Vol.2017-MUS-115 No.34
2017/6/17

顔から声への統計的対応付けに関する技術的諸検討

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，合成音声とともに擬人化エージェント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，声の話者性と顔の静的な個人性に表われる声・顔の印象に基づいて，顔から声への統計的対応付けを検討した．声・顔の印象を表す特徴量の抽出には Eigenvoice と CLNF (Constrained Local Neural Filed) を使用し，GMM (Gaussian Mixture Model) または CCA (Canonical Correlation Analysis) に基づいて顔の特徴量から声の特徴量を推定した．ここでは，あらかじめ複数人の被験者を集め，ある顔に印象的に対応すると思われる声を対応付ける主観実験を実施し，顔と声を対応付けたパラレルコーパスを作成し，CCA 及び GMM の学習に利用した．ただし，顔と声の変換写像は被験者に依存する可能性を考慮し，被験者依存の形で実験を行った．比較においては，声の特徴量が表す話者の音声を Eigenvoice Conversion により合成し，そのメルケプストラムひずみを利用した．結果として，GMM を用いた方が CCA を用いるよりも精度が高く，より確率的な写像が有効であることが示唆された．

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

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．今日，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]，Amazon の Amazon Echo [*4] などの音声インターフェースとして実現され普及し始めている．しかし，それらの使用方法は音声検索や簡単な設定を行うための命令形式であり，映画等で描かれているような自然かつ多様な会話は実現していないと言える．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る．既存の音声インターフェースをその形式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．多くの場合，半ば強引にそれらの組み合せを決めてしまうだろうが，そのときにも我々は無意識に自然な顔と声の組み合せを考えながら最終的な答えを出していると考えられる．その無意識の顔と声の対応付けが定量的かつ計算機利用可能な形式で取り出せていないため，顔と声の対応付けを自動で行うことはできず，音声インターフェースの多様化の足かせとなりうる．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，声と顔の対応関係を定量的かつ計算機応用可能な形でモデル化し，実際に応用することを目的とする．

今回は，声の印象を表す空間と顔の印象を表す空間を構成し，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，Canonical Correlation Analysis (CCA) に基づく写像を検討し，先行研究である Gaussian Mixture Model (GMM) に基づく写像と比較した．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，顔画像と音声を手動で対応付ける主観実験により収集した．

次章以降の本稿の構成を示す．第 2 章で関連研究を紹介し，第 3 章で提案手法である CCA 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空間の構成実

[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]
[*4] `https://www.amazon.com/Amazon-Echo-Bluetooth-Speaker-with-WiFi-Alexa/dp/B00X4WHP5E` [Accessed 28 May 2017]

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る．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 [1]．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．GMM の各分布がある音韻にそれぞれ対応していると考え，$\lambda^{(s)}$ の各分布の平均 $\boldsymbol{\mu}_m^{(s)}$ ($m = 1, 2, ..., M$) を連結したスーパーベクトル $\boldsymbol{\nu}^{(s)}$ を話者 s の声の特徴とする．

$\boldsymbol{\nu}^{(s)}$ ($s = 1, 2, ..., S$) 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．

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

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)[2] 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 [3]．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．

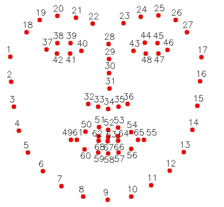

図 1　68 点の Face Landmarks

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

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

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

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 Landmarks がある [4]．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 [5]．CLNF は，式 (7) の Point Distribution Model (PDM) で Face Landmarks 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 Landmarks 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 Landmarks 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．

2.4 GMM に基づく声質変換法

声質変換 (Voice Conversion: VC) とは，入力話者の音響特徴量と出力話者の音響特徴量を対応させた変換モデルに基づいて入力話者の声質を出力話者の声質に変換する技術であり，ここでは変換モデルとして GMM を用いた手法を紹介する [6]．

時刻 t における入力・出力話者の静的・動的特徴量をそれぞれ $\boldsymbol{X}_t = [\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{\Delta x}_t^{\mathrm{T}}]$，$\boldsymbol{Y}_t = [\boldsymbol{y}_t^{\mathrm{T}}, \boldsymbol{\Delta y}_t^{\mathrm{T}}]$ とする．これ

らを連結した $\boldsymbol{Z}_t = [\boldsymbol{X}_t^{\mathrm{T}}, \boldsymbol{Y}_t^{\mathrm{T}}]^{\mathrm{T}}$ を用いて，式 (8) に従って GMM のパラメータ λ を結合確率密度 $p(\boldsymbol{Z}_t|\lambda)$ が最大となるように学習する．

$$\begin{aligned}\hat{\lambda} &= \arg\max_{\lambda} \prod_{t=1}^{T} p(\boldsymbol{Z}_t|\lambda) \\ &= \arg\max_{\lambda} \prod_{t=1}^{T} \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma_m}^{(Z)})\end{aligned} \tag{8}$$

$$\boldsymbol{\mu}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \ \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{9}$$

学習した GMM において，入力系列 $\boldsymbol{X} = [\boldsymbol{X}_1^{\mathrm{T}}, ..., \boldsymbol{X}_T^{\mathrm{T}}]^{\mathrm{T}}$ が与えられた時の出力系列 $\boldsymbol{Y} = [\boldsymbol{Y}_1^{\mathrm{T}}, ..., \boldsymbol{Y}_T^{\mathrm{T}}]^{\mathrm{T}}$ の条件付き確率 $p(\boldsymbol{Y}|\boldsymbol{X}, \lambda)$ を最大化することで，入力話者の音響特徴量を出力話者のそれに変換し声質変換を行う．

$$\hat{\boldsymbol{Y}} = \arg\max_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\lambda}) \tag{10}$$

静的特徴量から静的・動的特徴量への変換を考慮しつつ式 (10) を解くことで，所望の声質を持った静的音響特徴量を得ることができる．

2.5 Canonical Correlation Analysis

正準相関分析 (Canonical Correlation Analysis: CCA) は，同じ観測対象から測定された二つの信号から，それぞれの独自因子を除き，両者に共通する因子を見つけ出す分析手法である [7]．一つの観測信号を $\boldsymbol{x}_n$，もう一つの観測信号を $\boldsymbol{y}_n$ とし，両者が $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ $(n = 1, 2, ..., N)$ のようにペアを成しているとする．$\boldsymbol{x}_n$，$\boldsymbol{y}_n$ をそれぞれ以下の式で u_n，v_n へと線形変換する．

$$u_n = \boldsymbol{a}^{\mathrm{T}}(\boldsymbol{x}_n - E[\boldsymbol{x}]) \tag{11}$$

$$v_n = \boldsymbol{b}^{\mathrm{T}}(\boldsymbol{y}_n - E[\boldsymbol{y}]) \tag{12}$$

ただし，$E[\cdot]$ はサンプル平均を表す．u と v の相関係数 $\rho(\boldsymbol{a}, \boldsymbol{b})$ を最大にするような $\boldsymbol{a}$ と $\boldsymbol{b}$，すなわち $\hat{\boldsymbol{a}}$ と $\hat{\boldsymbol{b}}$ を求める．

$$\begin{aligned}\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}} &= \arg\max_{\boldsymbol{a},\boldsymbol{b}} \rho(\boldsymbol{a}, \boldsymbol{b}) \\ &= \arg\max_{\boldsymbol{a},\boldsymbol{b}} \frac{E[uv]}{\sqrt{E[u^2]}\sqrt{E[v^2]}} \\ &= \arg\max_{\boldsymbol{a},\boldsymbol{b}} \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{V}_{xy}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{V}_{xx}\boldsymbol{a}}\sqrt{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{V}_{yy}\boldsymbol{b}}} \\ &= \arg\max_{\boldsymbol{a},\boldsymbol{b}} \boldsymbol{a}^{\mathrm{T}}\boldsymbol{V}_{xy}\boldsymbol{b}\end{aligned} \tag{13}$$

ただし，$V_{xx} = E[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]$，$V_{xy} = E[\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}]$，$V_{yy} = E[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}]$ であり，式 (14) の制約条件を加えた．

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{V}_{xx}\boldsymbol{a} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{V}_{yy}\boldsymbol{b} = 1 \tag{14}$$

これは一般化固有値問題に帰着され，その固有値 λ に対応する固有ベクトルが $\hat{\boldsymbol{a}}$，$\hat{\boldsymbol{b}}$ となり，λ は相関係数 $\rho(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}})$ と一致する．式 (11) と式 (12) は一次元への射影だが，複数の固有値に対応する固有ベクトルを考えることで，多次元空間への射影を考えることができる．

3. 提案手法

先行研究 [8] にて我々は 2.4 節を応用した GMM に基づく顔と声の対応付けを行った．本稿では，2.5 節で紹介した CCA を用いて，ある顔に印象的に対応した声を推定することを考える．

顔の印象を表す特徴量を d_x 次元ベクトル $\boldsymbol{x}$，声の印象を表す特徴量を d_y 次元ベクトル $\boldsymbol{y}$ とする．まず，印象的に対応する顔と声の組み合せ N 対 $\{\boldsymbol{x}_i, \boldsymbol{y}_i : i = 1, 2, ..., N\}$ に対し CCA を行い，得られた各 $\hat{\boldsymbol{a}}$ を行ベクトルとする変換行列 $\boldsymbol{A}$ と各 $\hat{\boldsymbol{b}}$ を行ベクトルとする変換行列 $\boldsymbol{B}$ を得る．すなわち，CCA において固有値が n 個得られた場合，$\boldsymbol{A}$ は n 行 d_x 列の行列，$\boldsymbol{B}$ は n 行 d_y 列の行列となる．

次に，ある顔の特徴量 $\boldsymbol{x}$ に対応する声の特徴量 $\boldsymbol{y}$ を推定することを考える．変換行列 $\boldsymbol{A}$，$\boldsymbol{B}$ により，以下の線形変換を行う．

$$\boldsymbol{u} = \boldsymbol{A}(\boldsymbol{x} - E[\boldsymbol{x}]) \tag{15}$$

$$\boldsymbol{v} = \boldsymbol{B}(\boldsymbol{y} - E[\boldsymbol{y}]) \tag{16}$$

ここで，$E[\boldsymbol{x}]$ と $E[\boldsymbol{y}]$ は，CCA に用いた N 対のサンプル平均である．CCA により $\boldsymbol{u}$ と $\boldsymbol{v}$ の相関は最大となっているため，次の近似が成り立つとして差し支えない．

$$\boldsymbol{u} \propto \boldsymbol{v} \tag{17}$$

ただし，定数倍は $\boldsymbol{A}$ や $\boldsymbol{B}$ で吸収できることを考えると，次の近似が成り立つとしても差し支えない．

$$\boldsymbol{u} \simeq \boldsymbol{v} \tag{18}$$

式 (15) と (16) より，以下の近似が成り立つ．

$$\boldsymbol{A}(\boldsymbol{x} - E[\boldsymbol{x}]) \simeq \boldsymbol{B}(\boldsymbol{y} - E[\boldsymbol{y}]) \tag{19}$$

与えられた顔の特徴量 $\boldsymbol{x}$ に対応する声の特徴量 $\boldsymbol{y}$ を推定するため，式 (19) に基づいて最小二乗法を用いて最小ノルム解を求めると，

$$\begin{aligned}\hat{\boldsymbol{y}} &= \arg\min_{\boldsymbol{y}} ||\boldsymbol{A}(\boldsymbol{x} - E[\boldsymbol{x}]) - \boldsymbol{B}(\boldsymbol{y} - E[\boldsymbol{y}])||^2 \\ &= E[\boldsymbol{y}] + \boldsymbol{\Phi}\boldsymbol{\Sigma}^{+}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x} - E[\boldsymbol{x}])\end{aligned} \tag{20}$$

ただし，$\boldsymbol{\Theta}$，$\boldsymbol{\Sigma}$，$\boldsymbol{\Phi}$ は $\boldsymbol{B}$ の特異値分解で得られる行列である．

$$\boldsymbol{B} = \boldsymbol{\Theta}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}} \tag{21}$$

$\boldsymbol{\Sigma}$ と $\boldsymbol{\Sigma}^{+}$ の関係は $\boldsymbol{B}$ の特異値を対角成分に持つ対角行列 $\boldsymbol{D}$ を用いて以下で表される．

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}, \ \boldsymbol{\Sigma}^{+} = \begin{bmatrix} \boldsymbol{D}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \tag{22}$$

表 1 Iconified Face Feature

部位	近似図形	特徴量
眉毛	直線	両眉毛の距離 (ebd)，中心からの距離 (eby)，長さ (ebl)，水平からの角度 (θ)
目	楕円	両目の距離 (ed)，中心からの距離 (ey)，目の幅 (ew)，目の高さ (eh)
瞳	楕円	幅 (pw)
鼻	三角形	幅 (nw)，高さ (nl)
口	直線	中心からの距離 (my)，幅 (mw)
輪郭	楕円	幅 (fw)，高さ (fh)

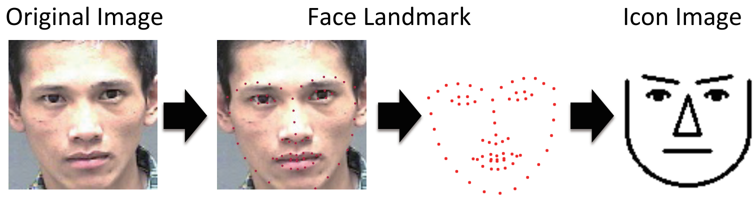

図 2 OpenFace で推定された Face Landmarks とそれに基づいて抽出された IFF を可視化したアイコン画像

表 2 IFF の主成分と寄与率の関係 (単位:%)

主成分数	1	2	3	4
累積寄与率	76.7	93.1	97.0	99.0

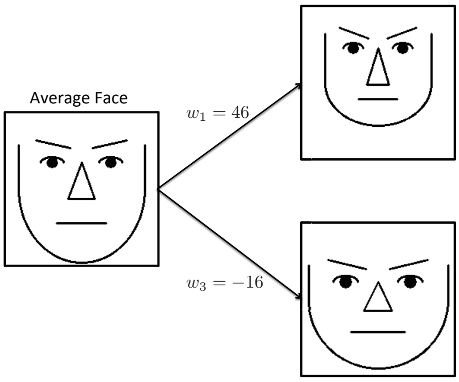

図 3 IFF に基づく固有空間上で重みを変化させた結果: w_i は第 i 主成分を表す

4. 顔の印象を表す空間の構成実験

先行研究 [8] で我々は，表 1 で定義される 15 個の特徴量を定義した．以後これらの特徴量を Iconified Face Feature (IFF) と呼ぶこととする．今回は，2.3 節の CLNF を用いて推定された 68 点の Face Landmarks の座標から IFF を求め，それらの主成分分析により固有空間を構成した．IFF を反映したアイコン画像を作成し，その空間が顔の印象を表しているかどうか確認した．

4.1 実験条件

MORPH[9] の顔画像約 54,000 枚に対し，CLNF が実装されている OpenFace[10] を用いて Face Landmarks 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た．また，得られる空間の多様性を確保するため，使用する画像の人種や性別は考慮しなかった．得られた Face Landmarks $\boldsymbol{p}_i = [x_i, y_i]$ $(i = 1, ..., 68)$ に対し，描かれるアイコン画像が左右対称になるような規則に基づいて IFF を抽出した．さらに，得られた IFF を主成分分析し固有空間を構成した．その固有空間が顔の印象を反映しているかどうか確認するため，各主成分の値を調節したときに描かれるアイコン画像から受ける印象が変化するかどうか調査した．

4.2 実験結果

OpenFace で推定された Face Landmarks とそれに基づいて推定された IFF を反映したアイコン画像を図 2 に示す．入力された画像の目鼻の位置と Face Landmarks の位置はほぼ合致しており，Face Landmarks の推定が正確に行われていると言える．また，アイコン画像を見ると，適度に人間の顔を抽象化した画像となっている．

各顔画像から得られた IFF 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 2 主成分の時に累積寄与率は 90%を超え，第 4 主成分の時には 99%となった．主成分の値を調整した場合のアイコン画像を図 3 に示す．第 1 主成分が大きくなるとあごがより丸くなり鼻も長くなったアイコン画像となった．また，第 3 主成分が小さくなると顔の幅が広くなり，比較的丸顔のアイコン画像となった．従って，IFF の主成分分析により得られた固有空間は，顔の印象を表していると言える [*5]．

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

*5 実験結果を `https://goo.gl/JT9jUM` で公開している

表 3　第 6 章の実験条件

声の印象を表す空間	
手法	Eigenvoice
データ	JNAS 男性話者 127 人
特徴量	6，18，25 次元
顔の印象を表す空間	
手法	IFF，IFF の主成分分析
データ	MORPH 顔画像約 54,000 枚
特徴量	IFF(15 次元)，IFF の主成分 (3，4 次元)
両空間の写像	
手法	GMM，CCA
データ	第 5 章で収集した被験者ごとのパラレルコーパス
備考	GMM の混合数は 2

表 4　Eigenvoice と累積寄与率の関係 (単位:%)

主成分数	1	2	6	18	25	33
累積寄与率	14.9	25.8	51.1	70.7	75.9	80.1

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 [11]，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 10 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[*6] 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 [8] 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．具体的には当該話者の Eigenvoice の重みの正負を利用して二分木を構成した．ただし，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である．そこで，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・声の印象を表す空間に射影する必要がある．

6. CCA 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．実験条件を表 3 に示す．[8] で行った実験から，Eigenvoice で構成された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．Eigenvoice と累積寄与率の関係を表 4 に示す．Eigenvoice を構成するために

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

表 5　手動または GMM・CCA に基づいて推定した話者間の平均 MCD [dB]

写像	顔の特徴量	声の特徴量	MCD
GMM	IFF の主成分 3 次元	Eigenvoice 6 次元	**2.44**
CCA	IFF の主成分 3 次元	Eigenvoice 6 次元	2.69
CCA	IFF の主成分 3 次元	Eigenvoice 18 次元	2.58
CCA	IFF の主成分 3 次元	Eigenvoice 33 次元	2.64
CCA	IFF の主成分 4 次元	Eigenvoice 6 次元	2.89
CCA	IFF(15 次元)	Eigenvoice 6 次元	3.77
CCA	IFF(15 次元)	Eigenvoice 25 次元	2.97
CCA	IFF(15 次元)	Eigenvoice 33 次元	2.84

WORLD[12] と SPTK[*7] 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．

顔の印象を表す空間として，第 4 章で得られた IFF の各特徴量を軸とする空間と IFF の主成分が張る固有空間の 2 つを比較した．

顔・声の印象を表す空間の間の写像として，2.4 節を応用した GMM に基づく写像 [8] と第 3 章の CCA に基づく写像を比較した．また，顔・声の特徴量や次元数を変化させたときに変換の様子が変化するかどうか検証した．GMM の学習または CCA には，第 5 章で得られたパラレルコーパスを利用した．ただし，[8] の結果より，顔から声への対応付けには個人性があることが示唆されたため，各被験者ごとに GMM の学習または CCA を行い，統計的対応付けと手動の対応付けを比較した．1 人の被験者につきパラレルデータは 44 組得られたが，その内 40 組を GMM の学習または CCA 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・CCA 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 (23) のメルケプストラムひずみ (Mel-cepstrum distortion: MCD) を用いて比較した．

$$\mathrm{MCD[dB]} = 10/\ln 10\sqrt{2\sum_{\mathrm{d}=1}^{24}\left(\mathrm{c}_{\mathrm{d}}^{(\mathrm{tar})}-\mathrm{c}_{\mathrm{d}}^{(\mathrm{ref})}\right)^2} \quad (23)$$

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・CCA に基づいて推定された話者の

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

間の MCD を全評価データセット・全被験者に関して平均した結果を表 5 に示す．顔・声の特徴量を変化させたときに，最も手動の結果と近かった（MCD が小さかった）のは，先行研究である IFF の主成分 3 次元と Eigenvoice 6 次元を用いた GMM に基づく対応付けであり，提案手法である CCA に基づく対応付けよりも優れていた．しかし，顔・声の特徴量の次元数を変化させたとき，GMM のパラメータ学習が上手くいかない場合が頻繁に起きたが，CCA は学習サンプル数以下の次元であれば上手く分析できていた．

入出力の次元数を変化させた時の CCA の結果について考えると，まず，IFF の主成分を使用した場合については，IFF の主成分の次元数を上昇させると精度が悪化した．また，Eigenvoice の次元数が 18 次元のときが精度が最も高かった．次に，IFF そのものを使用した場合については，Eigenvoice の次元数を向上させると精度が向上したものの，IFF の主成分を使用した場合よりも精度は悪かった．このことから，IFF には冗長性があり，対応付けにおいては IFF を次元圧縮して用いた方が効果的であると言える．

以上の結果から，CCA に基づく対応付けは，GMM を用いた場合と比べて入出力の次元数の変化に頑健であると言えるが，手動の対応付けとの比較においては GMM よりも精度が悪く，顔から声への対応付けを上手くモデル化できなかった．これは，式 (16) において，声の印象を表す空間から顔と声に共通する因子が従う空間への写像を考えているが，式 (20) では，共通因子と声の特徴量が一対一に対応するものとして変換を行っていることに起因すると考えられる．第 5 章の主観実験の結果を見ても，同じ声が複数の顔に割り当てられており，本来共通因子と声の特徴量は一対一ではなく，ある程度のばらつきを伴って対応していると言える．これは，同じ顔に複数の声が割り当てられたことから，顔の特徴量と共通因子においても同様であると言える．GMM に基づく対応付けにおいては，顔と声の特徴量が連結された特徴量が統計的に GMM を用いて表現されるため，このばらつきを上手く表現できている可能性がある．従って，CCA においても，共通因子が従う分布を考慮し，より確率的な対応付けを考える必要がある．

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，顔の特徴量から声の特徴量を統計的に推定す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 IFF または IFF を主成分分析した結果を用いた．変換写像として CCA に基づく変換法を検討し，先行研究である GMM に基づく対応付けと比較した．GMM の学習と CCA には，あらかじめ手動で対応付けた顔と声のパラレルコーパスを使用した．結果として，CCA に基づく対応付けは GMM に基づく対応付けよりも精度が悪かった．その原因として，提案手法は顔の特徴量と声の特徴量を一対一に変換しようとする写像であることが考えられる．しかし，GMM よりも CCA は入出力の次元数に頑健であることも示唆された．従って今後は，共通因子が従う分布を考慮した CCA に基づく対応付けを考える．また，今回は MORPH の顔画像の被写体が全て異なるという条件のもと IFF の抽出を行ったため，現在の方法では厳密に個人性を表現しているとは言えない．そこで，同じ被写体の顔画像複数枚から一つの IFF を抽出する方法についても検討する．

参考文献

[1] 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).

[2] 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).

[3] 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).

[4] 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).

[5] 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).

[6] 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).

[7] Hotelling, H.: Relations between two sets of variates, *Biometrika*, Vol. 28, No. 3/4, pp. 321–377 (1936).

[8] 大杉康仁，齋藤大輔，峯松信明：Eigenvoice と CLNF を用いた顔から声への統計的対応付けの検討，研究報告音声言語情報処理（SLP）(2017).

[9] 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).

[10] 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).

[11] 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).

[12] 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).