頭，目，胴体の動きを学習したモデルはリアルタイムで自然な視線の動きを生成できる．加えて，ゲーム業界の専門家を対象に実施されたユーザー調査の結果は，有名なゲーム会社の手続き型視線アニメーション システムと比較して，彼らの方法がより自然であると認識されることを示している．一方で，この研究ではキャラクターが他のキャラクターと対話していない状況で視線動作を提供するようにトレーニングされているため，会話している時の再現は不明である．また，目の結果は不十分だったと述べている．このことから，会話時かつより視線に注力した今回の研究は，Alex Klein らの発展として新規性があると考える．

3　提案手法

本研究では，LSTM(Long Short-Term Memory)[7]，GRU(Gated Recurrent Unit)[8]，RNN を用いた回帰モデルを構築した．この 3 種類の RNN を用いて対話中の聞き手の視線角度の予測を行う．

3.1　データセット

今回は，会話データコーパス収録実験のビデオデータを用いた．ビデオデータは，話し手役の健常な高齢者（69〜73 歳）x 4 名（男女 2 名ずつ）と，聴き手役の若者（22 歳）x 4 名（男女 2 名ずつ）が 15〜30 分画面越しに会話するものを用いた．話し手 1 人と聴き手 1 人の組み合わせのため，16 セッション分のデータとなる．ビデオデータから OpenFace[9] という顔の表情認識や顔の特徴抽出を行うためのオープンソースのツールキットを使用し，視線及び頭部動作にまつわる特徴量を抽出し学習を行った．抽出した特徴量は以下の表 1 の様になっている．

信頼度は，OpenFace がビデオデータから，視線及び頭部動作を検出できているかを表しており，前処理の段階で欠損値や外れ値の削除に用いた．

表 1: OpenFace から抽出された特徴量の一覧

カテゴリ	概要
信頼度	顔認識の正確さ (1 次元)
視線方向ベクトル	右目と左目の視線方向ベクトル (6 次元)
頭部情報	頭部位置 (3 次元)，頭部回転角度 (3 次元)
表情	まばたきの強度 (1 次元)

視線方向ベクトルはワールド座標での左右それぞれの目の視線方向ベクトルである人の目の動きは．対話相手の目の動きに影響を受けると考えられるため，話し手である高齢者の視線方向ベクトルを入力データ，聞き手である若者の視線方向ベクトルを出力データとして用いる．

頭部情報はカメラに対する頭の位置とカメラを原点とする頭部動作に関するデータである．石井ら [4] の研究から，複数の視線情報と頭部動作の情報の両方を用いた推定モデルが最も性能が高く，頑健であることが示されているので，入力データとして用いる．

3.2 学習アルゴリズム

RNN は時系列データや連続した情報を処理できるニューラルネットワークの一種であり，過去の情報を記憶しつつ新しい情報を処理できる．本研究では，時系列的な特徴量（複数フレームにわたる聞き手の特徴量）を入力データとして扱う必要がある．RNN は，短いシーケンスや単純な時系列データの処理に適しており，過去の視線情報を参照しながら視線の予測ができる．

LSTM は RNN の一種であり，RNN と同様にシーケンスデータや時系列データなど、時間的な依存関係を持つデータを処理するのに適している．一般的な RNN と比較して，LSTM は勾配消失問題を回避する設計を持っており，長期間の依存関係をモデル化する能力が高い．視線の動きが影響を受ける期間が長い場合，RNN よりも LSTM の方が有効である．

GRU は LSTM が持つ機能を統合した構造を持っている．LSTM と比較してシンプルな構造のため，計算が速くメモリ使用量が少ない．また，短期的なデータ処理においても高い性能を発揮する．LSTM ではモデルが複雑すぎるときに，GRU は有効である．

3.3 実験手順

上記の 4.1 節のデータをウィンドウサイズ 30(約 1 秒が 30 フレーム)，1 フレームごとにスライドィングウィンドウを行った．使用した特徴量は視線方向ベクトル (6 次元) と頭部情報 (6 次元)，表情 (1 次元) の 13 個となる．よって一回の入力データが 30×13 となる．この際に信頼度の値が 0.9 以下のデータを含んでいた場合削除した．全てのデータに正規化を施し，話し手ごとに交差検証を行った．出力データとして，聞き手の視線方向ベクトル (6 次元) を選択した．ネットワークの構造は，入力データのサイズに対応した 30×13 の入力層時系列データを処理するための 64 ユニットの RNN/LSTM/GRU 層，過学習を防ぐため Dropout 層，32 ユニットの ReLU 活性化関数を持つ全結合層，6 ユニット（注視角度の予測）を持つ出力層とした．加えて，出力を 1 フレームのものと 7 フレームのもので比較した．また，出力範囲を 0 から 1 に制限するため、シグモイド活性化関数を使用．損失関数は平均 2 乗誤差 (MSE)[10]，オプティマイザとして Nadam を選択した．

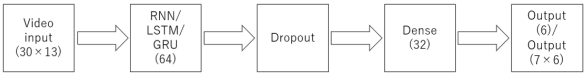

図 1: ネットワーク構造

3.4 実験結果と考察

表 2 の実験結果から，1 フレームを予測する場合 3 種類の RNN のいずれも MSE が 0.005 と低い値を示しており，高い精度で予測できていることがわかる．一方で，7 フレームの場合，MSE は 0.018 と約 3.6 倍に増加している．これは，時間が経つにつれて予測誤差が蓄積し，未来のフレームの予測が難しくなることを示唆している．また，1 フレームと 7 フレームの両方で，3 種類の RNN の MSE がほぼ同じ値を示している．このことから，今回のタスクでは 3 種類の RNN の性能差がほとんど見られなかったことがわかる．また，7 フレーム（約 0.23 秒）の範囲では長期的な依存関係がそれほど重要でない可能性がある．結果を踏まえ，特に 7 フレーム予測の精度向上のために考えられる改善策として，特徴量の追加を考えている．現在の特徴量は視線情報，顔の姿勢，瞬きを使用しているが，なぜその視線に動きになったかが分からない．そこで，視線の向いている先にあるもの，及び座標を加えることでより，視線の変化パターンを学習しやすくなる可能性がある．

表 2: 実験結果

アルゴリズム	出力フレーム数	結果 (MSE)
RNN	1	0.005
	7	0.018
LSTM	1	0.005
	7	0.018
GRU	1	0.005
	7	0.019

4 おわりに

本研究では，対話中の聞き手の視線を自然に生成する手法として，3 種類の RNN を用いた回帰モデルを構築し，会話データをもとに学習を行った．実験の結果 RNN と LSTM，GRU の性能差がほとんど見られず，短期の予測（1 フレーム）では高い精度であったが，長期の予測（7 フレーム）では改善の必要がある．今後は視線の向きだけはなく，その先にある物体及び座標のデータも加えた，より多様なデータセットを用いた学習による汎化性能，精度の向上を目指す．

参考文献

[1] 内閣官房孤独・孤立対策担当室: 人々のつながりに関する基礎調査， 内閣府， (2022).

[2] A. Kendon: Some functions of gaze-direction in social interaction, Acta Psychologica, Vol.26, pp.22-63, (1967).

[3] 内田 貴久， 高橋 英之， 伴 碧， 島谷 二郎， 吉川 雄一郎， 石黒 浩: ロボットによる傾聴を通じた自己開示の促進， 日本認知科学学会大会， (2017).

[4] 石井 亮， 大西 亮太， 中野 由紀子， 西田 豊明: 視線と頭部動作に基づくユーザの会話参加態度の推定, 情報処理学会論文， Vol. 52， No. 12， pp3626-3636， (2011).

[5] Alex Klein， Zerrin Yumak， Arjen Beij， A: Frank van der Stappen:Data-driven Gaze Animation using Recurrent Neural Networks， MIG '19: Motion， Interaction and Games， Newcastle upon Tyne， United Kingdom， (2019)

[6] J. L. Elman: Finding structure in time， Cognitive Science， Vol. 14， No. 2， pp. 179-211， (1990).

[7] S. Hochreiter and J. Schmidhuber: Long short-term memory， Neural Computation， Vol. 9， No. 8， pp. 1735-1780， (1997).

[8] K. Cho， B. van Merri?nboer， C. Gulcehre， D. Bahdanau， F. Bougares， H. Schwenk， and Y. Bengio: Learning phrase representations using RNN encoder-decoder for statistical machine translation， arXiv preprint arXiv:1406.1078， (2014).

[9] T. Baltrusaitis, P. Robinson, and L. P. Morency: OpenFace: An open source facial behavior analysis toolkit, 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp.1-10, (2016).

[10] G. E. P. Box and D. R. Cox: An analysis of transformations， Journal of the Royal Statistical Society: Series B (Methodological)， Vol.26， No.2， pp.211-252、 (1964).