End-to-End 学習を用いた音声からの表情アニメーション生成

Generation of Facial Animation from Voice using End-to-End Learning

大道 博文
Hirofumi Omichi

広島市立大学大学院 言語音声メディア工学研究室
Language and Speech Research Laboratory, Graduate School of Information Sciences, Hiroshima City University

概要 近年，インターネット上で CG アバターを介した他者とのコミュニケーションが普及しつつある．代表的な表情同期手法として Face Tracking が挙げられるが，表情を持たない収録済みの音声や合成音声から表情を作り出すことができない．そこで本研究では，音声のみを用いて表情アニメーションの自動生成を行うことを提案する．実験の結果，既存手法よりも提案手法の方が有効性を確認できた．

1 はじめに

近年，VR ゲームや Virtual YouTuber といった 2D や 3D アバターを介した他者とのコミュニケーションが普及しつつある．このようなアバターを通じてユーザの心理状態を伝達させるために，特定の感情を示す表情をアバターに表出させる方法や，自身の表情や動作をアバターと同期させる方法がよく用いられている．

代表的なアバターの表情表現として Face Tracking が挙げられる．この手法では人間の目や眉，口といった顔部位を座標点として捉え，それをアバターの顔に対応させて表現することができる．つまり，男性ユーザが女性アバターを演じることやその逆のことも可能となり，性別に関係なく使用できるという利点がある．しかし，Face Tracking の手法では人間の顔が必要であるため，顔画像を伴わない収録済み音声や合成によって作られた音声から表情アニメーションを作り出すことは困難である．

そこで本研究では，音声のみを用いて表情アニメーション生成を行う．本研究の概要図を図 1 に示す．具体的には音声の音響的特徴量を入力とし，表情動画から解析された表情パラメータを教師データとして学習モデルを設計する．評価として，既存手法と提案手法の Loss 値を比較していく．また，生成された表情アニメーションに対して自然な表情として知覚できるかについても議論する．

2 先行研究

本研究の目的は音声から表情アニメーションを生成することである．本節では音声を用いた表情アニメーション生成の先行研究を 2 つ紹介する．

[1]は音声のフォルマント情報から表情アニメーション生成を試みている．この研究で使用したネットワークは主に CNN で構成されており，フォルマント情報の特徴抽出や時系列の流れを CNN で分析している．しかし，この研究では時系列の流れをより意識した Recurrent Neural Network（RNN）や Long Short-Term Memory（LSTM）を用いた検討がされていない．

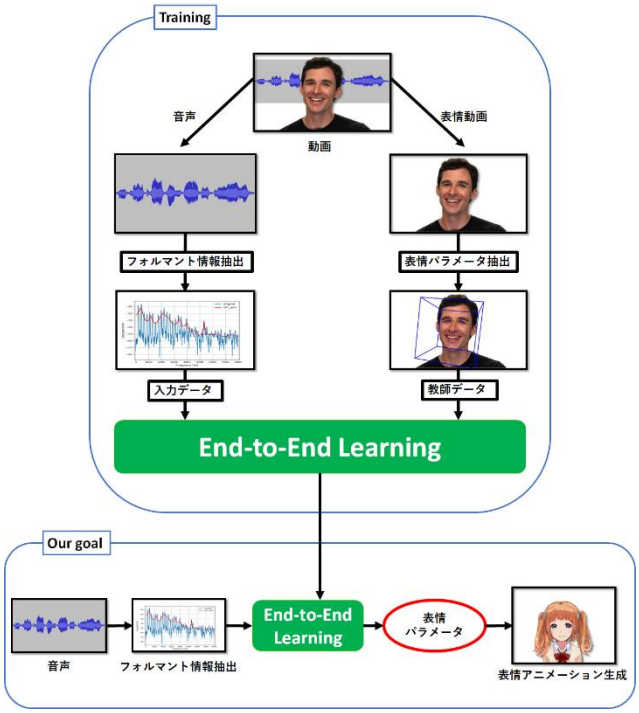

図 1: 本研究の概要図

したがって，本研究では音声の時系列の流れを意識したネットワークを構築する必要がある．

また，[2]は音声から解析されたスペクトログラム情報を用いた研究を行っている．この研究では深層学習モデルとして LSTM-RNN を使用しており，聴覚特性を考慮した比較実験をしている．ただ，この研究の課題にも指摘されているように深層学習の出力結果が滑らかではない，つまり表情として不自然であることを示唆している．したがって，本研究では生成された表情アニメーションが自然な表情として知覚できるかを検討していく．

3 提案手法

本研究では音声から表情アニメーションのパラメータを出力する End-to-End 学習モデルを提案する．

3.1 ネットワーク

前章で述べた先行研究は CNN と LSTM が単独で用いられており，それらの手法を組み合わせた試みが行われていない． したがって，本研究では[1]のネットワーク構造を基にして LSTM 層を加えた手法を提案する．改良したネットワーク構造をそれぞれ表 2，3，4 に示す．

表 2，3，4 の太字部分が本研究で提案する改良点である．それぞれの表に対して，改良した内容を順に説明する．

表 2: フォルマント分析のネットワーク構造

Layer type	Kernel	Stride	Outputs
Convolution	1×3	1×2	72×64×16
Convolution	1×3	1×2	108×64×8
Convolution	1×3	1×2	162×64×4
Convolution	1×3	1×2	243×64×2
Convolution	1×2	1×2	256×64×1
Convolution	**1×1**	**1×1**	**128×64×1**
Convolution	**1×1**	**1×1**	**64×64×1**
Convolution	**1×1**	**1×1**	**1×64×1**

表 3: 発音のネットワーク構造

Layer type	Input	Hidden	Outputs
LSTM	**1**	**256**	**1×1×256**

表 4: 出力のネットワーク構造

Layer type	Input	Hidden	Outputs
Fully connected	**256**	**-**	**128**
Fully connected	**128**	**-**	**2**

表 5: Happy の AU の一覧

感情	AU number	動作
Happy	6	頬を上げる
	12	口端を上げる

表 2 はフォルマント分析を行うネットワークであり，基本的に[1]と同様の構成要素である．しかし，その次のネットワークに LSTM 層を実装するためにチャネル数 256 から 1 まで次元圧縮を行う処理を加えた．

表 3 は発音に関するネットワークで[1]は 5 つの Convolution 層で構成されていた．しかし，ここでは表 2 の出力結果の時系列変化を分析する処理であるため，2 章で述べたように時系列の流れに重きをおいた LSTM 層を実装した．

表 4 は表情パラメータを出力するネットワークである．なお，本研究で用いる表情パラメータについては次節で説明する．

3.2 表情パラメータ

本研究ではアニメ調のアバターを用いた表情アニメーション生成を目的としているため，先行研究とは異なるアプローチで表情パラメータを考える．

まず，アバターの表情表現について[3]は Action Unit（AU）の組み合わせを用いた表情アニメーションの生成を試みている．この研究では複雑な感情を表現する際に幸福と怒りの AU を部分的に組み合わせることで快感情抑圧表現を定義している．本研究では複雑な感情は考慮しない．しかし，単一感情に対応する表情の定義づけという観点から AU は有効であると考えられる．

次に，教師データとして用いる表情動画から表情パラメータを取得する方法を説明する．本研究では AU を表情パラメータとして定義するため，OpenFace[4]を使用する．OpenFace は顔の座標点や視線推定，頭の姿勢推定，AU の強度を検出することができる．なお，アバターのコミュニケーションにおいて Happy の表情は最も重要であると考えたため，本研究では表情パラメータを Happy に関する AU に限定する． [5]から Happy に関する AU の一覧を表 5 に示す．

以上より，本研究では表 5 の AU の強度を表情パラメータとして定義する．

3.3 アバターの表情作成

アバターの表情作成は[3]と同様に Live2D Cubism [6]を用いる．また，感情を表現するアバターも同様にしずくを使用して，表 5 の AU の動きを手作業で作成した．

4 実験と考察

本章では表情アニメーション生成において，提案手法が既存手法より有効であることを示すために比較実験を行う．また，それぞれの推定結果をグラフ化して，表情が滑らかに推移しているかについて考察する．

4.1 使用するデータセット

本研究で使用する表情動画として RAVDESS[7]を用いる．RAVDESS には男性 12 名，女性 12 名のプロの俳優による 8 種類の感情音声ビデオが収録されている．感情は Neutral, Calm，Happy，Sad，Angry，Fearful，Disgust，Surprised である．本研究では表情のパラメータを Happy の AU に限定しているため，Happy の感情音声ビデオを使用する．そして，そのビデオから学習データとして 520 ミリ秒（520ms）の音声， OpenFace を用いて解析された AU の強度を教師データとした．なお，前処理としてサンプリング数を 16kHz に設定し，音声ごとのボリュームと AU の強度の正規化（Normalization）を行った．

4.2 実験 1（Training における Loss 値の推移）

本節では Training において既存手法と提案手法の Loss 値の推移を比較する実験を行う．この実験ではそれぞれの手法が与えられた訓練データを学習していること，つまり収束しているかを確認する．本実験では既存手法として[1]のネットワーク（CNN），LSTM1 層のネットワーク（LSTM）を用いて提案手法（CNN-LSTM）との比較を行う．Loss 関数は[1]から以下の 2 つを使用した．なおスカラーは提案手法の出力パラメータ数に調整した．

- Position term P(x):

$$\mathrm{P(x)} = \frac{1}{2}\sum_{i=1}^{2}\left(y^{(i)}(x) - \hat{y}^{(i)}(x)\right)^2$$

- Motion term M(x):

$$\mathrm{M(x)} = \frac{1}{2}\sum_{i=1}^{2}\left(m\left[y^{(i)}(x)\right] - m\left[\hat{y}^{(i)}(x)\right]\right)^2$$

Position term は訓練サンプルx を入力とし，期待される出力結果y と推定された出力結果$\hat{y}$の Mean Squared Error（MSE）関数であり，$y^{(i)}$の i は表情パラメータの添え字である．

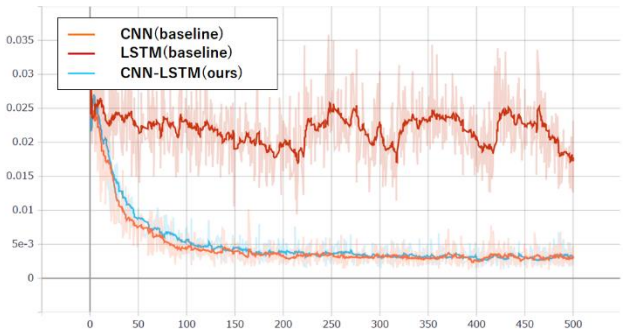

図 2: Training における Position term の推移

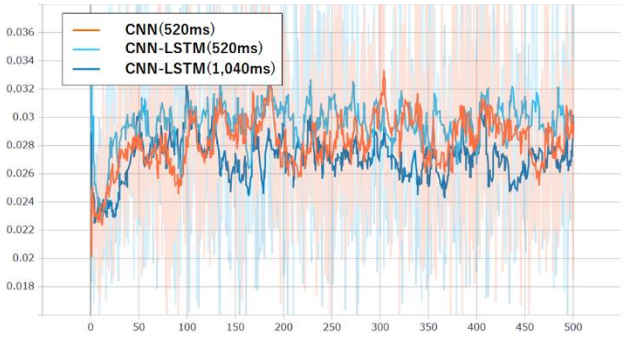

図 3: Validation における Position term の推移

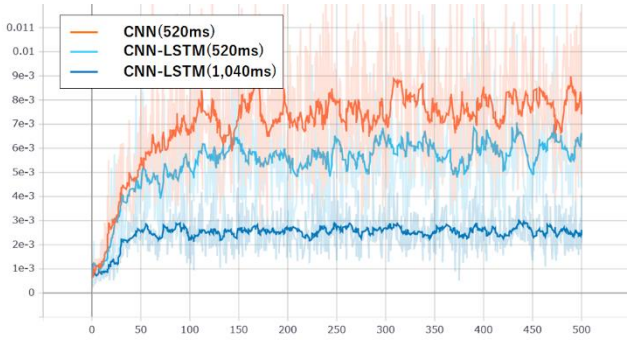

図 4: Validation における Motion term の推移

Motion term は Position term と同様に MSE 関数である．ただ，出力結果とその隣合うフレームの差分をとる $m[\cdot]$が定義されている点が異なる．

本実験では Position term の推移から，それぞれの手法の収束具合を判断する．バッチサイズは 32，epoch 数は 500，最適化手法を Adam に設定した．実験の結果を図 2 に示す．

図 2 から，既存手法の CNN と提案手法の CNN-LSTM が epoch500 で収束しているように見える．一方，もう 1 つの既存手法である LSTM については変動が大きいことから収束していないことがわかる．そこで実験 2 では LSTM を除外して実験を行う．

4.3 実験 2（Validation における Loss 値の推移）

本節では Validation において既存手法と提案手法の Loss 値の推移を比較した．本実験では実験 1 と同様に CNN，CNN-LSTM との比較を行う．さらに提案手法の CNN-LSTM の入力音声の長さを 1,040 ミリ秒（1,040ms）にした実験も行い，比較する．これはより長い音声を入力することで，

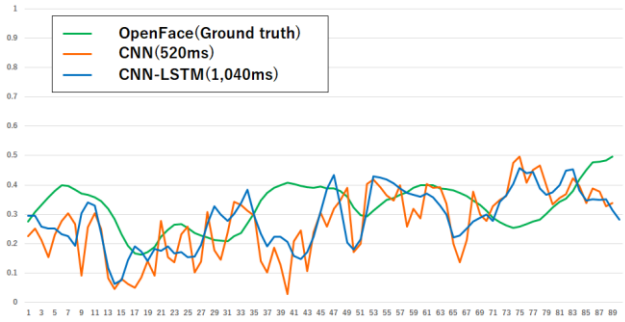

図 5: 各手法における AU6 の推移

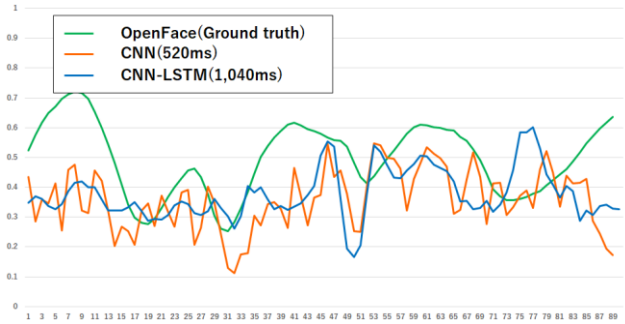

図 6: 各手法における AU12 の推移

広範囲の時系列の流れに着目した学習ができると考え，追加で実験をした．実験の結果を図 3，4 に示す．

まず図 3 から，どの手法の Loss 値もほぼ横ばいであることがわかる．したがって，Position term については提案手法の有効性を示すことができなかった．

次に図 4 から，CNN（520ms）と CNN-LSTM（520ms）の Loss 値の推移を見ると CNN-LSTM（520ms）の方が低いことがわかる．また，最も Loss 値が低い結果は CNN-LSTM（1,040ms）であった．したがって，Motion term については提案手法の有効性を示すことができた．

4.4 考察

実験 2 での Validation データを使用した推定結果をグラフ化する．比較対象は OpenFace から解析された AU の推移（Ground truth），既存手法（CNN），提案手法（CNN-LSTM（1,040ms））の 3 種類で比較した．表情パラメータの AU6，12 の推移を図 5，6 に示す．

図 5，6 の縦軸は AU の強度，横軸がフレーム数である．図 5，6 から，CNN（520ms）と CNN-LSTM（1,040ms）の AU の推移を比較すると CNN-LSTM（1,040ms）のほうが滑らかに推移していることがわかる．これは図 4 より，提案手法の Motion term が既存手法よりも低い結果になった影響だと考えられる．すなわち，表情アニメーションにおいて，提案手法のほうがより自然な表情が生成できたことを示している．しかし，提案手法と OpenFace（Ground truth）を比較すると，OpenFace（Ground truth）の AU の推移にあまり近似していないことがわかる．これは図 3 より，どの手法においても Position term が横ばいになっていたため，真値に近づくことが困難であったと考えられる．

5　おわりに

本研究では音声のみを用いた表情アニメーション生成手法を提案した．[1]のネットワーク構造を基にして時系列の流れに重きをおいた LSTM 層を実装し，改良を行った．

従来手法との比較実験の結果，Validation における Motion term が既存手法よりも提案手法のほうが低くなり，表情が滑らかに推移していることが明らかになった．このことは自然な表情アニメーション生成において重要な知見である．しかし，Position term についてはどの手法も真値に近づいた推定結果を得ることが出来なかった．

今後の課題としては訓練データ数をより多くして学習を行うことが挙げられる．また，生成した表情アニメーションに対して人手による印象評定実験を行う予定である．

6　参考文献

[1] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen: Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion, ACM Transactions on Graphics, Vol. 36, No. 4, pp. 94 (2017).

[2] 西村亮佑，酒田信親，富永登夢，土方嘉徳，原田研介，清川清: 深層学習を用いた入力音声に適した顔表情生成，第 23 回日本バーチャルリアリティ学会大会論文集，(2018).

[3] 大道博文，林柚季，目良和也，黒澤義明，竹澤寿幸: Action Unit の組み合わせを用いた快感情抑圧表現アニメーションの生成，第 33 回人工知能学会全国大会論文集，(2019).

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

[5] P. Ekman, V. W. Friesen, and J. C. Hager: Facial Action Cording System Investigator’s Guide, Network Information Research Corp. (2002).

[6] Live2D Cubism，株式会社 Live2D，https://www.live2d.com/，（2020 年 7 月 21 日アクセス）．

[7] S. R. Livingstore and F. 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, PloS one, Vol. 13, No. 5, p. e0196391 (2018).