J-021

テキスト読み上げ感情表情エージェントのための転移学習を用いたリップシンク表現の獲得

Acquisition of Lip-sync Expression Using Transfer Learning For Text-to-Speech Agents with Emotional Facial Expression

近藤 新太郎 †　原田 誠一 †　佐久間 拓人 †　加藤 昇平 †‡
Shintaro Kondo　Seiichi Harata　Takuto Sakuma　Shohei Kato

1　はじめに

近年，Alexa や Siri のような，人間とインタラクションする音声エージェントが広く普及している．これらのエージェントとの対話を通じて，ユーザーは情報を検索したり，音楽を再生するなど，インターネット上の様々なサービスを利用できる．また，企業側も，自社が運営しているサービスを簡単に利用するためのインタフェースを提供できる．エージェントと対話することへのモチベーションが増加すれば，これらの企業のサービス利用促進に繋がることから，エージェントの対話促進は重要な課題である．

しかし，現在普及しているエージェントは，スマートフォンやスマートスピーカーから音声のみを介してコミュニケーションすることが多く，人間の顔表情のように，感情を表出するための視覚ディスプレイを持たない．

我々は，エージェントが感情を表現するための視覚ディスプレイとして，エージェントの発話に，人間らしい感情を表出可能な顔表情を付与することを検討する．また，顔表情での感情表出を通じてエージェントの人間らしさを向上させ，エージェントをより親しみやすくすることで，人間の対話モチベーションを向上させ，エージェントのサービス利用促進を目的とした研究を行っている．

Zhou ら [1] の先行研究では，LSTM をはじめとした深層学習モデルによって，発話音声から任意の顔画像に対する読み上げ表情を作成する手法を提案した．しかし，Zhou らの手法は生成表情を感情伝達ディスプレイとすることを目的としていないため，生成表情の感情を任意に変化させることはできない．

近藤ら [2] の先行研究では，Zhou らの手法を参考に，エージェント向けの顔表情ディスプレイをテキストから任意の顔画像に対して，ポジティブ・ニュートラル・ネガティブの 3 感情の中から意図した感情を表出可能な読み上げ顔表情を生成する手法を提案した．しかし，学習に用いた感情付き発話動画データセットには含有される音素数が少なく，リップシンクを始めとした自然な顔顔表情生成能力を学習できないため，モデルの生成した顔表情の口元部分に不自然さが残った．

2　提案手法

本稿では，近藤らの先行研究で課題となっていた自然な発話顔表情の生成能力を獲得することを目的として，テキスト読み上げ顔表情としてより自然な，感情付き発話動画を生成するための深層学習モデルを提案する．

先行研究では，顔表情生成モデルを学習するために，感情付き発話動画を集めた RAVDESS データセット [3](以下，単に RAVDESS) を用いた．しかし，RAVDESS を始めとした感情表情発話データセットは文章の発話パターンを学習することを意図したモデルではないため，含有する発話パターンが少ない．また，感情付き発話動画はデータ収集が困難であり，データサイズも少ない場合が多い．そのため，感情表情発話データセットのみを用いて学習を行ったモデルは，自然な読み上げ顔表情の生成能力を獲得できる保証がないと考える．

そのため，本稿では，リップシンク表情生成の分野で用いられるデータセットを用いて自然な読み上げ顔表情を学習し，その後感情付き発話データセットを用いてネットワークを転移学習することで，より自然な読み上げ顔表情の生成能力と，感情を表出した表情の生成能力を両立することを意図したモデルを提案する．

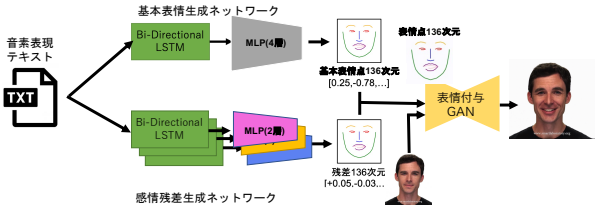


Fig. 1　提案モデル

本稿で提案するモデルの概形を図 1 に示す．ネットワーク内部では，入力として与えられる音素を，リップシンク知識のみで学習した基本表情生成ネットワーク（図 1 上段）と，リップシンク知識・感情付き発話知識で学習した感情残差ネットワーク (図 1 下段) の 2 つに入力し，顔パーツを 2 次元座標で表現した表情点を中間物として生成する．図上段のネットワーク (以下，基本表情生成ネットワーク) では，読み上げ表情として自然な基本表情点を生成する．ネットワーク下段 (以下，感情残差生成ネットワーク) では，基本表情を感情的な表情の座標上の差分を計算する．感情残差ネットワークについては，m 種類の感情を付与するネットワークを個別に学習し，表出したい感情によってネットワークを切り替えることで，異なる感情を表出可能となる．その後，基本表情点と感情残差の線形和をとることによって，最終的な表情点を生成する．生成された表情点は音素に対応した表情になっているので，音素の発話時間に合わせて線形補完する．最後に，表情点に対し GAN を用いて実在人間の顔表情を付与することで，表情点を任意の人物の発話動画へと変換する．この GAN は，Zakharov ら [4] の手法を用いて学習する．提案モデルにおいても GAN については先行研究と同じものを用いるため，本稿ではモデル内部の表情点生成に着目して学習し，結果を評価する．

3　実験

提案モデルに対し，リップシンクドメインの知識を有するデータセット，および感情付き発話データセットを用いて学習する．

3.1　データセット

リップシンクに関する知識を有する顔表情データセットは，単にリップシンク知識を有しているだけでなく，感情的な発話表情の知識を転移しやすいことが望ましい．

† 名古屋工業大学，Nagoya Institute of Technology
‡ 名古屋工業大学　情報科学フロンティア研究院，Frontier Research Institute for Information Science, Nagoya Institute of Technology

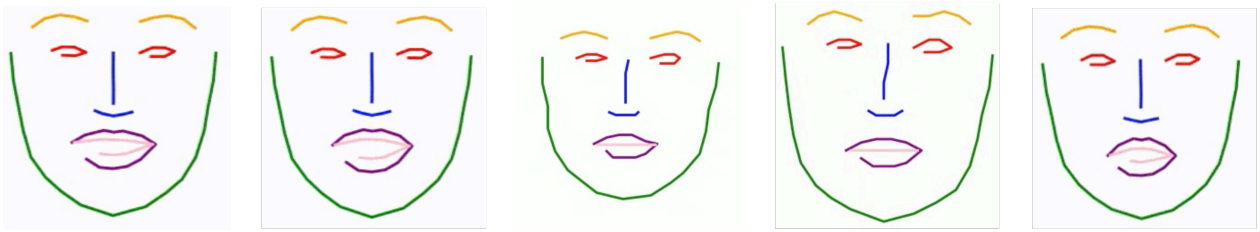


Fig. 2　先行研究

Fig. 3　提案手法

Fig. 4　RAVDESS

本稿では，Suwajanakorn ら [5] の先行研究でも用いられていた Obama 氏の Weekly Address 動画 (以下，単に Weekly Address とする) を，リップシンク知識を含有するデータセットとして用いた．RAVDESS データセットでは，各感情ごとの学習に利用可能なデータは 24 分程度だったが，Weekly Address は，17 時間分の英語発話データが含まれており，RAVDESS データセットに比べ豊富なリップシンク知識を含有していることが期待される．それだけでなく，発話者の顔がほとんど正面向きであり，顔表情の表出する感情の機微が少ないことなどから，後から感情的な発話表情の知識を転移しやすいため，感情的な発話表情に関する知識の転移学習がしやすいと考えられる．また，感情的な発話表情の知識を有するデータセットとして，近藤らの先行研究 [2] と同様に RAVDESS を用いた．今回は m=3，すなわちポジティブ，ニュートラル，ネガティブの 3 感情を表出するモデルを学習する．ポジティブ感情として Happiness，ネガティブ感情として Sad，ニュートラルとして Neutral 感情の表情をそれぞれ用いた．

3.2　提案モデルの訓練

まず，モデル内部の基本表情生成ネットワークのみを，Weekly Address を用いて学習する．損失関数には MSE を用いて 1000 エポック学習した．その後，基本表情生成ネットワークの学習パラメータを固定し，残差を含む表情点生成モデル全体を RAVDESS データセットを用いて学習した．基本表情生成ネットワークの学習と同様損失関数に MSE を用いて 100 エポック学習を行った．本稿で提案したモデルによる表情点を図 3，先行研究 [2] で提案されたモデルによる表情点を図**??**，データセットから抽出した表情点を図 4 に示す．

4　感性実験

これらの生成表情動画に対し，感性実験により，提案モデルが先行研究 [2] に比べより自然な表情を生成する能力を獲得したこと，および提案モデルが感情的な表情を生成可能であることを検証する．実験では，実験協力者に対しいくつかの表情点を見せ，表情点の発話動画の口元の自然さについて，「まったく自然に見えない」を 1，「とても自然に見える」を 7 とした 7 件法により評価させる．また，表情点の表出する感情について，ネガティブ，ニュートラル，ポジティブの 3 値分類により評価させる．実験協力者に見せる表情動画種類は以下の通りである．

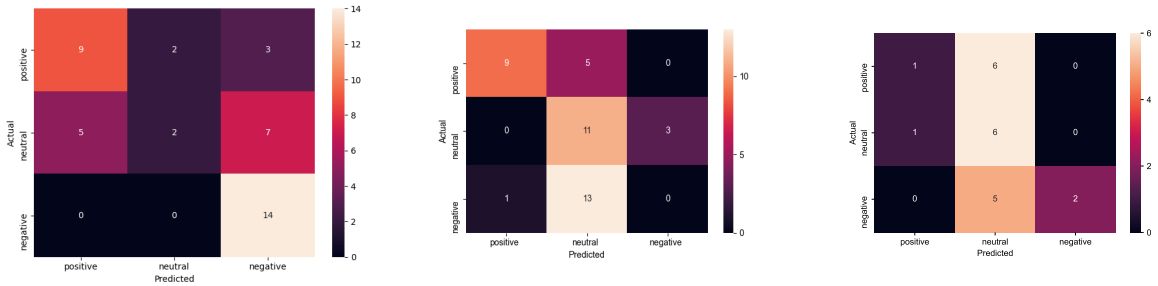
1. RAVDESS データセットから抽出した表情点
2. 提案モデルの基本表情生成ネットワークのみで生成した基本表情点 (「提案'」とする)
3. 提案モデルの感情残差ネットワークで感情を付与した表情点 (「提案」とする)
4. 近藤らの先行研究 [2] で提案された手法による生成表情点

実験協力者 7 名にこれらの動画を見せ，上記の質問に回答させた．

4.1　表情点そのものの表現能力に関する検証

Table. 1　RAVDESS データセットの感性評価 (平均値)

表情点の感情ラベル	口元の自然さ
Positive	4.36
Neutral	3.93
Negative	3.50



(a) RAVDESS　(b) 提案モデル　(c) 先行研究

Fig. 5　混同行列

まず，表情点そのものに，発話の自然さ，および感情に関する表現力があることを検証する．そのため，比較対象の中で最も表情の口元が自然で，かつ感情が強く表出されている RAVDESS データセットから抽出した表情点に対する回答結果を比較する．各感情ラベルに対する被験者の回答結果について，各感情に対する口元の自然さ，および表出感情の種類について，回答の平均値を表 1 に，RAVDESS データセットに対する各感情に対する分類結果の混同行列を図 5(a) に示す．まず，口元の自然さについては，7 段階中 3 から 4 程度の値となった．4.2 章の考察では，この値を基準に評価する．次に，表情の表出する感情について，ポジティブ，ネガティブのそれぞれについて，多くの人が意図した感情を正答することができた．(ポジティブ 100%，ネガティブ 64%) しかし，ニュートラルについては，ポジティブおよびネガティブという回答がニュートラルより多く，正解率も低い (14%) ことから，表情点では伝わりづらい感情である可能性がある．また，分類結果の分散に着目すると，ニュートラル・ネガティブについては回答の分散が多く，ニュートラル・ネガティブ特徴は表情点に現れにくい，あるいは人間が表情点から読み取ることが難しいことが考えられる．

4.2　提案手法の生成表情点の自然さについての検証

Table. 2　RAVDESS データセットの感性評価 (平均値)

表情点の種類	口元の自然さ
RAVDESS	3.93
FiT’	2.21
FiT	3.55
先行研究 [2]	3.52

次に，提案モデルが先行研究 [2] と比較して，より自然な発話を実現していることを確認する．そのため，RAVDEES，FiT’，FiT，および先行研究 [2] での表情点に対する発話の自然さへの回答結果を比較する．

各データセット・モデルに対する口元の自然さの回答の平均値を表 2 に，回答結果をプロットした混同行列を図 5 に示す．FiT，すなわち提案モデル全体で生成した結果の自然さについては，先行研究 [2] に比べわずかに向上が見られた．しかし，FiT’ はリップシンクに特化したネットワークであるため，FiT，先行研究 [2] に比べより口元が自然な表情点動画が生成されることが期待されるが，実際には FiT, 先行研究 [2] に比べ，FiT’ の方が自然さの評価が低いという結果となった．この原因として，FiT’ のモデルで生成した音素ごとの表情点画像の段階では，音素ごとの微細な動きまで捉えているものの，画像から動画を作成する段階で，他の生成モデルに合わせフレームレート (10fps) や動画時間 (1 発話あたり 3 秒) を均一にしたため，これらの微細な表現が失われてしまったためであると考えられる．逆に FiT および先行研究 [2] では，口元の動きがあいまいになっており，動画に変換しても違和感が少なかったのではないかと考えられる．今後，この仮説を検証するため，音素ごとの表情画像を高フレームレートかつ低速の動画に変換し，伝達可能な表現を増やした上で，各データセット・モデル間の口元の自然さに対する追加の検証を実施する予定である．

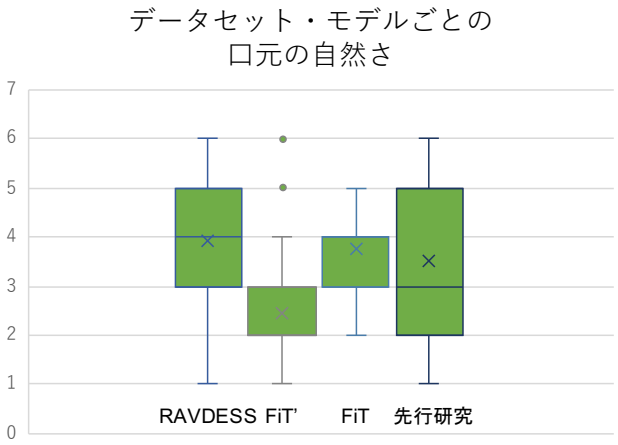


Fig. 6　各モデルの口元の自然さに関する回答結果

4.3　表情点の感情表現能力の検証

Table. 3　各データセット・モデルの感情分類結果 (正答率)

表情点の種類	Positive	Neutral	Negative
RAVDESS	100%	14.3%	64.3%
FiT	64.3%	78.6%	0.0%
先行研究 [2]	14.3%	85.7%	28.6%

最後に，提案モデルの表出表情に対する感情付与能力について，先行研究 [2] に比べ表出される感情がどの程度変化したかを確認する．そのため，RAVDESS，FiT, 先行研究 [2] の各データセット・モデルごとに，ポジティブ，ニュートラル，ネガティブな感情を表出した表情点を見せ，それらがどの感情を表しているか回答させた結果を比較する．各データセット・モデルに対する回答の正答率を表 3 に示す．結果から，ポジティブ感情の生成能力については，先行研究 [2] に比べ向上したといえる．しかし，ニュートラルおよびネガティブ感情の生成能力については，先行研究 [2] に比べわずかに悪化した．これら 2 ラベル

の回答については，4.1 章において RAVDESS データセットに対して行った調査 (図 5) でも，回答結果の分散が大きくなっており，表情点における感情の学習・判別が難しい可能性がある．よって，ニュートラル感情・ネガティブ感情については，表情点生成モデルの段階ではなく，GAN で実在表情に変換する段階で，感情表出能力を持たせることも検討する．

5 まとめ

本研究では，エージェントに人間らしい顔表情ディスプレイを付与することを目的として，先行研究 [2] で問題となっていた自然な読み上げ表情の知識を別のデータセットから転移学習することで，より自然な感情付き読み上げ表情点生成モデルを提案した．その結果，先行研究 [2] に比べ自然さの改善がみられた．また先行研究 [2] に比べ，ポジティブ感情をより強く表出されたことから，一部感情表出において有効性が確認できた．
今後の課題として，高フレームレートかつ低速の動画において読み上げ表情の自然さが向上するかについて検証する．また，表情点上でのニュートラル・ネガティブ感情の表出は困難である可能性が示唆されたため，これらの感情については GAN を用いて付与することを検討する．

6 学習結果の動画

6 章の感性実験で用いた動画を Youtube に掲載する．

- RAVDESS
  - ポジティブ感情
    - https://youtu.be/mx8vdxhiqz8
    - https://youtu.be/0nh8zlixuta
  - ニュートラル
    - https://youtu.be/QeVWJNumrjE
    - https://youtu.be/s3lK1E3-lQU
  - ネガティブ感情
    - https://youtu.be/-0xmo5LR-1M
    - https://youtu.be/cgOfoyRX81A
- FiT’
  - https://youtu.be/fi4Ipmf5Twc
  - https://youtu.be/IYWrmSjnI10
- FiT
  - ポジティブ感情
    - https://youtu.be/ctUNdHnQ2NA
    - https://youtu.be/eKimCekOknw
  - ニュートラル
    - https://youtu.be/7VSdiucIGSg
    - https://youtu.be/EPolUiDt4ig
  - ネガティブ感情
    - https://youtu.be/H9gvjwx9h8w
    - https://youtu.be/x2Lyk1B6L8g
- 近藤らの先行研究 [2]
  - ポジティブ感情
    - https://youtu.be/GaWpBgbvp0g
  - ニュートラル
    - https://youtu.be/c98B6LOhL20
  - ネガティブ感情
    - https://youtu.be/34IEBeUL7mo

参考文献

[1] Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E. and Li, D.: MakeltTalk: speaker-aware talking-head animation, *ACM Transactions on Graphics (TOG)*, Vol. 39, No. 6, pp. 1–15 (2020).
[2] 近藤新太郎, 佐久間拓人, 加藤昇平 : 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, 情報処理学会 第 83 回全国大会 (2021).
[3] Livingstone, S. R. and Russo, F. A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, *PloS one*, Vol. 13, No. 5, p. e0196391 (2018).
[4] Zakharov, E., Shysheya, A., Burkov, E. and Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9459–9468 (2019).
[5] Suwajanakorn, S., Seitz, S. M. and Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio, *ACM Transactions on Graphics (TOG)*, Vol. 36, No. 4, pp. 1–13 (2017).