第 23 回日本バーチャルリアリティ学会大会論文集（2018 年 9 月）

深層学習を用いた入力音声に適した顔表情生成

西村亮佑 [1)]，酒田信親 [2)]，冨永登夢 [1)]，土方嘉徳 [3)]，原田研介 [1)]，清川清 [2)]

1) 大阪大学大学院 基礎工学研究科（〒 560-0043 大阪府豊中市待兼山町 1-3）
2) 奈良先端科学技術大学院大学 先端科学技術研究科（〒 630-0101 奈良県生駒市高山町 8916-5）
3) 関西学院大学 商学部（〒 662-8501 兵庫県西宮市上ケ原一番町 1-155）

概要: 近年，VR で利用する高品質な 3DCG キャラクタに対する需要が高まっている．CG キャラクタの表情アニメーションは，CG コンテンツのリアリティを高めるうえで重要な要素である．表情アニメーションの生成手法は，カメラを用いた手法が一般的であるが，カメラの画角による撮影範囲の制限やカメラの光軸と顔の向きの関係によっては人間の表情を取得できないなど，カメラの性能や設置場所に依存する問題がある．一方で，音声を用いた手法は，これらの問題を解決する可能性がある．本稿では，音声データのみを入力とし，RNN を用いて音声に適した表情を出力するシステムを提案する．そして，このシステムを評価した結果，感情により声の周波数が変わるという知見から，あらかじめデータを前処理することで，先行研究に比べて出力の精度が向上することを確認した．

キーワード： 表情アニメーション，音声駆動，RNN，コミュニケーション

1. はじめに

近年，VR で利用する高品質な 3DCG キャラクタに対する需要が高まっている．CG キャラクタの表情アニメーションは，コンテンツのリアリティを高めるうえで不可欠である．現在，表情アニメーション生成の手法は，カメラを用いた手法が一般的であるが，カメラの画角による撮影範囲の制限や，カメラの光軸と顔の向きの関係によって顔認識ができないなど，カメラの性能や設置場所に依存する問題がある．

カメラを用いる手法に対し，音声を用いた手法は，これらの問題が生じないために，移動時などでもマイク一つで表情アニメーションが生成できる可能性がある．音声を用いる手法には，遮蔽物を回り込んでくる音の特性から音素の解析により口の形を再現するもの [1] や，深層学習により顔全体の表情を再現するもの [2] がある．これらは，アニメーション駆動が顔の一部だけであったり，音声がスピーチに限られており，コミュニケーションは考慮していない．

著者の目標は，音声のみの入力に対して，もっともらしい表情を生成するシステムを構築することである．この生成したもっともらしい表情が自然に見えるためには，口の動きだけでなく，頬の起伏や眉の動きも考慮する必要がある．そこで，リカレントニューラルネットワーク（RNN）を用いて表情を推測する手法 [3] をベースとし，コミュニケーションに用いることが可能なシステムの構築を行い，システムの評価を行う．

Ryosuke NISHIMURA, Nobuchika SAKATA, Tomu TOMINAGA, Yoshinori HIJIKATA, Kensuke HARADA, and Kiyoshi KIYOKAWA

2. 関連研究

本章では，画像データや音声データからユーザの表情を推測し，3D モデルに適用する研究を紹介する．

3D モデルに表情を適用させる際によく利用される方法は，ブレンドシェイプに基づく方法である [4]．ブレンドシェイプはあらかじめ複数の顔形状を用意しておき，それぞれの顔形状をブレンドすることで表情を生成する手法である．顔の動きを包括的に測定するための手法として，Facial Action Coding System（FACS）[5] がある．FACS に基づいて，あらかじめ用意する顔形状を決定することが多い．

ブレンドシェイプの各顔形状のブレンド具合の重みパラメータを求めるために，カメラを用いて顔特徴データを取得する．顔の 2D 特徴点の分布を表した画像データベースを用いて，顔画像のみから 3D 顔形状を推測する手法が存在する [6]．この手法はカメラのみで実行できる点が優れているが，カメラの性能や設置場所に依存する問題がある．

音声データのみを用いた表情アニメーションの生成には，音響特性を用いた手法 [1][7] と深層学習を用いた手法 [2][3] がある．音響特性を用いた手法は，音の発生源である口周りの表情の動きを推定するには十分であるが，目や眉の動きなど，動きが音に影響しない部分の表情は推定できない．それに対し，深層学習を用いた手法では，音声と表情を対応付けることができる．

3. 提案システム

本稿では，音声データのみを入力とし，RNN を用いて音声に適した表情を出力するシステムを提案する．学習の入力は音声解析データ，出力はブレンドシェイプの表情重みパラメータとする．システムの全体像を図 1 に示す．

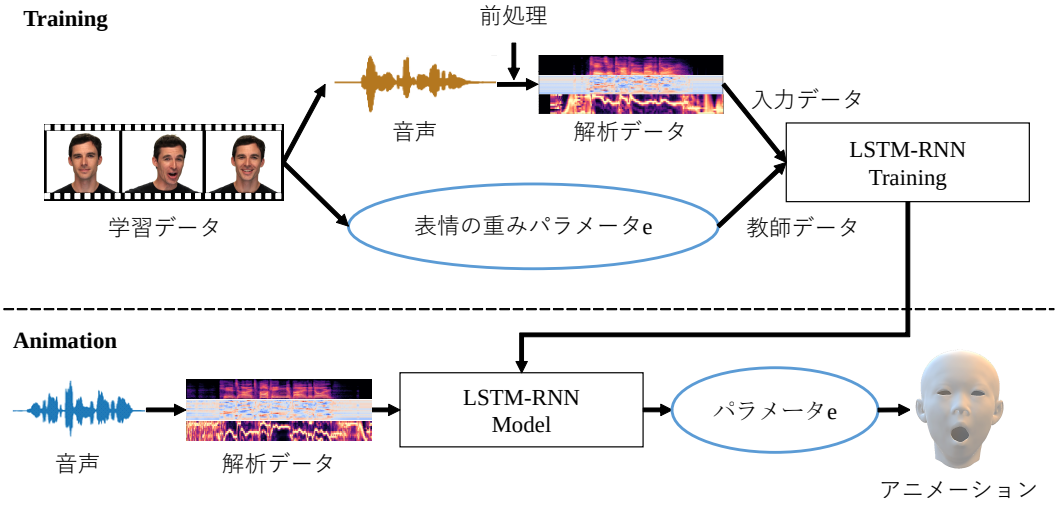

図 1: 提案システムの全体像

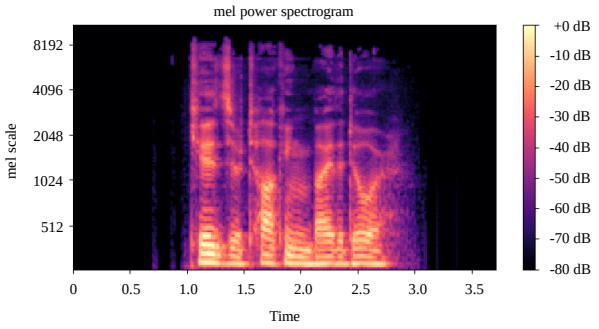

図 2: メルスケーリングされたスペクトログラム

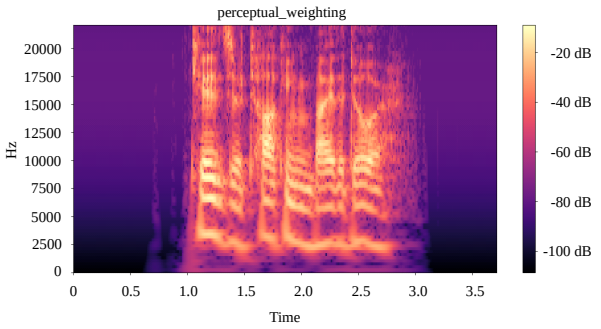

図 3: A 特性手法により重み付けされたスペクトログラム

3.1 音声データ

機械学習に用いる音声データの解析方法について説明する．Pham ら [3] は音声データをメルスケーリングされたスペクトログラム，メル周波数ケプストラム係数，クロマグラムの 3 つに解析した．さらに，ケプストラム係数の一回微分，及び二回微分も入力データとして用いた．

本研究では，音声データに対して A 特性手法を用いた知覚的重み付けの処理が機械学習に対して良好な結果をもたらすと考えている．A 特性手法とは人間の聴覚の特性を考慮に入れて，周波数に知覚的重み付けを行うものである．感情により声の周波数が変わるという知見 [8] から，機械学習への入力である音声データに対し，A 特性手法を用いた前処理を施すと機械学習の結果に影響を与えると考えた．メルスケーリングされたスペクトログラムと A 特性手法により重み付けされたスペクトログラムの出力例を，それぞれ図 2 と図 3 に示す．図 2 と比較すると，A 特性手法を用いたことで低音域と高音域が抑えられ，声の特徴がより強調されているのがわかる．

3.2 顔特徴データ

任意の人物の 2D 顔画像から 3D 顔形状を再構築し，顔画像の特徴点に最も一致した 3D 顔形状となるようなブレンドシェイプの表情重みパラメータを求める．本研究では，テンソルを用いた多重線形顔モデル [9] を参考にする．3-mode テンソルを，Morphace[10] を用いて構築する．Morphace は 150 人の表情を主成分分析したデータであり，パラメータを変えることで様々な 3D 顔形状を生成できる．このデータには真顔しか含まれていないため，メッシュの変換 [11] を適応して 47 個の表情を生成することで，テンソルを構築する．47 個の表情は FACS を参考に，手作業で生成する．

任意の人物の表情 $V(B, \boldsymbol{e})$ は以下のように表すことができる．

$$V(B, \boldsymbol{e}) = B_0 + \sum_{j=1}^{E} (B_j - B_0) \cdot e_j \qquad (1)$$

ここで，B_0 は真顔，B_j はブレンドシェイプに用いる顔形状，$e_j \in [0, 1]$ は各顔形状に対する重みである．頭部の回転 R や平行移動 T も加えた顔 S は以下のように表すことができる．

$$S = R \cdot V(B, \boldsymbol{e}) + T \qquad (2)$$

3.3 表情データ

式 (1) に示される各表情の重みパラメータ e_j が求めたいブレンドシェイプの表情重みパラメータである．手法は，[12] を参考にする．データには RAVDESS を用いる [13].

表 1: 既存手法 単位:[cm]

感情		ニュートラル	穏やか	幸せ	悲しみ	怒り	恐れ	嫌悪	驚き
三次元誤差	平均	0.817	0.665	0.753	0.740	0.947	0.727	0.830	0.836
	分散	2.232	1.712	2.648	2.223	3.160	2.530	2.686	2.926
二次元誤差	平均	1.570	1.248	1.347	1.315	1.798	1.337	1.426	1.625
	分散	3.246	2.566	3.936	3.250	5.152	4.035	3.622	4.798

表 2: 簡略手法 単位:[cm]

感情		ニュートラル	穏やか	幸せ	悲しみ	怒り	恐れ	嫌悪	驚き
三次元誤差	平均	0.751	0.711	0.846	0.809	0.989	0.815	0.902	0.868
	分散	1.949	1.638	2.769	2.233	3.034	2.588	2.706	2.946
二次元誤差	平均	1.439	1.368	1.561	1.502	1.898	1.545	1.601	1.690
	分散	2.918	2.320	3.864	3.125	4.688	3.863	3.384	4.593

表 3: A 特性手法 単位:[cm]

感情		ニュートラル	穏やか	幸せ	悲しみ	怒り	恐れ	嫌悪	驚き
三次元誤差	平均	0.778	0.641	0.721	0.717	0.875	0.708	0.794	0.795
	分散	2.040	1.488	2.184	1.813	2.478	2.159	2.163	2.535
二次元誤差	平均	1.480	1.205	1.331	1.300	1.694	1.330	1.400	1.570
	分散	2.919	2.112	3.159	2.543	3.992	3.259	2.769	4.119

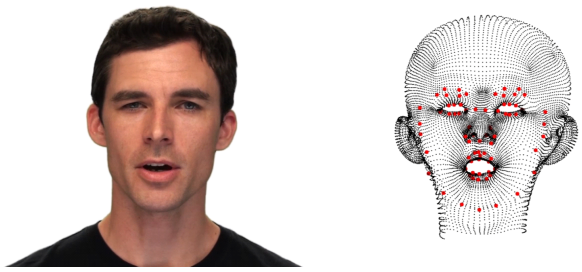

図 4: 2D 顔画像に一致させる最適化の結果例．左：2D 顔画像．右：得られた 3D 顔形状（赤点は特徴点）

このデータは，24 人（男女 12 人ずつ）が 8 つの感情で話す動画で構成されている．この動画の顔データから，顔の特徴点を 74 個抽出し [14]，特徴点に最も一致するような最適化の計算により，表情重みパラメータ $\boldsymbol{e}$ を求める．得られた最適化の結果例を図 4 に示す．

3.4 機械学習の構成

Pham ら [3] の手法に従って，機械学習を用いる．入力データは音声データ，教師データは表情の重み $\boldsymbol{e}$ と頭部の回転 R とする．学習モデルは，音声が時系列を持った信号であることから，LSTM-RNN を用いる．

本研究の LSTM-RNN のフレームワークでは，音声データ x_t を入力とし，表情の重みと頭部回転 y_t を出力とする．ここで，$t = 1, \cdots, T$ であり，T はビデオフレームの数である．任意の時間 t において，学習モデルは入力特徴ベクトル x_t から $y_t = (R_t, \boldsymbol{e}_t)$ を推定する．表情の重み $\boldsymbol{e}$ は $[0, 1]$ 内に制限されているため，R と $\boldsymbol{e}$ は 2 つのレイヤーに分割して出力する．回転の出力 y_R は単なる線形層，表情の重みの出力 y_e は ReLU の活性化を用いて出力が非負となるようにする．学習モデルは二乗誤差が最小になるように訓練する．

4. 実験

入力音声として，以下の 3 パターンを用いる．

- Pham ら [3] の既存手法：3 種類の解析データ
- 簡略手法：スペクトログラムのみ
- 提案手法：知覚的重み付けのされたスペクトログラム

学習モデルから得られる出力は表情パラメータである．評価として，顔の頂点データの三次元誤差と，特徴点の二次元誤差を用いる．それぞれの誤差は，ユークリッド距離により算出する．学習の構造は，Pham ら [3] が最も良い結果を出した構造を採用する．

5. 結果と考察

3 パターンの入力音声に対する出力結果を以下の表 1〜3 に示す．三次元誤差は 1 つの頂点，二次元誤差は 1 つの特徴点に対する誤差の平均である．

感情が含まれる音声に対しては，三次元誤差と二次元誤差のどちらの場合も，A 特性手法が既存手法よりも誤差が小さいことがわかる．これは，A 特性による知覚的重み付けにより，感情に含まれる周波数の特徴が強調されたため，機械学習でより良いモデルが得られたからだと考えられる．

ニュートラルの音声に対しては，スペクトログラムだけを用いたシンプルな簡略手法が最も誤差が小さくなった．こ

れは，ニュートラルの音声では，周波数の特徴が多く含まれていないからだと考えられる．

各感情を比較すると，どの手法においても怒りに対する誤差が最も大きくなっている．怒りの表情には，眉間にしわが寄ることや，目を見開くこと，声を強く発することで喉の筋が浮かび上がるなどの特徴が出る．本研究の表情推定の手法は 74 個の特徴点にしか注目していないため，これらの表情の特徴をとらえきれているとは言えない．特徴点に加え，影などの情報をデータに含めることで，怒りに対する誤差が小さくなる可能性がある．また，怒った音声でも表情はあまり変化していないデータも含まれていたことも，誤差が大きくなった原因だと考えられる．表情の推定方法を改良することや，データセットを選別することで，より良い結果が得られると期待される．

6. おわりに

本稿では，音声の入力により，表情アニメーションを生成するシステムの提案を行った．リカレントニューラルネットワーク（RNN）を用いて表情を推測する手法 [3] をベースとし，音声データに A 特性手法による知覚的重み付けを施すことで，表情推定精度が向上することを確認した．現状の問題点として，深層学習の出力による表情が滑らかでない点，学習データの質が悪い点，精度の検証方法があいまいな点が挙げられる．これらの問題を解決するために，システムを再度構築し，被験者実験によるシステムの評価を今後の研究方針とする．

謝辞 本研究は JSPS 科研費 JP18H04116，JP18H03273 の助成を受けたものです．

参考文献

[1] Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., Hodgins, J. and Matthews, I.: A deep learning approach for generalized speech animation, *ACM Transactions on Graphics (TOG)*, Vol. 36, No. 4, p. 93 (2017).

[2] Karras, T., Aila, T., Laine, S., Herva, A. and Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion, *ACM Transactions on Graphics (TOG)*, Vol. 36, No. 4, p. 94 (2017).

[3] Pham, H. X., Cheung, S. and Pavlovic, V.: Speech-driven 3d facial animation with implicit emotional awareness: a deep learning approach, *The 1st DALCOM workshop, CVPR* (2017).

[4] Lewis, J. P., Anjyo, K., Rhee, T., Zhang, M., Pighin, F. H. and Deng, Z.: Practice and Theory of Blendshape Facial Models., *Eurographics (State of the Art Reports)*, Vol. 1, No. 8 (2014).

[5] Ekman, P. and Friesen, W.: Facial action coding system: a technique for the measurement of facial movement, *Palo Alto: Consulting Psychologists* (1978).

[6] Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M. and Theobalt, C.: State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications, *Computer Graphics Forum*, Vol. 37, No. 2, Wiley Online Library, pp. 523–550 (2018).

[7] Edwards, P., Landreth, C., Fiume, E. and Singh, K.: JALI: an animator-centric viseme model for expressive lip synchronization, *ACM Transactions on Graphics (TOG)*, Vol. 35, No. 4, p. 127 (2016).

[8] Murray, I. R. and Arnott, J. L.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *The Journal of the Acoustical Society of America*, Vol. 93, No. 2, pp. 1097–1108 (1993).

[9] Cao, X., Wei, Y., Wen, F. and Sun, J.: Face alignment by explicit shape regression, *International Journal of Computer Vision*, Vol. 107, No. 2, pp. 177–190 (2014).

[10] Paysan, P., Knothe, R., Amberg, B., Romdhani, S. and Vetter, T.: A 3D face model for pose and illumination invariant face recognition, *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, Ieee, pp. 296–301 (2009).

[11] Pawaskar, C., Ma, W.-C., Carnegie, K., Lewis, J. P. and Rhee, T.: Expression transfer: A system to build 3D blend shapes for facial animation, *Image and Vision Computing New Zealand (IVCNZ), 2013 28th International Conference of*, IEEE, pp. 154–159 (2013).

[12] Pham, H. X., Pavlovic, V., Cai, J. and Cham, T.-j.: Robust real-time performance-driven 3D face tracking, *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 1851–1856 (2016).

[13] 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).

[14] Ren, S., Cao, X., Wei, Y. and Sun, J.: Face alignment at 3000 fps via regressing local binary features, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692 (2014).