Vol.2017-SLP-115 No.3
2017/2/17

EigenvoiceとCLNFを用いた顔から声への統計的対応付けの検討

大杉 康仁[1,a)]　齋藤 大輔[1]　峯松 信明[1]

概要：音声インターフェースを擬人化する場合に合成音声の話者の顔を提示する方法があるが，どのような声や顔を選択すべきかという問題が生じる．本研究では，声・顔の印象に基づいてそれらを統計的に対応付けることを目的とし，特に声の話者性と顔の静的な個人性に着目した．特徴量抽出にはEigenvoiceとCLNF (Constrained Local Neural Filed) を用い，GMM (Gaussian Mixture Model) に基づいて顔の特徴量から声の特徴量を推定した．GMMは，一人の被験者が手動で対応付けた声と顔のパラレルコーパスを用いて学習した．手動もしくはGMMに基づいて推定された特徴量はEigenvoiceに基づく話者空間上の一点に対応するため，Eigenvoice Conversionにより音声を合成し，手動またはGMMに基づいて推定された話者の類似度をメルケプストラムひずみで評価した．平均メルケプストラムひずみは2.32dBであり，先行研究の特徴量を用いた場合に比べわずかに劣る結果となったものの，入力する顔画像によってはCLNFが有効に働く場合もあった．

キーワード：声・顔の印象，Eigenvoice，Eigenvoice Conversion，CLNF，GMMに基づく声質変換法

1. はじめに

コンピュータとの会話は，ロボットとの面と向かって行うものや音声のみのものまで幅広くSF小説や映画で描かれており，人間がコンピュータと会話したいという願望が表現されていると考えられる．実際，音声のみでのやり取りは，AppleのSiri[*1]やMicrosoftのCortana[*2]，docomoのしゃべってコンシェル[*3]などの音声インターフェースで実現され普及し始めている．しかし，映画等で描かれている会話は人間どうしの自然な会話に近く，現実の音声インターフェースはまだその域には達していないと考えられる．自然な会話に近づける方法の一つに音声だけでなくその話者の顔も提示する方法があり，しゃべってコンシェルは有名人やアニメのキャラクター自身の声と顔を同時に提示している．既存の音声インターフェースを話者の顔も提示するものへ拡張することを考えると，既に声質が決まっているものに話者の顔を割り当てる方法や，既に話者の顔が決まっているものをある声質でしゃべらせる方法があり，どちらも適切な声質と顔の組み合せを決める必要が生じる．これには大変な労力を要することは容易に想像できるが，その困難さの原因は，声と顔の対応関係が明らかになっていないことにあると考える．そこで，本研究では，適切な声と顔の組み合せを決める上で重要となるそれぞれの印象，特に音声の話者性と顔の静的な個人性に着目し，声と顔の関係を明らかにすることを目的とする．

人間がある声を聞いた時や顔を見た時に受ける印象を評価する手法に，言葉でその印象を表現する方法がある．例えば，高椋らは，人間がある音声を聞いて評価する過程を階層構造を持つモデルで表現し，そのモデルを検証する中で声の印象を言葉を用いて評価した [1], [2], [3], [4]．また，永田らは，「プロレスラー」や「東京大学の学生」などの職業またはグループについて，所属する人の平均顔を職業名やグループ名と結びつけることで顔の印象に関する分析を行った [5]．これらの先行研究を参考に，Semantic Differential法 (SD法)[6] により人の手で声または顔を言葉により評価し，その評価値を元に声と顔の印象的対応付けを行うことも可能である．しかし，評価者の負担を考慮すると対となる単語間の尺度間隔を細かく設定できず，微妙な印象の変化を上手く評価できない可能性がある．また，選択作業の中で評価者の抱く印象が変化する可能性も捨て

[1] 東京大学大学院工学系研究科電気系工学専攻
[a)] yasuhito.ohsugi@gavo.t.u-tokyo.ac.jp

[*1] http://www.apple.com/jp/ios/siri/ [Accessed 19 January 2017]
[*2] https://www.microsoft.com/ja-jp/windows/cortana [Accessed 19 January 2017]
[*3] https://www.nttdocomo.co.jp/iphone/service/entertainment/shabette_concier/index.html [Accessed 19 January 2017]

ⓒ 2017 Information Processing Society of Japan

きれない．

そこで今回は，声の印象を表す固有空間と顔の印象を表す固有空間を構成し，両空間を統計的に対応付けることを検討した．特に，顔画像から最適な話者を推定することを検討した．統計的対応付けとして，声質変換（二話者間の音響空間の写像を推定する手法）で広く利用されている混合正規分布 (Gaussian Mixture Model: GMM) に基づく対応付け（写像推定）を利用した．この時必要となる二つの特徴量空間に対するパラレルコーパスを，音声と顔画像を手動で対応付ける主観実験により収集した．ただし，[7] で顔のから声への対応付けには顔を見る人に依存する部分が大きいことが示唆されたため，一人が手動で対応付けたパラレルコーパスのみを収集した．

次章以降の本稿の構成を示す．第 2 章で関連研究である Eigenvoice，Eigenface，Constrained Local Neural Field を紹介する．第 3 章で提案手法である GMM に基づく統計的対応付けについて述べる．第 4 章で顔の印象を表す固有空間の構成実験について，第 5 章でパラレルコーパスの収集方法について述べ，第 6 章で提案手法の有効性を確認する実験について述べ，最後に第 7 章で本稿をまとめる．

2. 関連研究

2.1 Eigenvoice

固有声 (Eigenvoice) は，話者非依存の音声認識モデルを特定の話者に依存したモデルに少数のパラメータで適応させる手法の一つとして考案された [8]．以下では，複数の話者から少数の基底を学習し固有空間を構成する点に注目して紹介する．

まず，全 S 人の話者の音声データを使用して話者非依存の M 混合の GMM($\lambda^{(0)}$) を得る．次に，話者 s ($s = 1, 2, ..., S$) の音声データのみを使用して $\lambda^{(0)}$ の各分布の平均のみを更新することで，話者 s に依存した GMM($\lambda^{(s)}$) を得る．$\lambda^{(s)}$ の各分布の平均 $\boldsymbol{\mu}_m^{(s)}$ ($m = 1, 2, ..., M$) を連結し，スーパーベクトル $\boldsymbol{\nu}^{(s)}$ を作成する．

S 人の話者のスーパーベクトルに対する主成分分析 (Principal Component Analysis: PCA) を行い，第 $K(< S)$ 主成分までを Eigenvoice $\boldsymbol{e}(i)$ ($i = 1, 2, ..., K$) と定義する．これらと $\boldsymbol{\nu}^{(s)}$ の平均 $\boldsymbol{b}^{(0)}$ の線形和によりスーパーベクトル $\boldsymbol{\nu}^{(s)}$ は近似される．

$$\boldsymbol{\nu}^{(s)} \simeq \sum_{i=1}^{K} w^{(s)}(i)\boldsymbol{e}(i) + \boldsymbol{b}^{(0)} \tag{1}$$

ここで，$w^{(s)}(i)$ は話者 s の重みを表す．$\boldsymbol{b}^{(0)}$，$\boldsymbol{e}(i)$ について，分布 m に対応するベクトルを $\boldsymbol{b}_m^{(0)}$，$\boldsymbol{e}_m(i)$ とすると，式 (1) の分布 m に対応する部分は式 (2) で表される．

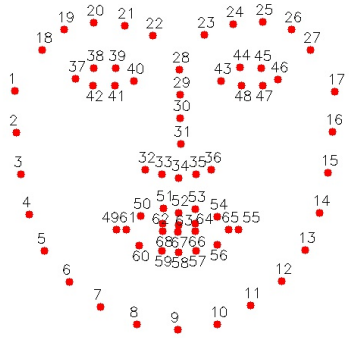

図 1　68 点の Face Landmark

$$\begin{aligned}\boldsymbol{\mu}_m^{(s)} &= \sum_{i=1}^{K} w^{(s)}(i)\boldsymbol{e}_m(i) + \boldsymbol{b}_m^{(0)} \\ &= [\boldsymbol{e}_m(1), ..., \boldsymbol{e}_m(K)][w(1)^{(s)}, ..., w(K)^{(s)}]^{\mathrm{T}} + \boldsymbol{b}_m^{(0)} \\ &= \boldsymbol{B}_m\boldsymbol{w}^{(s)} + \boldsymbol{b}_m^{(0)}\end{aligned} \tag{2}$$

重み $\boldsymbol{w}^{(s)}$ は K 個の Eigenvoice で張られる固有空間のある一点の座標に対応する．指定した重みの声質を持つ音声は，Eigenvoice Conversion (EVC)[9] により作成可能である．

2.2 Eigenface

顔画像を固有空間の一点に写像し定量的に評価する手法の一つに Eigenface がある [10]．縦横それぞれ N 個の画素情報がある 2 次元顔画像を N^2 次元のベクトルで表す．M 個の顔画像 $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, ..., \boldsymbol{\Gamma}_M$ に対し，式 (3) により平均顔 $\boldsymbol{\Psi}$ を求め，各顔画像と平均顔画像との差を式 (4) で表す．式 (5) で得られる，M 個の顔画像に関する共分散行列 $\boldsymbol{C}$ を用いた主成分分析を行い，第 $M'(< M)$ 主成分までを選択し，それらを Eigenface $\boldsymbol{E}(i)$ ($i = 1, 2, ..., M'$) とする．

$$\boldsymbol{\Psi} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\Gamma}_m \tag{3}$$

$$\boldsymbol{\Phi}_m = \boldsymbol{\Gamma}_m - \boldsymbol{\Psi} \tag{4}$$

$$\boldsymbol{C} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\Phi}_m\boldsymbol{\Phi}_m^T \tag{5}$$

ある画像 $\boldsymbol{\Gamma}_m$ は，重み $w^{(m)}(i)(i = 1, 2, ..., M')$ を用いて式 (6) で近似される．

$$\boldsymbol{\Gamma}_m \simeq \sum_{i=1}^{M'} w^{(m)}(i)\boldsymbol{E}(i) + \boldsymbol{\Psi} \tag{6}$$

2.3 Constrained Local Neual Field

顔認識や顔検出の指標の一つに図 1 に示す顔の輪郭や目鼻の位置を表す 68 個の Face Landmark がある [11]．これらの特徴点を検出する方法の一つに Constrained Local Neural Fields (CLNF) がある [12]．CLNF は，式 (7)

の Point Distribution Model (PDM) で Face Landmark の位置を表し，そのパラメータ $\boldsymbol{p} = \{s, \boldsymbol{t}, \boldsymbol{R}_{2D}, \boldsymbol{q}\}$ を Face Landmark 周辺の画像情報を用いて推定する手法である．

$$\boldsymbol{x}_i = s\boldsymbol{R}_{2D}(\overline{\boldsymbol{x}}_i + \boldsymbol{\Phi}_i\boldsymbol{q}) + \boldsymbol{t} \tag{7}$$

ここで，$\boldsymbol{x}_i = [x_i, y_i]^{\mathrm{T}}$ $(i = 1, ..., N)$ は i 番目の Face Landmark の画像上の位置を表す．PDM は，平均ベクトル $\overline{\boldsymbol{x}}_i$ と基底行列 $\boldsymbol{\Phi}_i$ で構成される三次元空間を，回転行列 $\boldsymbol{R}_{2D}$ で回転させさらに二次元平面へ射影し，拡大係数 s と平行移動ベクトル $\boldsymbol{t}$ で対象である顔の Face Landmark の位置を表すモデルである．すなわち，$\overline{\boldsymbol{x}}_i$ は 3 次元だが $\boldsymbol{R}_{2D}$ は 2×3 行列，$\boldsymbol{x}_i$ と $\boldsymbol{t}$ は 2 次元である．$\overline{\boldsymbol{x}}_i$ と $\boldsymbol{\Phi}_i$ は三次元フレームモデルの主成分分析により計算され，$\boldsymbol{q}$ は各 Face Landmark について同一である．

本稿では，OpenFace[13] を用いて $\boldsymbol{p}$ を推定した．

3. 提案手法

本提案手法では，GMM に基づく声質変換法 [14] を応用し，ある顔に印象的に対応した声を GMM に基づいて推定する．声・顔の特徴量はそれぞれの印象を表す固有空間の座標であり，入力特徴量を d_x 次元ベクトル $\boldsymbol{x}$，出力特徴量を d_y 次元ベクトル $\boldsymbol{y}$ とし，これらの連結ベクトル $\boldsymbol{z} = [\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{y}^{\mathrm{T}}]^{\mathrm{T}}$ が混合数 M の GMM に従うものとする．式 (8) に示すように，モデルパラメータ λ を結合確率密度 $p(\boldsymbol{z}|\lambda)$ が最大となるように学習する．

$$\lambda = \arg\max_{\lambda} p(\boldsymbol{z}|\lambda) \tag{8}$$

$$p(\boldsymbol{z}|\lambda) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma_m}^{(z)}) \tag{9}$$

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \ \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{10}$$

学習した λ において，式 (11) に示すように尤度を最大化することで，入力特徴量から出力特徴量への変換を行う．

$$\begin{aligned} \boldsymbol{y} &= \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}, \lambda) \\ &= \arg\max_{\boldsymbol{y}} \sum_{m=1}^{M} p(m|\boldsymbol{x}, \lambda) p(\boldsymbol{y}|\boldsymbol{x}, m, \lambda) \end{aligned} \tag{11}$$

$$p(m|\boldsymbol{x}, \lambda) = \frac{\alpha_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m^{(xx)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_n^{(xx)}, \boldsymbol{\Sigma}_n^{(xx)})} \tag{12}$$

$$p(\boldsymbol{y}|\boldsymbol{x}, m, \lambda) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{E}_m^{(y)}, \boldsymbol{D}_m^{(y)}) \tag{13}$$

ただし，$\boldsymbol{E}_m^{(y)}$ と $\boldsymbol{D}_m^{(y)}$ は以下で表される．

$$\boldsymbol{E}_m^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)}) \tag{14}$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)} \tag{15}$$

式 (11) は式 (16) の補助関数 $Q(\boldsymbol{y}, \hat{\boldsymbol{y}})$ を繰り返し最大化することで解くことが可能であり，補助関数を最大化するような $\hat{\boldsymbol{y}}$ は式 (17) で表される．

$$Q(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{m=1}^{M} p(m|\boldsymbol{x}, \boldsymbol{y}, \lambda) \log p(\hat{\boldsymbol{y}}, m|\boldsymbol{x}, \lambda) \tag{16}$$

$$\hat{\boldsymbol{y}} = \left(\overline{\boldsymbol{D}^{(y)-1}}\right)^{-1} \overline{\boldsymbol{D}^{(y)-1} \boldsymbol{E}^{(y)}} \tag{17}$$

$$\overline{\boldsymbol{D}^{(y)-1}} = \sum_{m=1}^{M} \gamma_m \boldsymbol{D}_m^{(y)-1} \tag{18}$$

$$\overline{\boldsymbol{D}^{(y)-1} \boldsymbol{E}^{(y)}} = \sum_{m=1}^{M} \gamma_m \boldsymbol{D}_m^{(y)-1} \boldsymbol{E}_m^{(y)} \tag{19}$$

$$\gamma_m = p(m|\boldsymbol{x}, \boldsymbol{y}, \lambda) \tag{20}$$

4. 顔の印象を表す固有空間の構成実験

[7] では，表 1 に基づく特徴量を主成分分析することで，顔の印象を表す固有空間を得た．しかし，これらの特徴量の抽出法は，対象となる画像の各ピクセルの輝度値に依存する部分が大きく，正確に顔の特徴量を抽出できていない可能性がある．より正確に顔の特徴を捉えるため，2.3 節の CLNF を用いて推定された 68 点の Face Landmark の座標を連結したベクトル $\boldsymbol{\Gamma}_m$ を用いた主成分分析により固有空間を構成した．

4.1 実験条件

MORPH[15] の顔画像約 54,000 枚に対し，CLNF が実装されている OpenFace[13] を用いて Face Landmark 68 点を推定した．MORPH には，同一人物ではあるが撮影時期が異なる画像が存在するが，全ての画像にそれぞれ異なる人物が写っているものとし，使用する画像の人種や性別は考慮しなかった．m 番目の画像の Face Landmark $\boldsymbol{x}_i^{(m)} = [x_i^{(m)}, y_i^{(m)}]$ $(i = 1, ..., 68)$ を $\boldsymbol{\Gamma}_m = [x_1^{(m)}, ..., x_{68}^{(m)}, y_1^{(m)}, ..., y_{68}^{(m)}]$ で表し，それらに関する主成分分析により固有空間を構成した．

4.2 実験結果

OpenFace で推定された Face Landmark の例を図 2 に示す．入力された画像の目鼻の位置と Face Landmark の位置はほぼ合致しており，Face Landmark の推定が正確に行われていると言える．ただし，(ii) の画像であごの輪郭が描画されていないのは，その位置の Face Landmark が画面外の位置に推定されたためであり，(iii) の画像ではそれを考慮し，鼻の中心が画像中心となるよう平行移動させて描画している．主成分分析においてはそのような平行移動は行っておらず，推定された結果をそのまま用いた．

約 54,000 枚の顔画像に関する主成分分析を行ったときの主成分数と寄与率の関係を表 2 に示す．第 3 主成分の時に累積寄与率は 80%を超え，第 9 主成分の時には 99%を

表 1　顔の部位に基づく特徴量

部位	近似図形	特徴量
眉毛	直線	両眉毛の距離，中心からの距離，長さ，水平からの角度
目	楕円	両目の距離，中心からの距離，目の幅，目の高さ
瞳	楕円	幅
鼻	三角形	幅，高さ
口	直線	中心からの距離，幅
輪郭	楕円	幅，高さ

表 2　CLNF を用いた場合の主成分と寄与率の関係 (単位:%)

主成分数	1	2	3	4				
寄与率	40.2	30.6	15.2	4.22				
累積寄与率	40.2	70.8	86.1	90.3				
主成分数	5	6	7	8	9	10	11	12
寄与率	3.83	2.89	1.01	0.605	0.400	0.232	0.190	0.189
累積寄与率	94.1	97.0	98.0	98.6	99.0	99.3	99.4	99.6

超えた．得られた固有空間が顔の印象を反映しているか確認するため，手動で重みを変化させた時に Face Landmark はどのように描画されるかを調べたところ，寄与率の大きな第 4 主成分までは，顔の三次元的な回転に対応していた．例えば，図 3 のように第 1 主成分が大きい時には Face Landmark が全体的に反時計回りに回転した．また，第 2 主成分が大きい時には顔が上を向いているような配置に Face Landmark が変化した．ところが，寄与率が小さい第 5 主成分から第 12 主成分では顔の印象の変化が見られた．例えば，図 3 のように第 12 主成分を小さくすると顔の輪郭は広がり目鼻は逆に中心に集まるように Face Landmark は変化した．また，第 5 主成分や第 7 主成分が変化すると，顔の向きはほぼ一定で顔の幅が変化した．回転は本研究で扱う顔の印象とは独立であるため，以降では第 5 主成分から第 12 主成分を顔の印象を表現するパラメータとして利用した．

5. 手動に基づく声・顔のパラレルコーパスの収集

本研究で対象としているのは，声の印象と顔の印象が合致しているパラレルコーパスであり，話者本人の声と顔の組み合せとは限らない．そこで，ある顔に適切だと思われる音声を選択する主観実験を行い，両者のパラレルコーパスを得た．ただし，被験者と提示する音声・顔の性別により知覚モデルが異なる可能性があることから [3]，被験者・提示顔画像・選択対象の音声の性別を全て男性に固定した．被験者は 20 代男性 1 名であり，提示した顔画像は MORPH のアジア系男性顔画像 44 枚，音声は JNAS[*4] の男性話者 127 人の音素バランス文を使用した．ただし，提示した音声は各話者が発話したサブセットの一文目のみである．被験者の負担を減らすため，あらかじめ話者を [7] で作成した Eigenvoice の固有空間の座標に基づき二分木を用いて分類し，その木を辿らせ最終的に最も適切な音声を選択する方法を採用した．二分木を構成する基準として当該話者の Eigenvoice の重みの正負を利用した．

この主観実験で得られるパラレルコーパスは，顔画像と最終的に選択された音声ファイルが結びつけられたものとなるため，提案手法における対応付けにおいては，顔画像と音声をそれぞれ顔・声の固有空間に射影する必要がある．

6. GMM に基づく顔から声への対応付け実験

6.1 実験条件

第 3 章の提案手法を用いて，顔画像から印象的に対応する話者を統計的に導くことを検討する．

[7] で行った実験から，Eigenvoice で構成された固有空間は話者の声の印象を反映しているものと考えることができるため今回もこの空間を使用した．すなわち，JNAS の男性話者 127 人を用いて得られる Eigenvoice で構成された話者空間を利用した．ただし，特徴量抽出には WORLD[16] と SPTK[*5] を使用し，話者空間内の一点に対応する話者の音声合成には別の男性話者を入力とする EVC を用いた．

顔の印象を表す固有空間として，先行研究 [7] に基づく空間と第 4 章で得られた CLNF に基づく空間の両方を使用し比較した．前者の空間について，表 1 の特徴量を抽出するために OpenCV[*6] を使用し，顔画像には MORPH の顔画像約 21,000 枚を使用し．人種と性別は考慮せず各画像に異なる人物が写っているものとした．ここで，二つの固有空間を構成するために用いた顔画像の数が異なるのは，OpenCV によって特徴量を検出できなかった画像が存在したためである．ただし，表 1 で定義した顔の特徴量を持つ顔写真を作成することは困難であるため，顔の各部位をそ

[*4] http://research.nii.ac.jp/src/JNAS.html [Accessed 19 January 2017]

[*5] http://sp-tk.sourceforge.net/ [Accessed 19 January 2017]

[*6] http://opencv.jp/ [Accessed 19 January 2017]

ⓒ 2017 Information Processing Society of Japan

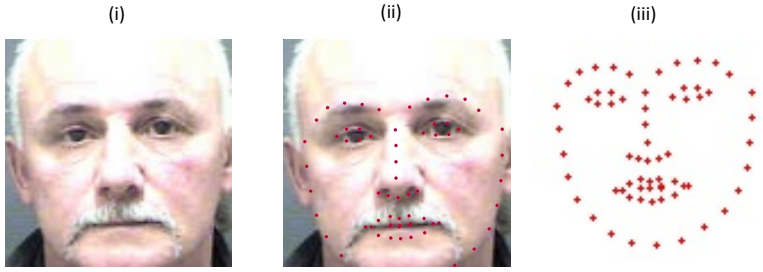

図 2　OpenFace で推定された Face Landmark の例: (i) 入力顔画像 (ii) 顔画像に Face Landmark を描画した画像 (iii)Face Landmark のみを描画した画像

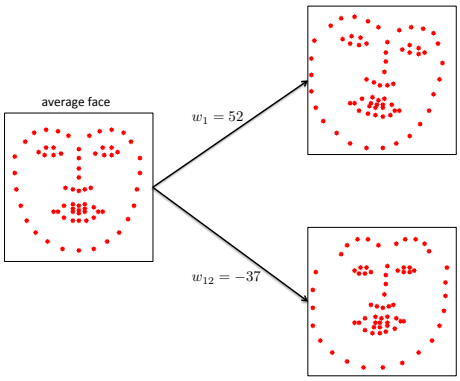

図 3　CLNF に基づく固有空間上で重みを変化させた結果: w_i は第 i 主成分を表す

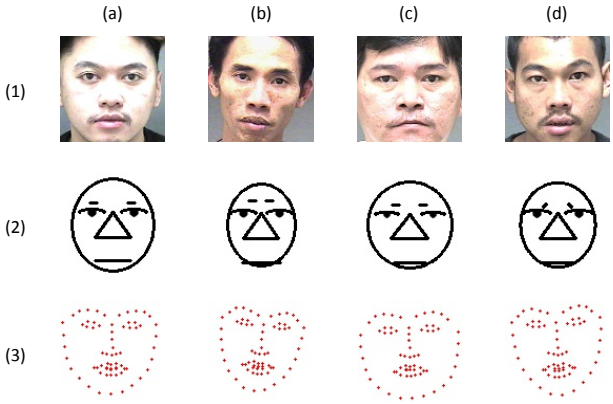

図 4　評価データセット 1 の顔画像: (1)MOPRH の画像から OpenCV を用いて検出した顔画像 (2) 表 1 の特徴量を表現したアイコン画像 (3)OpenFace で推定した Face Landmark

れぞれ割り当てた近似図形で近似したアイコン画像を作成した．

GMM の混合数を 2，出力特徴量を話者空間の低次 6 次元ベクトル $\boldsymbol{y}$ とした．入力特徴量 $\boldsymbol{x}$ について，表 1 に基づく固有空間を用いる場合は低次 3 次元を，CLNF に基づく固有空間を用いる場合は，第 5 主成分から第 12 主成分を表す 8 次元を用いた．それぞれの累積寄与率は 75.8% と 9.35% であった．第 5 章で得られたパラレルデータ 44 組の内，40 組を GMM の学習に用い 4 組を評価に用いるクロス・バリデーションで提案手法の有効性を検証した．評価は，手動もしくは GMM に基づいて推定された話者空間の座標から，EVC を用いて音声を合成し，両者を式 (21) のメルケプストラムひずみ (Mel-cepstrum distortion: MCD) を用いて比較した．

$$MCD[dB] = 10/\ln 10 \sqrt{2 \sum_{d=1}^{24} \left(c_d^{(tar)} - c_d^{(ref)} \right)^2} \quad (21)$$

ただし，$c_d^{(tar)}$ は GMM に基づいて推定された話者の音声のメルケプストラムであり，$c_d^{(ref)}$ は手動で推定された話者の音声のメルケプストラムである．MCD が小さい程，比較対象の二つの音声は似通っていると言える．合成した音声は，JNAS のサブセット J の音声 53 文であり，それらのメルケプストラムひずみの平均を二人の話者の類似度とした．

6.2　実験結果

手動または GMM に基づいて推定された話者間の MCD を表 3 に示す．それぞれの平均 MCD を比較すると，先行研究 [7] で使用した表 1 の特徴量を用いた場合の方が CLNF を用いた場合よりもわずかながら MCD が小さかったが，評価データセットによっては CLNF を用いた方が MCD が小さい場合もあった．例えば，図 4 に示すデータセット 1 の各画像について MCD の値を比較すると，表 4 のように (c) の顔画像に対しては MCD の値が CLNF の方がかなり小さかった．c(c) の顔画像は比較的広い顔幅を持っており，狭い顔幅を持つ (b) の画像に比べ，MCD の値が先行研究または CLNF の両方で大きくなった．

このように，今回の実験においては，先行研究で用いられた特徴量の方が有効に働く傾向があったと言える．しかし，使用した特徴量の各固有空間における累積寄与率を比較すると，先行研究の特徴量のそれは 75.8% であったが，CLNF に基づく特徴量のそれは 9.35% であり，CLNF を使用した場合は，構成した固有空間の情報を十分に対応付けに反映できなかったにも関わらず，両者の対応付けの MCD は非常に近い値となった．今回寄与率の小さい主成分を使用したのは，顔の印象に独立な頭部の回転が CLNF に基づく固有空間に大きな影響を与えていたためであり，頭部の回転の影響が少ない固有空間を構成すれば，より多くの情報を対応付けに活かすことができる可能性がある．

ⓒ 2017 Information Processing Society of Japan

表 3　手動または GMM に基づいて推定した話者間の MCD [dB]: 先行研究は表 1 に基づく固有空間を用いた場合を示し，CLNF は第 4 章の固有空間を用いた場合を示す．

評価データセット ID	1	2	3	4	5	6	7	8	9	10	11	平均
先行研究 [7]	3.48	1.81	2.13	**1.61**	**1.68**	**1.81**	**1.86**	**2.70**	1.80	**1.46**	**1.34**	**1.97**
CLNF	**3.04**	**1.65**	**1.95**	2.41	1.89	2.18	2.12	2.89	**1.77**	3.33	2.27	2.32

表 4　評価データセット 1 に関する MCD [dB]

顔画像	(a)	(b)	(c)	(d)	平均
先行研究 [7]	4.76	**1.31**	5.93	1.93	3.48
CLNF	**3.90**	2.10	**4.76**	**1.39**	**3.04**

7. まとめ

本研究では，声の話者性と顔の静的な個人性に着目し，声・顔の印象に基づいてそれらを統計的に対応付けることを目的とした．声の特徴量には Eigenvoice を，顔の特徴量には表 1 に基づく特徴量または CLNF に基づく特徴量を用い，GMM に基づいて顔の特徴量から声の特徴量を推定した．GMM 学習には一人の被験者が手動により対応付けた声と顔のパラレルコーパスを使用し，その人の対応付けを統計的に表現することを試みた．Eigenvoice Conversion により音声を合成し MCD で手動または GMM に基づいて推定された話者の類似度を評価したところ，表 1 を用いた場合は 1.97dB，CLNF を用いた場合は 2.32dB であったことから，今回の実験では，表 1 の特徴量の方が対応付けには有効に働く傾向が見られた．しかし，評価顔画像によっては，CLNF の方が有効に働く場合もあり，両者の平均 MCD が近しいことから，CLNF を用いた場合の対応付けについてさらに調査する必要がある．また，今回使用したパラレルコーパスは一人の被験者によるものであるため，今回の実験系はこの被験者の対応付けに依存していることに注意する必要がある．よって，今後は複数の人間の対応付けを個別に表現しそれらを比較することを検討する．その結果を用いて，複数の人間の対応付けに共通点があるか否かを調査し，顔から声への印象的対応付けに固定観念が存在するかどうかを検証する．

参考文献

[1] 高椋琴美，保田千津子，谷田泰郎：音響特徴量と声の印象に関する分析，日本音響学会講演論文集，pp. 445–448 (2013).
[2] 高椋琴美，　東優，谷田泰郎：声の印象を表現する単語による認知構造モデルの検討，日本音響学会講演論文集，pp. 451–454 (2014).
[3] 高椋琴美，谷田泰郎：声の印象と音響特徴量の関係性評価と対話応用への検討，日本音響学会講演論文集，pp. 379–482 (2014).
[4] 高椋琴美，谷田泰郎：声の印象評価にみられる評価者の個性の影響，日本音響学会講演論文集，pp. 399–402 (2015).
[5] 永田明徳，金子正秀，　原島博：平均顔を用いた顔印象分析，電子情報通信学会論文誌 A，Vol. 80, No. 8, pp. 1266–1272 (1997).
[6] Osgood, C. E.: Semantic differential technique in the comparative study of cultures, *American Anthropologist*, Vol. 66, No. 3, pp. 171–200 (1964).
[7] 大杉康仁，齋藤大輔，峯松信明：声・顔の固有空間と GMM に基づく両空間の印象的対応付けに関する検討，研究報告音楽情報科学 (MUS)，Vol. 2016, No. 56, pp. 1–6 (2016).
[8] Kuhn, R., Junqua, J.-C., Nguyen, P. and Niedzielski, N.: Rapid speaker adaptation in eigenvoice space, *Speech and Audio Processing, IEEE Transactions on*, Vol. 8, No. 6, pp. 695–707 (2000).
[9] 戸田智基，大谷大和，鹿野清宏：固有声に基づく声質変換法，電子情報通信学会技術研究報告. SP, 音声， Vol. 106, No. 221, pp. 25–30 (2006).
[10] Turk, M. and Pentland, A. P.: Face recognition using eigenfaces, *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pp. 586–591 (1991).
[11] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S. and Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge, *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 397–403 (2013).
[12] Baltrusaitis, T., Robinson, P. and Morency, L.-P.: Constrained local neural fields for robust facial landmark detection in the wild, *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 354–361 (2013).
[13] Baltru, T., Robinson, P., Morency, L.-P. et al.: OpenFace: an open source facial behavior analysis toolkit, *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp. 1–10 (2016).
[14] Toda, T., Black, A. W. and Tokuda, K.: Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory, *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 15, No. 8, pp. 2222–2235 (2007).
[15] Ricanek Jr, K. and Tesafaye, T.: Morph: A longitudinal image database of normal adult age-progression, *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, IEEE, pp. 341–345 (2006).
[16] 森勢将雅，西浦敬信，河原英紀：高品質音声分析変換合成システム WORLD の提案と基礎的評価–基本周波数・スペクトル包絡制御が品質の知覚に与える影響，聴覚研究会資料，Vol. 41, No. 7, pp. 555–560 (2011).

ⓒ 2017 Information Processing Society of Japan