Vol.2017-CVIM-208 No.27
2017/9/16

# 読唇技術における顔画像に対する輝度補正効果の検証

窪川 美智子[1,a)]　齊藤 剛史[1,b)]

**概要**：読唇技術では，発話中の口唇の動きを適切に表現することが重要であり，個人の肌の色の違いなどによる影響を受けない特徴を抽出できることが望まれる．しかし，これまでの読唇技術では，これらの違いが考慮されていない．本論文では，色の違いを軽減するために，顔画像に対する 2 通りの輝度補正手法を提案する．認識処理には既存の再帰型ニューラルネットワークを用いる．OuluVS と CENSREC-1-AV の二つの公開データベースを用いた認識実験を通じて，提案手法の効果を検証する．

**キーワード**：読唇，顔画像，輝度補正，コントラスト変換

# Intensity Correction Effect on Facial Image for Lip Reading

Michiko Kubokawa[1,a)]　Takeshi Saitoh[1,b)]

***Abstract:*** In the lip reading technology, it is important to extract the representation of the mouth motion during the utterance, and it is desirable to be unaffected by the appearance of the speaker, such as his/her skin color. However, these differences were not considered in the traditional approaches. This paper proposes two intensity correction methods: contrast adjustment and histogram equalization, for face image to reduce the variability in appearance. The proposed method is evaluated on two publicly available databases: OuluVS and CENSREC-1-AV, and the experiments show improved recognition accuracy.

***Keywords:*** Lip reading, facial image, intensity correction, contrast adjustment.

## 1. はじめに

視覚情報から発話内容を推定する読唇技術では，発話中の口唇の動きを適切に表現することが重要であり，光源による見えの違いや個人の肌の色などの外観に影響を受けない特徴を抽出できることが望まれる．読唇で用いられている顔画像の多くは，光源環境が整備されたスタジオなどで撮影されているため，光源により見た目が影響を受けることはあまりない．一方，性別あるいは人種により肌の色に違いが生じたり，体調により顔色が変化したりすることがある．従来研究ではこれらの違いが考慮されていない．本論文では，色の違いを軽減するために，顔画像に対する 2 通りの輝度補正手法を提案し，その効果を検証する．

読唇分野では数年前までは，画像ベース，モーションベース，幾何特徴ベース，モデルベースの 4 種類のいずれかのアプローチで特徴量を求め，Hidden Markov Model（HMM）により認識する方法が多く提案されてきた [1]．

一方，近年では，深層学習が音声認識や画像認識など様々な分野で高い精度を達成しており，読唇分野でも利用されている．Noda らは Convolutional Neural Network（CNN）を用いて視覚特徴量を抽出し，HMM を用いて認識する手法を提案している [2]．高島らは視聴覚音声認識を目的とし，Convolutional Bottleneck Network を用いて特徴量を抽出し，HMM により認識する手法を提案している [3]．齊藤らは発話シーンのフレーム画像を連結した画像 Concatenated Frame Image（CFI）を提案し，CFI に対する Data Augmentation 法，および CFI を用いた CNN による読唇手法を提案している [4]．Chung と Zisserman はテレ

[1] 九州工業大学
Kyushu Institute of Technology, Kawazu 680–4, Iizuka, Fukuoka 820–8502, Japan
[a)] michiko.kubokawa@slab.ces.kyutech.ac.jp
[b)] saitoh@ces.kyutech.ac.jp

ビ番組より大規模発話シーンを収集し，数百単語を効果的に学習・認識可能な CNN を開発した [5]．Iwasaki らは，顔特徴点より求まるモーションベース特徴量と Autoencoder より求まる画像ベース特徴量を組み合わせ，Gated Recurrent Unit を用いて認識する手法を提案している [6]．これらの手法は，深層学習を用いることによって高い認識精度を達成している．

個人の肌の色などの見た目の違いは，多人数の発話シーンを用いて学習することで影響を軽減できる．しかし，前処理などにより事前に影響を軽減することで認識精度は向上すると期待される．そこで本論文では，認識処理の前処理として，顔特徴点を用いて顔画像に対する 2 通りの輝度補正手法を提案する．認識処理には既存の再帰型ニューラルネットワークを用いる．二つの公開データベースを用いた認識実験を通じて，提案手法の効果を検証する．

## 2. コントラスト変換

本論文では，輝度補正手法として，画像処理の基礎的な手法であるコントラスト変換を用いる．ここでは線形変換を用いた方法とヒストグラム平坦化の両手法を簡潔に説明する．

### 2.1 線形変換を用いたコントラスト変換

原画像の画素値 $a$ に対して，線形変換を用いて新しい画素値 $a'$ を求める．原画像の最小画素値と最大画素値をそれぞれ $a_{low}$，$a_{high}$，$a_{high} \neq a_{low}$ とする．変換後の画素値の範囲を $[a_{min}, a_{max}]$ とすると，変換関数は次式で表される．

$$a' = a_{min} + (a - a_{low}) \frac{a_{max} - a_{min}}{a_{high} - a_{low}}.$$

上記関数は，画像内のノイズとなる少数の極端な最小画素値あるいは最大画素値の強い影響を受ける．この問題を避けるために，四つの値 $q_{low}$，$q_{high}$，$a'_{low}$，$a'_{high}$ を設ける．$a'_{low}$ は，画像 $I$ 内の全画素値の累積分布が全体の $q_{low}$ となる画素値であり，$a'_{high}$ は，$I$ 内の全画素値の累積分布が全体の $q_{high}$ となる画素値である．ただし，$0 \leq q_{low}$，$q_{high} \leq 1$,$q_{low} + q_{high} \leq 1$ とする．$a'_{low}$ よりも小さい画素値，あるいは $a'_{high}$ よりも大きい画素値はそれぞれ $a_{min}$，$a_{max}$ に割り当てる．その他の中間の画素値は，$[a_{min}, a_{max}]$ 間を線形で割り当てる．以上は次式で表される．

$$a' = \begin{cases} a_{min} & \text{for } a \leq a'_{low} \\ a_{min} + (a - a_{low}) \frac{a_{max} - a_{min}}{a_{high} - a_{low}} & \text{for } a'_{low} < a < a'_{high} \\ a_{max} & \text{for } a \geq a'_{low} \end{cases}$$

### 2.2 ヒストグラム平坦化を用いたコントラスト変換

累積ヒストグラム H を次式で定義する．

$$H(i) = \sum_{j=0}^{i} h(j), \qquad \text{for } 0 \leq i < K,$$

ただし，$h(j)$ は画素値 $j$ における画像内の画素数であり，$K$ は階調数である．このとき，$a'$ は次式で表される．

$$a' = \left\lfloor H(a) \cdot \frac{K-1}{MN} \right\rfloor,$$

$M \times N$ は画素数であり，$\lfloor x \rfloor$ は floor 関数である．

### 2.3 カラー画像に対するコントラスト変換

本論文では濃淡画像のみでなく，カラー画像も対象とする．カラー画像に対するコントラスト変換には，R，G，B の各成分を独立に処理する方法があるが，この場合，色相が変化してしまう．そこで本論文では RGB 色空間から HSV 色空間に変換し，彩度 S と明度 V にコントラスト変換を適用する．

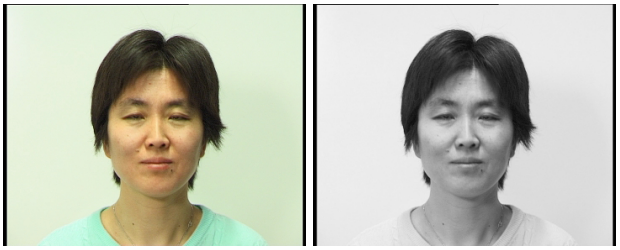
濃淡画像に対しては，濃度値に対してコントラスト変換を適用する．

## 3. 提案手法

### 3.1 顔画像に対するコントラスト変換

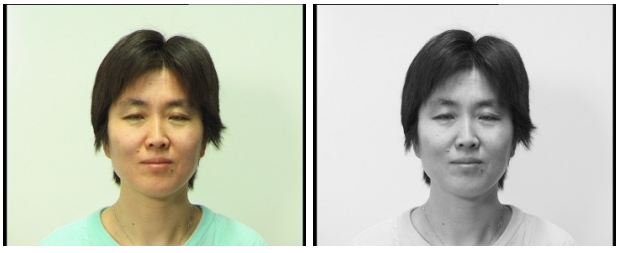
一般的な顔画像では，図 1 に示すように顔領域よりも背景領域の占める割合が多い．このような画像に対して，**2** で述べたコントラスト変換を適用すると，背景領域の影響を強く受けてしまう．図 1 の 2 画像に対して，**2.1** および **2.2** で説明した二通りのコントラスト変換を適用した結果を図 2 に示す．図 2(a)(b) に示す線形変換を用いたコントラスト変換後の画像は，図 1(a)(b) とほぼ同じである．一方，ヒストグラム平坦化を用いたコントラスト変換に関して，入力画像と異なり図 2(c)(d) は顔領域の画素値が一様になっている．このことを確認するため，画像内の全画素より求まるヒストグラム $H_{whole}$ を図 3 に示す．図中，横軸は画素値，縦軸は画素数を対数で示している．青色破線は S，赤色破線は V の $H_{whole}$ である．V，すなわち赤色破線に着目すると，右側（高画素値）の分布が大きい．これは背景領域による分布である．一方，左側（低画素値）は S，V 共に分布が大きい．これは人物の髪領域による分布である．$H_{whole}$ の S は，$a_{low} = 0$，$a_{high} = 255$，$a'_{low} = 1$，$a'_{high} = 254$ であり，$H_{whole}$ の V は，$a_{low} = 0$，$a_{high} = 255$，$a'_{low} = 1$，$a'_{high} = 251$ であった．このため，線形変換を用いたコントラスト変換を適用しても，画像はほとんど変わらない．

前述の問題を解決するため，本論文では顔領域内の画素値より求まるヒストグラム $H_{face}$ を用いたコントラスト変換法を提案する．まず，原画像から顔特徴点を検出する．検出された特徴点をもとに図 4 に示すような顔領域を得る．この顔領域より $H_{face}$ を求め，$a'_{low}$ と $a'_{high}$ を計算する．図 1(a) に対して求めた $H_{face}$ を図 3 の実線で示す．青色実線は S，赤色実線は V である．$H_{face}$ と $H_{whole}$ を比較すると，$H_{face}$ は背景領域および髪領域を含めていないため，高画素値と低画素値の分布が小さい．$H_{face}$ の S

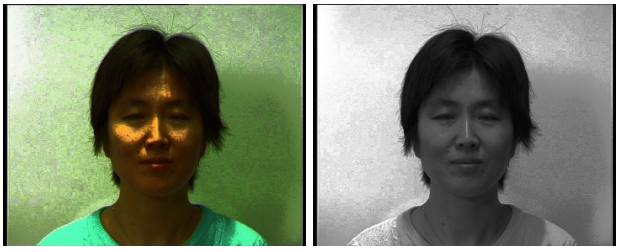


(a) カラー画像　(b) 濃淡画像

**図 1**　入力画像



(a) 線形変換（カラー画像）　(b) 線形変換（濃淡画像）



(c) ヒストグラム平坦化（カラー画像）　(d) ヒストグラム平坦化（濃淡画像）

**図 2**　画像全体に基づくコントラスト変換

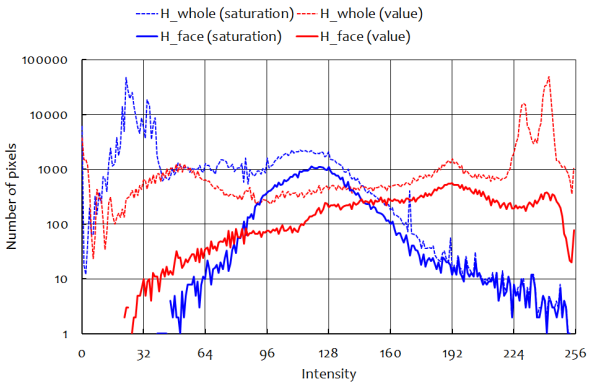



**図 3**　ヒストグラム

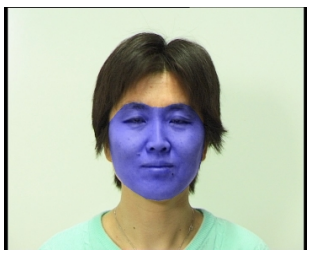
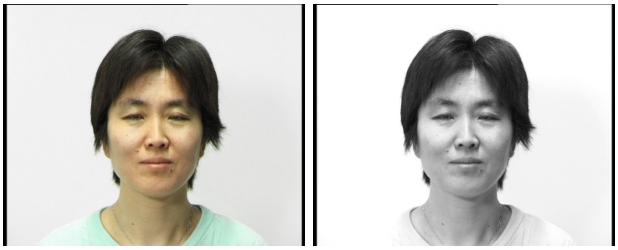
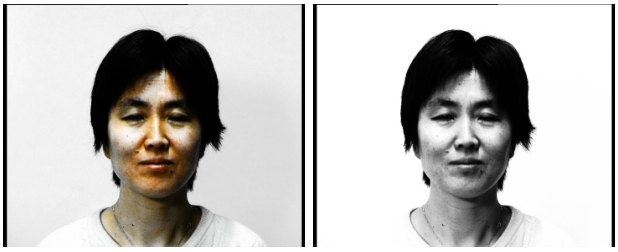


**図 4**　特徴点より得られる顔領域



(a) 線形変換（カラー画像）　(b) 線形変換（濃淡画像）



(c) ヒストグラム平坦化（カラー画像）　(d) ヒストグラム平坦化（濃淡画像）

**図 5**　顔領域に基づくコントラスト変換

は，$a_{low} = 33$，$a_{high} = 255$，$a'_{low} = 72$，$a'_{high} = 220$ であり，$H_{face}$ の V は，$a_{low} = 9$，$a_{high} = 255$，$a'_{low} = 49$，$a'_{high} = 249$ であった．これらの値を利用して，線形変換およびヒストグラム平坦化を用いたコントラスト変換を適用した結果を図 5 に示す．図 2 と比べると，顔領域内で適切にコントラスト変換が適用されていることを確認できる．

### 3.2　発話シーンに対するコントラスト変換

本論文は読唇を対象としているため，入力データは単一フレーム画像でなく，時系列画像である．この場合，全フレーム画像に対してコントラスト変換を適用する必要がある．ここで，処理対象の発話シーンは固定環境下で撮影されている．そのため，コントラスト変換に必要な $a'_{low}$，$a'_{high}$ やヒストグラムを，全てのフレーム画像で求めずに初期フレーム画像のみで求める．初期フレーム以外のフレーム画像では，初期フレーム画像で求めたパラメータを利用してコントラスト変換を適用する．

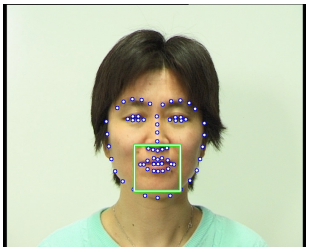
### 3.3　口唇 ROI の抽出

読唇分野では，顔画像全体でなく口唇周辺の ROI を利用するアプローチが多い．本論文でも口唇 ROI を抽出する手段を採る．

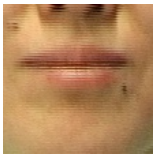
口唇 ROI は顔特徴点に基づいて抽出する．左右の両目間の距離を $d_{eye}$，鼻特徴点座標を $(x_{nose}, y_{nose})$ とすると，口唇 ROI の左上座標を $(x_{nose} - d_{eye} \times S/2, y_{nose} - d_{eye} \times S/8)$ として，口唇 ROI のサイズを $d_{eye}S \times d_{eye}S$ 画素とする．$S = 0.8$ における口唇特徴点を図 6 に示す．

### 3.4　自己符号化器による特徴抽出

自己符号化器（Autoencoder）は入力データと出力デー



(a) 特徴点と口唇 ROI



(b) ROI 画像

**図 6** 特徴点を用いた ROI 抽出

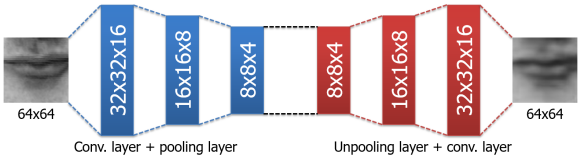



**図 7** Stracture of stacked convolutional autoencoder (SCAE).

タが同じになるようにニューラルネットワークを学習させるものであり，次元削減や特徴表現の目的で利用されている．一般的に隠れ層の次元数は入力層と出力層の次元数よりも小さい．本論文では，Masci らによって提案された，Autoencoder を積み上げて多層化した stacked convolution autoencoder（SCAE）[7] を適用する．SCAE は畳み込み層（CL），プーリング像（PL），アンプーリング層（UPL）から構成されており，畳み込みの特徴である重みを共有するという特徴をもつ．

図 7 に本論文で用いる SACE の構造を示す．入力画像のサイズは 64 × 64 画素であり，符号化プロセスでは三つの CL と PL のペアで構成されている．256 次元のボトルネック層を認識処理の特徴として用いる．

### 3.5 再帰型ニューラルネットワークによる認識

再帰型ニューラルネットワーク（RNN）は，音声や言語，動画像などの系列データを扱うニューラルネットワークである．RNN はネットワークの内部に閉路をもち，この構造により，入力情報を記憶し，振る舞いを動的に変化させることができる．RNN では長期の時系列データに対して誤差逆伝播を適用させた場合，深い構造となり，勾配の消失問題が起こることが知られている．この問題を解決するために Long short-term memory（LSTM）が提案されている．LSTM はユニットの内部にメモリセルをもち，入力や出力やメモリセルなどを制御するためにゲートの構造をもつ．様々 RNN ユニットが提案されており，本研究ではの 3 種類（LSTM original [8]，LSTM forget [9]，LSTM peep [10]）の LSTM と Gated reccurent unit（GRU）[11] を用いる．

## 4. 評価実験

### 4.1 データセット

読唇向けの公開データセットには，AVLetters [12]，CUAVE [13]，Grid [14]，OuluVS [15]，OuluVS2 [16]，CENSREC-1-AV [17] などがある．発話内容や話者数などはデータセットによって異なる．本実験では，OuluVS と CENSREC-1-AV の二つのデータセットを用いて提案手法を評価した．

OuluVS は，2009 年に Zhao らによって公開された無償のデータセットである [15]．発話内容は英語 10 文（“Hello”，“Excuse me”，“I am sorry”，“Thank you”，“Good bye”，“See you”，“Nice to meet you”，“You are welcome”，“How are you”，“Have a good time”）である．画像サイズは 720 × 576 画素，フレームレートは 25fps，話者数は 20 名（男性 17 名，女性 3 名）である．発話シーンは話者の正面から撮影されており，背景はおおよそ白色である．

CENSREC-1-AV は，情報処理学会音声言語情報処理研究会雑音下音声認識評価ワーキンググループによって 2010 年に公開された日本語の発話シーンである [17]．カラー画像と近赤外線画像が用意されており，発話内容は数字（“イチ”，“ニ”，“サン”，“ヨン”，“ゴ”，“ロク”，“ナナ”，“ハチ”，“キュウ”，“ゼロ”，“マル”）を 1〜7 個連続して発話している．画像サイズは 720 × 480 画素，フレームレートは 29.97fps，このデータセットは，学習用として 42 名（男性 22 名，女性 20 名）から収録された 3,234 発話シーン，テスト用として 51 名（男性 25 名，女性 26 名）から収録された 1,963 発話シーンから構成されている．発話シーンは話者の正面から撮影されており，背景はおおよそ青色である．

### 4.2 実験条件

読唇分野では特定話者認識実験（SD）と不特定話者認識実験（SI）がある．SD は学習データとテストデータが同一話者における認識実験であり，SI は学習データの中にテストデータの話者が含まれていない場合における認識実験である．SD より SI の方が難しいタスクであり，本実験では SI に取り組む．OuluVS の話者数は 20 名であり，本実験では leave-one-person-out で評価する．一方，CENSREC-1-AV は学習用データとテスト用データに分けられており，本実験ではそのまま学習用データとテスト用データに用いる．また，OuluVS に収録されている発話シーンは 10 文のいずれかである．一方，CENSREC-1-AV は 1〜7 桁の連続数字発話シーンであり，サンプルによって発話内容が異なる．そこで，本実験では 1 桁の数字発話シーンのみを用いて，11 クラスの分類問題とした．この場合，学習用データとテスト用データはそれぞれ 924，561 で

あった．

本実験では，RNN モデルの学習およびテストには Google が提供する機械学習ライブラリ TensorFlow [*1]を用いた．RNN モデルを学習する際の最適化手法には実験的に良好な結果を得られた RMSProp を用いた．また学習率の初期値を 0.01 とした．

### 4.3 前処理

**3.1** で述べた提案手法を適用するための特徴点検出として，本実験では，機械学習ライブラリ Dlib[*2]に実装されている `get_frontal_face_detector` 関数を用いて 68 点の顔特徴点を検出した．**2.1** で述べた線形変換を用いたコントラスト変換において，ヒストグラムの両端をカットするためのパラメータ $q_{low}$，$q_{high}$ をそれぞれ 0.005，0.995 とした．口唇 ROI の抽出では，経験的に $S = 0.8$ を与えた．Autoencoder の入力画像サイズは $64 \times 64$ 画素とし，ボトルネック層の 256 次元の値を RNN の入力データとして用いた．

### 4.4 実験結果

本実験では，以下に列挙する 6 種の ROI 画像を生成した．

(1) コントラスト変換を適用しないカラー画像（C-NP）．
(2) カラー画像に対して線形変換を用いたコントラスト変換を適用した画像（C-CA）．
(3) カラー画像に対してヒストグラム平坦化を用いたコントラスト変換を適用した画像（C-HE）．
(4) コントラスト変換を適用しない濃淡画像（G-NP）．
(5) 濃淡画像に対して線形変換を用いたコントラスト変換を適用した画像（G-CA）．
(6) 濃淡画像に対してヒストグラム平坦化を用いたコントラスト変換を適用した画像（G-HE）．

二つのデータベースに対して提案手法を用いて認識実験を実施した結果を表 1 に示す．二つのデータベースの認識精度を比較すると，CENSREC-1-AV より OuluVS の方が高い精度を得ている．これは，OuluVS2 の方がデータ数が多く，また発話内容に関しても OuluVS2 の方が複雑でシーン長も長いためと推測する．両データベースで精度の違いは生じているものの，RNN モデルとして GRU，ROI として C-CA を用いた場合にそれぞれのデータベースで最高認識率を得ている．濃淡画像よりカラー画像の方が認識精度が高い傾向にある．ヒストグラム平坦化を用いたコントラスト変換は，コントラスト変換を適用しない場合とほぼ同じ認識精度であった．

C-NP と C-CA における話者毎の認識結果の分布を図 8 に示す．OuluVS では，C-NP よりも C-CA の認識率が高い話者は 20 名中 9 名であったが，CENSREC-1-AV では，51 名中 40 名であった．OuluVS の話者は白人種が多い．一方，CENSREC-1-AV の話者は日本人のみであるが，肌の色が濃い話者が多い．このため，CENSREC-1-AV では提案手法を適用することで肌の色の違いが軽減され認識精度が向上したと推測する．

次に OuluVS に関して，提案手法と他 4 手法を比較する．表 2 に文献に記載されていた各手法の認識精度を示す．提案手法よりも高い精度を得ている手法がある．本論文の提案手法は，認識処理の前処理に位置づけられる．そのため，

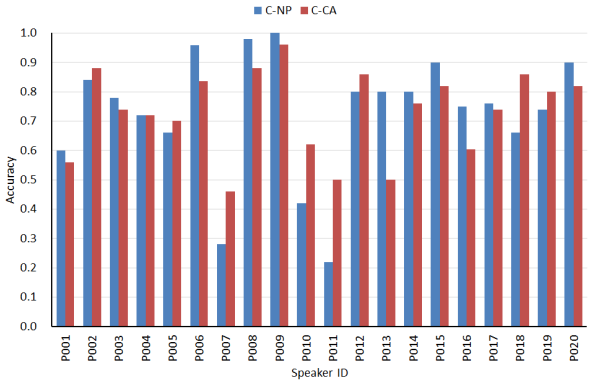
**表 1** 認識結果

(a) OuluVS

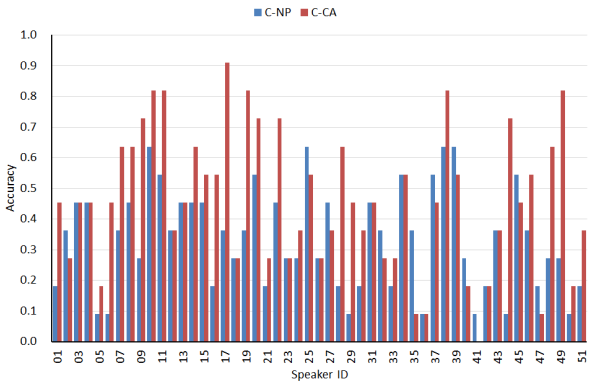
| RNN model | LSTM original | | LSTM forget | | LSTM peep | | GRU |
|---|---|---|---|---|---|---|---|
| # of layer | 1 | 2 | 1 | 2 | 1 | 2 | 1 |
| C-NP | 0.537 | 0.686 | 0.558 | 0.712 | 0.541 | 0.714 | 0.729 |
| C-CA | 0.539 | 0.694 | 0.484 | 0.721 | 0.518 | 0.686 | **0.731** |
| C-HE | 0.521 | 0.667 | 0.552 | 0.638 | 0.554 | 0.626 | 0.689 |
| G-NP | 0.511 | 0.656 | 0.531 | 0.685 | 0.516 | 0.677 | 0.679 |
| G-CA | 0.535 | 0.686 | 0.534 | 0.676 | 0.532 | 0.682 | 0.716 |
| G-HE | 0.499 | 0.597 | 0.461 | 0.659 | 0.507 | 0.610 | 0.706 |

(b) CENSREC-1-AV

| RNN model | LSTM original | LSTM forget | LSTM peep | GRU |
|---|---|---|---|---|
| # of layer | 1 | 1 | 1 | 1 |
| C-NP | 0.257 | 0.267 | 0.266 | 0.335 |
| C-CA | 0.364 | 0.353 | 0.358 | **0.462** |
| C-HE | 0.239 | 0.271 | 0.250 | 0.424 |
| G-NP | 0.216 | 0.196 | 0.201 | 0.433 |
| G-CA | 0.177 | 0.271 | 0.362 | 0.283 |
| G-HE | 0.139 | 0.125 | 0.234 | 0.148 |




(a) OuluVS




(b) CENSREC-1-AV

**図 8** 話者毎の認識率

[*1] https://www.tensorflow.org/
[*2] http://dlib.net/

表 2　他手法との比較結果（OuluVS）

| method | accuracy [%] |
|---|---|
| LBP-TOP/SVM [15] | 62 |
| MKL-fusion/LVM [18] | 81.3 |
| shape+HOG+LBP/RFMA [19] | 89.7 |
| PLSD/KELM [20] | 68.75 |
| **ours (C-CA+autoencoder/GRU)** | 73.1 |

他手法との比較はあまり大きな意味をもたない．他手法に対して本論文で提案するコントラスト変換を適用することで，認識精度の改善が期待できる．

## 5. Conclusion

本論文では，個人の肌の色などの見た目の違いが認識精度に影響をあたえることを軽減するために，顔特徴点を用いた顔画像に対する 2 通りの輝度補正手法を提案した．二つの公開データベース，OuluVS と CENSREC-1-AV を用いて提案手法を評価した．その結果，カラー画像に対して線形変換を用いた方がコントラスト変換を適用することで認識精度が向上した．これより提案手法の有効性を示した．

本論文で得た認識精度は既存手法よりも高い精度を達成していない．しかし，提案手法は認識処理の前処理に位置づけられる．そのため今後の課題としては，既存手法に対して提案手法を適用することで有効性を確認することが挙げられる．また提案手法は読唇のみでなく，表情認識などにも効果があると期待される．そのため，提案手法を読唇以外の問題に適用して効果を検証することも今後の課題である．

**謝辞**　本研究の一部は，JSPS 科研費 15K12601 および 16H03211 の助成によるものである．

### 参考文献

[1] Zhou, Z., Zhao, G., Hong, X. and Pietikainen, M.: A review of recent advances in visual speech decoding, *Image and Vision Computing*, Vol. 32, pp. 590–605 (2014).
[2] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G. and Ogata, T.: Lipreading using Convolutional Neural Network, *INTERSPEECH*, pp. 1149–1153 (2014).
[3] Takashima, Y., Kakihara, Y., Aihara, R., Takiguchi, T., Araki, Y., Mitani, N., Omori, K. and Nakazono, K.: Audio-visual speech recognition using convolutive bottleneck networks for a person with severe hearing loss, *IPSJ Transaction on Computer Vision and Applications*, Vol. 7, pp. 64–68 (2015).
[4] Saitoh, T., Zhou, Z., Zhao, G. and Pietikainen, M.: Concatenated Frame Image Based CNN for Visual Speech Recognition, *ACCV 2016 Workshops, LNCS 10117*, pp. 277–289 (2017).
[5] Chung, J. S. and Zisserman, A.: Lip Reading in the Wild, *Asian Conference on Computer Vision*, pp. 87–103 (2017).
[6] Iwasaki, M., Kubokawa, M. and Saitoh, T.: Two Features Combination with Gated Recurrent Unit for Visual Speech Recognition, *IAPR Conference on Machine Vision Applications (MVA 2017)*, pp. 300–303 (2017).
[7] Masci, J., Meier, U., Cireşan, D. and Schmidhuber, J.: Stacked Convolutional Auto-encoders for Hierarchical Feature Extraction, *21th International Conference on Artificial Neural Networks*, pp. 52–59 (2011).
[8] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (online), DOI: 10.1162/neco.1997.9.8.1735 (1997).
[9] Gers, F. A., Schmidhuber, J. A. and Cummins, F. A.: Learning to Forget: Continual Prediction with LSTM, *Neural Computation*, Vol. 12, No. 10, pp. 2451–2471 (online), DOI: 10.1162/089976600300015015 (2000).
[10] Gers, F. A., Schraudolph, N. N. and Schmidhuber, J.: Learning Precise Timing with Lstm Recurrent Networks, *Journal of Machine Learning Research*, Vol. 3, pp. 115–143 (online), DOI: 10.1162/153244303768966139 (2003).
[11] Chung, J., Gulcehre, C., Cho, K. and Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv preprint arXiv:1412.3555* (2014).
[12] Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S. and Harvey, R.: Extraction of visual features for lipreading, *IEEE Trans. Pattern Anal. & Mach. Intell.*, Vol. 24, No. 2, pp. 198–213 (2002).
[13] Patterson, E. K., Gurbuz, S., Tufekci, Z. and Gowdy, J. N.: Moving-talker, 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).
[14] Cooke, M., Barker, J., Cunningham, S. and Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition, *The Journal of the Acoustical Society of America*, Vol. 120, No. 5, pp. 2421–2424 (2006).
[15] Zhao, G., Barnard, M. and Pietikainen, M.: Lipreading with local spatiotemporal descriptors, *IEEE Transactions on Multimedia*, Vol. 11, No. 7, pp. 1254–1265 (2009).
[16] Anina, I., Zhou, Z., Zhao, G. and Pietikainen, M.: OuluVS2: a multi-view audiovisual database for non-rigid mouth motion analysis, *IEEE International Conference on Automatic Face and Gesture Recognition (FG)* (2015).
[17] Tamura, S., Miyajima, C., Kitaoka, N., Yamada, T., Tsuge, S., Takiguchi, T., Yamamoto, K., Nishiura, T., Nakayama, M., Denda, Y., Fujimoto, M., Matsuda, S., Ogawa, T., Kuroiwa, S., Takeda, K. and Nakamura, S.: CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition, *International Conference on Auditory-Visual Speech Processing (AVSP)* (2010).
[18] Zhou, Z., Hong, X., Zhao, G. and Pietikäinen, M.: A compact representation of visual speech data using latent variables, *IEEE Trans. Pattern Anal. & Mach. Intell.*, Vol. 36, No. 1 (2014).
[19] Pei, Y., Kim, T.-K. and Zha, H.: Unsupervised random forest manifold alignment for lipreading, *IEEE International Conference on Computer Vision (ICCV)*, pp. 129–136 (2013).
[20] Lu, L., Zhang, X., Xu, X. and Shang, D.: Video analysis using spatiotemporal descriptor and kernel extreme learning machine for lip reading, *Journal of Electronic Imaging*, Vol. 24, No. 5, pp. 053023–053023 (2015).