7ZF-06

効果的なインタラクションのための表情生成モデルの提案および性能評価

近藤新太郎 † 原田誠一 ‡ 佐久間拓人 ‡ 加藤昇平 ‡§

† 名古屋工業大学 工学部 情報工学科
‡ 名古屋工業大学 大学院工学研究科工学専攻
§ 名古屋工業大学 情報科学フロンティア研究院

1 はじめに

近年, 人とインタラクションをしながら接客・施設案内などの業務を遂行するエージェントが普及している．エージェントがより良く情報伝達を行えれば，業務品質そのものの向上につながるため，エージェントの情報伝達手法の研究は重要である．

Mehrabian ら [1] によると，視覚・聴覚・言語の各情報が同時に与えられるインタラクションにおいて，感情や態度の情報が矛盾したメッセージが与えられた場合，受け手の重視する情報の割合は視覚情報が 55%，聴覚情報が 38%、言語情報が 7%であるとされる．よってエージェントのインタラクションにおいても，伝達する言語情報と視覚情報の感情的な矛盾を少なくすることで，インタラクションの質が大きく向上することが期待される．

しかし, 特にスマートスピーカーなどの一部エージェントにおいて，伝達される情報は言語のみである場合が多い．その上，言語情報はインタラクション中に動的に発生しうる．そのようなエージェントにおいて顔表情付きインタラクションを実現するためには，言語情報のみからの動的な感情抽出・感情表出を行えるシステムが必要である．

Zhou ら [2] は，入力として発話音声を受け取り，音声に対応する読み上げ顔表情を生成するアーキテクチャを提案した．このアーキテクチャでは，目元の開き方・口角の上がり方といった，話者の話し方を再現した顔表情生成が可能である．しかし，この手法では，入力・出力の感情の一致については考慮していない．また，テキストを入力としたモデルについても考慮していない．

著者ら [3] は，Zhou らの手法を改良し，テキスト入力から感情付き顔表情を生成するモデルを提案した．提案モデルでは，生成表情と感情ラベル付けがされた教師データ表情との平均二乗誤差を用いて顔表情生成モデルを学習し，テキストに表現された感情を顔表情に反映することを試みた．しかし，生成表情に対し意図した感情を十分に付与できなかった．

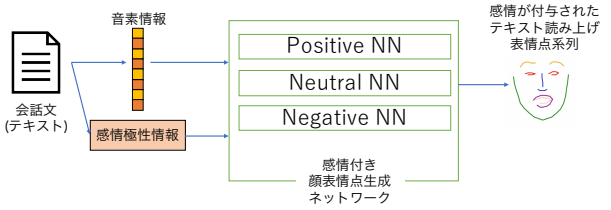


図 1: 顔表情点生成モデル

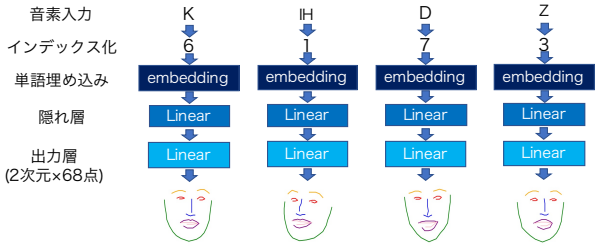


図 2: 顔表情点生成ネットワーク

2 *EmotionLoss* の導入

本稿では，先行研究 [3] の生成表情に感情を付与することを目的とし，生成表情に意図した感情が表出されていることに着目した *EmotionLoss* を導入する．

図 1 で示される顔表情点生成モデルは，テキストから読み取った感情極性情報に応じて内部の感情付き表情点生成ネットワークを使い分けることで，入力テキストの感情情報を反映した表情を生成する．生成ネットワークは**図 2** の構成となっており，各モデルはそれぞれ対応する感情で発話された動画を用いて学習する．本稿ではこのネットワークの学習に，感情ラベル付き発話動画データセットとして RAVDESS[4] を用いた．データセットは各発話ごとに切り出し，68 個の表情点に変換する．モデルの損失関数には以下の式で表される平均二乗誤差（*MSELoss*）を用いることで，モデルが生成する表情点は，データセットの表情点に近い値を出力できようになる．

$$MSELoss = \frac{1}{A}\sum_{i}^{A}((G_i - T_i)^2), \qquad (1)$$

$$G_i, T_i \in \mathbb{R}^2.$$

ここで，A は 1 フレームに含まれる表情点の個数 68，モデルが生成した A 個の表情点を G，教師データの A 個の表情点を T とする．

平均二乗誤差を用いることで，モデルが生成する表情点は，データセットの表情点に近い値を出力できる

*Proposal and Performance Evaluation of Expression Generation Models for Effective Interaction, Shintaro KONDO †Seiichi HARATA ‡Takuto SAKUMA ‡and Shohei KATO‡§
†Dept. of Computer Science, Nagoya Institute of Technology
‡Dept. of Engineering, Graduate School of Engineering, Nagoya Institute of Technology
§Frontier Research Institute for Information Science, Nagoya Institute of Technology †‡Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
{kondo,harata,sakuma,shohey}@katolab.nitech.ac.jp

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

ようになる．しかし，平均二乗誤差のみを用いて十分学習したモデルによる出力は，単に教師データの表情に対し座標が近い表情点ではあるものの，生成表情にデータセットの感情が付与されていることは保証されない．

そこで，生成表情に対し意図した感情を付与するため，生成表情と教師データの表出感情の誤差を表現した *EmotionLoss* を導入する．この損失関数では，生成表情および教師データの表情を感情分類モデルに入力し，Softmax 関数によって出力される分類結果のカテゴリカル分布が近づくよう損失関数を設定する．表情分類器は，表情生成の教師データに用いたものと同じ RAVDESS データセット [4] で学習する．

教師データ表情の感情を 3 クラス分類した結果のカテゴリカル分布を P とし，同様に生成データ表情の，感情を 3 クラス分類した結果のカテゴリカル分布を Q とする．生成表情と教師データ表情の表出感情を近づけるということは，P，Q のカルバック・ライブラー情報量 D_{KL} を最小化することと同義である．モデルの学習に際し P が不変であることから，

$$\begin{aligned}\min_Q D_{KL}(P||Q) &= \min_Q \sum_c P(c)\ln\frac{P(c)}{Q(c)}, \\ &= \sum_c P(c)\ln P(c) \\ &\quad + \min_Q \sum_c (-P(c)\ln Q(c)), \qquad (2) \\ &\quad c \in \{positive, negative, neutral\}\end{aligned}$$

と表せるため，P に対する Q の交差エントロピー最小化問題に帰着される．この交差エントロピーを *EmotionLoss* とする．

$$EmotionLoss = \sum_c (-P(c)\ln Q(c)). \qquad (3)$$

この交差エントロピーと，教師データ・生成データに対する *MSELoss*(式 1) との加重和を，感情付き表情生成における損失関数として定義する．

$$\begin{aligned}Loss &= MSELoss \qquad (4) \\ &\quad + \lambda \cdot EmotionLoss, \\ &\quad \lambda > 0.\end{aligned}$$

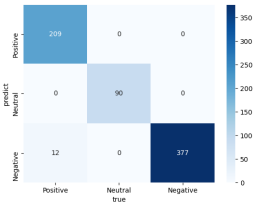
3 実験

EmotionLoss を導入しない場合 ($\lambda = 0$) と，*MSELoss* と *EmotionLoss* を同じ比率で導入した場合 ($\lambda = 1$) のそれぞれについて，表情分類器を用いて分類を行う．

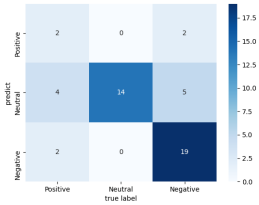
図 3 に感情分類結果の混同行列，**表 1** に分類精度，macro F1-Score を示す．$\lambda = 0$ に対し，$\lambda = 1$ では分類結果が精度・F1-Score の双方で向上していることから，*EmotionLoss* を導入することで，生成モデルが意図した感情が付与された表情点を出力できるようになった可能性が示唆された．

表 1: 精度・macro F1-Score

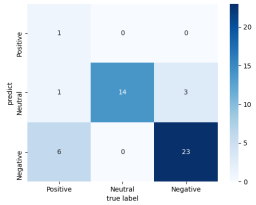
	教師表情	$\lambda = 0$	$\lambda = 1$
精度	98.2%	72.9%	79.1%
macro F1-Score	0.985	0.632	0.644



(a) 教師表情での分類結果



(b) $\lambda = 0$



(c) $\lambda = 1$

図 3: 教師表情点・生成表情点の混同行列

4 おわりに

本稿では，テキストの感情極性情報を反映した表情生成を目的とし，*MSELoss* と *EmotionLoss* の加重和を用いた新たな損失関数を提案した．また実験において，提案手法を用いることで，出力感情に意図した感情が付与されることを示唆した．今後は，より最適な λ の値について調査し，EmotionLoss の導入で生成結果に有意な差が現れるかを検証する．また，λ を固定せず，学習に応じて動的に変化させることで，λ を固定した場合と比べて生成結果に感情が付与できるかについても検証する．

参考文献

[1] Albert Mehrabian. ”Silent messages”, Vol. 8, No. 152, pp. 40-56. Belmont, CA: Wadsworth, 1971.

[2] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakisand,Dingzeyu Li. ”MakeltTalk: speaker-aware talking-head animation.”, Vol. 39, No. 6, pp. 1-15. ACM Transactions on Graphics (TOG), 2020.

[3] 近藤 新太郎, 佐久間 拓人, 加藤 昇平. ”任意の顔画像に対して感情極性を反映したテキスト読み上げ表情生成モデルの提案”. 電気・電子・情報関係学会第 34 回東海支部連合大会, 2020.

[4] Steven R Livingstone,and Frank A Russo. ”The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English”, Vol.13, No. 5, pp. e0196391, Public Library of Science San Francisco, CA USA, 2018.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.