修士論文

顔の印象と声の印象の統計的対応付けと
それに基づく顔声変換

2018年 2月 1日

指導教員　峯松 信明 教授

電気系工学専攻

37-166438　大杉 康仁

内容梗概

古くからSF小説や映画で描かれてきたロボット・ヴァーチャルエージェントと人間との会話は今日チャットボットなどの文字対話システムやスマートスピーカーなどの音声対話システムによって一部実現され普及している．しかし，その会話は命令形式で入力・発話し簡単な機能を実現するだけにとどまっており，映画等で描かれているような自然かつ多様な会話はまだ実現していないと言える．それらの会話しか実現できておらず簡単な命令を実行させる以外の用途には使われていない原因の一つとして，文字対話システムや音声対話システムのエージェントのキャラクター性が薄いという問題がある．その問題を解決する方法の一つに，文字・音声・画像という複数のメディアを用いてエージェントを容易に想像できるようにする方法があり，今後はそのような音声対話システムが主流になると考えられる．既存の顔（声質）が決まっている対話システムをその形式に拡張するにはそのエージェントの声質（顔）を決める必要があるが，顔・声の対応関係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に最適な組合せを自動的に導くことは困難であり，対話システムの多様化の足かせとなりうる．そこで本研究では，顔と声の適切な組み合せを決める上で重要となるそれぞれの印象，特に顔の静的な個人性と音声の話者性に着目し，顔と声の対応関係をモデル化し実際に応用することを目的とする．本論文では，顔の印象を表す特徴量から対応する声の印象を表す特徴量へと統計的に変換しそれに基づいて音声を合成する顔声変換を検討した．初めに顔の特徴量として，(1) 顔の輪郭や目鼻の位置を表すFace Landmark・(2) 顔のパーツ形状に着目したIconified Face Feature (IFF)・(3) 画像圧縮と生成に利用可能なVariational Autoencoder (VAE)・(4) 人種や性別といったラベル情報を利用したConditional VAE (CVAE)の4通りの特徴量を検討した．次に声の特徴量として，話者の音韻分布の形状に基づく特徴量であるEigenvoiceを検討した．最後に統計的写像として，(a)Gaussian Mixture Model (GMM)・(b)Canonical Correlation Analysis (CCA)・(c)prpbabilistic CCA (pCCA)・(d)mixure of pCCA (mPCCA) をそれぞれ用いて顔の特徴量から声の特徴量へと変換する実験を行いそれぞれの統計モデルを比較した．その際，モデル学習・評価に必要な顔と声が対応付けられたパラレルコーパスも収集した．パラレルコーパスはアジア系男性の顔画像を使用したコーパス（アジア系コーパス）と日本人男性の顔画像を使用したコーパス（日本人コーパス）の2通りを収集した．結果として，コーパスの種類に関わらず，顔の印象を表す特徴量から声の印象を表す特徴量への変換には確率分布を仮定した変換法（GMM・pCCA・mPCCA），特にpCCAが有効に働くことが示された．IFFに基づく特徴量と3次元まで圧縮したCVAEが顔の印象を表す特徴量として有効である可能性とパラレルコーパス内の人種を統一することで声の特徴量への変換精度が向上する可能性が示唆された．

目次

図目次

表目次

第1章

序論

1.1 研究の背景

古くからSF小説や映画では人間がロボットやヴァーチャルエージェントと会話するシーンが描かれてきた．会話の形式はface-to-faceで行うものや文字のみ・音声のみのやり取りまで幅広く，そのシーンには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，文字のみでのやり取りはカスタマーセンターのチャットボット等の文字対話システムで実現され，音声のみでのやり取りはSiri [1]・Cortana [2]・しゃべってコンシェル [3]・Amazon Echo [4]・Google Home [5] などの様々な音声対話システム・音声インターフェースとして実現され普及している．特にAmazon EchoやGoogle Homeなどのスマートスピーカーの市場は今後さらに拡大すると見込まれる．しかし現在は，文字対話システムはFAQに答えるために，音声対話システムは音声検索や照明・テレビ等家電のコントロールなど簡単な機能をハンズフリーの形で行うために使用されることが多く，その際には人間が「提供されるサービスについて教えて.」「明日の天気は？」「電気をつけて.」「テレビを消して.」などの命令形式で入力・発話する会話が多いと考えられ，映画等で描かれているような自然かつ多様な会話はまだ実現していないと言える．単純な会話で簡単な命令を実行させる以外の用途には使われていない原因の一つとして，文字対話システムや音声対話システムのヴァーチャルエージェントのキャラクター性が薄く人間どうしのような会話を想定しづらいという点があると考えられる．

自然な会話に近づける方法の一つに，文字だけ・音声だけでなくその話者の顔も提示することで使用者（人間）がエージェントを容易に想像できるようにする方法がある．実際に，エージェントの顔や上半身画像が付与されている文字対話システム [6] や3Dモデルで全身がモデル化されたヴァーチャルエージェントを持つ音声対話システム [7] もあり，文字・音声・画像といった複数のメディアを組み合わせることでエージェントのキャラクター性を向上させようとする動きもあり，今後自然な会話の実現，特に映画で描かれるようなフランクかつ友人どうしのように機知に

[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]
[5] https://store.google.com/product/google_home?hl=ja [Accessed 19 January 2018]
[6] http://www.relia-group.com/ [Accessed 19 January 2018]
[7] https://www.nitech.ac.jp/mei/index.html [Accessed 19 January 2018]

富んだ会話を実現させるために複数のメディアを使用した音声対話システムが主流になると考えられる．

既存の対話システムをその形式へ拡張することを考えると，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（エージェントの顔画像が付与された文字対話システムの拡張）や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（音声対話システムの拡張）があり，どちらも顔と声（質）の適切な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，そのときにも無意識に顔と声の自然な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．この無意識の対応付けとは，論文著者欄などの写真でしか見た事のない人と初めて会うときや電話で話した相手に初めて会うときに,「この人はこのような声（顔）をしているのではないか」「このような顔（声）をしているのだからあのような声（顔）であってほしい」と頭の中で予想したときの顔・声の対応関係であり，アニメや洋画の吹き替えの声優を決めたり小説の登場人物を想像したりするときに無意識ではあるが日常的に触れているもしくは行っているものであると考えられる．しかし現状では，この対応付け（顔・声の関係性）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に，対話システム開発者がエージェントの声質と容姿の適切な組合せを簡単に決める事ができないという問題が生じており，これは対話システムの多様化の足かせとなりうる．

1.2　研究の目的

既存の文字対話システムや音声対話システムを複数メディアを用いた音声対話システムに自動的に拡張することができれば，システムの設計がより簡単になり多様な対話システムが開発されやすくなると考えられる．そのためには，人間が無意識に行っている顔・声の自然な対応付けを定量的かつ計算機利用可能な形式で取り出す必要がある．この対応付けとは顔から受ける印象と声を聞いたときの印象が合致するような対応付けであると考えられる．

そこで本研究では，顔・声の印象，特に顔の静的な個人性と音声の話者性に着目し，顔と声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．本論文では特に，顔の印象を表す特徴量と対応する声の印象を表す特徴量との間の統計的な対応付け（写像）について検討し，さらにそれに基づいて顔の特徴量から声の特徴量へと変換する顔声変換について検討する．初めに顔・声の印象の特徴量化について検討し，次に両者の写像について検討する．その際に統計モデルの学習に必要な顔・声が対応付けられたパラレルデータも収集する．

1.3　本論文の構成

本論文は全 5 章で形成される．第 1 章（本章）では研究の背景・目的と本論文の構成について述べ，第 2 章では顔・声の関係性や顔・声の特徴量化，異なる信号の変換・分析手法に関する関連研究について述べる．第 3 章で顔の印象を表す特徴量から声の印象を表す特徴量へと統計的に変換する顔声変換を提案し，第 4 章で提案手法を実験を通じて検証する．第 5 章で実験結果から得た成果と課題について述べ本論文をまとめる．

第2章

関連研究

2.1 はじめに

本研究では，顔から声への印象に基づく統計的対応付けとそれに基づいて顔の特徴量から声の特徴量へ変換する顔声変換について検討する．まず，2.2 節で顔・声の印象の関連性について述べ，2.3 節・2.4 節で顔・声の特徴量化についてそれぞれ述べる．最後に 2.5 節で顔・声の特徴量間の変換・分析に応用可能な異なる二つの信号に関する変換・分析法について述べる．

2.2 顔・声の関係性の調査

本研究では，顔の特徴量から声の特徴量へと統計的に変換する顔声変換について検討する．その顔声変換により，ある顔画像が与えられた時にその顔と印象が合致するような声質を持つ音声を得ることができる．しかし，顔画像から受ける印象と音声の声質から受ける印象の間に関連が無ければ，両者の統計的な対応付けは無意味なものとなる．そこで本節では，心理実験に基づいて顔と声の関係性，特に同一人物の顔・声の関係性について調査した研究を紹介する．

2.2.1 顔と声には共通する情報があるか

顔の印象・声の印象と体内に分泌されるホルモンとの関係性についてそれぞれ調査されている．例えば Penton-Voak らは，男性ホルモンの一種であるテストステロンの分泌量が多い男性は分泌量が少ない男性よりも他者から見て「男性らしい」と判断されたと報告している [1]．また，Abitbol らはテストステロンの量が声の男性らしさに影響を与えると報告している [2]．このように，顔の印象と声の印象は同一のホルモンから影響を受けている可能性がある．

顔の印象と声の印象の関係性についても調査されている．例えば，Collins らは女性の顔と声の魅力度を被験者に別々に評価させる主観実験を行い，その結果顔から受ける魅力度と声から受ける魅力度は似た値となった [3]．また Smith らは，同一人物の顔と声を別々に被験者に提示し，それぞれの男性らしさ・女性らしさを評価させる主観実験を行った [4]．結果として，顔から受けた男性らしさ・女性らしさと声から受けた男性らしさ・女性らしさに強い正の相関が見られた．

さらに，顔で表現された感情と声で表現された感情の間の関連性を調査した研究もある．田中らは喜びを表現する顔の動画と怒りを表現する音声を組合せた刺激を被験者に提示し，刺激から受ける感情を推定させる主観実験を行った [5]．その結果，顔が表現する感情に声が表現する感情が影響を及ぼすことがあり，その度合いには被験者の文化差が表われることも確認された．

表 2.1: 心理実験に基づく先行研究：提示形式の「連続」は選択肢を含む刺激の全てが順番に提示される方式であり,「同時」は刺激が提示された後に選択肢が同時に提示される方式である．

提示形式	動画（無音）	静画
連続	[4,6,7]	[4,8]
同時	-	[8,9]

これらの研究から，顔・声の印象の中でも男性らしさ・女性らしさ・魅力度・感情表現には関連があり，特に男性らしさ・女性らしさにはホルモンが影響を及ぼすことが確認された．よって，顔の印象と声の印象の間には少なくとも何らかの関係性があることが示唆された．次の節では，その関係性を明示的に使う事無く，同一人物の顔または声から他方を推定する心理実験に基づく研究を紹介する．

2.2.2　同一人物の顔または声から他方を推定できるか

これまで，同一人物の顔と声について，顔を見てその人の声を推定する，または声を聞いてその人の顔を推定するという心理実験はいくつか行われてきた．以下,「顔を見てその人の声を推定する」実験を **F-V 実験**,「声を聞いてその人の顔を推定する」実験を **V-F 実験**と呼ぶ．これらの実験の多くは二択問題を被験者に解かせる方式を取っており，その選択肢の内一つは必ず正解となっている．例えば F-V 実験の場合，選択肢の一つの音声が顔画像（動画）の被写体本人の音声でありもう一つの選択肢の音声は別人の音声となる．そのような心理実験に基づいて顔と声の関連性を調査した研究を刺激の提示形式と提示する顔が動画か静画かで分類した結果を表 2.1 に示す．ただし提示形式の内「連続」は，例えば F-V 実験において顔の動画または静画が提示された後に選択肢の一つ目の音声が流れ次に二つ目の音声が流れる，といったように選択肢を含む刺激の全てが順番に提示される方式である．また,「同時」は刺激が提示された後に選択肢が同時に提示される方式である．選択肢の一つは必ず正解である（同一人物から得られた顔と音声の組である）ことから，チャンスレベルである 50%以上の精度で回答できた場合に顔または声から他方を推定可能であると考えられる．しかし，Kamachi らは顔の刺激が静画ではなく動画でなければ推定は可能ではないとしている一方で [6]，Krauss らや Mavica らは静画でも推定は可能であるとしており矛盾する結果が得られている [8,9]．これは，それぞれの研究で独自のコーパス・提示方法を用いているため異なる結果が得られてしまったと考えられる．

そこで，Smith らは統一されたコーパスを用いて提示形式・顔の刺激の種類を変えた時に推定精度がどのように変化するか検証した [10]．Smith らはまず連続提示方式で動画と静画を比較し，静画を用いてもチャンスレベルを超えることはあったが動画を用いた場合の方が推定精度は高かったと報告した．次に顔の静画または無音の動画と音声を組合せ，一つの有声動画を作成しそれを選択肢とする提示方式を新たに提案し検証した．すなわち，F-V 実験においては選択肢の有声動画の被写体の顔は同一で音声は異なり，逆に V-F 実験においては有声動画の音声は同一で被写体の顔は異なる．このときも静画を用いた場合でチャンスレベルを超えることはあったものの，無音の動画と音声を組み合わせて作成した選択肢を提示した方が推定精度は高かった．最後に Smith らは静画を使用した場合の推定についてより検討するために，V-F 実験において音声刺激を提示した後に，選択肢である顔の静画二枚を同時に提示する方法で検証した．これは [8] と [9] でも取られた方法である．結果として推定精度は 61.0%となりチャンスレベルを超えたことが確認された．これらの結果から，Smith らは静画・動画のどちらを使用しても F-V・V-F 実験はチャンスレ

ベル以上の精度で推定可能な場合があり，静画・動画のどちらも顔・声に共通する情報を持ちうると結論付けた．動画を用いた場合，顔の動きの情報も推定に用いることができたため推定精度が向上した可能性がある．また，動画・静画のどちらを用いても選択肢を順番に提示する方法を取った場合はその提示順が推定精度に影響を及ぼすことがあり，特に静画は提示順に敏感であった．さらに，推定精度について，被験者に起因する分散よりも実験刺激に起因する分散の方が大きかったことから，F-V・V-F 実験のどちらも提示する刺激によって推定精度が変化する可能性が示唆された．

以上の先行研究から，F-V・V-F 実験のどちらも実験条件を丁寧に整えればチャンスレベル以上の精度で推定することが可能であると考えられる．しかし，F-V・V-F 実験を行う際の被験者の立場に立つと，被験者が選んだ選択肢は「この顔の持ち主はこの声だろう」（F-V 実験）「この声の持ち主はこの顔だろう」（V-F 実験）といった考えの元選ばれた答えであり，提示された顔・声の印象に基づく選択であると考えられる．すなわち，被験者が選んだ選択肢が正解であったか（同一人物の顔または声であったか）に関わらず，被験者は印象に基づく対応付けを自然に行った可能性が考えられる．本研究では，このような印象に基づく対応付けについて調査しそれを統計モデルを用いて計算機利用可能な形で取り出すことを検討する．

2.3　顔画像の特徴量化

本研究で扱う顔・声の印象に基づく統計的対応付けのためには，顔・声の印象を表す特徴量をそれぞれ設計することが必要となる．そこで本節では，まず一般に顔認識で用いられる特徴量について簡単に紹介した後，本研究で実際に検討した手法について詳しく紹介する．

2.3.1　顔認識において用いられる特徴量

一般的に，画像中の人物が誰であるか推定するには，まず画像中のどこに顔があるのかを検出し得られた顔画像から特徴量を抽出する必要があり，その後，画像にラベルが付与されている場合はそれを活用し，画像分類・同定を行うための分類器を学習する [11]．本節では，顔画像から特徴量を抽出する手法のうち，ID や属性を推定する過程を用いる方法・教師なし学習に基づく方法・輪郭や目鼻の位置に着目した方法について紹介する．

まず，各顔画像が誰かを表す ID が付与されたデータベースを用いて ID を顔画像から推定するような Deep Neural Network (DNN) を構成し，その中間層の内低次元の層で得られる特徴量 (bottleneck feature) を用いる研究がある．例えば，[12] では約 4000 人の ID を推定する DNN を，[13] では約 800 万人の ID を推定する DNN を構成し，それらの bottleneck feature をその顔画像の特徴量として定義している．特徴量を分類するには従来のクラスタリング手法が用いられることが多く，[14] では最近傍法が用いられており，[13] では bottleneck feature がユークリッド空間にあるものとしユークリッド距離が顔画像の類似度に直結するような制約を加えている．

また，画像中の被写体はどのような外見をしているか具体的に記述することに着目した研究もある．すなわち,「眼鏡をかけている」「細い目をしている」などの属性 (Facial Attributes) を画像から推定する研究である．[15] では，画像のどの部分 (ピクセル) が目鼻に対応しているかを推定した後その情報を入力として属性を推定している．また，[16] では写真ではなくイラストから髪の色・性別・何を持っているかなどのタグを推定している．一枚の画像について「眼鏡をかけている」などの絶対的な特徴を推定するのではなく，二枚の画像から「画像 1 は画像 2 より目が大きい」といった相対的な特徴を推定する研究もある [17].

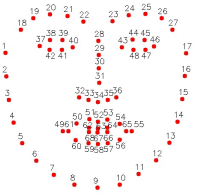


図 2.1: 68 点の Face Landmark

一方，ID や属性などの教師データを用いない教師なし学習に基づく研究も多く行われている．例えば，Turk らは画素値を主成分分析 (Principal Component Analysis: PCA) して画像の特徴を少数の基底で表現する Eigenface を提案した [18] （付録 A）．また，Autoencoder (AE) [19]・Variational Autoencoder (VAE) [20] を用いた顔画像の圧縮・復元により顔の特徴量を抽出する研究もある [21–24]．特に Rosca らは，Generative Adversarial Network (GAN) [25] と VAE を組合せて画像の生成分布を求める手法を提案した [24]．

これまで紹介した手法は画像中の顔が撮影されている部分のピクセル値全てを用いて特徴量を推定する手法であったが，顔の輪郭や目鼻の位置に重点的に着目した特徴量として Face Landmark がある [26, 27]．Face Landmark の位置関係から顔の向き・表情・視線の推定にも使用されている．Face Landmark の点数を増やす事で顔の形状をより細かく表現できるが，51 点や 68 点など数十点の数を用いる場合が多い．手動により推定する方法もあるが [28]，現在は機械学習を用いて推定する方法が主流である [29]．

以上のように，顔認識に用いられる特徴量には様々なものがある．ID や属性を推定する DNN の bottleneck feature を用いる手法では，ID や属性に対応した画像の特徴を上手く捉えることができるが，bottleneck feature の各次元が画像中のどの要素に対応しているのか分析しにくいという問題がある．一方 AE・VAE を用いた特徴量や Face Landmark は，前者は画像の生成分布を扱い後者は輪郭・目鼻の位置を直接扱っていることから，特徴量の分析が比較的容易であり主観的な特徴である顔の印象が画像によってどのように変化するか検証しやすいと考えられる．従って，本研究では Face Landmark や AE・VAE を用いた特徴量について検討し，次節からはそれらの抽出方法をいくつか紹介する．

2.3.2　Face Landmark

顔認識や顔検出の指標の一つに顔の輪郭や目鼻の位置を表す Face Landmark がある [26]．Face Landmark の数は様々だが，68 点の場合の Face Landmark を図 2.1 に示す．検出された Face Landmark の位置関係から顔の向きや目線の方向などを推定する事が可能である．Face Landmark を検出する方法の一つに Constrained Local Neural Fields (CLNF) がある [29]．CLNF は，式 (2.1) 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．

$$\boldsymbol{x}_i = s\boldsymbol{R}_{2D}(\boldsymbol{X}_i) + \boldsymbol{t} \tag{2.1}$$

$$\boldsymbol{X}_i = \overline{\boldsymbol{X}_i} + \boldsymbol{\Phi}_i \boldsymbol{q} \tag{2.2}$$

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 2 次元ベクトルである．まず，人間の顔が本来三次元空間に位置するものであることを利用しあらかじめ複数の 3 次元 Face Landmark に関する PCA を行い，得られた平均 $\overline{\boldsymbol{X}_i}$ と基底行列 $\boldsymbol{\Phi}_i$ を用いて三

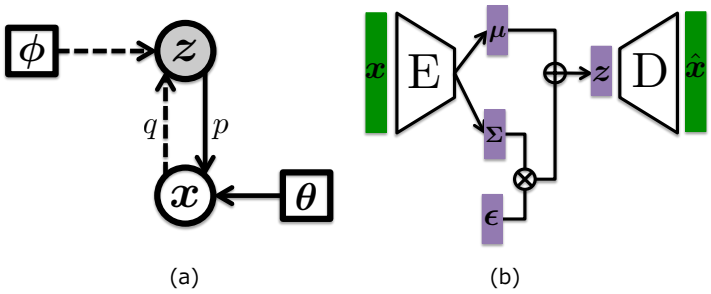


図 2.2: VAE の構成 : (a) グラフィカルモデル．白色の丸ノードは観測変数を，灰色の丸ノードは潜在変数を表し，四角形のノードはパラメータを表す．実線は真の生成確率 $p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z})$ を表し，点線は真の事後確率 $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ の変分近似 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ を表す．(b) ネットワークの構造．E は Encoder を，D は Decoder をそれぞれ表す．

次元空間上の Face Landmark の位置 $\boldsymbol{X}_i$ を式 (2.2) で表す．ただし，$\boldsymbol{q}$ は各 Face Landmark について同一である．次に，$\boldsymbol{X}_i$ を $\boldsymbol{R}_{2D}$ で回転しさらに三次元から二次元へ射影し，s と $\boldsymbol{t}$ で拡大縮小と平行移動を行うことで二次元画像中の Face Landmark の位置 $\boldsymbol{x}_i$ を得る．このように，PDM は N 点の点群である Face Landmark に対し人間の顔の形状としての制約を加えている．CNLF では，特徴点があると考えられる範囲 (パッチ) に対し層の浅い Neural Network (NN) を適用し Face Landmark の存在確率を求め．それに基づいて $\boldsymbol{p}$ を求める．ただし，NN を通して得られる確率には雑音が混ざっているためパッチ内部の信頼度を特徴点算出に用いる．このとき，NN のパラメータは手動で Face Landmark の位置が与えられたデータを用いてあらかじめ学習しておく．

2.3.3　Variational Autoencoder (VAE)

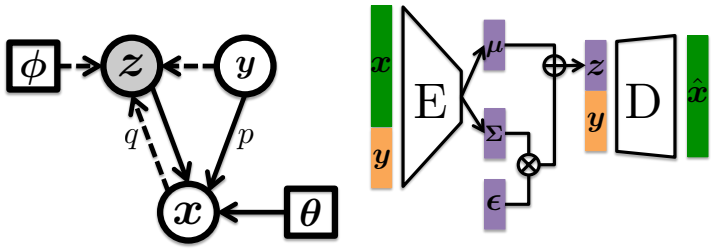
Autoencoder とは，m 次元の情報を n 次元の情報へ圧縮（拡張）する手法の一つであり，m 次元から n 次元への圧縮（拡張）を行う Encoder と n 次元から m 次元への復元を行う Decoder を主に NN を用いて同時に学習する手法である [19]．Variational Autoencoder は Encoder・Decoder が確率分布を表す Autoencoder であり，そのグラフィカルモデルを図 2.2(a) に示す [20]．入力 $\boldsymbol{x}$ に対し圧縮（拡張）先の潜在変数 $\boldsymbol{z}$ を仮定し，事後分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ を Encoder で生成分布 $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ を Decoder で表す．このとき，事前分布 $p_\theta(\boldsymbol{z})$ は標準正規分布であり，$q_\phi(\boldsymbol{z}|\boldsymbol{x})$ は対角な共分散行列を持つ正規分布とする場合が多く，$\boldsymbol{x}$ が画像のとき式 (2.7) の $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ は多変量ベルヌーイ分布に従うものとする場合が多い．すなわち，NN で構成される Encoder(·) と Decoder(·) を用いて次のように表される．

$$\boldsymbol{z} = \text{Encoder}(\boldsymbol{x}) \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.3}$$

$$\hat{\boldsymbol{x}} = \text{Decoder}(\boldsymbol{z}) \sim p_\theta(\boldsymbol{x}|\boldsymbol{z}) \tag{2.4}$$

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{2.5}$$

図 2.2(b) に示すように，Encoder の NN は $\boldsymbol{x}$ を入力，$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を出力とし，Decoder の NN は $\boldsymbol{z}$ を入力，$\hat{\boldsymbol{x}}$ を出力とする構造を持つ．ただし，$\boldsymbol{z}$ は $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ を用いて $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\epsilon}$ のようにサンプリングされる．これらの NN のパラメータは以下の損失関数 L_{vae} を最小化することで求めら



(a)　　(b)

図 2.3: CVAE の構成：(a) グラフィカルモデル．白色の丸ノードは観測変数を，灰色の丸ノードは潜在変数を表し，四角形のノードはパラメータを表す．実線は真の生成確率 $p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})p_\theta(\boldsymbol{y})p_\theta(\boldsymbol{z})$ を表し，点線は真の事後確率 $p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$ の変分近似 $q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$ を表す．(b) ネットワークの構造．E は Encoder を，D は Decoder をそれぞれ表す．

れる．

$$L_{vae} = L_{rec} + L_{kl} \tag{2.6}$$
$$L_{rec} = -E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \tag{2.7}$$
$$L_{kl} = D_{kl}[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})] \tag{2.8}$$

2.3.4　Conditional VAE (CVAE)

Conditional VAE (CVAE) は，ラベル付けされた小規模なデータを元にラベル付けされていない大規模なデータについても特徴量を抽出しようとする半教師有り学習の一つである [30]．そのグラフィカルモデルは図 2.2(a) の通りであり，入力 $\boldsymbol{x}$ とラベル $\boldsymbol{y}$ に対し潜在変数 $\boldsymbol{z}$ を仮定する．このとき，Encoder(·) と Decoder(·) には図 2.2(b) のように $\boldsymbol{x}$ と $\boldsymbol{y}$ の両方が入力され，それぞれ以下の確率分布を表す．

$$\boldsymbol{z} = \text{Encoder}(\boldsymbol{x},\boldsymbol{y}) \sim q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{2.9}$$
$$\hat{\boldsymbol{x}} = \text{Decoder}(\boldsymbol{y},\boldsymbol{z}) \sim p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) \tag{2.10}$$
$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}) \tag{2.11}$$

CVAE も式 (2.6) と同様に式 (2.12) の損失関数を最小化することで学習される．

$$L_{cvae} = L_{rec}^{(cvae)} + L_{kl}^{(cvae)} \tag{2.12}$$
$$L_{rec}^{(cvae)} = -E_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})] \tag{2.13}$$
$$L_{kl}^{(cvae)} = D_{kl}[q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})||p_\theta(\boldsymbol{z})] \tag{2.14}$$

2.4　音声話者の特徴量化

本研究で扱う顔・声の印象に基づく統計的対応付けのためには，顔・声の印象を表す特徴量をそれぞれ設計することが必要となる．そこで本節では，まず一般に話者認識で用いられる特徴量について簡単に紹介した後，本研究で実際に検討した手法について詳しく紹介する．

2.4.1　話者認識において用いられる特徴量

音声にはその発話内容（言語情報）以外にも感情・発話意図・話者の性別など様々な情報が含まれている [31]．音声に含まれる情報の内「誰が発話したか」という情報を音声から推定する問題が話者認識であり，それは入力音声からの特徴量抽出とその識別という二つのプロセスに大きく分ける事ができる [32, 33]．この内，本節では前者の特徴量抽出について扱う．

近年話者認識について様々な研究がなされているが，音声波形そのものが特徴量として用いられることは少なく，まず音声波形から言語情報や話者情報を含むと考えられる声道成分をメル周波数ケプストラム係数 (Mel-Frequency Cepstrum Coefficient: MFCC) やメルケプストラムとして抽出し [34]，さらにそれらから言語情報を分離する手法がとられることが多い．中でも，様々な発話内容を含む音声から話者がどのような発音をしているかを表す音韻分布を推定し，その分布形状から話者情報を抽出することが検討されてきた．例えば，Reynolds らは複数発話から抽出された MFCC が混合正規分布 (Gaussian Mixture Model: GMM) に従うものとして話者の音韻空間を表しその最尤基準による識別を提案した [32]．ここで，GMM とは複数の正規分布に対し全積分が 1 となるような重みを付与した確率モデルである．確率変数 d 次元ベクトル $\boldsymbol{x}$ が混合数 M の GMM に従うとき，m 番目の分布の平均ベクトル $\boldsymbol{\mu}_m$・分散共分散行列 $\boldsymbol{\Sigma_m}$ を用いて $\boldsymbol{x}$ の確率分布 $p(\boldsymbol{x})$ は式 (2.15) で表される．

$$p(\boldsymbol{x}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{2.15}$$

$$\sum_{m=1}^{M} \alpha_m = 1 \tag{2.16}$$

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_m|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_m)^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_m)\right) \tag{2.17}$$

同一話者の複数発話から抽出された MFCC 系列を用いて推定された GMM は話者情報が統一されているため，各分布が音素（ある言語においてその理解者が区別可能な最小単位）を表していると期待できる．そこで，GMM のパラメータ $\lambda = \{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \alpha_m\}_{m=1}^{M}$ のうち $\boldsymbol{\mu}_m$ に着目し，式 (2.18) のようにそれらを連結した特徴量 GMM super vector (GMM-SV) が登場した [35].

$$\boldsymbol{\nu} = [\boldsymbol{\mu}_1^{\mathrm{T}}, \boldsymbol{\mu}_2^{\mathrm{T}}, ..., \boldsymbol{\mu}_M^{\mathrm{T}}]^{\mathrm{T}} \tag{2.18}$$

ただし，話者認識は事前に登録された複数話者の中から入力音声の発話者を推定する問題として解かれることが多く，入力音声が短時間かつ限られた音素しか含まないという制限が生じる．さらに，話者ごとに GMM を学習してしまうと GMM の各分布が表すと期待される音素は話者によって異なる可能性がある．そこで，事前に大量の不特定話者の音声データから一つの GMM を Universal Background Model (UBM) として学習しておき，求めたい話者の音韻分布を UBM を事前分布とした最大事後確率基準による推定（MAP 推定）を行うことで，小規模なデータから音素環境が統一された話者 GMM-SV を得る手法が提案された [36]．GMM-SV はそのままでは次元数が大きく冗長であるため，因子分析 (Factor Analysis: FA) や PCA によって次元圧縮されることが多く，前者は i-vector として [37]，後者は Eigenvoice として [38] それぞれ提案されている．特に i-vector は近年話者情報を表す特徴量として標準的に用いられている．

一方，2.3 節と同様に DNN を用いて MFCC 等の音響特徴量から話者情報を抽出することも検討されている．例えば，Snyder らは音声の短時間部分（フレーム）に関する処理を行う部分と発

話全体に関する処理を行う部分を持つ DNN の bottleneck feature を話者特徴量とする手法を提案し，短時間発話での話者認識性能が i-vector と同程度得られたと報告している [39]．しかし，このような DNN の bottleneck feature の各次元がどのような話者の声質を反映しているか分析しにくく，本研究の対象である声の印象に関する特徴量を抽出することが困難となる可能性がある．一方で，i-vector と Eigenvoice は話者の音韻空間のモデル化を行うため話者の声質の分析が比較的容易である可能性がある．そこで，本研究では GMM-SV に基づく i-vector と Eigenvoice について検討することとし，次節からそれぞれを詳細に紹介する．

2.4.2　i-vector

i-vector は UBM の GMM-SV と話者 GMM の GMM-SV の差に関する因子分析から得られる話者特徴量の一つであり，話者認識における標準的な特徴量として現在用いられている [37]．すなわち，UBM の GMM-SV を $\boldsymbol{\nu}^{(0)}$，対象の発話を用いた MAP 推定で得られた話者 s の GMM-SV を $\boldsymbol{\nu}^{(s)}$ としたとき，式 (2.19) の $\boldsymbol{w}^{(s)}$ が i-vector である．

$$\boldsymbol{\nu}^{(s)} = \boldsymbol{\nu}^{(0)} + \boldsymbol{T}\boldsymbol{w}^{(s)} \tag{2.19}$$

ただし，入力発話が小規模の場合は $\boldsymbol{\nu}^{(s)}$ には話者の違いだけでなく収録環境の違い（チャネル情報）も含まれるため，$\boldsymbol{T}$ は話者情報・チャネル情報を表す写像行列となる．多くの場合 GMM 学習に用いる音響特徴量は MFCC であり，異なる話者 x・y の類似度は式 (2.20) のコサイン類似度で表現される．

$$\mathrm{score}(\boldsymbol{w}^{(x)}, \boldsymbol{w}^{(y)}) = \frac{\boldsymbol{w}^{(x)} \cdot \boldsymbol{w}^{(y)}}{|\boldsymbol{w}^{(x)}||\boldsymbol{w}^{(y)}|} \tag{2.20}$$

2.4.3　Eigenvoice

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であり，話者の音韻空間を GMM で表現しその主成分分析により話者空間を得ることができる [38]．i-vector と同様に，複数話者の音声データから UBM を学習し，話者ごとの音声データを用いた適応により話者 GMM を学習する．話者 s の GMM-SV $\boldsymbol{\nu}^{(s)}$ $(s = 1, 2, ..., S)$ に対する PCA により，話者 s の Eigenvoice は式 (2.21) の $w_i^{(s)}$ と定義される．

$$\boldsymbol{\nu}^{(s)} \simeq \sum_{i=1}^{K} w_i^{(s)} \boldsymbol{b}(i) + \boldsymbol{b}(0) \tag{2.21}$$

$$\boldsymbol{b}(0) = \frac{1}{S} \sum_{s=1}^{S} \boldsymbol{\nu}^{(s)} \tag{2.22}$$

ただし $K < S$ であり，$\boldsymbol{b}(i)$ は PCA における固有ベクトルを，$\boldsymbol{b}(0)$ は式 (2.22) のように平均ベクトルを表す．$\boldsymbol{b}(0) \cdot \boldsymbol{b}(i)$ について分布 m に対応するベクトルを $\boldsymbol{b}_m(0) \cdot \boldsymbol{b}_m(i)$ とすると式 (2.21) の分布 m に対応する部分は式 (2.24) で表される．

$$\boldsymbol{\mu}_m^{(s)} = \sum_{i=1}^{K} w_i^{(s)} \boldsymbol{b}_m(i) + \boldsymbol{b}_m(0) \tag{2.23}$$

$$= \boldsymbol{B}_m \boldsymbol{w}^{(s)} + \boldsymbol{b}_m^{(0)} \tag{2.24}$$

ただし，$\boldsymbol{B}_m$ と $\boldsymbol{w}^{(s)}$ は以下である．

$$\boldsymbol{B}_m = [\boldsymbol{b}_m(1), ..., \boldsymbol{b}_m(K)] \tag{2.25}$$

$$\boldsymbol{w}^{(s)} = [w_1^{(s)}, ..., w_K^{(s)}]^{\mathrm{T}} \tag{2.26}$$

$$\boldsymbol{w}^{(s)} = [w_1^{(s)}, ..., w_K^{(s)}]^{\mathrm{T}} \tag{2.27}$$

[38] では GMM 学習における音響特徴量として不可逆的処理を含む Perceptual Linear Prediction (PLP) [40] に基づく特徴量を用いていたが，戸田らは音響特徴量としてメルケプストラムを用いさらに 2.5.1 節で述べる GMM に基づく声質変換と組み合わせた Eigenvoice Conversion (EVC) を提案した [41]（付録 B）．EVC により指定した重み $\boldsymbol{w}$ を持つ話者の音声を合成可能となるため，$\boldsymbol{w}$ の違いを音声を聞く事で検証できるだけでなく，声質変換の分野で用いられる式 (2.28) のメルケプストラムひずみ (Mel-cepstral distortion: MCD) によって客観的に評価することも可能である．

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}\left(c_d^{(s_1)} - c_d^{(s_2)}\right)^2} \tag{2.28}$$

ここで $c_d^{(s_1)}$・$c_d^{(s_2)}$ はそれぞれ話者 s_1・s_2 のメルケプストラムの d 次元目の要素であり，D はメルケプストラムの次元数を表す．

2.4.4　i-vector と Eigenvoice の違い

2.4.2 節の i-vector は UBM と話者 GMM の差をモデル化し，2.4.3 節の Eigenvoice は複数話者の GMM 間の差をモデル化している点で異なるが，共に複数話者の音声から学習した UBM から適応して得られる話者 GMM-SV に関する PCA として捉えることができる [33]．従って本研究では，話者空間での違い（重み $\boldsymbol{w}$ の違い）を実際に音声を聞いて知覚する事ができ，かつ話者の声質の差を MCD で定量的に評価することができる Eigenvoice について検討する．

2.5　異なる信号の変換・分析手法

本研究で扱う顔・声の印象に基づく統計的対応付及びそれに基づく顔声変換のためには，2.3 節で述べたような顔の特徴量と 2.4 節で述べたような声の特徴量の間の関係性を分析しそれに基づいて変換することが必要となる．すなわち，顔の特徴量空間から声の特徴量空間への適切な写像を求める必要がある．そこで本節では，ある信号から別の信号への変換・分析を行うための手法について紹介する．

2.5.1　Gaussian Mixture Model (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，ここでは変換写像として GMM を用いた手法を紹介する [42]．

時刻 t における入力・出力話者の音声の静的・動的特徴量をそれぞれ $\boldsymbol{X}_t = [\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{\Delta x}_t^{\mathrm{T}}]$，$\boldsymbol{Y}_t = [\boldsymbol{y}_t^{\mathrm{T}}, \boldsymbol{\Delta y}_t^{\mathrm{T}}]$ $(t = 1, 2, ..., T)$ 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 (2.30) 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．

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{t=1}^{T} p(\boldsymbol{Z}_t|\lambda) \tag{2.29}$$

$$= \arg\max_{\lambda} \prod_{t=1}^{T} \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma_m}^{(Z)}) \tag{2.30}$$

$$\boldsymbol{\mu}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \ \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{2.31}$$

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．

$$\hat{\boldsymbol{Y}} = \arg\max_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\lambda}) \tag{2.32}$$

$$= \arg\max_{\boldsymbol{Y}} \prod_{t=1}^{T} \sum_{m=1}^{M} p(m|\boldsymbol{X}_t, \hat{\lambda}) p(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \hat{\lambda}) \tag{2.33}$$

ここで、フレーム t の入力話者特徴量 $\boldsymbol{X}_t$ がある分布 m から出力される確率は式 (2.34) で表され，分布 m から $\boldsymbol{X}_t$ が出力される場合に，出力話者特徴量 $\boldsymbol{Y}_t$ が出力される確率は式 (2.35) で表される．

$$p(m|\boldsymbol{X}_t, \hat{\lambda}) = \frac{\alpha_m \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(XX)}, \boldsymbol{\Sigma}_m^{(XX)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_n^{(XX)}, \boldsymbol{\Sigma}_n^{(XX)})} \tag{2.34}$$

$$p(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \hat{\lambda}) = \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{E}_{m,t}^{(Y)}, \boldsymbol{D}_m^{(Y)}) \tag{2.35}$$

ただし、$\boldsymbol{E}_{m,t}^{(Y)}$ と $\boldsymbol{D}_m^{(Y)}$ は以下で表される．

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} (\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}) \tag{2.36}$$

$$\boldsymbol{D}_m^{(Y)} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} \boldsymbol{\Sigma}_m^{(XY)} \tag{2.37}$$

式 (2.32) は式 (2.38) 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 (2.39) で表される．

$$Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{t=1}^{T} \sum_{m=1}^{M} p(m|\boldsymbol{X}_t, \boldsymbol{Y}_t, \lambda) \log p(\hat{\boldsymbol{Y}}_t, m|\boldsymbol{X}_t, \lambda) \tag{2.38}$$

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^{\mathrm{T}} \overline{\boldsymbol{D}^{(Y)-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^{\mathrm{T}} \overline{\boldsymbol{D}^{(Y)-1} \boldsymbol{E}^{(Y)}} \tag{2.39}$$

$$\overline{\boldsymbol{D}^{(Y)-1}} = diag\left[\overline{\boldsymbol{D}_1^{(Y)-1}}, ..., \overline{\boldsymbol{D}_T^{(Y)-1}}\right] \tag{2.40}$$

$$\overline{\boldsymbol{D}^{(Y)-1} \boldsymbol{E}^{(Y)}} = \left[\overline{\boldsymbol{D}_1^{(Y)-1} \boldsymbol{E}_1^{(Y)}}^{\mathrm{T}}, ..., \overline{\boldsymbol{D}_T^{(Y)-1} \boldsymbol{E}_T^{(Y)}}^{\mathrm{T}}\right] \tag{2.41}$$

$$\overline{\boldsymbol{D}_t^{(Y)-1}} = \sum_{i=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)-1} \tag{2.42}$$

$$\overline{\boldsymbol{D}_t^{(Y)-1} \boldsymbol{E}_t^{(Y)}} = \sum_{i=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)-1} \boldsymbol{E}_{m,t}^{(Y)} \tag{2.43}$$

$$\gamma_{m,t} = P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t, \lambda) \tag{2.44}$$

ただし，$\boldsymbol{W}$ は静的特徴量系列 $\boldsymbol{y} = [\boldsymbol{y}_1^{\mathrm{T}}, ..., \boldsymbol{y}_T^{\mathrm{T}}]^{\mathrm{T}}$ を静的・動的特徴量系列 $\boldsymbol{Y} = [\boldsymbol{Y}_1^{\mathrm{T}}, ..., \boldsymbol{Y}_T^{\mathrm{T}}]^{\mathrm{T}}$ に変換する行列である．

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{2.45}$$

2.5.2 Canonical Correlation Analysis (CCA)

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 [43]．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．$\boldsymbol{x}_n$ と $\boldsymbol{y}_n$ をそれぞれ以下の式でスカラー量 u_n と v_n へと線形変換する．

$$u_n = \boldsymbol{a}^{\mathrm{T}}(\boldsymbol{x}_n - E[\boldsymbol{x}]) \tag{2.46}$$

$$v_n = \boldsymbol{b}^{\mathrm{T}}(\boldsymbol{y}_n - E[\boldsymbol{y}]) \tag{2.47}$$

ただし，$E[\cdot]$ はサンプル平均を表す．u_n と v_n が $\boldsymbol{x}_n$ と $\boldsymbol{y}_n$ との間の共通因子として捉え，u と v の相関係数 $\rho(\boldsymbol{a}, \boldsymbol{b})$ を最大にするようなベクトル $\boldsymbol{a}$ と $\boldsymbol{b}$，すなわち $\hat{\boldsymbol{a}}$ と $\hat{\boldsymbol{b}}$ を求める．

$$\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}} = \arg\max_{\boldsymbol{a}, \boldsymbol{b}} \rho(\boldsymbol{a}, \boldsymbol{b}) \tag{2.48}$$

$$= \arg\max_{\boldsymbol{a}, \boldsymbol{b}} \boldsymbol{a}^{\mathrm{T}} \boldsymbol{\Sigma}_{xy} \boldsymbol{b} \tag{2.49}$$

ただし，$\boldsymbol{\Sigma}_{xx} = E[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]$，$\boldsymbol{\Sigma}_{xy} = E[\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}]$，$\boldsymbol{\Sigma}_{yy} = E[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}]$ であり，式 (2.50) の制約条件を加えている．

$$\boldsymbol{a}^{\mathrm{T}} \boldsymbol{\Sigma}_{xx} \boldsymbol{a} = \boldsymbol{b}^{\mathrm{T}} \boldsymbol{\Sigma}_{yy} \boldsymbol{b} = 1 \tag{2.50}$$

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 (2.46) と式 (2.47) は u_n と v_n がスカラー量なので一次元への射影を表しているが，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．

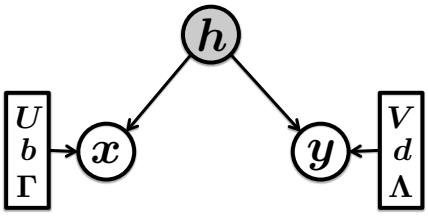


図 2.4: pCCA のグラフィカルモデル：白色の丸ノードは観測変数を，灰色の丸ノードは潜在変数を表し，四角形のノードはパラメータを表す

2.5.3　probabilistic CCA (pCCA)

確率的正準相関分析（probabilistic CCA: pCCA）は，2.5.2 節の CCA に確率的な潜在変数を仮定した分析法である [44]．pCCA のグラフィカルモデルは図 2.4 のようになり，通常の CCA における共通因子を確率的潜在変数 $\boldsymbol{h}$ で表している．観測変数 $\boldsymbol{x}$，$\boldsymbol{y}$ と潜在変数 $\boldsymbol{h}$ が従う確率分布を以下のように定義する．

$$p(\boldsymbol{h}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{2.51}$$

$$p(\boldsymbol{x}|\boldsymbol{h}) = \mathcal{N}(\boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}, \boldsymbol{\Gamma}) \tag{2.52}$$

$$p(\boldsymbol{y}|\boldsymbol{h}) = \mathcal{N}(\boldsymbol{V}\boldsymbol{h} + \boldsymbol{d}, \boldsymbol{\Lambda}) \tag{2.53}$$

よって，同時確率 $p(\boldsymbol{x}, \boldsymbol{y})$ は式 (2.54) となる．

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.54}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{d} \end{bmatrix} \tag{2.55}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} + \boldsymbol{\Gamma} & \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} \\ \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}} & \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{\Lambda} \end{bmatrix} \tag{2.56}$$

Bach らは，モデルパラメータ $\boldsymbol{\theta} = \{\boldsymbol{U}, \boldsymbol{b}, \boldsymbol{\Gamma}, \boldsymbol{V}, \boldsymbol{d}, \boldsymbol{\Lambda}\}$ を以下のように解析的に求める方法を提案した [44]．

$$\boldsymbol{U} = \boldsymbol{\Sigma}_{xx}\boldsymbol{A}\boldsymbol{M} \tag{2.57}$$

$$\boldsymbol{V} = \boldsymbol{\Sigma}_{yy}\boldsymbol{B}\boldsymbol{M} \tag{2.58}$$

$$\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} \tag{2.59}$$

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \tag{2.60}$$

$\boldsymbol{b}$ と $\boldsymbol{d}$ はそれぞれ，$\boldsymbol{x}$，$\boldsymbol{y}$ のサンプル平均で，$\boldsymbol{\Sigma}_{xx}$ と $\boldsymbol{\Sigma}_{yy}$ はサンプルの分散行列で求められる．ただし，$\boldsymbol{A}$ と $\boldsymbol{B}$ は CCA で得られた各 $\hat{\boldsymbol{a}}$，$\hat{\boldsymbol{b}}$ を行ベクトルとする変換行列であり，$\boldsymbol{M}$ は，CCA で得られた固有値が対角に並んだ行列を $\boldsymbol{P}$ としたときに $\boldsymbol{M} = \boldsymbol{P}^{\frac{1}{2}}$ で表される．

2.5.4　mixture of probabilistic CCA (mPCCA)

mixture of probabilistic CCA (mPCCA) は，pCCA の混合モデルに基づく分析法である [45]．mPCCA のグラフィカルモデルは図 2.5 のようになる．潜在変数 $\boldsymbol{z}$ は混合を表すベクトルであり，

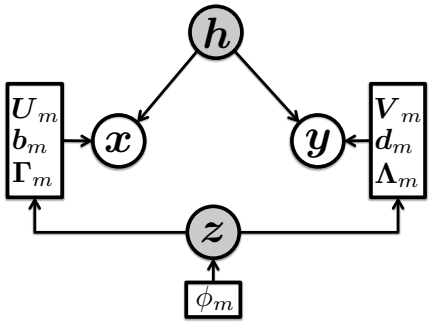


図 2.5: mPCCA のグラフィカルモデル : 白色の丸ノードは観測変数を，灰色の丸ノードは潜在変数を表し，四角形のノードはパラメータを表す

ここでは $\boldsymbol{z}$ が m 次元目の要素 z_m を 1 とする one-hot ベクトルである確率 $p(z_m = 1)$ を単に $p(z_m)$ と表すこととする．それぞれの変数は以下の分布に従う．

$$p(\boldsymbol{h}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{2.61}$$

$$p(\boldsymbol{x}|\boldsymbol{h}, z_m) = \mathcal{N}(\boldsymbol{U}_m\boldsymbol{h} + \boldsymbol{b}_m, \boldsymbol{\Gamma}_m) \tag{2.62}$$

$$p(\boldsymbol{y}|\boldsymbol{h}, z_m) = \mathcal{N}(\boldsymbol{V}_m\boldsymbol{h} + \boldsymbol{d}_m, \boldsymbol{\Lambda}_m) \tag{2.63}$$

$$p(z_m) = \phi_m \tag{2.64}$$

従って，同時確率は以下のように表される．

$$p(\boldsymbol{x}, \boldsymbol{y}) = \int \sum_m p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}, z_m) \mathrm{d}\boldsymbol{h} \tag{2.65}$$

$$= \int \sum_m \{p(\boldsymbol{x}|\boldsymbol{h}, z_m) p(\boldsymbol{y}|\boldsymbol{h}, z_m) p(\boldsymbol{h}) p(z_m)\} \, \mathrm{d}\boldsymbol{h} \tag{2.66}$$

$$= \sum_m \phi_m \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{2.67}$$

ただし，$\boldsymbol{\mu}_m$ と $\boldsymbol{\Sigma}_m$ は以下で表される．

$$\boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{b}_m \\ \boldsymbol{d}_m \end{bmatrix} \tag{2.68}$$

$$\boldsymbol{\Sigma}_m = \begin{bmatrix} \boldsymbol{U}_m\boldsymbol{U}_m^{\mathrm{T}} + \boldsymbol{\Gamma}_m & \boldsymbol{U}_m\boldsymbol{V}_m^{\mathrm{T}} \\ \boldsymbol{V}_m\boldsymbol{U}_m^{\mathrm{T}} & \boldsymbol{V}_m\boldsymbol{V}_m^{\mathrm{T}} + \boldsymbol{\Lambda}_m \end{bmatrix} \tag{2.69}$$

混合 m に関するモデルパラメータは，$\boldsymbol{\Theta}_m = \{\boldsymbol{U}_m, \boldsymbol{b}_m, \boldsymbol{\Gamma}_m, \boldsymbol{V}_m, \boldsymbol{d}_m, \boldsymbol{\Lambda}_m, \phi_m\}$ となる．

混合数 M の mPCCA のパラメータ $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_m\}_{m=1}^M$ の学習は，N 対の学習サンプル $\boldsymbol{X} = [\boldsymbol{x}_1^{\mathrm{T}}, ..., \boldsymbol{x}_N^{\mathrm{T}}]^{\mathrm{T}} \cdot \boldsymbol{Y} = [\boldsymbol{y}_1^{\mathrm{T}}, ..., \boldsymbol{y}_N^{\mathrm{T}}]^{\mathrm{T}}$ を用いて式 (2.71) で表され EM アルゴリズムによって求めることができる．

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{X}, \boldsymbol{Y}) \tag{2.70}$$

$$= \arg\max_{\boldsymbol{\Theta}} \int \sum_{m=1}^{M} p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{H}, z_m) \mathrm{d}\boldsymbol{H} \tag{2.71}$$

ただし，$\boldsymbol{H} = [\boldsymbol{h}_1^{\mathrm{T}}, ..., \boldsymbol{h}_N^{\mathrm{T}}]^{\mathrm{T}}$ である．

学習された $\hat{\boldsymbol{\Theta}}$ を用いて，ある $\boldsymbol{x}$ から対応する $\boldsymbol{y}$ への変換は式 (2.73) で表される．

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \ln p(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{\Theta}}) \tag{2.72}$$

$$= \arg\max_{\boldsymbol{y}} \ln \left[\int \sum_{m=1}^{M} p(\boldsymbol{y}, \boldsymbol{h}, z_m|\boldsymbol{x}, \hat{\boldsymbol{\Theta}}) \mathrm{d}\boldsymbol{h} \right] \tag{2.73}$$

これも EM アルゴリズムを用いて解くことができ，M ステップにおける更新式は式 (2.74) である．

$$\hat{\boldsymbol{y}} = \left(\sum_m \boldsymbol{G}_m \right)^{-1} \sum_m \boldsymbol{G}_m \left(\boldsymbol{\psi}_m + \boldsymbol{d}_m \right) \tag{2.74}$$

ただし，$\boldsymbol{G}_m$ と $\boldsymbol{\psi}_m$ は以下であり，$\langle \cdot \rangle$ は期待値を表す．

$$\boldsymbol{G}_m = g_m \boldsymbol{\Lambda}_m^{-1} \tag{2.75}$$

$$\boldsymbol{\psi}_m = \boldsymbol{V}_m \left\langle \boldsymbol{h}|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}, z_m \right\rangle \tag{2.76}$$

$$g_m = \left\langle z_m|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old} \right\rangle \tag{2.77}$$

ここで，$\hat{\boldsymbol{y}}^{old}$ は直前の M ステップで得られた出力 $\hat{\boldsymbol{y}}$ を表す．

第3章

提案手法

3.1 はじめに

本研究では，顔の印象を表す特徴量と声の印象を表す特徴量を統計的に対応付け，それに基づいて顔の特徴量から声の特徴量へと変換する顔声変換について検討する．図 3.1 に示すように提案手法は大きく三つの処理に分けることができる．一つ目は顔の印象を表す特徴量の抽出であり，二つ目は声の印象を表す特徴量の抽出であり，三つ目は顔の特徴量と声の特徴量の間の写像の学習と評価である．以下でそれぞれの処理の検討について述べるが，ここで本研究で扱う顔・声の印象について再考する．ある顔や声から受ける印象を人間が表現するときに使う最も簡単な方法は「彫りが深い」「丸顔だ」「太い」「優しい」などの言葉（印象語）を用いる方法である．例えば高椋らは「高い/低い」「大きい/小さい」などの簡単な言葉・「明るい/暗い」「暖かい/冷たい」などのより抽象的な言葉・「繊細」「素っ気ない」「優しい」「可愛い」などの主観的な言葉を用いて音声の印象を評価した [46–49]．また，個人の顔の印象ではないものの，永田らは「プロレスラー」や「東京大学の学生」などの所属する職業名またはグループ名をその人が所属する集団に共通する印象として採用し，顔の印象に関する分析を行った [50]．主観的な言葉も含む印象語で顔・声の印象を評価するときには，[49] のように数多くの印象語を用いたアンケート調査とその因子分析により低次元の特徴量に圧縮されることが多い．我々がある顔・声の印象を実際に言葉で表すとき（例えば友人に恋人の顔・声を伝えるときなど）を考えると，用いる言葉は人それぞれでも表す印象の方向性は限られているはずであり，因子分析などを通じて低次元の特徴量に圧縮することは妥当であると言える．このことから，顔・声の印象は比較的低次元で表現可能であるという仮定を立てることができ，これは印象語を用いた分析だけでなく，顔画像・音声から機械学習により得られる定量的な特徴量を用いた分析にも適用できると考えられる．そこで，本研究では顔・声の印象を表す特徴量の次元数は比較的低次元であると仮定し特徴量の選定基準の一つとする．次節以降で提案手法の各処理について述べる．

3.2 顔の印象を表す特徴量の抽出

顔の特徴量の抽出として 2.3 節で紹介した Face Landmark・VAE・CVAE を検討する．VAE・CVAE については，Encoder で得られる $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ のうち $\boldsymbol{\mu}$ の方が潜在変数の性質をより表すと考え，$\boldsymbol{\mu}$ のみを特徴量として用いる．Face Landmark については，複数の点が互いに関係している可能性を考慮し PCA を行い Face Landmark の主成分を特徴量として用いる．しかし，本研究で扱う顔の印象（顔の静的な個人性）は複数の Face Landmark の組合せから成る顔のパーツの形状

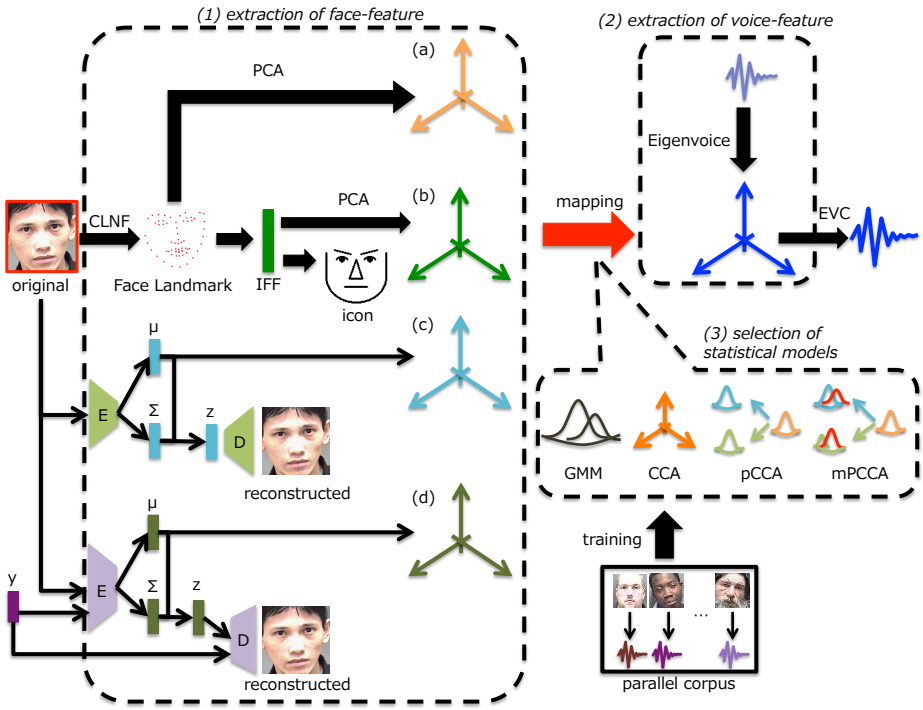


図 3.1: 提案手法の概要：original の画像から顔の印象を表す特徴量を抽出し声の印象を表す特徴量へと変換する．(1) 顔の特徴量として (a)Face Landmark の主成分・(b)IFF の主成分・(c)VAE の $\boldsymbol{\mu}$・(d)CVAE の $\boldsymbol{\mu}$ を，(2) 声の特徴量として Eigenvoice を，(3) 写像として GMM・CCA・pCCA・mPCCA をそれぞれ検討する．

や配置に起因すると考えられるが，Face Landmark の PCA では結果的にいくつかの主成分が目鼻等の顔のパーツに対応する可能性はあるもののそれらに明示的に着目できないという問題点がある．そこで本研究では，顔のパーツの形状・配置を表す 3.2.1 節の Iconified Face Feature (IFF) を提案しその有効性も検討する．

3.2.1　Iconified Face Feature (IFF)

顔の静的な個人性を反映していると考えられる顔のパーツの形状・配置を明示的に扱うため，表 3.1 のように顔のパーツを簡単な図形で近似しその形状を表す 15 次元の特徴量 Iconified Face Feature (IFF) を提案する．IFF は近似図形を描くことでアイコン画像として可視化することができ，その様子を図 3.2 に示す．本論文では 68 点の Face Landmark $\boldsymbol{f}_i$ $(i = 1, 2, ..., 68)$ から以下の規則に基づいて IFF を抽出する方法を提案する．ただし，Face Landmark が検出された顔画像は画像サイズが統一されており，顔が画面いっぱいに撮影されているものとする．また，Face Landmark の座標は画像の左上を原点とする絶対座標で表されており一方で IFF は相対的な特徴量であるため，鼻の先端付近を表す $\boldsymbol{f}_{31}$ を画像中心に移動させる正規化も行う．IFF の抽出においてはアイコン画像が左右対称になるように平均や重心を用いた正規化を行い，距離を計測するときにはユークリッド距離を用いる．Face Landmark の各点のインデックスは図 2.1 に従う．

$\boldsymbol{f}_{38}, \boldsymbol{f}_{39}, \boldsymbol{f}_{41}, \boldsymbol{f}_{42}$ の最小外接円の中心を **rc**・半径を pw_r とし，$\boldsymbol{f}_{44}, \boldsymbol{f}_{45}, \boldsymbol{f}_{47}, \boldsymbol{f}_{48}$ の最小外接円の中心を **lc**・半径を pw_l とする．これらと式 (3.1) の鼻の先端 **noseTop** を用いて，目を表す ed・

表 3.1: Iconified Face Feature (IFF)：() 内は図 3.2 の各変数を表す.

部位	近似図形	特徴量
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)
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)
輪郭	半楕円と直線	幅 (fw)，高さ (fh)

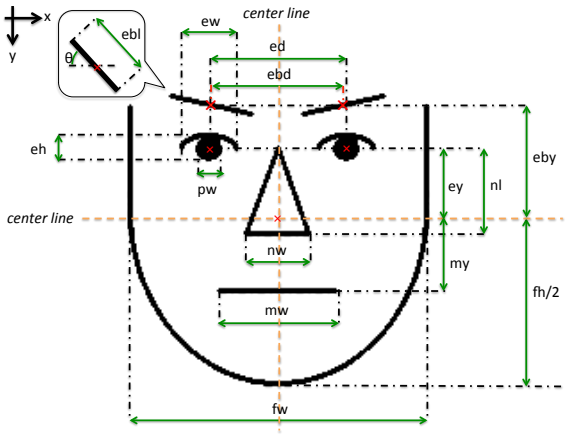


図 3.2: IFF の各変数とそれらを可視化したアイコン画像：緑の線は IFF の各変数を，オレンジの点線は中心線を，赤のバツ印は中心をそれぞれ表す．顔の輪郭の内あごの部分は半楕円で表しその他は直線で表している．

ey・ew・eh と瞳を表す pw は以下の式で求められる．

$$\mathbf{noseTop} = \frac{1}{3}(\mathbf{rc} + \mathbf{lc} + \frac{1}{2}(\boldsymbol{f}_{28} + \boldsymbol{f}_{29})) \tag{3.1}$$

$$\mathrm{ed} = ||\mathbf{rc} - \mathbf{lc}|| \tag{3.2}$$

$$\mathrm{ey} = ||\boldsymbol{f}_{31} - \mathbf{noseTop}|| \tag{3.3}$$

$$\mathrm{ew} = \frac{1}{2}\left(||\boldsymbol{f}_{37} - \boldsymbol{f}_{40}|| + ||\boldsymbol{f}_{43} - \boldsymbol{f}_{45}||\right) \tag{3.4}$$

$$\mathrm{pw} = \frac{1}{2}(2\mathrm{pw}_r + 2\mathrm{pw}_l) = \mathrm{pw}_r + \mathrm{pw}_l \tag{3.5}$$

$$\mathrm{eh} = \mathrm{pw} \tag{3.6}$$

ここで，表 3.1 では瞳の形状を楕円としているが，今回は簡単のため真円を仮定し eh と pw は等しいとした．また，pw_r と pw_l は半径を表すが eh と pw は直径を表すことに注意されたい．

鼻を表す nw と nl は以下の式で求められる．

$$\mathrm{nw} = ||\boldsymbol{f}_{32} - \boldsymbol{f}_{36}|| \tag{3.7}$$

$$\mathrm{nl} = 2||\frac{1}{2}(\boldsymbol{f}_{32} + \boldsymbol{f}_{36}) - \mathbf{noseTop}|| \tag{3.8}$$

口を表す mw と my は以下の式で求められる．

$$\mathrm{mw} = ||\boldsymbol{f}_{49} - \boldsymbol{f}_{55}|| \tag{3.9}$$

$$\mathrm{my} = 2||\boldsymbol{f}_{31} - \frac{1}{2}(\boldsymbol{f}_{49} + \boldsymbol{f}_{55})|| \tag{3.10}$$

右の眉の中心 **rbc**・左の眉の中心 **lbc** をそれぞれ式 (3.11)・(3.12) で表すと，眉毛を表す ebd・eby・ebl・θ は以下の式で求められる．

$$\mathbf{rbc} = \frac{1}{2}(\boldsymbol{f}_{19} + \boldsymbol{f}_{22}) \tag{3.11}$$

$$\mathbf{lbc} = \frac{1}{2}(\boldsymbol{f}_{23} + \boldsymbol{f}_{26}) \tag{3.12}$$

$$\mathrm{ebd} = ||\mathbf{rbc} - \mathbf{lbc}|| \tag{3.13}$$

$$\mathrm{eby} = ||\frac{1}{2}(\mathbf{rbc} + \mathbf{lbc}) - \boldsymbol{f}_{31}|| \tag{3.14}$$

$$\mathrm{ebl} = \frac{1}{2}\left(||\boldsymbol{f}_{22} - \boldsymbol{f}_{19}|| + ||\boldsymbol{f}_{26} - \boldsymbol{f}_{23}||\right) \tag{3.15}$$

$$\theta_r = \frac{\arccos(\boldsymbol{e} \cdot (\boldsymbol{f}_{22} - \boldsymbol{f}_{19}))}{||\boldsymbol{f}_{22} - \boldsymbol{f}_{19}||} \tag{3.16}$$

$$\theta_l = \frac{\arccos(\boldsymbol{e} \cdot (\boldsymbol{f}_{26} - \boldsymbol{f}_{23}))}{||\boldsymbol{f}_{26} - \boldsymbol{f}_{23}||} \tag{3.17}$$

$$\theta = \frac{\theta_r + \theta_l}{2} \tag{3.18}$$

ただし，式 (3.16) と式 (3.17) において $\boldsymbol{e} = [1, 0]^{\mathrm{T}}$ である．

顔の輪郭を表す fw と fh は以下の式で求められる．

$$\mathrm{fw} = ||\frac{1}{3}(\boldsymbol{f}_1 + \boldsymbol{f}_2 + \boldsymbol{f}_3) - \frac{1}{3}(\boldsymbol{f}_{15} + \boldsymbol{f}_{16} + \boldsymbol{f}_{17})|| \tag{3.19}$$

$$\mathrm{fh} = 2||\boldsymbol{f}_{31} - \frac{1}{3}(\boldsymbol{f}_8 + \boldsymbol{f}_9 + \boldsymbol{f}_{10})|| \tag{3.20}$$

以上のように，式 (3.2) から式 (3.10)・式 (3.13) から式 (3.15)・式 (3.18) から式 (3.20) により Face Landmark から IFF への変換が行われ，本論文では各変数を式 (3.21) の順に並べたベクトル $\boldsymbol{\Gamma}$ を IFF として扱う．

$$\boldsymbol{\Gamma} = [\mathrm{ed}, \mathrm{ey}, \mathrm{ew}, \mathrm{eh}, \mathrm{pw}, \mathrm{nw}, \mathrm{nl}, \mathrm{my}, \mathrm{mw}, \mathrm{fw}, \mathrm{fh}, \mathrm{ebd}, \mathrm{eby}, \mathrm{ebl}, \theta]^{\mathrm{T}} \tag{3.21}$$

3.3　声の印象を表す特徴量の抽出

2.4 節で話者認識に用いられる特徴量として DNN の bottlenech feature・i-vector・Eigenvoice を紹介した．これらの内，DNN の bottleneck feature は各次元がどのような話者の声質を反映しているか分析しにくく，本研究の対象である声の印象に関する特徴量を抽出することが困難となる可能性がある．また，2.4.4 節で述べたように i-vector と Eigenvoice には GMM-SV の PCA を行う点で共通している．そこで，EVC により音声を実際に合成しその音声を聞く事で実際に音声の印象を確認できることから，本研究では Eigenvoice を声の印象を表す特徴量として利用できるか検討する．

3.4　統計的顔声変換

本節では顔の特徴量 $\boldsymbol{x}$ から声の特徴量 $\boldsymbol{y}$ へと変換する写像について述べる．写像として 2.5 節で紹介した GMM・CCA・pCCA・mPCCA を検討し，次節からそれらを用いた変換法について順に述べる．ただし，$\boldsymbol{x}$・$\boldsymbol{y}$ の次元数はそれぞれ d_x・d_y とし，統計モデルの学習に必要な顔・声のパラレルデータは $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$ とする．

3.4.1　GMM に基づく顔声変換法

連結ベクトル $\boldsymbol{z}_i = [\boldsymbol{x}_i^{\mathrm{T}}, \boldsymbol{y}_i^{\mathrm{T}}]^{\mathrm{T}}$ が混合数 M の GMM に従うものとする．モデルパラメータ $\hat{\lambda}$ を式 (3.22) に従い結合確率密度 $p(\boldsymbol{z}|\lambda)$ が最大となるように学習する．

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{N} p(\boldsymbol{z}_i|\lambda) \tag{3.22}$$

$$= \arg\max_{\lambda} \prod_{i=1}^{N} \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}_i; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma_m}^{(z)}) \tag{3.23}$$

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \ \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{3.24}$$

学習した $\hat{\lambda}$ を用いて式 (3.25) に示すように尤度を最大化することで顔の特徴量 $\boldsymbol{x}$ から声の特徴量 $\boldsymbol{y}$ へと変換する．

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}, \lambda) \tag{3.25}$$

$$= \arg\max_{\boldsymbol{y}} \sum_{m=1}^{M} p(m|\boldsymbol{x}, \lambda) p(\boldsymbol{y}|\boldsymbol{x}, m, \lambda) \tag{3.26}$$

$$p(m|\boldsymbol{x}, \lambda) = \frac{\alpha_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m^{(xx)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_n^{(xx)}, \boldsymbol{\Sigma}_n^{(xx)})} \tag{3.27}$$

$$p(\boldsymbol{y}|\boldsymbol{x}, m, \lambda) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{E}_m^{(y)}, \boldsymbol{D}_m^{(y)}) \tag{3.28}$$

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

$$\boldsymbol{E}_m^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)}) \tag{3.29}$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)} \tag{3.30}$$

式 (3.25) は式 (3.31) 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 (3.32) で表される．

$$Q(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{m=1}^{M} p(m|\boldsymbol{x}, \boldsymbol{y}, \lambda) \log p(\hat{\boldsymbol{y}}, m|\boldsymbol{x}, \lambda) \tag{3.31}$$

$$\hat{\boldsymbol{y}} = \left(\overline{\boldsymbol{D}^{(y)-1}}\right)^{-1} \overline{\boldsymbol{D}^{(y)-1}\boldsymbol{E}^{(y)}} \tag{3.32}$$

$$\overline{\boldsymbol{D}^{(y)-1}} = \sum_{m=1}^{M} \gamma_m \boldsymbol{D}_m^{(y)-1} \tag{3.33}$$

$$\overline{\boldsymbol{D}^{(y)-1}\boldsymbol{E}^{(y)}} = \sum_{m=1}^{M} \gamma_m \boldsymbol{D}_m^{(y)-1}\boldsymbol{E}_m^{(y)} \tag{3.34}$$

$$\gamma_m = p(m|\boldsymbol{x}, \boldsymbol{y}, \lambda) \tag{3.35}$$

3.4.2　CCA に基づく顔声変換法

パラレルデータ $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$ を用いた CCA で得られた $\hat{\boldsymbol{a}} \cdot \hat{\boldsymbol{b}}$ を行ベクトルとする変換行列をそれぞれ $\boldsymbol{A} \cdot \boldsymbol{B}$ とする．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．

顔の特徴量 $\boldsymbol{x}$ から声の特徴量 $\boldsymbol{y}$ へと変換するとき，変換行列 $\boldsymbol{A} \cdot \boldsymbol{B}$ を用いて以下の線形変換を行う．

$$\boldsymbol{u} = \boldsymbol{A}(\boldsymbol{x} - E[\boldsymbol{x}]) \tag{3.36}$$

$$\boldsymbol{v} = \boldsymbol{B}(\boldsymbol{y} - E[\boldsymbol{y}]) \tag{3.37}$$

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．

$$\boldsymbol{u} \propto \boldsymbol{v} \tag{3.38}$$

さらに，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差し支えない．

$$\boldsymbol{u} \simeq \boldsymbol{v} \tag{3.39}$$

これと式 (3.36)・式 (3.37) より以下の近似が成り立つ．

$$\boldsymbol{A}(\boldsymbol{x} - E[\boldsymbol{x}]) \simeq \boldsymbol{B}(\boldsymbol{y} - E[\boldsymbol{y}]) \tag{3.40}$$

式 (3.40) を満たす $\boldsymbol{y}$ を最小二乗法を用いて求めると最小ノルム解は式 (3.42) のようになる．

$$\hat{\boldsymbol{y}} = \arg\min_{\boldsymbol{y}} ||\boldsymbol{A}(\boldsymbol{x} - E[\boldsymbol{x}]) - \boldsymbol{B}(\boldsymbol{y} - E[\boldsymbol{y}])||^2 \tag{3.41}$$

$$= E[\boldsymbol{y}] + \boldsymbol{\Phi}\boldsymbol{\Sigma}^{+}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x} - E[\boldsymbol{x}]) \tag{3.42}$$

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

$$\boldsymbol{B} = \boldsymbol{\Theta}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}} \tag{3.43}$$

$\boldsymbol{\Sigma}$ 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．

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}, \ \boldsymbol{\Sigma}^{+} = \begin{bmatrix} \boldsymbol{D}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \tag{3.44}$$

3.4.3　pCCA・mPCCA に基づく顔声変換法

pCCA は mPCCA の混合数が 1 であるモデルとして捉え，2.5.3 節で紹介した解析解を初期値とする EM アルゴリズムを適用しパラメータ $\boldsymbol{\theta}$ の学習を行う．すなわち，顔・声のパラレルデータを用いて pCCA または mPCCA のパラメータ $\hat{\boldsymbol{\Theta}}$ を式 (2.71) に従い学習する．

学習された $\hat{\boldsymbol{\Theta}}$ を用いて，ある顔の特徴量 $\boldsymbol{x}$ から声の特徴量 $\hat{\boldsymbol{y}}$ へと式 (2.73) に基づいて変換する．ただし，[45] では式 (2.73) を解く EM アルゴリズムの初期値を GMM に基づいて推定された $\hat{\boldsymbol{y}}$ としていたが，本研究では潜在変数 $\boldsymbol{h}$ を明示的に考慮するため式 (2.74) の $\boldsymbol{G}_m$ と $\boldsymbol{\psi}_m$ について以下の初期値を与える方法を提案する．

$$\boldsymbol{G}_m = \langle z_m | \boldsymbol{x} \rangle \, \boldsymbol{\Lambda}_m^{-1} \tag{3.45}$$

$$\boldsymbol{\psi}_m = \boldsymbol{V}_m \langle \boldsymbol{h} | \boldsymbol{x}, z_m \rangle \tag{3.46}$$

第4章

実験

4.1 はじめに

本研究では，顔の印象を表す特徴量と声の印象を表す特徴量を統計的に対応付け，それに基づいて顔の特徴量から声の特徴量へと変換する顔声変換について検討する．第3章で述べたように，本提案手法は以下の三つのステップに分けることができる．

1. 顔の印象を表す特徴量の抽出
2. 声の印象を表す特徴量の抽出
3. 両者の統計的対応付けおよびそれに基づく顔声変換

まずステップ1について4.2節で述べ，次にステップ2について4.3節で述べる．さらに，ステップ3の統計モデルの学習に必要な印象が合致した顔・声のパラレルデータの収集について4.4節で述べる．最後に，ステップ3について4.5節で述べる．

4.2 顔の印象を表す特徴量の抽出に関する実験

4.2.1 概要

本提案手法の一つ目のステップである顔の印象を表す特徴量の抽出について本節で述べる．特徴量としてFace Landmark・IFF・VAE・CVAEに基づく特徴量を検討しそれぞれについて順番に述べる．ただし，顔画像の検出にはOpenCV[1]を用い，画像のサイズが縦100横100になるよう拡大縮小を行った．

4.2.2 Face Landmarkを用いた特徴量抽出

本節では，顔画像からFace Landmarkを抽出しPCAにより顔の印象を表す特徴量を得られるかどうか実験した．

i) 実験条件

顔画像データベースとしてMORPH [51]を用いた．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

[1] `http://opencv.jp/` [Accessed 19 January 2017]

表 4.1: Face Landmark 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

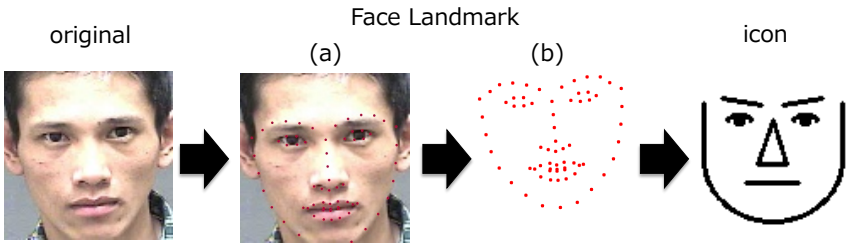


図 4.1: Face Landmark と IFF を可視化したアイコン画像 : (a)Face Landmark を入力画像の上に描画した画像 (b)Face Landmark のみを描画した画像

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．MORPH の顔画像 54,147 枚に対し，CLNF が実装されている OpenFace [52] を用いて 68 点の Face Landmark $\boldsymbol{f}_i = [x_i, y_i]^{\mathrm{T}}$ $(i = 1, 2, ..., 68)$ を推定し，それらを連結したベクトル $\boldsymbol{\Gamma} = [x_1, ..., x_{68}, y_1, ..., y_{68}]^{\mathrm{T}}$ を用いて PCA を行った．

ii)　実験結果

OpenFace で推定された Face Landmark の例を図 4.1 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，(a) の画像であごの輪郭が全て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，(b) 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．PCA においてはそのような平行移動は行わず推定された結果をそのまま用いた．

主成分数と寄与率の関係を表 4.1 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 [2]．例えば，図 4.2 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 4.2 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，本研究では Face Landmark に基づく特徴量として第 5 主成分から第 12 主成分を利用した．

[2]実験結果を `https://goo.gl/Z3jhW7` で公開している

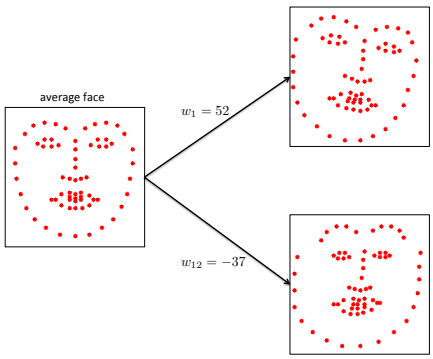


図 4.2: Face Landmark の主成分の変化：w_i は第 i 主成分を表す

表 4.2: IFF の主成分と寄与率の関係 (単位:%)

主成分数	1	2	3	4
寄与率	76.7	16.4	3.96	1.93
累積寄与率	76.7	93.1	97.0	99.0

4.2.3　IFF を用いた特徴量抽出

前節で検討した Face Landmark の主成分は顔の回転成分の寄与率が大きい一方で本研究の対象である印象に関わる成分の寄与率が小さくそれらは重視されなかった．そこで本節では，明示的に目鼻等の顔のパーツに着目した 3.2.1 節の IFF の有効性について検討する．

i)　実験条件

前節で MORPH の顔画像 54,147 枚から抽出した Face Landmark に対し 3.2.1 節の規則に基づき IFF への変換を行いそれらを用いた PCA を行った．

ii)　実験結果

OpenFace で推定された Face Landmark 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 4.1 に示す．アイコン画像を見ると適度に入力顔画像を抽象化した画像となっている．また，主成分数と寄与率の関係を表 4.2 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 4.3 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 の主成分は顔の印象を表す特徴量として利用可能と言える [3]．

4.2.4　VAE を用いた特徴量抽出

4.2.2・4.2.3 節で検討した Face Landmark・IFF は顔の輪郭や目鼻の位置に重点的に着目した特徴量であるが，画像中の顔が撮影されている部分のピクセル値全てを用いた画像圧縮・復元法の一つである VAE を用いて顔の印象を表す特徴量が得られるか実験で検討する．

[3]実験結果を `https://goo.gl/4NCvFQ` で公開している

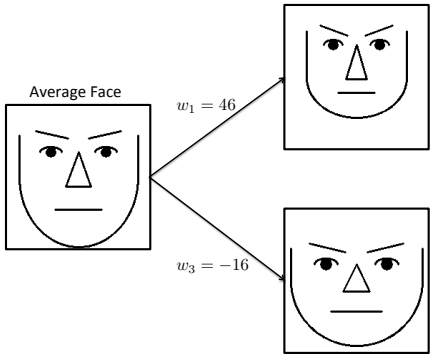


図 4.3: IFF の主成分の変化: w_i は第 i 主成分を表す

表 4.3: Neural Network (NN) に関して本論文で用いる略称

略称	内容
full connected	全結合層
conv	畳み込み層
deconv	逆畳み込み層
BN	バッチ内正規化 [53]
F	畳み込み層・逆畳み込み層におけるフィルタサイズ
S	畳み込み層・逆畳み込み層におけるスライドサイズ
P	畳み込み層・逆畳み込み層におけるパディングサイズ

i)　実験条件

MORPH の顔画像を入力とする VAE を構成しその性能を検証した．VAE の Encoder・Decoder には [23] を参考に表 4.4 に示す Convolutional Nerural Network (CNN) をそれぞれ用いた．ただし，表内の略称は表 4.3 の通りである．CNN のパラメータは表 4.5 の通りであり，学習には顔画像 50,000 枚を用い評価には 4,147 枚を用いた．第 3 章で述べたように顔の印象は比較的低次元で表現できると仮定できるため，潜在変数 $\boldsymbol{z}$ の次元数 $d_z = 10$ とした．Encoder の出力の一つ $\boldsymbol{\Sigma}$ は対角であることを仮定しているため $\boldsymbol{z}$ の各次元は独立であると考えられる．よって，式 (4.1) に従い作成された $\boldsymbol{z}$ を Decoder に入力することで $\boldsymbol{z}$ の各次元の分析を行った．

$$\boldsymbol{z}^{(i)} = [0, 0, ..., 0, \hat{z_d}^{(i)}, 0, ..., 0] \tag{4.1}$$

$$\hat{z_d}^{(i)} = z_d^{(min)} + \frac{i}{N}(z_d^{(max)} - z_d^{(min)}) \tag{4.2}$$

ここで，$\boldsymbol{z}^{(i)}$ は d 次元目のみ $\hat{z_d}$ でそれ以外は 0 となるベクトルであり，$z_d^{(max)}$ と $z_d^{(min)}$ はそれぞれ評価セットにおける d 次元の最大値・最小値である．N は線形補間の分割数であり i はそのインデックス $(i = 0, 1, ..., N)$ である．今回は $N = 10$ とした．

ii)　実験結果

図 4.4 に示すように回転や性別・人種など入力画像の特徴を捉えた画像を再構成できた．$\boldsymbol{z}$ の各次元を操作した時の変化を付録 C の図 C.1 に示す．第 2・3・5 次元などいくつかの次元は性差・人種・回転などに対応していたがその他の次元を変化させることで印象の異なる顔画像が作成で

表 4.4: VAE の構造：NN-type は Neural Network の種類を表し，表内の略称は表 4.3 に従う．

Encoder	
NN-type (F, S, P)	output size (ch × w × h)
input	3 × 100 × 100
conv (4, 2, 1) + BN + max pooling (3, 1, 1)	32 × 50 × 50
conv (4, 2, 1) + BN + max pooling (3, 1, 1)	64 × 25 × 25
conv (4, 2, 1) + BN + max pooling (3, 1, 1)	128 × 12 × 12
conv (4, 2, 1) + BN + max pooling (3, 1, 1)	256 × 6 × 6 (9,216)
full connected	d_z
Decoder	
NN-type (F, S, P)	output size (ch × w × h)
input	d_z
full connected	9,216 (256 × 6 × 6)
deconv (4, 2, 1) + BN	128 × 12 × 12
deconv (4, 2, 1) + BN	64 × 24 × 24
deconv (4, 2, 1) + BN	32 × 48 × 48
deconv (4, 2, 1) + BN	3 × 96 × 96
deconv (7, 1, 1)	3 × 100 × 100

表 4.5: CNN の学習パラメータ

パラメータ	値
最適化手法	Adam [54]
バッチサイズ	100
Encoder の最終層の活性化関数	なし（恒等写像）
Decoder の最終層の活性化関数	sigmoid
その他の層の活性化関数	ReLU [55]

きることから，本研究の対象である顔の印象を表す特徴量として利用できる可能性がある．4.2.2 節と 4.2.3 節では顔の印象に関係する主成分を寄与率を指標として選別したが，VAE には $\boldsymbol{z}$ の各次元の重要度を示す指標はなく，CNN の学習条件によって各次元が表す特徴が変化する可能性があるため，本論文では VAE で得られる $\boldsymbol{z}$ を顔の印象を表す特徴量としてそのまま用いることとした．

4.2.5　CVAE を用いた特徴量抽出

前節で VAE を検討したが，結果として潜在変数 $\boldsymbol{z}$ の空間に性差や回転などの成分が残ってしまうという問題が生じた．そこで，MORPH に付与されているラベルの内人種・性別に関する情報を用いた CVAE を学習し，$\boldsymbol{z}$ の空間からそれらの成分を取り除くことができるか実験し検討した．

図 4.4: VAE・CVAE で再構成した画像：() 内の数字は $\boldsymbol{z}$ の次元数を表す．M は男性，F は女性，W は白人，B は黒人，A はアジア系，H はヒスパニック，O はその他の人種を表す．赤で囲まれた一番右の列は 4.4.3 節で述べる森島データベースの日本人画像でありその他の画像は MORPH の画像である．[4]

i)　実験条件

MORPH の顔画像を入力する CVAE を構成しその性能を検証した．Encoder・Decoder には表 4.6 に示した構造の CNN をそれぞれ用いた．4.2.4 節の条件と同じ $d_z = 10$ とし実験した．さらに，4.2.4 節の結果から性別・人種に関わる次元がいくつかあったことから，CVAE でそれらの情報をラベルとして与えることで顔の印象に関わる成分を 10 次元以下に抑えられる可能性があると考えられ，次元数をさらに削減した $d_z = 3, 6$ の場合についても実験した．また，ラベル $\boldsymbol{y}$ は性別（男性/女性）と人種（黒人，白人，アジア系，ヒスパニック，その他）の組合せを表す 10 次元の one-hot ベクトルとした．その他の CNN の学習条件は 4.2.4 節の条件と同じである．また，式 (4.1) に従い $\boldsymbol{z}$ の各次元の分析も行った．ただし CVAE では Decoder にラベル $\boldsymbol{y}$ を入力する必要があるが，式 (4.1) の $\boldsymbol{z}^{(i)}$ にはラベル情報は付与されていないため 10 種類のラベルをそれぞれ与えて Decoder に入力した．

ii)　実験結果

図 4.4 に示すように d_z が大きいほど精度よく画像を再構成できていたが，$d_z = 3, 6$ の場合でも入力画像の特徴を捉えた画像を再構成できていた．次元数が $d_z = 3, 6, 10$ のときラベルを黒人男性として $\boldsymbol{z}$ の各次元を変化させた様子を付録 C の図 C.2・図 C.3・図 C.4 に示す．性別・人種は統一されたまま顔の印象に対応する成分と回転成分が $\boldsymbol{z}$ に現れた．このうち，回転成分は $d_z = 3$ のときは第 3 次元に，$d_z = 6$ のときには第 2・3 次元に，$d_z = 10$ のときには第 2・3・8 次元に主に表れた．顔の印象に対応する成分には丸顔から面長へと変化する成分や鼻の大きさが変化する成分などがあった．このように，回転成分は残ったものの，$\boldsymbol{z}$ には性別と人種の情報を除いた顔の印象に関係する成分が含まれることが確認されたため，前節と同様に顔の印象を表す特徴量として CVAE で得られる $\boldsymbol{z}$ をそのまま用いることとした．

表 4.6: CVAE の構造：NN-type は Neural Network の種類を表し，表内の略称は表 4.3 に従う．

Encoder	
NN-type (F, S, P)	output size (ch × w × h)
input	3 × 100 × 100
conv (4, 2, 1) + BN + max pooling (3, 1, 1)	32 × 50 × 50
conv (4, 2, 1) + BN + max pooling (3, 1, 1)	64 × 25 × 25
conv (4, 2, 1) + BN + max pooling (3, 1, 1)	128 × 12 × 12
conv (4, 2, 1) + BN + max pooling (3, 1, 1)	256 × 6 × 6 (9,216)
$\boldsymbol{y}$ 追加 ($d_y = 10$)	9,226
full connected	d_z
Decoder	
NN-type (F, S, P)	output size (ch × w × h)
input	d_z
$\boldsymbol{y}$ 追加 ($d_y = 10$)	$d_y + d_z$
full connected	9,216 (256 × 6 × 6)
deconv (4, 2, 1) + BN	128 × 12 × 12
deconv (4, 2, 1) + BN	64 × 24 × 24
deconv (4, 2, 1) + BN	32 × 48 × 48
deconv (4, 2, 1) + BN	3 × 96 × 96
deconv (7, 1, 1)	3 × 100 × 100

4.3　声の印象を表す特徴量の抽出に関する実験

4.3.1　概要

本提案手法の二つ目のステップである声の印象を表す特徴量の抽出について本節で述べる．特徴量として Eigenvoice について検討した．

4.3.2　Eigenvoice の抽出

2.4.3 節で述べた Eigenvoice が声の印象を表す特徴量として利用可能かどうか実験で検証した．

i)　実験条件

JNAS[5] の音素バランス文を使用し，サブセット A〜I を読んだ男性話者 127 人を用いて Eigenvoice を学習した．ただし，Eigenvoice に基づく空間が話者の声質を反映しているかどうか音声を聞いて確認するため，別のある男性話者をピボット話者とする EVC の枠組みで Eigenvoice を学習した．EVC は GMM に基づく声質変換（2.5.1 節）と Eigenvoice を組み合わせた手法であり（付録 B），複数の話者の音声データを用いて一人の入力話者から複数の話者への声質変換（一対多変換）を可能にする手法である [6]．このとき出力話者の GMM-SV は Eigenvoice・基底ベクトル・バイアスベクトルで表されるため，Eigenvoice が変化することで様々な GMM が表現でき，それに

[5] `http://research.nii.ac.jp/src/JNAS.html` [Accessed 19 January 2017]
[6] 複数の話者から一人の話者への声質変換（多対一変換）も可能である．

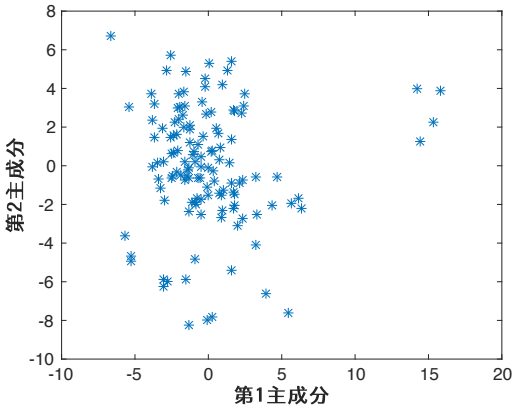


図 4.5: 男性話者 127 人の Eigenvoice：第 i 主成分は i 番目の Eigenvoice を表す

表 4.7: Eigenvoice と寄与率の関係 (単位:%)

主成分数	1	2	3	4	5	6	33
寄与率	14.9	10.9	9.66	6.74	5.17	3.68	0.48
累積寄与率	14.9	25.8	35.5	42.2	47.4	51.1	80.1

従う音韻空間を持つ話者を表現できる．音響特徴量として，サンプリング周波数 16kHz・量子化ビット数 16bit の音声から 1 次から 24 次のメルケプストラム係数を WORLD [56] と SPTK[7] を使用し抽出した．GMM の混合数は 256 とした．

ii)　実験結果

得られた第 1・2 主成分（一番目・二番目の Eigenvoice）を図 4.5 に，Eigenvoice と累積寄与率の関係を表 4.7 に示す．構成された固有空間上の 1 点に対応する話者の音声データを実際に聞いたところ，第 1 主成分が大きくなる程声は太くなり，小さくなる程声は細くなった．また，累積寄与率が 50% を超える第 6 主成分までの重み（6 次元までの Eigenvoice）を手動で変化させ EVC により音声を作成したところ異なる声質を持つ音声が作成できた．従って，Eigenvoice は声の印象を表す特徴量として利用可能であると言える．

4.4　手動に基づく顔・声のパラレルコーパスの収集

4.4.1　概要

本研究では 4.2 節で検討した顔の特徴量から 4.3 節で検討した声の特徴量へと統計的に変換することを検討している．統計モデルを学習するためには顔の印象と声の印象が合致しているパラレルコーパスが必要であるが，そのパラレルデータは話者本人の顔と声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顔と声のパラレルコーパスを収集した．コーパス収集において，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 [48]，提示顔画像の被写体・音声の発話者・主観実験の被験者の性別を全

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

表 4.8: アジア系男性顔画像を用いたパラレルコーパスの収集条件

	データ	使用数
顔画像	MORPH アジア系男性	44 枚
音声	JNAS 男性話者	127 人
被験者	日本人の 20 代男性	10 人

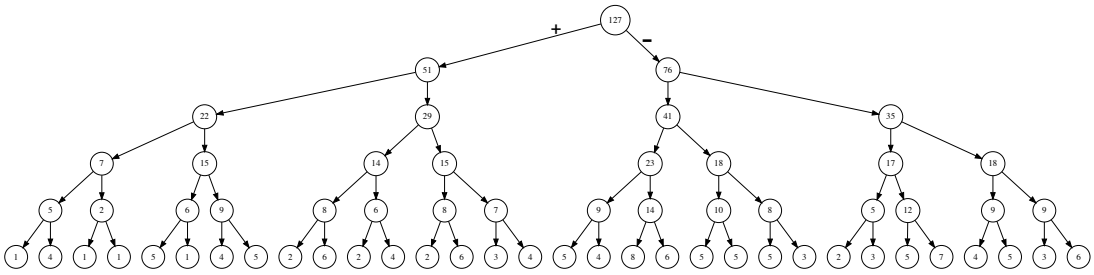


図 4.6: JNAS 音声話者 127 人の Eigenvoice に基づく二分木：サブセット J を読み上げていない男性話者を使用．各ノード内の数字はノードに属する話者数を表す．

て男性に固定した．さらにそれぞれの人種も顔と声の対応関係に影響を及ぼす可能性があると考え，日本語母語話者の日本人男性である著者がコーパスを分析しやすいよう人種を日本人に固定することを検討した．そこで，音声を日本語母語話者が発話した日本語音声に固定し，主観実験の被験者を日本人に固定した．しかし，日本人男性画像を含んだデータベースはほとんど一般に公開されておらず入手が困難であったため，まず一般に公開されている顔画像データベースのうちアジア系男性の画像を含むものを用いてコーパスの収集を行った．その後，早稲田大学の森島繁生先生より主に日本人の画像で構成されたデータベースを提供していただいたためそちらを用いたコーパスの収集も行った．以下ではそれぞれの詳細な収集方法について説明する．

4.4.2　アジア系男性の顔画像を用いたパラレルコーパスの収集

i)　実験条件

顔画像データベース MORPH は，約 13,000 人の画像約 5,5000 枚が収録されたデータベースであり，被写体それぞれに人種・性別・年齢のラベルが付与されている．この MORPH に収録されているアジア系男性の顔画像 44 枚を使用し，主観実験により顔と声が対応付けられたパラレルコーパスを収集した．行った主観実験は，提示された顔画像に対し被験者が最適だと思う音声を選択するという実験である．顔画像・音声・被験者の数を表 4.8 に示す．顔画像と音声についてそれぞれ以下の処理を行った．

提示顔画像について　4.2 節で使用した MORPH の顔画像（縦 100 横 100 の画像）を用いた．ただし，被験者に提示する際には分かりやすいように縦 200 横 200 になるよう拡大して表示した．

提示音声について　JNAS の話者の内サブセット J を読み上げていない男性 127 人を採用し，各話者が発話したサブセットの一文目のみを提示した．被験者の負担を減らすため，あらかじめ話

者を 4.3 節で作成した Eigenvoice 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とった．具体的には当該話者の Eigenvoice の重みの正負を利用して図 4.6 に示す深さ 5 の二分木を構成した．ここで深さ d の分類には，d 次元目の Eigenvoice が正の時は左のノードに非正の場合は右のノードに分類する方法をとった．

これらにより得られた顔画像・音声を用いて以下の手順で主観実験を行った．

1. 2 つの音声の内，提示された顔画像により相応しいと思う音声を選択する．
2. 手順 1 を 5 回（二分木の葉ノードに到達するまで）繰り返す．
3. 葉ノードに属する話者の内，提示された顔画像に最も相応しいと思う音声を一つ選択する．
4. 手順 1 から 3 までを全ての画像に対して行う．

ただし，手順 1 では各ノードに属する話者の内ランダムに選択された話者の音声を流した．

ii)　実験結果

各被験者について 44 対のパラレルデータが得られた．全被験者・全顔画像について分析すると，$44 \times 10 = 440$ 対のパラレルデータの内 23 対（5.2%）で選択された話者や 15 対（3.4%）で選択された話者など複数回選択された話者が多く 127 人全員は選択されなかった（全被験者・全顔画像について 112 人の音声話者が少なくとも一度選択された）．ただし，手順 1 の選択も含めると 127 人全員が少なくとも一度は選択されていた．各被験者について分析すると複数の顔画像に対し同じ話者の音声を選択することがあり，それは被験者ごとに傾向が異なった．また，各顔画像について分析すると同じ顔に同じ話者が選択される（複数の被験者の意見が一致する）ことがあり，本研究の対象となる顔の印象に合致するような声の選択が可能であることが示唆された．

以上のように，各被験者に強く依存するものの顔・声の印象が対応付けられたパラレルデータを得る事ができた．以降このパラレルコーパスを**アジア系コーパス**と呼ぶ事とする．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・声の印象を表す空間に射影する必要がある．しかし，被験者から顔画像の人種や年齢が対応付けに影響を与えることがあったという意見があり，コーパス収集時の条件の内人種・年齢を統一することの重要性が浮き彫りになった．そこで，次節で日本人の 20 代男性の顔画像を用いたパラレルコーパスの収集を行いその問題について検討した．

4.4.3　日本人男性の顔画像を用いたパラレルコーパスの収集

前節で収集したパラレルコーパスはアジア系男性の画像を使用しており，顔画像・音声話者・被験者間で人種の統一が完全に取れているとは言えない．これは日本人男性の画像が収録された顔画像データベースはほとんど一般に公開されていないことから，日本人が属するアジア圏の男性画像で代用したためである．しかし前節の主観実験の結果から人種の違いや年齢の違いが対応付けに影響を及ぼすことが示唆され，前節の実験条件では不十分である可能性が生じた．そこで，早稲田大学の森島繁生先生より日本人顔画像を提供していただき，それらを用いた主観実験を行い人種・年齢も統一したパラレルコーパスを収集した．

表 4.9: 日本人男性顔画像を用いたパラレルコーパスの収集条件

	データ	使用数
顔画像	森島データベースの 20 代男性	71 枚
音声	JNAS の 20 代男性話者	73 人
被験者	日本人の 20 代男性	17 人

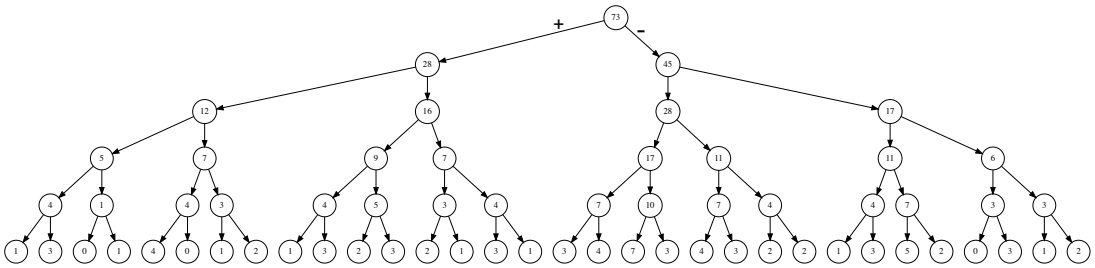


図 4.7: JNAS 音声話者 73 人の Eigenvoice に基づく二分木：サブセット J を読み上げていない 20 代男性話者を使用．各ノード内の数字はノードに属する話者数を表す．

i)　実験条件

早稲田大学森島研究室 [8] では人間の顔の経年変化に関する研究が行われておりその分析のために収集された顔画像を提供していただいた．以下ではこのデータベースを森島データベースと呼ぶ．顔画像は自然光下で撮影された RGB 正面画像であり被写体はほとんどが日本人男性である．画像それぞれに年齢ラベルも付与されている．森島データベースの内 20 代（20 歳から 29 歳）の顔画像 71 枚を使用し，4.4.2 節の実験手順と同一の手順で顔と声が対応付けられたパラレルコーパスを収集した．顔画像・音声・被験者の数を表 4.9 に示す．

提示顔画像について　4.4.2 節と同様に OpenCV で顔検出した後縦 100 横 100 に拡大・縮小した．

提示音声について　JNAS の話者の内サブセット J を読み上げていない 20 代男性 73 人を採用し，各話者が発話したサブセットの一文目のみを提示した．4.4.2 節と同様にこれらの話者を図 4.7 に示す深さ 5 の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とった．4.4.2 節の実験では手順 1 でランダムに選んだ話者の音声を提示したが，その方法では左右のノードの境界付近の音声が選択された場合は適切な判断がされない可能性があるため，各ノード内の話者の Eigenvoice の平均に最も近い話者をそのノードの代表話者として固定しその音声を提示した．

ii)　実験結果

各被験者について 71 対のパラレルデータが得られた．全被験者・全顔画像について分析すると，$71 \times 17 = 1,207$ 対のパラレルデータの内，126 対（10%）で選択された話者や 106 対（8.8%）で選択された話者など複数回選択された話者が多く，手順 1 の選択を含めても 73 人全員は選択され

[8] `http://www.mlab.phys.waseda.ac.jp/` [Accessed 22 November 2017]

表 4.10: 統計的対応付けに関する実験条件

顔の印象を表す特徴量（4.2 節参照）				
データ	MORPH 顔画像 54,147 枚			
手法	Face Landmark の PCA	IFF の PCA	VAE	CVAE
次元数	8	3	10	3・6・10
声の印象を表す特徴量（4.3 節参照）				
データ	JNAS の男性話者 127 人			
手法	Eigenvoice			
次元数	6			
両空間の写像（第 3 章参照）				
データ	4.4.2・4.4.3 節で収集したパラレルコーパス			
手法	GMM	CCA	pCCA	mPCCA
変換式	式 (3.25)	式 (3.42)	式 (2.73)	
備考	GMM・mPCCA の混合数は 2，潜在変数 $\boldsymbol{h}$ の次元数は $d_h = 3$			

なかった（全被験者・全顔画像について 71 人の音声話者が少なくとも一度選択された）．アジア系の画像を用いた前節と同様に，各被験者についての分析では複数の顔画像に対し同じ話者の音声を選択することがあり，それは被験者ごとに傾向が異なった．また，各顔画像についての分析では同じ顔に同じ話者が選択される（複数の被験者の意見が一致する）ことがあった．

以上から，顔画像の被写体・音声の発話者・主観実験の被験者が全て日本人の 20 代男性に統一されたパラレルコーパスが収集できた．以降このパラレルコーパスを**日本人コーパス**と呼ぶ事とする．さらにある顔に印象的に対応する音声（声質）は被験者に強く依存することが再確認された．

4.5　統計的顔声変換に関する実験

4.5.1　概要

本研究では，顔の印象を表す特徴量と声の印象を表す特徴量を統計的に対応付けそれに基づいて顔声変換を行うことを目的としている．本節では 4.2 節で抽出した Face Landmark・IFF・VAE・CVAE に基づく顔の印象を表す特徴量と 4.3 節で抽出した Eigenvoice に基づく声の印象を表す特徴量を統計的に対応付け，前者から後者へ変換することを検討する．統計モデルとして GMM・CCA・pCCA・mPCCA を検討し，その学習には 4.4 節で収集したアジア系男性・日本人男性の顔画像を用いたパラレルコーパスをそれぞれ用いた．

4.5.2　アジア系コーパスを用いた顔声変換実験

i)　実験条件

4.4.2 節で収集したアジア系コーパスを用いて GMM・CCA・pCCA・mPCCA を学習し，顔の特徴量から声の特徴量へと統計的に変換することを検討した．特徴量の次元数や GMM・mPCCA の混合数等の実験条件を表 4.10 に示す．また 4.4 節で収集したパラレルコーパスは，顔と声という異なるメディアを結びつけたデータであり，コーパスの規模が小さく，複数の顔に対して同一

の話者が選択されることが多かったため，確率モデルを安定して学習するにはいくつかの制約条件が必要であり，以下でその制約条件について述べる．

GMM の学習において，対数尤度の上昇幅を閾値として EM アルゴリズムの収束を確認したが，対数尤度が正になってしまった場合・ある分布から出力されるサンプルが存在しないと事後確率に基づいて判断された場合（混合数を削減すべき場合）・式 (3.24) の $\boldsymbol{\Sigma}_m^{(z)}$ の正定値性が崩れ逆行列を持たなくなった場合には 1・2 イタレーション前のモデルを採用することで GMM の特徴（正規分布を混合したモデルであること）を損なわず混合数を保持した学習を行った．また，GMM 学習の際には式 (3.24) の $\boldsymbol{\Sigma}_m^{(z)}$ の全成分を計算に用いたが，GMM に基づく顔声変換では，式 (3.30) の $\boldsymbol{D}_m^{(y)}$ の対角成分のみを計算に用いた．これは．本提案手法で扱う入出力のメディアは異なるため少なくとも $\boldsymbol{\Sigma}_m^{(xy)}$ と $\boldsymbol{\Sigma}_m^{(yx)}$ は対角行列として扱うことができず結果として $\boldsymbol{\Sigma}_m^{(z)}$ の全ての成分を考慮した GMM 学習が必要だが，式 (3.30) は $\boldsymbol{y}$ のみの分布（単一メディアの分布）の分散共分散行列を表すためである [9]．

pCCA・mPCCA の学習において，$\boldsymbol{\Gamma}_m$ と $\boldsymbol{\Lambda}_m$ はそれぞれ $\boldsymbol{x}$・$\boldsymbol{y}$ の単一メディアの分布を表すため対角行列を仮定し，それらの各要素が極端に小さくなり混合数が保持できなくなることを防ぐためそれぞれの初期値の 0.01 倍を下限値とした．また，mPCCA に基づく顔声変換において式 (2.77) の g_m について以下の近似を用いた．

$$g_m = \begin{cases} 1, & \text{if } m = \arg\max_m g_m \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

これは，入力特徴量 $\boldsymbol{x}$ と直前の M ステップで得られた $\hat{\boldsymbol{y}}^{old}$ に基づく事後確率を表す g_m が最大の混合のみを考えることを意味する．

4.4.2 節で顔・声の対応付けは被験者（顔画像を見た人）に依存する可能性が示唆されたため，各被験者ごとのパラレルコーパスを用いて被験者依存の形で統計的対応付けと手動の対応付けを比較した．アジア系コーパスでは一人の被験者につきパラレルデータは 44 対得られたが，その内 40 対を学習データ，4 対を評価データとするデータセットを 11 セット作成しクロス・バリデーションで提案手法の有効性を検証した．評価は，手動もしくは統計モデルに基づいて変換された話者空間の座標 (Eigenvoice) から EVC を用いて音声を合成し，両者を式 (4.4) のメルケプストラムひずみ (Mel-cepstral distortion: MCD) を用いて比較した．

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(c_d^{(est)} - c_d^{(ref)}\right)^2} \tag{4.4}$$

ただし，$c_d^{(est)}$ は統計モデルに基づいて変換された話者の音声のメルケプストラムであり，$c_d^{(ref)}$ は手動で変換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．

ii)　実験結果

実験結果を表 4.11 に示す．変換手法を比較すると，どの顔の特徴量を用いても CCA が最も変換精度が低く pCCA が最も高い変換精度となったため，GMM・pCCA・mPCCA のように確率

[9] 2.5.1 節の GMM に基づく声質変換法では，入力 $\boldsymbol{X}$ と出力 $\boldsymbol{Y}$ が同じメディアである音響特徴量であるため $\boldsymbol{\Sigma}_m^{(XX)}$・$\boldsymbol{\Sigma}_m^{(XY)}$・$\boldsymbol{\Sigma}_m^{(YX)}$・$\boldsymbol{\Sigma}_m^{(YY)}$ は全て対角行列として仮定されることが多い．

表 4.11: アジア系コーパスを用いたときの平均メルケプストラムひずみ（単位：dB）: () の中は顔の特徴量の次元数を表す．太字は各変換写像を比較したときに最も変換精度が高かったものを示す．

写像	Face Landmark(8)	IFF(3)	VAE(10)	CVAE(3)	CVAE(6)	CVAE(10)
GMM	2.90	2.36	3.13	2.40	2.63	3.13
CCA	3.46	2.66	3.36	2.62	3.18	3.36
pCCA	**2.35**	**2.30**	**2.34**	**2.28**	**2.32**	**2.35**
mPCCA	2.61	2.39	2.52	2.46	2.50	2.56

分布を仮定した写像が有効であると言える．GMM と mPCCA を比較すると，Face Landmark・VAE($d_z = 10$)・CVAE($d_z = 10$) では mPCCA の方が変換精度が高かったがそれ以外の特徴量では近しい精度であった．これは，式 (2.67) で同時確率 $p(\boldsymbol{x}, \boldsymbol{y})$ が GMM で表されるように mPCCA は GMM に潜在変数 $\boldsymbol{h}$ を仮定したモデルと解釈できることから両者の表現力が同等であったためである可能性が考えられる．また，表現力の高い混合モデルである GMM・mPCCA よりも単一モデルである pCCA の方が変換精度が高かった理由として，モデル学習に使用したパラレルデータ数が 40 対と少なかったために混合モデルのパラメータを十分に学習できなかった可能性がある．しかし付録 D の図 D.1 のように mPCCA を用いたときに，顔の特徴量の分布形状は似通っているものの声の特徴量の分布には急峻な分布が現れ，さらに急峻な分布に関する混合重みが小さくなるような被験者・データセットの組合せが複数存在したことから，入力の分布形状と出力の分布形状が異なる可能性も考えられ，これは GMM・mPCCA では入出力の分布形状が等しい（混合数が等しい）という仮定に反する場合があることを示唆している．

顔の特徴量について比較すると，IFF または $d_z = 3$ とした CVAE を用いた場合がどの変換手法を用いても変換精度が高かった．pCCA を用いた場合のみを比較すると各特徴量の結果は全て近しい値をとっているものの，顔の特徴量の次元数が低くなる程 MCD の値は減少するため，3.1 節で述べた顔の印象は比較的低次元で表現可能であるという仮定に矛盾しない結果となった．また，同じ次元数である IFF と $d_z = 3$ とした CVAE を比較するとどの変換手法でも大きな差がなかった．

以上の結果から，顔の印象を表す特徴量から声の印象を表す特徴量へと変換する際には確率分布を仮定した変換法，特に pCCA を用いた変換法が有効に働くことが示された．また，顔の印象を表す特徴量として IFF や $d_z = 3$ とした CVAE に基づく特徴量が有効である可能性が示唆されたが，パラレルコーパスの規模が小さいことと CVAE において次元数が低くなるほど変換精度が上がったことから，各特徴量の性質が変換精度に与える影響よりも次元数が与える影響の方が大きかった可能性も無視できない．さらに，本節で使用したパラレルコーパスはアジア系男性画像に関するものであり，顔・声の人種の不一致がこれらの結果に影響を与えた可能性は排除できない．そこで次節ではアジア系コーパスよりも規模が大きく人種と年齢が統一された日本人コーパスを用いた実験を行いそれらの可能性について検証した．

4.5.3　日本人コーパスを用いた顔声変換実験

i)　実験条件

4.4.3 節で収集した日本人コーパスを用いて GMM・CCA・pCCA・mPCCA を学習し，顔の特徴量から声の特徴量へと統計的に変換することを検討した．特徴量の次元数や GMM・mPCCA の混合数等の実験条件は前節の実験条件と同じく表 4.10 の通りである．また，前節と同様に各被

表 4.12: 日本人コーパスを用いたときの平均メルケプストラムひずみ（単位：dB）: () の中は顔の特徴量の次元数を表す．太字は各変換写像を比較したときに最も変換精度が高かったものを示す．

写像	Face Landmark(8)	IFF(3)	VAE(10)	CVAE(3)	CVAE(6)	CVAE(10)
GMM	2.13	**1.89**	2.28	**1.88**	2.04	2.24
CCA	2.81	2.15	2.75	2.20	2.77	2.78
pCCA	**1.95**	1.90	**1.96**	1.89	**1.93**	**1.96**
mPCCA	2.08	1.98	2.11	1.98	2.08	2.12

験者ごとのパラレルコーパスを用いて被験者依存の形で統計的対応付けと手動の対応付けを比較した．ただし，日本人コーパスでは一人の被験者につきパラレルデータは 71 対得られたが，その内 64 対を学習データ，7 対を評価データとするデータセットを 10 セット作成しクロス・バリデーションで提案手法の有効性を検証した．その他の GMM・pCCA・mPCCA の学習における条件と評価方法も前節と同様である．

ii)　実験結果

実験結果を表 4.12 に示す．変換手法の比較・顔の特徴量の比較を行ったところ，前節と同様の傾向が見られた．すなわち，確率分布を仮定した変換法が有効に働き，IFF と $d_z = 3$ とした CVAE に基づく特徴量が有効である可能性が示唆された．IFF と $d_z = 3$ とした CVAE に基づく特徴量を用いた場合では，0.01 ポイントだけ GMM の方が pCCA よりも MCD が小さかったが有意な差とは言えず pCCA と GMM はほぼ同等の性能と考えた方が妥当である．また，学習データ数が前節の 40 対から 64 対に増えたが，混合モデル（GMM・mPCCA）と単一モデル（pCCA）の間の差は前節の結果から大きく縮まることはなかった．さらに，付録 D の図 D.2 のように mPCCA を用いたときに，アジア系コーパスを用いた場合と同様に顔の特徴量の分布形状と声の特徴量の分布形状が異なる場合が複数確認された．

表 4.11 のアジア系コーパスを用いた結果と表 4.12 の日本人コーパスを用いた結果を比較すると，日本人コーパスを用いた方が全体的に変換精度は高くなっている．アジア系コーパスの収集方法と日本人コーパスの収集方法が異なるため安易に比較はできないが，顔画像の被写体・音声話者・主観実験被験者の人種・性別・年齢を統一することで顔・声の対応関係にある制約がかかり両空間の写像がより簡単なものになったために変換精度が向上した可能性が考えられる．

以上から，顔の印象を表す特徴量から声の印象を表す特徴量へと変換する際には確率分布を仮定した変換法が有効に働くことが再確認され，顔の印象を表す特徴量として IFF に基づく特徴量が有効である可能性とパラレルコーパス内の人種を統一することで全体的な精度向上が見込める可能性が示唆された．学習に用いるパラレルデータ数を増加して実験を行い，混合モデルよりも単一モデルの方が精度よく変換可能であることが再確認されたが，データ数の増加分が 24 対のみ（1.6 倍）と大きくはなかったために混合モデルのパラメータを十分に学習できなかった可能性も残ってしまった．また，統計モデル学習に用いるパラレルデータ数が限られている（数十対程度）場合では混合モデル（GMM・mPCCA）よりも単一モデル（pCCA）が有効である可能性も再確認された．

第5章

結論

5.1 まとめ

本研究では，近年急速に普及してきた音声対話システムやスマートスピーカーなどのヴァーチャルエージェントのキャラクター性を向上させる手法の一つである音声話者の顔と音声を同時に提示する方法に着目し，エージェントの顔画像が与えられた文字対話システムをそのような顔付きの音声対話システムに拡張する場合において顔画像からその顔に相応しい声質へと変換することを検討した．すなわち，顔の印象（顔の静的な個人性）と声の印象（音声の話者性）に着目し，顔の印象を表す特徴量から対応する声の印象を表す特徴量へと統計的に変換する顔声変換について検討した．

顔の特徴量として (1) 顔の輪郭・目鼻の位置を表す Face Landmark・(2) より顔のパーツ形状に着目した IFF・(3) 画像圧縮または画像生成に適用可能な VAE・(4) 人種や性別といったラベル情報で制約を課す CVAE の 4 通りの特徴量を検討した．声の特徴量として話者の音韻分布の形状に基づく特徴量である Eigenvoice を検討した．統計的写像として GMM・CCA・pCCA・mPCCA をそれぞれ検討し，モデル学習・評価に必要な顔と声が対応付けられたパラレルコーパスも収集した．パラレルコーパスはアジア系男性の顔画像を使用したコーパス（アジア系コーパス）と日本人男性の顔画像を使用したコーパス（日本人コーパス）の 2 通りを収集した．

結果として，顔の印象を表す特徴量から声の印象を表す特徴量へと変換する際には確率分布を仮定した変換法が有効に働くことが示された．特に，統計モデル学習に用いるパラレルデータ数が限られている（数十対程度）場合では混合モデル（GMM・mPCCA）よりも単一モデル（pCCA）が有効であった．また，顔の印象を表す特徴量として IFF に基づく特徴量と 3 次元まで圧縮した CVAE が有効である可能性とパラレルコーパス内の人種を統一することで全体的な精度向上が見込める可能性が示唆された．さらに，顔の特徴量の分布形状と声の特徴量の分布形状が異なる可能性も示唆された．以上の検討を通して，本論文では顔の印象を表す特徴量と声の印象を表す特徴量との統計的対応付けとそれに基づく顔声変換の枠組みを提案・構築した．

5.2 今後の課題

本研究の課題として，パラレルコーパスの規模が小さいことが挙げられる．統計モデル学習に用いるパラレルデータ数が少なかったために混合モデル（GMM・mPCCA）のパラメータが十分に学習できなかった結果単一モデルである pCCA の方が精度が高かった可能性があり，パラレルコーパスの規模を拡張したときに変換精度がどのように変化するか検証する必要がある．しかし，

パラレルコーパス収集方法にあたり本研究で行った主観実験は一人の被験者につき 2 時間程度の時間がかかってしまう非常に被験者負担が大きい実験であった．そこで，今後はパラレルコーパスの収集方法も含めてコーパス拡張について検討する必要がある．また，本論文では顔の特徴量から声の特徴量へと統計的に変換し評価する枠組みを提案したが，その枠組みを利用した顔と声の結びつき・関係性に関する調査もすべきだと考える．特に，pCCA または mPCCA における $\boldsymbol{h}$ や $\boldsymbol{z}$ がどのような物理的意味を持つのかについて検討していきたい．

謝辞

本研究を進めるにあたって学部 4 年生に研究室に所属してから修士課程を終えるまで日頃より多大なるご指導ご鞭撻を賜りました指導教官である峯松信明教授に深く感謝いたします．峯松先生からは現実の諸問題をどのように技術で解決していくか，問題に直面し困っている人に我々技術者がどのように貢献できるかについて考え続ける姿勢を学びました．また，本研究を進める上で細かい実装方法や機械学習に関する理論的な質問にも快く答えてくださった齋藤大輔講師にも深く感謝いたします．齋藤先生からは様々な手法の理論的背景やそれらの間の関連性・物理的側面について学びました．さらに，研究室での生活を送る上で必要不可欠な計算機資源のサポートを手厚くしていただけたおかげで不自由無く研究を進める事ができました．

日頃の研究生活を様々な面から支えてくださった高橋登技官と秘書の池上恵氏にも厚く御礼申し上げます．高橋さんが研究室設備の整備をしてくださったおかげで不自由無く様々な設備を使用することができました．池上さんには度重なる主観実験の申請についてご協力していただき大変感謝しております．

研究室の先輩・同期・後輩の皆さんに感謝いたします．先輩方には音声に関する様々な知識や研究者としての姿勢を学びました．同期の皆とは同じ修士課程を過ごす中での困難や楽しさを分かち合えたと思っています．後輩の皆さんがいたからこそ先輩としての自覚が生まれ研究や学問に対するモチベーションを向上させることができたと思っています．研究室での生活を快く過ごせたのは皆様のおかげです．ありがとうございました．

また，日本人顔画像データベースを提供してくださった早稲田大学の森島繁生先生に厚く御礼申し上げます．森島先生のおかげで本研究の課題の一つであったコーパスの条件をそろえることができました．

最後に，今まで私を支えてくださった家族・友人に深く感謝いたします．友人の皆さんと話すだけで安らぎ，モチベーションを維持することができました．家族には帰る場所がある安心感を与えてくださり，上京してから今まで東京での生活を様々な面で支えてくださいました．本当にありがとうございました．

2018 年 2 月 1 日
大杉 康仁

参考文献

[1] Ian S Penton-Voak and Jennie Y Chen. High salivary testosterone is linked to masculine male facial appearance in humans. *Evolution and Human Behavior*, Vol. 25, No. 4, pp. 229 – 241, 2004.

[2] Jean Abitbol, Patrick Abitbol, and Béatrice Abitbol. Sex hormones and the female voice. *Journal of Voice*, Vol. 13, No. 3, pp. 424 – 446, 1999.

[3] Sarah A. Collins and Caroline Missing. Vocal and visual attractiveness are related in women. *Animal Behaviour*, Vol. 65, No. 5, pp. 997 – 1004, 2003.

[4] Harriet M. J. Smith, Andrew K. Dunn, Thom Baguley, and Paula C. Stacey. Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology*, Vol. 14, No. 1, p. 1474704916630317, 2016.

[5] 田中章浩. 顔と声による情動の多感覚コミュニケーション. 認知科学, Vol. 18, No. 3, pp. 416–427, 2011.

[6] Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. ‘putting the face to the voice‘: Matching identity across modality. *Current Biology*, Vol. 13, No. 19, pp. 1709 – 1714, 2003.

[7] Lorin Lachs and David B. Pisoni. Crossmodal source identification in speech perception. *Ecological Psychology*, Vol. 16, No. 3, pp. 159–187, 2004.

[8] L. W. Mavica and E. Barenholtz. Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 39, No. 2, pp. 307–312, 2013.

[9] Robert M Krauss, Robin Freyberg, and Ezequiel Morsella. Inferring speakers ’ physical attributes from their voices. *Journal of Experimental Social Psychology*, Vol. 38, No. 6, pp. 618 – 625, 2002.

[10] Harriet M. J. Smith, Andrew K. Dunn, Thom Baguley, and Paula C. Stacey. Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics*, Vol. 78, No. 3, pp. 868–879, Apr 2016.

[11] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, Vol. 35, No. 4, pp. 399–458, dec 2003.

[12] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

[13] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[14] C. Ding and D. Tao. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia*, Vol. 17, No. 11, pp. 2049–2058, Nov 2015.

[15] Mahdi M Kalayeh, Boqing Gong, and Mubarak Shah. Improving facial attribute prediction using semantic segmentation. *arXiv preprint arXiv:1704.08740*, 2017.

[16] Masaki Saito and Yusuke Matsui. Illustration2vec: a semantic vector representation of illustrations. In *SIGGRAPH Asia 2015 Technical Briefs*, p. 5. ACM, 2015.

[17] Nawaf Almudhahka, Mark Nixon, Jonathon Hare, et al. Automatic semantic face recognition. 2017.

[18] Matthew Turk and Alex P Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pp. 586–591, 1991.

[19] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Vol. 313, No. 5786, pp. 504–507, 2006.

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2014.

[21] Angshul Majumdar, Richa Singh, and Mayank Vatsa. Face verification via class sparsity based supervised encoding. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 39, No. 6, pp. 1273–1280, 2016.

[22] Rongbing Huang, Chang Liu, Guoqi Li, and Jiliu Zhou. Adaptive deep supervised autoencoder based image reconstruction for face recognition. *Mathematical Problems in Engineering*, Vol. 2016, , 2016.

[23] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pp. 1133–1141. IEEE, 2017.

[24] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani,

M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. 2014.

[26] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 397–403, 2013.

[27] Dhananjay Rathod, Vinay A, Shylaja SS, and S Natarajan. Facial landmark localization-a literature survey. *International Journal of Current Engineering and Technology*, Vol. 4, No. 3, pp. 1901–1907, 2014.

[28] 向田茂, 蒲池みゆき, 尾田政臣, 加藤隆, 吉川左紀子, 赤松茂, 千原國宏. 操作性を考慮した顔画像合成システム: Futon-顔認知研究のツールとしての評価. 電子情報通信学会論文誌 A, Vol. 85, No. 10, pp. 1126–1137, 2002.

[29] T. Baltrusaitis, P. Robinson, and L. P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 354–361, Dec 2013.

[30] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

[31] 中川聖一（編）. 音声言語処理と自然言語処理. コロナ社, 2013.

[32] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72–83, Jan 1995.

[33] 鈴木颯. テンソル分解に基づく音声表現とその言語識別・話者識別への応用. Master's thesis, 東京大学 大学院工学系研究科 電気系工学専攻, 2016.

[34] 小林隆夫. 音声のケプストラム分析, メルケプストラム分析. 電子情報通信学会技術研究報告. DSP, ディジタル信号処理, Vol. 98, No. 261, pp. 33–40, sep 1998.

[35] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, Vol. 13, No. 5, pp. 308–311, May 2006.

[36] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, Vol. 10, No. 1, pp. 19–41, 2000.

[37] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798, May 2011.

[38] Roland Kuhn, Jean-Claude Junqua, Patrick Nguyen, and Nancy Niedzielski. Rapid speaker adaptation in eigenvoice space. *Speech and Audio Processing, IEEE Transactions on*, Vol. 8, No. 6, pp. 695–707, 2000.

[39] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Proc. Interspeech 2017*, pp. 999–1003, 2017.

[40] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, Vol. 87, No. 4, pp. 1738–1752, 1990.

[41] 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.

[42] Tomoki Toda, Alan W Black, and Keiichi Tokuda. 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.

[43] Harold Hotelling. Relations between two sets of variates. *Biometrika*, Vol. 28, No. 3/4, pp. 321–377, 1936.

[44] Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[45] Hidetsugu Uchida, Daisuke Saito, and Nobuaki Minematsu. Acoustic-to-articulatory mapping based on mixture of probabilistic canonical correlation analysis. *Interspeech*, 2017.

[46] 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.

[47] 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, Mar 2014.

[48] 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, Sep 2014.

[49] 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, Mar 2015.

[50] 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.

[51] Karl Ricanek Jr and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pp. 341–345. IEEE, 2006.

[52] T. Baltrušaitis, P. Robinson, and L. P. Morency. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, March 2016.

[53] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37 of *Proceedings of Machine Learning Research*, pp. 448–456, 07–09 Jul 2015.

[54] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[55] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Vol. 15, pp. 315–323, 2011.

[56] 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.

発表文献

国際会議

[1] Yasuhito Ohsugi, Daisuke Saito, Nobuaki Minematsu. Experimental study of impression-based statistical mapping between speakers' faces and their voices. In *Program of The 5th Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan*, 2016.

国内研究会・全国大会

[2] 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. 情報処理学会音楽情報科学研究会（SIGMUS）, 2016.

[3] 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. 情報処理学会音声言語情報処理研究会（SIG-SLP）, 2017.

[4] 大杉康仁, 齋藤大輔, 峯松信明. 顔から声への統計的対応付けに関する技術的諸検討. 情報処理学会音楽情報科学研究会（SIGMUS）, 2017.

[5] 大杉康仁, 齋藤大輔, 峯松信明. 確率的潜在変数を仮定した顔から声への統計的対応付けの検討. 日本音響学会秋季講演論文誌, 2017.

学位論文

[6] 大杉康仁. 声・顔の固有空間と GMM に基づく両空間の印象的対応付けの検討. 東京大学工学部電子情報工学科卒業論文, 2016.

付録 A

Eigenface

A.1 概要

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 [18]．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 $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, ..., \mathbf{\Gamma}_M$ に対し，式 (A.1) により平均顔 $\mathbf{\Psi}$ を求め，各顔画像と平均顔画像との差を式 (A.2) で表す．式 (A.3) の共分散行列 $\boldsymbol{C}$ 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．

$$\mathbf{\Psi} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{\Gamma}_m \tag{A.1}$$

$$\mathbf{\Phi}_m = \mathbf{\Gamma}_m - \mathbf{\Psi} \tag{A.2}$$

$$\boldsymbol{C} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{\Phi}_m \mathbf{\Phi}_m^T \tag{A.3}$$

ある画像 $\mathbf{\Gamma}_m$ は，重み $w^{(m)}(i)(i = 1, 2, ..., M')$ を用いて式 (A.4) で近似される．

$$\mathbf{\Gamma}_m \simeq \sum_{i=1}^{M'} w^{(m)}(i) \boldsymbol{E}(i) + \mathbf{\Psi} \tag{A.4}$$

A.2 実験

i) 実験条件

顔画像データベース MORPH [51] の約 54,000 枚の顔画像を用いて実際に Eigenface を抽出した．顔画像は縦 100 横 100 の白黒画像に統一し，式 (A.4) の $\mathbf{\Gamma}_m$ を 10,000 次元のベクトルとした．

ii) 実験結果

累積寄与率が 80% を超えた第 13 主成分までを Eigenface とし，それらと平均顔を図 A.1 に示す．図 A.2 のように Eigenface に重みを付けて平均顔に加算すると，合成された画像は人間の顔として判別可能な範囲が減った画像となってしまった．これは，Eigenface が白黒顔画像の濃淡値のみを対象とした主成分分析（PCA）であるため画像に写っている顔の特徴よりも撮影環境や画像の劣化度などの画像としての特徴が優先されてしまったためであると考えられる．従って，Eigenface は顔の印象を表す特徴量としては不適切であると考えられる．

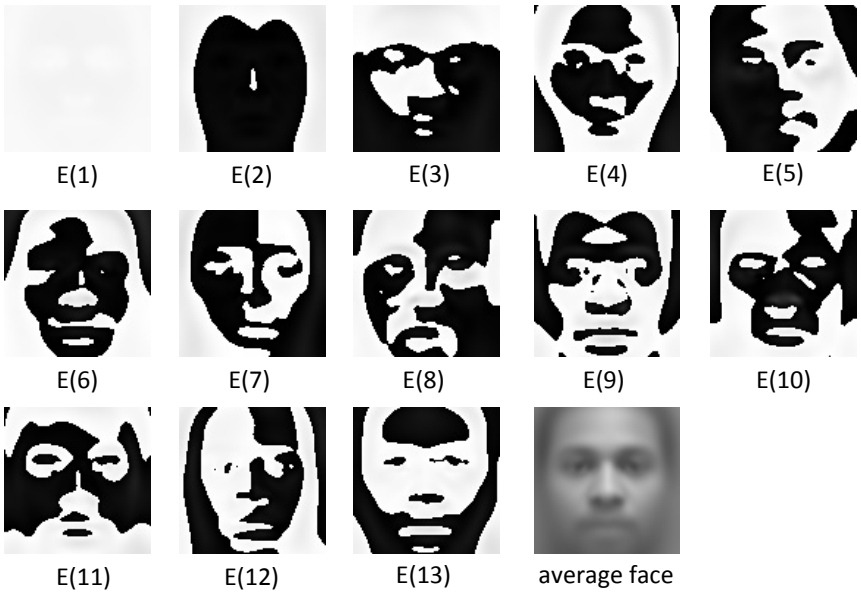


図 A.1: 低次 13 次元の Eigenface と平均顔 (average face): $E(i)$ は i 番目の Eigenface を表す.

図 A.2: Eigenface を用いた新規画像の作成: 式 (A.4) で $M' = 3$ とし，$[w(1), w(2), w(3)] = [0.9, 0.21, 0.16]$ とした場合

付録B

Eigenvoice Conversion (EVC)

B.1 概要

一人の話者から複数の話者への声質変換を考えた時，GMM に基づく声質変換では，GMM の学習には同一発話内容の入出力音声対から成るパラレルデータが必要であるため，同じ口調で決められた文を発声することは文の数が少ない場合でも話者にとって大変な負荷となる．さらに，所望の入出力話者ごとに GMM を構成する必要があるため事前に音声を収録可能な話者の声質しか表現できない．一方，固有声に基づく声質変換法 (Eigenvoice Conversion: EVC) は，予め収録した多数の入出力話者のパラレルデータにより任意の話者の声質を実現可能な手法の一つである [41]. EVC には，一人の入力話者の声質から複数の出力話者の声質へ変換する一対多変換法と複数の入力話者の声質から一人の出力話者の声質へ変換する多対一変換法があるが，以下では一対多変換法について説明する．

B.2 一対多声質変換のためのパラメータ学習

前項と同様に入力話者の音響特徴量と出力話者の特徴量を連結したベクトルが GMM に従うものとするが，m 番目の分布の出力話者の平均ベクトル $\boldsymbol{\mu}_m^{(Y)}$ を各列が Eigenvoice である行列 $\boldsymbol{B}_m$ を用いて式 (B.1) のように表す．

$$\boldsymbol{\mu}_m^{(Y)} = \boldsymbol{B}_m \boldsymbol{w}^{(Y)} + \boldsymbol{b}_m^{(0)} \tag{B.1}$$

このとき構成される GMM を EV-GMM と呼び，学習パラメータ $\lambda^{(EV)}$ は重みベクトル $\boldsymbol{w}^{(Y)}$・各分布の重み α_m・入力の平均ベクトル $\boldsymbol{\mu}_m^{(X)}$・出力のバイアスベクトル $\boldsymbol{b}_m^{(0)}$・出力の基底ベクトルから成る行列 $\boldsymbol{B}_m$・共分散行列 $\boldsymbol{\Sigma}_m^{(X,Y)}$ の 6 つである．これらの内，$\boldsymbol{w}^{(Y)}$ は出力話者に依存し，α_m・$\boldsymbol{\mu}_m^{(X)}$・$\boldsymbol{b}_m^{(0)}$・$\boldsymbol{B}_m$・$\boldsymbol{\Sigma}_m^{(X,Y)}$ は全出力話者において共通である．EV-GMM の学習は大きく 3 つの段階に分けられる．

Step1 ある話者を入力として全ての話者を出力とした不特定話者 GMM($\lambda^{(0)}$) を学習する．入力話者の音響特徴量 $\boldsymbol{X}_t$ と出力話者 s の音響特徴量 $\boldsymbol{Y}_t^{(s)} (s = 1, 2, ..., S)$ を連結し $\boldsymbol{Z}_t^{(s)}$ を得る．

$$\lambda^{(0)} = \arg\max_{\lambda} \prod_{s=1}^{S} \prod_{t=1}^{T_s} p(\boldsymbol{Z}_t^{(s)} | \lambda) \tag{B.2}$$

ここで，T_s は，話者 s の音声データのフレーム数である．式 (B.2) は，全ての出力話者に関して結合確率密度を最大化する点で式 (2.30) と異なる．

Step2　$\lambda^{(0)}$ のパラメータの内，式 (B.1) における $\boldsymbol{\mu}_m^{(Y)}$ のみを更新し，S 人の出力話者それぞれに応じた出力話者依存 GMM $\lambda^{(s)}$ を学習する．

$$\lambda^{(s)} = \arg\max_{\lambda} \prod_{t=1}^{T_s} p(\boldsymbol{Z}_t^{(s)}|\lambda) \tag{B.3}$$

Step3　それぞれの出力話者の各分布の平均を連結して GMM-SV$\boldsymbol{\nu}^{(s)}$ を作り，全出力話者のスーパーベクトルに関する主成分分析を行うことで $\boldsymbol{b}_m^{(0)}$ と $\boldsymbol{B}_m$ を求める．式 (B.4) により，全話者の $\boldsymbol{\nu}^{(s)}$ の平均から $\boldsymbol{b}_m^{(0)}$ を求める．

$$\boldsymbol{b}^{(0)} = [\boldsymbol{b}_1^{(0)\mathrm{T}}, \boldsymbol{b}_2^{(0)\mathrm{T}}, ..., \boldsymbol{b}_M^{(0)\mathrm{T}}]^{\mathrm{T}} = \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\nu}^{(s)} \tag{B.4}$$

$\boldsymbol{v}^{(s)} = \boldsymbol{\nu}^{(s)} - \boldsymbol{b}^{(0)}$ とし，行列 $\boldsymbol{V} = [\boldsymbol{v}^{(1)}\ \boldsymbol{v}^{(2)}\ ...\ \boldsymbol{v}^{(S)}]$ を特異値分解することで，話者に依存しない基底行列 $\boldsymbol{U} = [\boldsymbol{B}_1^{\mathrm{T}}\ \boldsymbol{B}_2^{\mathrm{T}}\ ...\ \boldsymbol{B}_M^{\mathrm{T}}]^{\mathrm{T}}$ と話者に依存する重み行列 $\boldsymbol{W} = [\boldsymbol{w}^{(1)}\ \boldsymbol{w}^{(2)}\ ...\ \boldsymbol{w}^{(S)}]$ を得る．ここで，$\boldsymbol{U}$ は $\boldsymbol{V}$ の左特異ベクトルを列とする行列であり，$\boldsymbol{W}$ は $\boldsymbol{V}$ の特異値行列 $\boldsymbol{A}$ と右特異ベクトルを列とする行列 $\boldsymbol{D}$ の転置行列の積である．

$$\boldsymbol{V} = \boldsymbol{U}\boldsymbol{A}\boldsymbol{D}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{W} \tag{B.5}$$

よって，各話者の GMM-SV$\boldsymbol{\nu}^{(s)}$ は式 (B.6) で表される．

$$\boldsymbol{\nu}^{(s)} = \boldsymbol{U}\boldsymbol{w}^{(s)} + \boldsymbol{b}^{(0)} \tag{B.6}$$

Step1 で得られた $\lambda^{(0)}$ と Step3 で得られた $\boldsymbol{B}_m$，$\boldsymbol{b}_m^{(0)}$ から EV-GMM($\lambda^{(EV)}$) を構成する．また，$\boldsymbol{U}$ および $\boldsymbol{B}$ は直交行列であるため式 (B.7) が成り立つ．

$$\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} = \boldsymbol{A}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{A} = \boldsymbol{A}^2 \tag{B.7}$$

$\boldsymbol{A}^2$ の対角成分は話者に関する共分散行列 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$ の固有値であり，各固有値に対応する $\boldsymbol{U}$ の列が $\boldsymbol{V}$ の固有ベクトルである．固有値が大きい固有ベクトルを Eigenvoice として選択することで，式 (B.6) における重みベクトル $\boldsymbol{w}^{(s)}$ の次元数を減らすことが可能である．

B.3　一対多声質変換法

学習した EV-GMM を用いて，特定の重みベクトル $\boldsymbol{w}^{(tar)}$ で表される声質の音声を得るには，式 (B.1) を式 (B.8) に置換し 2.5.1 節と同様に最尤基準で声質変換を行う．このとき，式 (2.36) の $\boldsymbol{E}_{m,t}^{(Y)}$ は式 (B.10) で置換される．

$$\boldsymbol{\mu}_m^{(tar)} = \boldsymbol{B}_m\boldsymbol{w}^{(tar)} + \boldsymbol{b}_m^{(0)} \tag{B.8}$$

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_m^{(tar)} + \boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)-1}(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}) \tag{B.9}$$

$$= \boldsymbol{B}_m\hat{\boldsymbol{w}} + \boldsymbol{b}_m^{(0)} + \boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)-1}(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}) \tag{B.10}$$

B.4　話者の重みの推定

重みベクトル $\boldsymbol{w}^{(tar)}$ は手動で決定することも可能だが，所望出力話者の音声が存在する場合はその音声の音響特徴量ベクトル系列を $\boldsymbol{Y}^{(tar)}$ とし式 (B.11) に基づき $\boldsymbol{w}^{(tar)}$ を推定することも可能である．

$$\boldsymbol{w}^{(tar)} = \arg\max_{\boldsymbol{w}} \int p(\boldsymbol{X}, \boldsymbol{Y}^{(tar)}|\lambda^{(EV)})\mathrm{d}X = \arg\max_{\boldsymbol{w}} p(\boldsymbol{Y}^{(tar)}|\lambda^{(EV)}) \tag{B.11}$$

$\boldsymbol{w}^{(tar)}$ の推定は式 (B.12) の補助関数を最大化することで行い，その結果得られる $\boldsymbol{w}$ の推定式は式 (B.13) で表される．

$$Q(\boldsymbol{w}, \hat{\boldsymbol{w}}) = \sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\boldsymbol{Y}_t^{(tar)}, \lambda^{(EV)}) \log p(\boldsymbol{Y}_t^{(tar)}, m|\hat{\lambda}^{(EV)}) \tag{B.12}$$

$$\hat{\boldsymbol{w}} = \left(\sum_{m=1}^{M} \overline{\gamma_m}^{(tar)} \boldsymbol{B}_m^T \boldsymbol{\Sigma}_m^{(YY)-1} \boldsymbol{B}_m\right)^{-1} \sum_{m=1}^{M} \boldsymbol{B}_m^T \boldsymbol{\Sigma}_m^{(YY)-1} \overline{\boldsymbol{Y}}_m^{(tar)} \tag{B.13}$$

ただし，$\overline{\gamma_m}^{(tar)}$ と $\overline{\boldsymbol{Y}}_m^{(tar)}$ は以下で表される．

$$\overline{\gamma_m}^{(tar)} = \sum_{t=1}^{T} p(m|\boldsymbol{Y}_t^{(tar)}, \lambda^{(EV)}) \tag{B.14}$$

$$\overline{\boldsymbol{Y}}_m^{(tar)} = \sum_{t=1}^{T} p(m|\boldsymbol{Y}_t^{(tar)}, \lambda^{(EV)})(\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \tag{B.15}$$

付録C

VAE・CVAEの分析結果

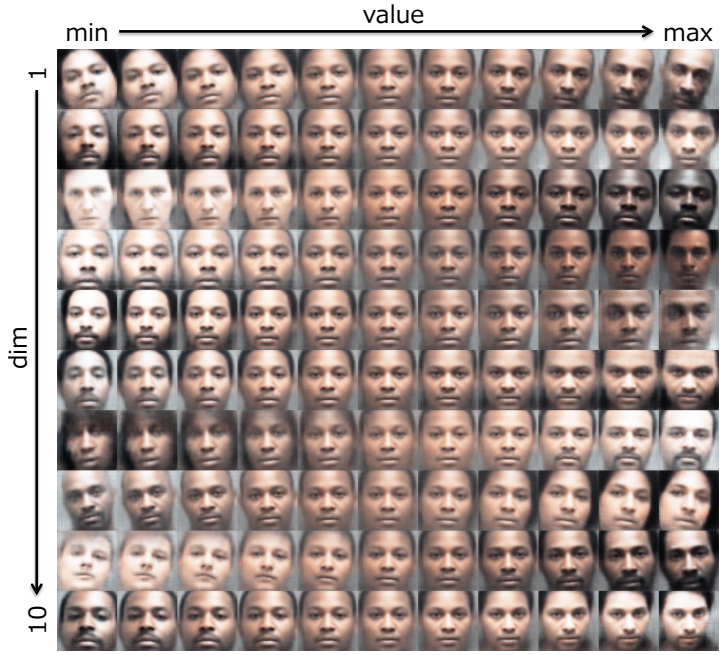


図 C.1: VAE の $\boldsymbol{z}$ の各次元を操作した時の Decoder 出力の変化

4.2.4 節で学習した VAE と 4.2.5 節で学習した CVAE について式 (4.1) を用いて $\boldsymbol{z}$ の各次元を操作した時の変化の様子をそれぞれ図 C.1 と図 C.2・図 C.3・図 C.4 に示す.

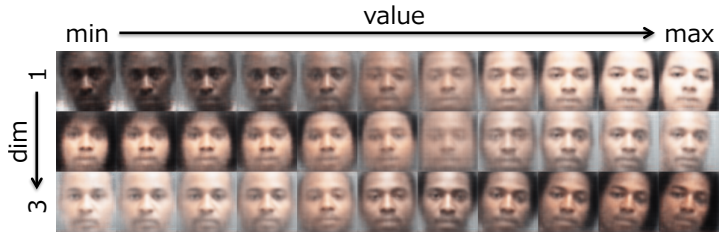


図 C.2: CVAE の $\boldsymbol{z}$ ($d_z = 3$) の各次元を操作した時の Decoder 出力の変化：ラベル $\boldsymbol{y}$ は黒人男性のものを入力した．

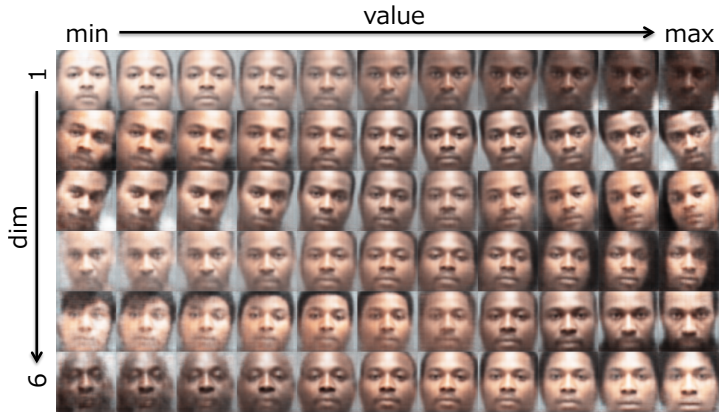


図 C.3: CVAE の $\boldsymbol{z}$ ($d_z = 6$) の各次元を操作した時の Decoder 出力の変化：ラベル $\boldsymbol{y}$ は黒人男性のものを入力した．

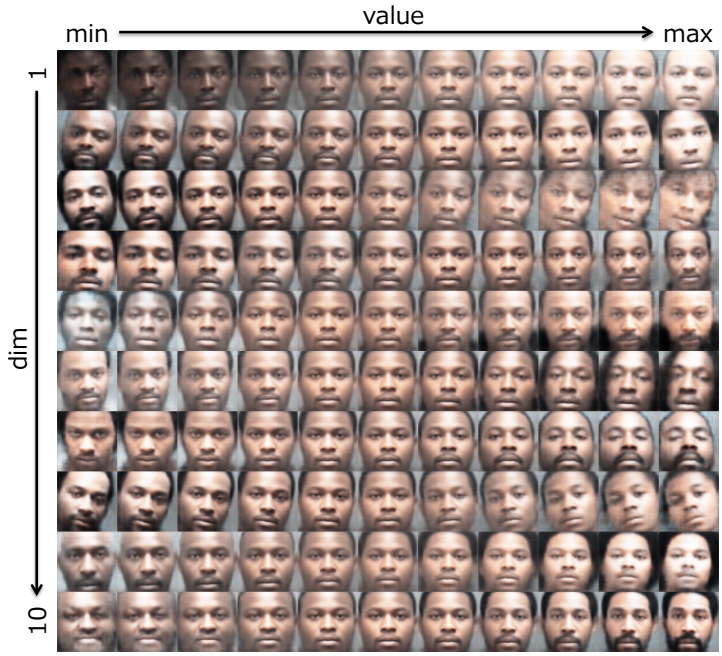


図 C.4: CVAE の $\boldsymbol{z}$ ($d_z = 10$) の各次元を操作した時の Decoder 出力の変化：ラベル $\boldsymbol{y}$ は黒人男性のものを入力した．

付録D

統計的顔声変換の結果の例

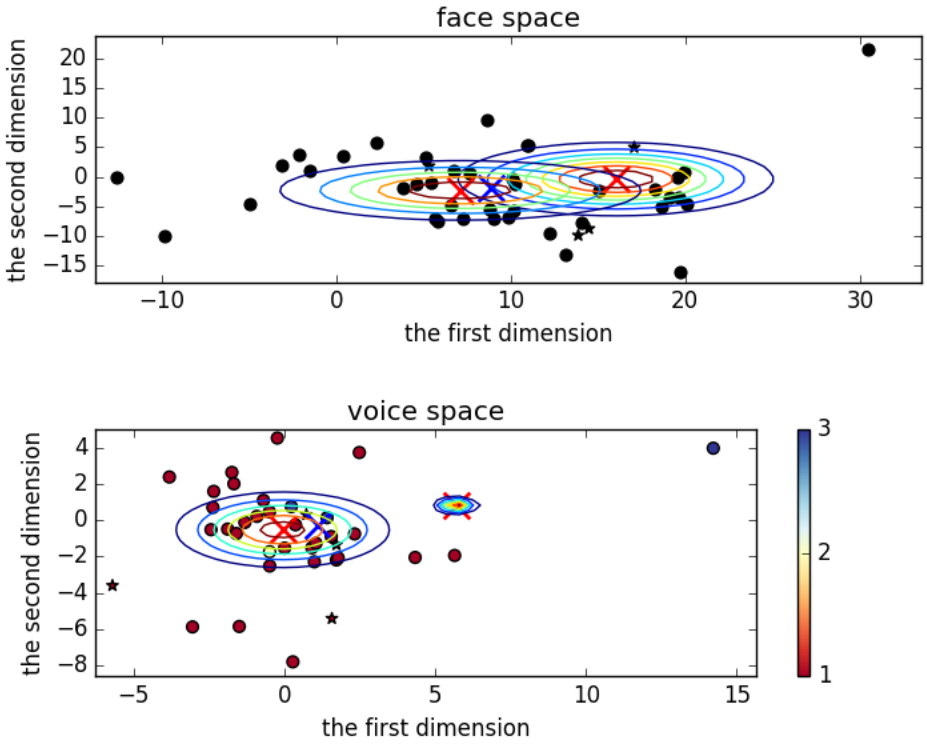


図 D.1: アジア系コーパスを用いた pCCA・mPCCA の例：上のグラフが顔の特徴量の分布を，下のグラフが声の特徴量のグラフを表す．下のグラフのカラーバーが選択回数を表しており，丸印の点が学習データを，星印の点が評価データを表す．青いバツ印が pCCA における平均 $\boldsymbol{b} \cdot \boldsymbol{d}$ を表し，赤いバツ印と等高線は mPCCA における式 (2.62) と式 (2.63) で $\boldsymbol{h} = \boldsymbol{0}$ とした場合の分布を表す．この例においては，声の特徴量の分布の内急峻なものに関する混合重みは 0.200 であった．

4.5 節において，アジア系・日本人コーパスを用いた mPCCA の各分布を可視化した例を図 D.1・図 D.2 に示す．共に顔の特徴量としては IFF を用いているが，それぞれデータセットと被験者は異なる．

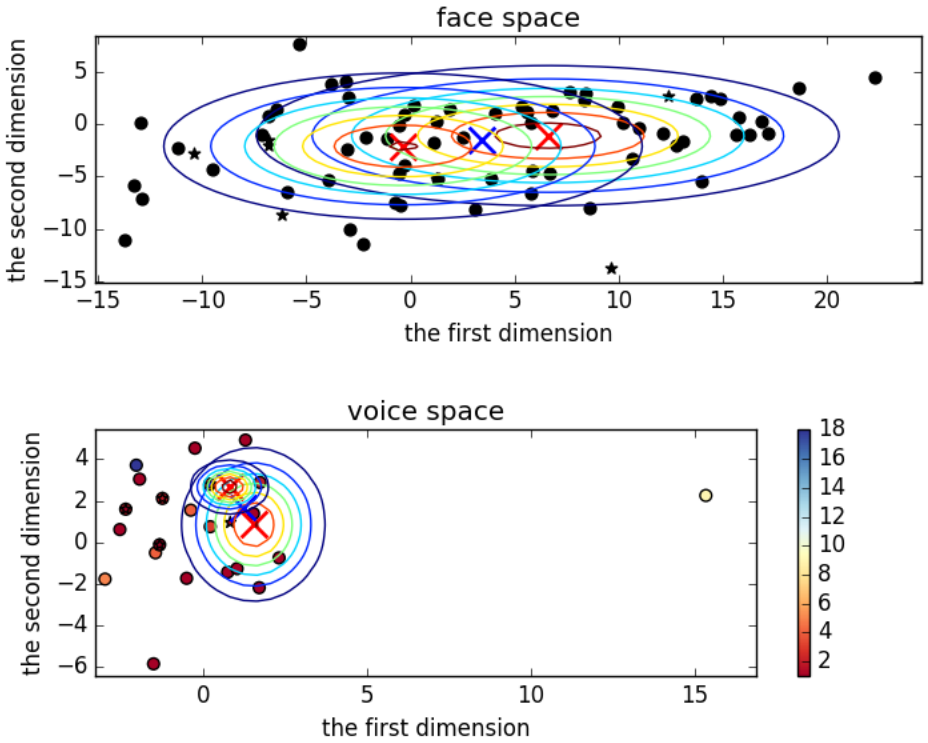


図 D.2: 日本人コーパスを用いた pCCA・mPCCA の例：上のグラフが顔の特徴量の分布を，下のグラフが声の特徴量のグラフを表す．下のグラフのカラーバーが選択回数を表しており，丸印の点が学習データを，星印の点が評価データを表す．青いバツ印が pCCA における平均 $\boldsymbol{b} \cdot \boldsymbol{d}$ を表し，赤いバツ印と等高線は mPCCA における式 (2.62) と式 (2.63) で $\boldsymbol{h} = \boldsymbol{0}$ とした場合の分布を表す．この例においては，声の特徴量の分布の内急峻なものに関する混合重みは 0.266 であった．