H-027

表情特徴を用いた読唇
Lip Reading using Facial Expression Features

白方 達也† 齊藤 剛史†
Tatsuya Shirakata Takeshi Saitoh

1. はじめに

私たちの生活において発話を用いたコミュニケーションは重要であるが，発話障害者や聴覚障害者にとっては困難である．また，近年普及が進んでいる音声認識は高騒音環境下や公共の場所においての使用は困難である．そこで，音声データを用いず，視覚データのみから音声内容を推定する読唇技術は，これらの問題を解決する次世代インターフェースとして期待されている．従来の読唇手法は，口唇周辺領域より特徴を抽出している．しかし，発話中は唇のみが動くのではなく，唇周辺の皮膚も動くことは明らかである．通常の読唇手法は無表情の発話シーンを利用するが，実際の会話においては無表情ではなく，発話内容に応じて表情が変化する．そこで，本稿では，口唇周辺のみに限定せず，顔全体より抽出できる表情特徴を導入する読唇手法を提案する．OuluVS，CUAVE，および CENSREC-1-AV の三つの公開データベースを用いて評価実験を行い，提案手法の有効性を検証する．

2. 関連研究

これまで読唇分野では，様々な手法が提案されている．特徴抽出においては（1）画像ベース，（2）モーションベース，（3）幾何特徴ベース，（4）モデルベースの大きく四つに分けられる[1].

従来は隠れマルコフモデル（Hidden Markov Model; HMM）などの機械学習を用いて認識していた[2]．近年は，深層学習を用いた研究もされるようになっており，精度も向上している．深層学習では与えられたデータから人間の手では抽出が難しい情報でも非常に細かい部分の情報が抽出できる．主な深層学習のモデルとして，畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）挙げられる．Saitoh らは発話シーンのフレーム画像を連結した画像フレーム連結画像（Concatenated Frame Image）を入力し，時系列情報を考慮した手法を提案し，公開データベース OuluVS2 を用いて評価した[3]．Chung と Zisserman はいくつかの CNN の構造を提案し，それらのモデルが高精度であることを示した．また，The Oxford-BBC Lip Reading in the Wild（LRW）[4]のデータセットを構築した．

3. 提案手法

図 1 に提案手法の概要を示す．本稿の手法は既存の手法に基づいている．これは，画像ベースとモーションベースの特徴を統合する手法であり，この構造により，他の特徴量を容易に統合できる．このことから，本稿ではこの手法をベースラインとした．

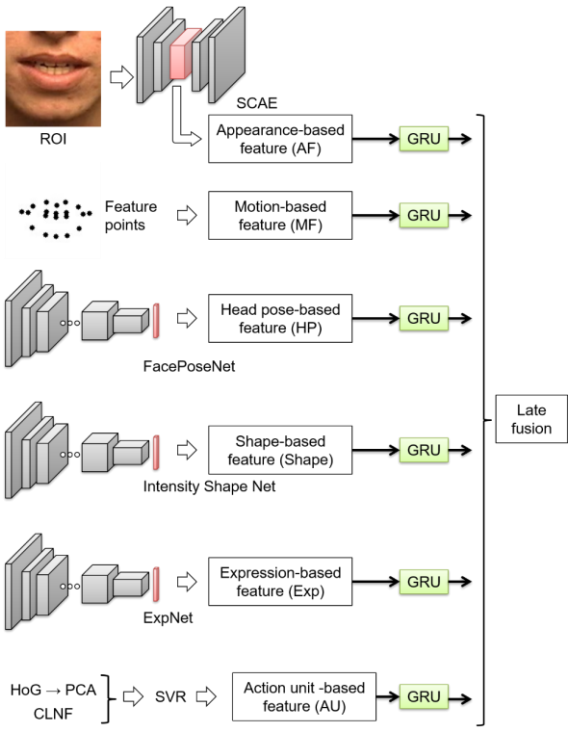

図 1 提案手法の概要

3.1 正規化処理

認識に用いる発話シーンは全て同じ環境で撮影されたものではないため，発話シーンによって顔の位置や大きさが異なる．この違いを解消するために，正規化処理を行い，全ての発話シーンを同じ条件に統一する．まず，入力顔画像に顔検出器が適応され，図 2 の赤線の長方形で示すように顔が検出され，顔 ROI を抽出する．次に，顔の特徴点検出を行い，68 点の顔特徴点（$P_i, i = 0,1,\cdots,67$）を検出する[5]．ここで求めた特徴点をもとに$d_{eye} = |P_{36} - P_{45}|$と$\theta = \angle P_{36}P_{45}$の二つを計算する．$P_{36}$と$P_{45}$はそれぞれ左目尻と右目尻の特徴点座標，$d_{eye}$は二つの目の間の距離，$\theta$は二つの目の間の角度を表す．$d_{eye}$が200[pixel]になるようにスケール変換を行い，θが0°となるように回転処理を適用する．

† 九州工業大学 Kyushu Institute of Technology

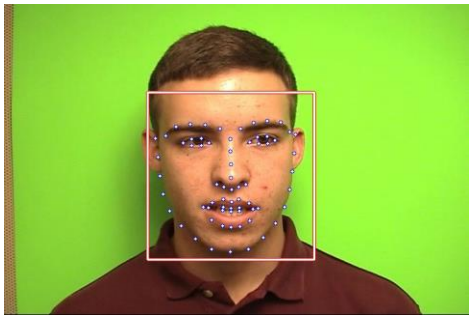

図 2 顔検出と特徴点検出の結果

3.2 特徴量

ここでは本稿で用いた六つの特徴量を説明する．

3.2.1 Appearance-based Feature (AF)

一般に画像ベースの特徴は口唇領域の ROI を抽出し，計算される．このことにより，唇の内側の歯や舌の情報も含めることができ，これらは読唇において有用な情報である．

深層学習が普及する以前，多くの画像ベースの特徴はグレースケール画像を使用し，特徴ベクトルとして直接使用されるか，主成分分析（PCA）や離散コサイン変換（DCT）などの画像処理を適用し使用されていた[1]．深層学習が普及すると，auto-encoder を使用して，画像ベースの特徴が計算されるようになった[6]．auto-encoder は，逆伝搬を適用する教師なし学習アルゴリズムであり，ターゲット値$\hat{x}$が入力値xと等しくなるように設定される．このネットワークはデータに固有であり，データから自動的に学習される．

Iwasaki らは Convolutional Layer，Pooling Layer，Unpooling Layer の三つの層で構成される stacked convolution auto-encoder（SCAE）[7]を用いた[6]．[6]では Convolutional Layer が 7 層の構造であるモデルを用いており，入力画像のサイズは64×64[pixels]である．本稿では 4 層目の Convolutional Layer の出力である 256 次元の値を AF として用いる．

3.2.2 Motion-based Feature (MF)

AF は性別，肌の色や照明条件などの色の違いの影響を受ける．よって，いくつかのモーションベースの特徴が提案されている．Shiraishi と Saitoh はオプティカルフローを用いたモーションベースの特徴を提案した[8]．ただし，オプティカルフローは読唇の動きに大きく関係しない頭部の動きを感知するため，認識に影響を及ぼす．そこで Iwasaki ら[6]は，顔の特徴点に基づくモーションベースの特徴を提案した．

各特徴点の現在のフレームと次のフレームの間の差分値は$d_*(i,f) = P_*(i,f) - P_*(i,f+1)$で定義される．$i$は特徴点の番号，$f$はフレーム番号，$P_*(i,f)$は$f$フレームの$i$番目の特徴点の座標である．また，$*$は$x$または$y$である．この特徴は距離値ではないため，正または負の値がある．本稿では唇の 20 個の特徴点のみの 40 次元の値を MF として用いる．

3.2.3 Head pose-based Feature (HP)

頭の姿勢は唇の動きとは無関係に変化するが，多くの人は頭を動かしながら話す．そこで，会話の内容や相手によって頭の動きが変化することが想定される．従来の読唇手法では，頭の姿勢は考慮されていない．本稿では，頭の姿勢を特徴として定義し，その有効性を検証する．

現在，様々な頭部姿勢推定の手法が提案されている．Chang らは画像強度から直接 CNN ベースの 6 自由度（6DoF）の頭部姿勢推定の手法である FacePoseNet（FPN）を提案した[9]．FPN は AlexNet[10]の構造を用いて，初期化した重みは[11]によって提供される．彼らの手法は，6DoF 3D の頭部姿勢$\mathrm{h} = (r_x, r_y, r_z, t_x, t_y, t_z)$を定義する．$(r_x, r_y, r_z)$はオイラー角（pitch, yaw, roll），$(t_x, t_y, t_z)^T$は 3D 頭部姿勢である．本稿では ExpNet によって実装された FPN を用いて，HP として 6 次元の値を用いる．

3.2.4 Shape-based Feature (Shape)

Tran らは CNN ベースの 3D 顔変形可能モデル（3DMM）を提案した[12]．これは，101 層の ResNet[13]の構造を用いており，入力画像から直接 3DMM の形状とテクスチャを回帰する．

3DMM は AF のように 2 次元ではなく，3 次元に基づく特徴である．本稿では Tran らのモデルを用いて，Shape として 99 次元の値を用いる．

3.2.5 Expression-based Feature (Exp)

一般に表情推定は分類問題として扱われ，画像または動画を怒り，恐怖，幸せ，悲しみなどの顔の表情クラス[14]や Action Units（AU）[15]に分類する．一方，Chang らは回帰問題として研究した[16]．

[16]では ResNet101 の構造を用いる ExpNet が提案された．ExpNet は顔画像の強度に直接適用され，3D 表現係数の 29D ベクトルを回帰し，顔の特徴点検出を必要とせず，3D 表現係数を直接推定する手法である．本稿では Exp として 29 次元の値を用いる．

3.2.6 Action unit-based Feature (AU)

Facial Action Coding System （FACS）は顔の外観によって人の顔の動きを分類するシステムである．FACS を用いると，ほぼ全ての表情をコード化し，表情を生み出した特定の AU に分解できる．

Baltrusaitis らは深層学習ベースの hand-craft ベースのアプローチではない顔の AU のリアルタイムの特徴量推定手法を提案した[15]．顔のアライメントがなされた後，二つの特徴が抽出される．一つ目は，主成分分析を Histograms of Oriented Gradients（HOG）に適用することで得られる画像ベースの特徴である．二つ目は，非剛体形状パラメータと Constrained Local Neural Field（CLNF）のパラメータによって得られる幾何特徴である．AU の推定には Support Vector Regression（SVR)が用いられる．Baltrusaitis らの手法では 17AU が推定され，本稿では AU として 17 次元の値を用いる．

3.3 認識手法

Recurrent Neural Networks（RNN）は，時系列データに対するニューラルネットワークであり，内部に閉路の構造を持つ．この構造により，前時刻の情報を記憶し，動的なふるまいをすることができる．このため，多層型パーセプトロンや CNN などの伝搬型ネットワークとの違いを生んでいる．しかし，ユニットの計算や誤差逆伝搬などは，他のネットワークと同じような構造を持つ．本稿では認識手法として RNN の一つである Gated Recurrent Unit（GRU）[17]を用いる．また，特徴量の統合には late-fusion を用いる．

4. 評価実験

4.1 データベース

ここでは本稿で用いたデータベースについて説明する．

4.1.1 CUAVE[18]

CUAVE（The Clemson University Audio-Visual Experiments）は Patterson らによって公開された無償のデータベースである．収録内容は，10 数字（“zero”，“one”，“two”，“three”，“four”，“five”，“six”，“seven”，“eight”，“nine”）を 1 話者につき 5 回発話したものとなっている．話者数は 36 名（19M＋17F），画像サイズは720 × 480[pixels]，フレームレートは 29.97fps である．サンプル画像を図 3 に示す．

4.1.2 OuluVS[19]

OuluVS は Zhao らによって公開された無償のデータベースである．英語 10 文（“Hello”，“Excuse me”，“I am sorry”，“Thank you”，“Goodbye”，“See you”，“Nice to meet you”，“You are welcome”，“How are you”，“Have a good time”）である．話者数は 20 名（17M＋3F），画像サイズは720 × 576[pixels]，フレームレートは 25fps である．サンプル画像を図 4 に示す．

4.1.3 CENSREC-1-AV[20]

CENSREC-1-AV は大西らによって公開された日本語のデータベースである．収録内容は，連続数字 1～7 桁を発話したシーンである．話者数は 93 名（47M＋46F），画像サイズは81 × 55 [pixels]，フレームレートは 29.97fps である．本稿では非公開の顔全体画像（720 × 480 [pixels]）を用いて，さらに 1 桁だけ発話されたもののみを用いる．この場合，3,234 個の学習データ，1,963 個のテストデータがある．

図 3 CUAVE（S01）のサンプル画像

図 4 OuluVS（P002）のサンプル画像

4.2 実験条件

本実験では CUAVE，OuluVS，CENSREC-1-AV の三つのデータベースを用いて，提案手法の認識精度を評価した．

CUAVE と OuluVS は leave-one-person-out 法で，CENSREC-1-AV は学習データとテストデータが定義されているため，hold-out 法で評価した．顔検出と特徴点検出には dlib[1]を用いた．三つの特徴量 HP，Shape，Exp は ExpNet[2]を適用し，抽出した．AU は OpenFace2.03[3][21]を用いた．深層学習フレームワーク TensorFlow の Keras を用いて，GRU モデルで学習とテストを行い，六つの特徴量を組み合わせて 11 の認識実験を行った．

4.3 実験結果

本実験の結果を表 1 に示す．この表の 2～7 列は各実験で用いられる特徴量であり，最後の 3 列は各データベースの平均認識率を示す．特徴量の括弧内の数字は次元数を示す．上の 6 条件は各特徴量を単独で用いる場合である．HP と Shape は唇の動きが特徴量の値に含まれていないため，認識精度は低くなっている．MF と AF を用いる従来手法である条件 7 は認識精度が高い．しかし，提案手法である条件 10 と 11 は従来手法よりも高い認識精度を得た．MF は唇周りの特徴点の動きを取得し，AF は唇周りとその周辺の皮膚を取得する．一方で，顔の表情特徴は，顔のパーツと肌の動きが暗黙的に含まれることから，MF と AF を補完できたことにより，認識精度の向上につながったと考えられる．

三つのデータベースを比較すると，OuluVS が他のデータベースより高い認識精度となった．全てのデータベースには 10 個のクラスがある．また，OuluVS は日常会話の短いフレーズの発話シーンであり，CUAVE と CENSREC-1-AV は数字の発話シーンである．数字の発話シーンの音の数は少なく，口の動きは単調であるが，フレーズの発話シーンは数字の発話シーンよりも複雑である．CUAVE と CENSREC-1-AV のフレームレートは同じだが，OuluVS のフレームレートは異なっている．しかし，全てのデータベースで発話の長さと言語が異なっているため，フレームレートの影響はないものと考える．

5. おわりに

従来手法では表情特徴を用いてなかったが，本稿では表情特徴を用いた新たな手法を提案した．三つの公開データベースで評価実験を行い，提案手法の有効性を検証した．その結果，全てのデータベースで認識精度が向上することを確認できた．また，HP と Shape の特徴量は読唇において有効でないことも確認できた．

提案手法により認識精度は向上したが，まだ向上の余地は残っている．また，無表情のデータベースに対してではなく，表情がある実際の発話シーンに対しても本手法が有効であるかを検証する必要がある．本稿では，三つのデータベースを用いたが，言語とフレームレートの違いについては十分に言及できていない．これらは今後の課題でもある．

[1] http://dlib.net/
[2] https://github.com/fengju514/Expression-Net
[3] https://github.com/TadasBaltrusaitis/OpenFace

表 1 実験結果

条件	特徴量						認識率		
	MF（40）	AF（256）	HP（6）	Shape（99）	Exp（29）	AU（17）	OuluVS[%]	CUAVE[%]	CENSREC-1-AV[%]
1[14]	○	-	-	-	-	-	73.2	76.4	54.3
2[14]	-	○	-	-	-	-	79.1	74.9	72.5
3	-	-	○	-	-	-	13.4	19.9	10.0
4	-	-	-	○	-	-	13.7	27.3	13.1
5	-	-	-	-	○	-	67.4	67.4	23.7
6	-	-	-	-	-	○	69.8	71.2	59.2
7[14]	○	○	-	-	-	-	83.1	79.9	74.3
8	○	○	○	-	-	-	79.4	75.5	24.0
9	○	○	-	○	-	-	79.3	75.1	74.5
10	○	○	-	-	○	-	85.6	83.1	74.5
11	○	○	-	-	-	○	**86.6**	**83.4**	**77.1**

謝辞

本研究は，JSPS 科研費 16H03211，17H01840，19KT0029 の助成によるものである．

参考文献

[1] Z. Zhou, G. Zhao, X. Hong, and M. Pietikainen. A review of recent advance in visual speech decoding. *Image and Vision Computing*, Vol. 32, pp. 590-605, 2014.

[2] T. Saitoh. Efficient face model for lip reading. In *12th International Conference on Auditory-Visual Speech Processing (AVSP2013)*, pp. 227-232, 8 2013.

[3] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikainen. Concatenated frame image based cnn for visual speech recognition. In *ACCV2016. workshop: Multi-view Lip-reading/Audio-visual Challenges (MLAC2016)*, 11 2016.

[4] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision (ACCV)*, 2016.

[5] V. Kazemi and J. Sullivan. One millisecond face alignment with anensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[6] M. Iwasaki, M. Kubokawa, and T. Saitoh. Two features combination with gated recurrent unit for visual speech recognition. In *IAPR Conference on Machine Vision Applications (MVA)*, 2017, pp. 300–303.

[7] J. Masci, U. Meier, D. Cires¸an, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *21th International Conference on Artificial Neural Networks*, 2011, pp. 52–59.

[8] J. Shiraishi and T. Saitoh. Optical flow based lip reading using non rectangular ROI and head motion reduction. In *11th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[9] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Faceposenet: Making a case for landmark-free face alignment. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[11] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4838–4846.

[12] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5163–5172.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] S. Li and W. Deng. Deep facial expression recognition: A survey. *arXiv:1804.08348*, 2018.

[15] T. Baltru š aitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *11th IEEE Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[16] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. ExpNet: Landmark-free, deep, 3D facial expressions. In *13th IEEE Conference on Automatic Face and Gesture Recognition (FG)*, 2018.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[18] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. Movingtalker, speaker-independent feature study, and baseline results using the cuave multimodal speech corpus. *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 1189–1201, 2002.

[19] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[20] S. Tamura, C. Miyajima, N. Kitaoka, T. Yamada, S. Tsuge, T. Takiguchi, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura. CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition. In *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2010.

[21] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *13th IEEE Conference on Automatic Face and Gesture Recognition (FG)*, 2018.