博 士 論 文

音声信号と調音状態の一対多の関係の分析及びその応用に関する検討

指導教官　党 建武 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

錦戸 信和

2011 年 3 月 24 日

要 旨

音声生成過程において，声道を形成する各調音器官の様々な位置や形状，すなわち調音状態から同一の音素カテゴリーに含まれる音響特徴量を持つ音声信号を生成することが可能である。この音声信号と調音状態の一対多の関係は，音声信号から調音状態を逆推定する場合に不良設定問題，つまり入力音声に対して逆推定の解となる調音状態が無数に存在する一対多の問題を生じさせる。従来の研究は，形態学的制約条件，動的制約条件，生理学的制約条件を導入し解の多様性を減少させることで一対多の問題の解決を試みているが，完全な解決には至っていない。ここで，腹話術や調音の補償動作の観測実験から，音声信号と一対多の関係にある調音状態には，自然な発話を行う際に観測され得る“自然調音状態”と，生理学的に可能な状態であるが自然な発話を行う際に観測され得ない“不自然調音状態”が含まれると考えられる。この不自然調音状態は，自然調音状態と同様の音響特性を持つ音声信号の生成が可能であるが通常の発話の際には観測され得ないため，逆推定の解の調音状態に多様性をもたらす要因となる。しかしながら，多くの不自然調音状態は従来の逆推定の制約条件を満たすため，逆推定の推定候補に含まれる不自然調音状態を取り除くことが，一対多の問題を解決するために必要と考えられる。

本研究では，音声信号から調音状態を逆推定する際に推定候補に含まれる不自然調音状態を除去することを目的とする。そのために，発話の計算モデルを用いて人間が発話可能な多数の多様な調音状態を生成し，生成した調音状態の分析に基づく分布構造の可視化により，音声信号と一対多の関係にある人間が発話可能な調音状態の分布の全体像を明らかにすることを試みた。さらに，調音状態の分布構造に基づき自然調音状態と不自然調音状態を識別する手法を提案し，提案手法を音声信号から調音状態を逆推定するシステムに適用することにより，逆推定の推定候補に含まれる不自然調音状態の除去を試みた。

上記の結果から，本研究では以下の成果が得られた。

1. 音声信号と一対多の関係にある人間が発話可能な調音状態の分布の全体像を，自然調音状態と不自然調音状態を含む非線形空間上のクラスタ構造として示した。
2. 調音状態のクラスタ構造に基づき自然調音状態と不自然調音状態を識別する手法を提案し，自然調音状態に対して 97%，不自然調音状態に対して 99%の精度で識別可能なことを示した。
3. 提案した識別手法を音声信号から調音状態の逆推定に適用した新たな逆推定システムを構築した。
4. 構築したシステムによる逆推定の結果，推定候補に含まれる不自然調音状態を 9 割除去可能なことを示した。

本研究で得られた上記の成果は，音声信号から調音状態の逆推定に大きく寄与するだけでなく，聴覚障害者や語学の学習者のための理想的な発話訓練システムの実現に貢献できると考えられる。また，本研究で得られた知見は，人間の音声生成機構の解明や，音声合成の研究に大きく寄与できると考えられる。

Abstract

In process of speech production, a speech sound within the same category can be produced by various articulations with different positions or configurations of speech organs. Such a relationship between speech sound and articulation brings an ill-posed problem, namely the one-to-many problem, in inverse estimation of articulation from speech sound. In other words, there may be countless solutions from an inverse estimation for a given speech sound. In previous studies, the one-to-many relationship has been reduced to some extent by employing morphological, dynamic and physiological constraints. Nevertheless, the one-to-many problem is far from being solved. Here, observations from ventriloquism or articulatory compensation indicate that there exist two types of basic articulations that can produce speech sounds with the same category. Both of them can be physiologically realized by humans. But, one kind of articulation appears in natural speech, while the other does not appear in natural speech. The former is referred to as "natural articulation", and the latter is "unnatural articulation". The unnatural articulations cause the one-to-many problem in the inverse estimation since some unnatural articulations satisfy the above constraints. Therefore, it is necessary to exclude unnatural articulations from candidates of the inverse estimation for solving the one-to-many problem.

The purpose of this study is to exclude unnatural articulations from candidates of the inverse estimation of articulation. For this purpose, we generated a great variety of possible articulations using a physiological articulatory model, and visualized the articulatory structure based on analyzing the articulations generated. Moreover, we proposed a method for discriminating between natural and unnatural articulations, and excluded the unnatural ones from candidates in the inverse estimation by applying the proposed method to the inverse estimation system.

The following outcomes are indicated from this study.

1. the structure of articulations in the one-to-many relationship was clarified by clusters in non-linear space.
2. a method for discriminating between natural and unnatural articulations was proposed, and natural articulations could be discriminated with accuracy more than 97%, and unnatural articulations could be discriminated with accuracy more than 99%.
3. the inverse estimation system for articulations was constructed with applying the proposed method for the inverse estimation.
4. 90% of unnatural articulations from inverse estimation was excluded.

The above outcomes are not only benefit for the inverse estimation of articulation from speech sound, but also help for people with hearing difficulties, or for acquiring a second language.

目 次

図 目 次

表 目 次

第 1章

序論

1.1 はじめに

人から人へ情報を伝達する際に，音声は広く用いられている手段である。その音声による情報伝達の源である人間の音声生成過程は，大まかに次のとおりである [1]。まず，脳において発話内容が想起され，その内容に対応した音声学的特徴系列が生成される。その後，生成された音声学的特徴系列に応じた運動指令系列が各音声器官を動かす筋に送られ，運動指令系列により各音声器官の運動が制御された結果，音声が生成される。この音声生成過程において制御される音声器官には，舌や下顎，口唇，軟口蓋，喉頭などの調音または発声器官が含まれ，音声を生成するためにはこれらの複数の音声器官による協調運動の制御が必要となる。ここで，上下肢や手指などの人間の随意運動制御において，制御対象を滑らかに素早く運動させるためには，制御対象の運動目標（運動軌跡）からもとの運動指令を求める逆ダイナミクスの計算が必要となることが，シミュレーションや観測実験により示されている [2]。音声器官の運動も随意運動であり，音声生成においても運動目標から運動指令を求める逆ダイナミクスの処理が行われていると考えられるが，その具体的な機構はよく分かっていない。また，音声生成の場合は，音声器官を制御するための運動目標もまだ明らかになっていない [3]。そのため，音声信号から舌の位置や形状などの調音状態を逆推定するための研究は，音声生成機構を解明する上で，また音声生成における運動目標を明らかにする意味でも重要な研究である。さらに，調音状態の逆推定を実現することができれば，推定結果を可視化することにより，聴覚障害者の発話訓練や外国語の学習のための理想的なシステムの構築が可能となる [4]。

しかしながら，同一の音素カテゴリーに含まれる音響特性を持つ音声信号と，その音声信号を生成可能な調音状態との間には一対多の関係 [5] があり，そのために音声信号から調音状態の逆推定は不良設定問題となる。つまり，音声信号から調音状態を逆推定する場合，一対多の関係から入力音声に対して理論的に無数の調音状態が推定候補となり，単純に解を一意に定めることはできない。これまで多くの研究者が逆推定における一対多の問題に取り組み，音声信号から調音状態を逆推定するための努力が行われてきた [6 - 11]。その結果，空間的（形態学的）[6]，動的 [7, 8] 及び生理学的制約条件 [11] を導入し一対多の関係性を抑えることにより

調音状態の推定候補の多様性を減らすこと成功しているが，問題の完全な解決には至っていない。

ここで，腹話術師を被験者とした観測実験の結果に基づき同じ音声信号を生成可能な調音状態について考える。なお，ここでいう同じ音声信号とは，音声信号から抽出された音響特徴量が同じ音素カテゴリーに含まれることを意味する。伊福部は，腹話術により発話された音声と通常の音声の音響特徴量が一致することを確認している [12]。この観測結果は，同じ音声信号を生成できる人間が発話可能な調音状態には，2種類の状態が含まれることを示している。一つは，自然な発話を行う際に観測され得る調音状態であり，これを“自然調音状態”と呼ぶこととする。もう一つは，自然調音状態と同じ音声信号を生成可能であり，生理学的に発話可能な状態であるが自然な発話を行う際に観測され得ない調音状態であり，これを“不自然調音状態”と呼ぶこととする。観測実験における通常の発話は自然調音状態に，腹話術は不自然調音状態に該当する。

この自然調音状態と不自然調音状態は音声信号に対して一対多の関係を持つが，人間は日常において自然調音状態のみを用いる。これは，日常の発話において腹話術の状態が現れないことから明白であり，人間が言語を獲得する際に調音状態の取捨選択が行われた結果，自然調音状態が獲得されたと考えられる。つまり，音声生成過程において，自然調音状態は“獲得された調音状態”，不自然調音状態は“淘汰された調音状態”と捉えることができる。従って，調音状態の逆推定における一対多の問題を解決するためには，自然調音状態と不自然調音状態を識別し，推定候補から不自然調音状態を除去する必要がある。

不自然調音状態の除去は，同じ音声信号を生成し人間が発話可能な調音状態の分布の全体像に基づくことにより，適切に行うことができると考えられる。しかしながら，従来の研究では，不自然調音状態に関する詳細な分析は行われておらず，自然調音状態と不自然調音状態の分布の関係も明らかにされていない。また，調音状態の分布の全体像を把握するためには，人間が発話可能な多数の多様な調音状態の分析が必要となるが，そのような調音状態を被験者を用いた観測から得ることは困難である。よって，本研究では，まず音声信号と一対多の関係にあり人間が発話可能な調音状態の分布を明らかにするために，人間の発話運動を精度良く再現可能な計算モデルを用いて系統的に多様な調音状態を生成する。そして，

その調音状態を被験者を用いて観測された発話器官の調音運動や音声信号に基づき分析することにより，調音状態の分布構造を明らかにする。さらに，得られた調音状態の分布構造に基づき，音声信号から調音状態の逆推定の推定候補に含まれる不自然調音状態を取り除く。

次節では本研究の背景として，従来の音声信号と調音状態との一対多の関係に関する研究及び，音声信号から調音状態の逆推定に関する研究について述べる。

1.2 本研究の背景

1.2.1 音声信号と調音状態の一対多の関係に関する研究

音声生成において，声道を形作る各調音器官の位置や形状，すなわち調音状態が決まれば，同一の音源に基づく音声信号は一意に決まる。しかしながら，逆に同一の音声信号を生成可能な調音状態は無数に存在する。このような音声信号と調音状態との一対多の関係は古くから知られている。Schroeder は声道を理想的な音響管と仮定し，ホルマント周波数のみから声道の断面積関数を一意に決められないことを明らかにした [5]。また Atal らは，異なる声道形状から生成された音響信号がほぼ等しいホルマント周波数と振幅を持つことを計算シミュレーションにより示した [6]。しかしながら，これらの計算シミュレーションで扱われている調音状態は，人間が発話可能な状態であることが保証されていない。

音声信号と調音状態の一対多の関係は，計算シミュレーションだけでなく，実際に被験者を用いた観測によっても示されている。伊福部は，腹話術師が普通に発話した音声のホルマント周波数と，腹話術を用いて発話した音声のホルマント周波数がほぼ等しいことを確認した [12]。また Lindblom らは，バイトブロックにより下顎を固定した状態で発話されたスウェーデン語母音のホルマント周波数が，自然なホルマント周波数の範囲内に含まれることを示している [13]。しかしながら，これらの研究では，被験者の許可がないため詳細な分析が行えない場合があり，また調音状態を観測するための大規模な測定機器の必要性や被験者への負担などの制限からデータ数はわずかである。さらに，それぞれの研究は個別の事象に対する分析で終わっており，音声生成過程における音声信号と一対多の関係に

ある調音状態という大局的な視点での分析は行われていない。なお，前述の通り腹話術による発話は不自然調音状態に該当し，バイトブロックを伴う発話も不自然調音状態に該当する。

1.2.2 音声信号から調音状態の逆推定に関する研究

音声信号から調音状態の逆推定に関する研究は，古くから行われている。Atal らは，声道断面積関数のパラメータとその声道断面積関数から求めた音響パラメータセットに基づき一対多の問題に対する空間的制約を示した [6]。Schroeter と Sondhi は，調音運動の逆推定に幾何学的調音モデルに基づき構築した調音音響対コードブックを用いた [7]。このコードブックを用いることにより形態学的制約が導入され，さらに調音運動の軌跡を最適化することにより動的制約も取り入れられている。鈴木らは，調音音響対コードブックを調音音響同時観測データに基づき構築し，調音運動の逆推定を行った [8]。また，Hiroya と Honda は，観測データに基づきパラメータ学習した確率モデルを用いて，調音運動の逆推定を行っている [9]。

一方，白井と誉田は，幾何学的調音モデルのパラメータを音声から直接推定することにより，調音運動の逆推定を行っている [10]。実測値の分析結果に基づき調音モデルの定数及びパラメータの変動範囲を定めることにより，形態学的制約が考慮されている。さらに，調音パラメータを逆推定する際の評価関数にパラメータの連続性に関する項を含めることにより，動的制約が加えられている。また，Dang と Honda は部分 3 次元生理学的発話機構モデル [14] を構築し，モデルのパラメータの逆推定を行った [11]。この発話機構モデルは，成人男性 1 名の MRI 画像に基づき構築され，発話機構の形状学的及び生理学的側面を忠実に再現している。従って，発話機構モデルに空間的，動的及び生理学的制約が内包されているため，モデルのパラメータを推定することにより制約条件が推定結果に効率的に反映される。

上記の研究では一対多の問題に対して推定結果の正誤にしか着目されておらず，音声信号に対して一対多の関係にある調音状態の分布に関する詳細な分析は行われていない。これは，同一の音素カテゴリーに含まれる音声信号を生成しうる人間が発話可能なすべての状態を観測することは困難なためと考えられる。

ここで，音声信号と一対多の関係にある調音状態空間において上記の制約条件

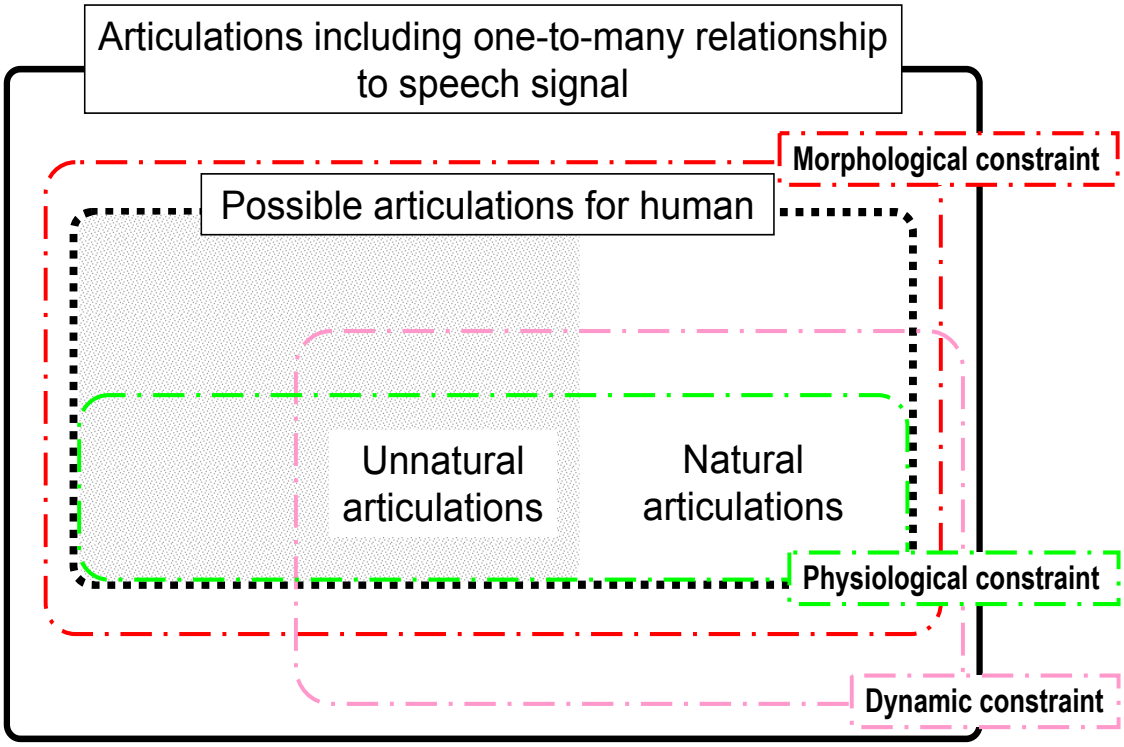

図 1.1: 音声信号と一対多の関係にある調音状態空間における逆推定の制約条件を満たす範囲

を満たす範囲を図 1.1 に示す。図 1.1 において，実線の四角形は音声信号と一対多の関係にある調音状態空間を，点線の四角形はその一部である人間が生理学的に発話可能な調音状態の範囲を表し，点線内の色の付いた部分は不自然調音状態の範囲を，色の付いていない部分は自然調音状態の範囲を表している。また，一点鎖線の四角形は各制約条件を満たす範囲を表している。図 1.1 に示されている通り，従来の制約条件をすべて満たす調音状態には，自然調音状態だけでなく不自然調音状態も含まれる。これは，不自然調音状態は人間が発話可能であり，連続音素の発話も可能なためである。しかしながら，不自然調音状態は，自然調音状態と同じ音素カテゴリーに含まれる音響特性を持つ音声信号を生成可能であるにもかかわらず，通常の発話では観測されない。よって，調音状態の逆推定の候補に含まれる不自然調音状態を取り除くことは，逆推定における一対多の問題を解決するための新たな制約条件となり得る。

なお，音声信号から調音状態を逆推定する場合，音声認識の技術を用いて音素

を認識し，その音素に対応する調音状態を提示する方法も考えられる [15]。この方法は前述の方法に比べ簡易であり，システムの規模や計算量を小さく抑えることが可能である。しかしながら，この方法で提示可能な調音状態は発話者自身の実際の状態を正しく反映していないため，音声生成機構の解明という観点で適切な方法ではない。また，調音状態の逆推定の技術を発話訓練システムなどに応用するためには，推定結果が実際に音声を発話した際の状態を正しく反映していることが重要と考えられる。

1.3 本研究の目的

本研究の目的は，音声信号から調音状態を逆推定する際の一対多の問題を解決するために，推定候補に含まれる不自然調音状態を取り除くことである。そのために，まず人間が発話可能な多数の多様な調音状態を計算機モデルを用いて生成し，生成した調音状態を分析することにより，人間が発話可能な調音状態の分布構造を明らかにすることを試みる。さらに，調音状態の分布構造に基づき，自然調音状態と不自然調音状態を識別する手法を提案し，提案手法を音声信号から調音状態の逆推定システムに適用することにより逆推定の推定候補から不自然調音状態を取り除くことを試みる。なお，本研究では母音を逆推定の対象とする。また，本研究で扱う自然調音状態及び不自然調音状態は，観測された音声信号と発話器官の調音運動に基づき定めた定常部において同一の音素カテゴリーに含まれる音響特性を持つ音声信号を生成可能とする。

1.4 本論文の構成

2 章以降は次のように構成される。まず，2 章では生理学的発話機構モデルを用いた人間が生理学的に発話可能な調音状態の生成及び，観測された音声信号の音響特性に基づく日本語 5 母音の調音状態の選定方法について述べ，その結果を示す。また，2 章以降で生成された調音状態の分析に用いられる，X-ray microbeam system により同時観測された発話器官の調音運動と音声信号についても述べる。3 章では，生成された調音状態の中から選定された 5 母音の調音状態を，観測され

た発話器官の調音運動に基づき自然調音状態と不自然調音状態に分類する方法について述べ，その結果を示す。さらに，自然調音状態と不自然調音状態の分布間の重複度について検討し，重複度を減少させる非線形空間上で特徴量間の類似性を考慮した次元圧縮を行うことにより，調音状態の分布構造を明らかにする。また，日本語5母音の取り得る不自然調音状態の形状とその傾向についても示す。4章では，調音状態の分布構造に基づき自然調音状態と不自然調音状態を識別する手法を提案し，識別精度について検討する。5章では，逆推定の推定候補から不自然調音状態を除去するため，提案した調音状態の識別手法を逆推定に適用した新たな調音状態の逆推定システムを構築する。さらに，構築したシステムを用いて音声信号から調音状態の逆推定実験を行い，逆推定の結果の調音状態に対する不自然調音状態の除去の精度を検証する。6章では，本研究を通して得た成果について考察し，最後に7章で，本研究の成果についてまとめ，今後の課題に対する対応方針を述べる。

第 2章

日本語5母音を生成可能な調音状態の作成

2.1 はじめに

従来の研究では，不自然調音状態の詳細な分析がされておらず，また自然調音状態と不自然調音状態を含む調音状態の分布も明らかにされていない。従って，調音状態の逆推定において推定候補に含まれる不自然調音状態を取り除くためには，まず音声信号と一対多の関係にある調音状態の分布を分析し，不自然調音状態を含む調音状態の分布の全体像を把握する必要がある。また，音声信号と一対多の関係にある調音状態の分布の分析には，同一の音素カテゴリーに含まれる音声信号に対する人間が調音可能な多数の多様な状態を必要とするが，実際に被験者の観測によりそのような調音状態を得ることは困難である。よって，本章では人間の発話運動を精度良く再現可能な部分 3 次元生理学的発話機構モデル [16] を用いて，日本語 5 母音の音声信号を生成可能な多数の多様な調音状態の作成を目指す。その手順として，まず発話機構モデルを用いて系統的に調音状態を生成し，各状態に基づき音声を合成する。さらに，観測された音声信号に基づき日本語 5 母音のカテゴリーに含まれる合成音声を抽出することにより，抽出された合成音声に対応する調音状態を選定する。選定された調音状態は，人間が日本語 5 母音の音声を生成可能な調音状態といえる。なお，合成した音声と収録音声を比較することにより，実際の状況を反映した 5 母音の調音状態の選定が可能となる。

2.2 日本語 5 母音の調音運動と音声信号の観測

2 章で行なう日本語 5 母音の調音状態の選定や， 3 章で行なう自然調音状態と不自然調音状態の分類は，母音区間に含まれる発話器官の調音運動と音声信号に基づいて行う。よって，本研究で用いる発話器官の調音運動と音声信号について述べる。さらに，それぞれの信号に対する母音発話中の定常部を表す母音区間を定める。なお，母音区間の発話器官の調音運動と音声信号は 5 章で行う調音状態の逆推定実験の実験データにも用いられる。

発話器官の調音運動は，発話機構モデルの目標話者を被験者とする X-ray microbeam system によるペレット位置の観測信号 [17] とし，本研究では，正中矢状断面の下顎のペレット位置（LJ）及び舌上 4 点のペレット位置（T1~ T4）の計 5

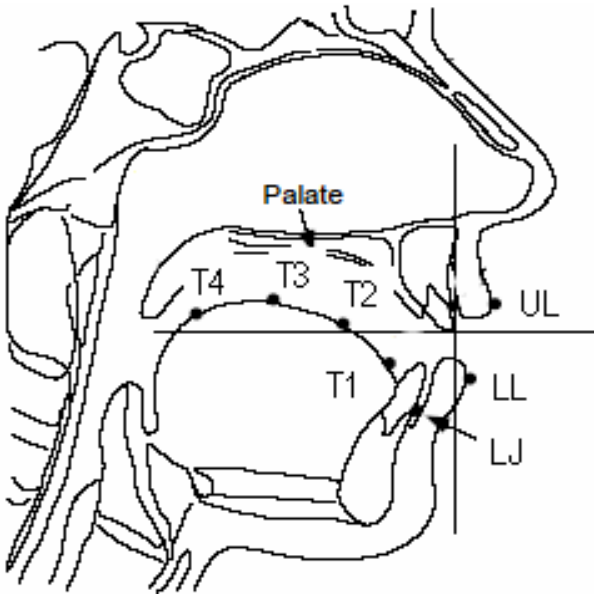

図 2.1: X-ray microbeam system におけるペレット位置

箇所の観測信号を用いる。具体的な各ペレット位置を図 2.1 に示す。LJ は下側歯列の歯と歯茎の境に位置する。また，T1~ T4 に関しては，舌の先端から 1cm 程度後方に T1，装着可能な最後方に T4 が位置し，T1 と T4 の間を等間隔に分ける 2 箇所に T2 及び T3 が位置する。なお，観測信号のサンプリング周波数は 146Hz とする。

収録された音声信号は，EMU-4545 マイクロホンを用いて，X-ray microbeam system によるペレット位置の観測と同期して収録された音声信号 [17] とし，サンプリング周波数は 16kHz とする。

観測における音声資料は，日本語の複数の単母音，VV 音節，CVC 音節，単語，文章であり，単独発話及び連続音声中の発話が含まれている。また，音声資料は 1 秒間のモーラ数の平均が 5.88，標準偏差が 1.34 の話速で発話された。

発話器官の調音運動に対する母音区間は，信号中の母音の中心位置の前後合わせて 12 個のサンプリングデータを含む範囲（75.3ms）とする。母音の中心位置は，Okadome と Honda の基準 [18] に基づき，母音ごとに特定のペレット位置の速度が 0 となる箇所とする。ただし，そのような箇所が見当たらない場合は，音声信号のスペクトログラムの目視により中心位置を定める。また，音声信号に対する母

音区間は，観測信号に対する母音区間に含まれるサンプリングデータの中心 6 個それぞれに対して，音声信号を 1 フレーム（34ms）ずつ切り出した計 6 フレームとする。なお，母音区間のすべてのフレームから求めた Mel frequency Cepstrum Coefficients（MFCC）[19] の平均から標準偏差の 2 倍の範囲を超えるフレームに対応するサンプリングデータは，母音区間から取り除く。

観測された発話器官の調音運動と音声信号に対して母音区間を定めた結果，5 母音合わせて 5892 個の母音区間に含まれる発話器官の調音運動のサンプリングデータと，2946 フレームの母音区間に含まれる音声信号のフレームデータを得た。次節以降，これらの母音区間に含まれる発話器官の調音運動と音声信号を観測データとして用いる。

2.3 部分 3 次元生理学的発話機構モデル

人間が発話可能な多数の多様な調音状態を生成するためには，人間の発話を精度良く再現可能なモデルを用いる必要がある。Dang と Honda により提案されている部分 3 次元生理学的発話機構モデル [16] は，日本人成人男性 1 名を被験者として MRI システムにより観測された画像に基づき，舌，下顎，舌骨及び声道壁により構成されており（図 2.2 ），舌と下顎の筋構造は MRI 画像及び解剖学的知見に基づき構築されている（図 2.3 ）。ただし，水平断面上の左右方向の構造は，正中矢状断面を中心に左右 2cm 幅のみとなっている。また，磁気センサシステムにより観測された発話における舌尖，舌背，下顎の最大速度と，モデルを用いたシミュレーションにおける同観測点の最大速度との比較により，下顎に関してわずかな誤差を含むが，舌尖と舌背の最大速度はほぼ一致することが示されている [16]。従って，このモデルは人間の生理学的構造が反映されており，また人間の発話動作を精度良く再現可能なことから，本研究ではこの部分 3 次元生理学的発話機構モデルを用いて調音状態の生成を行なう。

なお，舌の筋構造には，3 種類の外舌筋（オトガイ舌筋（Genioglossus: GG），舌骨舌筋（Hyoglossus: HG），茎突舌筋（Styloglossus:SG））と 4 種類の内舌筋（上縦舌筋（Superior Longitudinalis: SL），下縦舌筋（Inferior Longitudinalis: IL），横舌筋（Transversus: T），垂直舌筋（Verticalis: V））及び，2 種類の口腔底筋

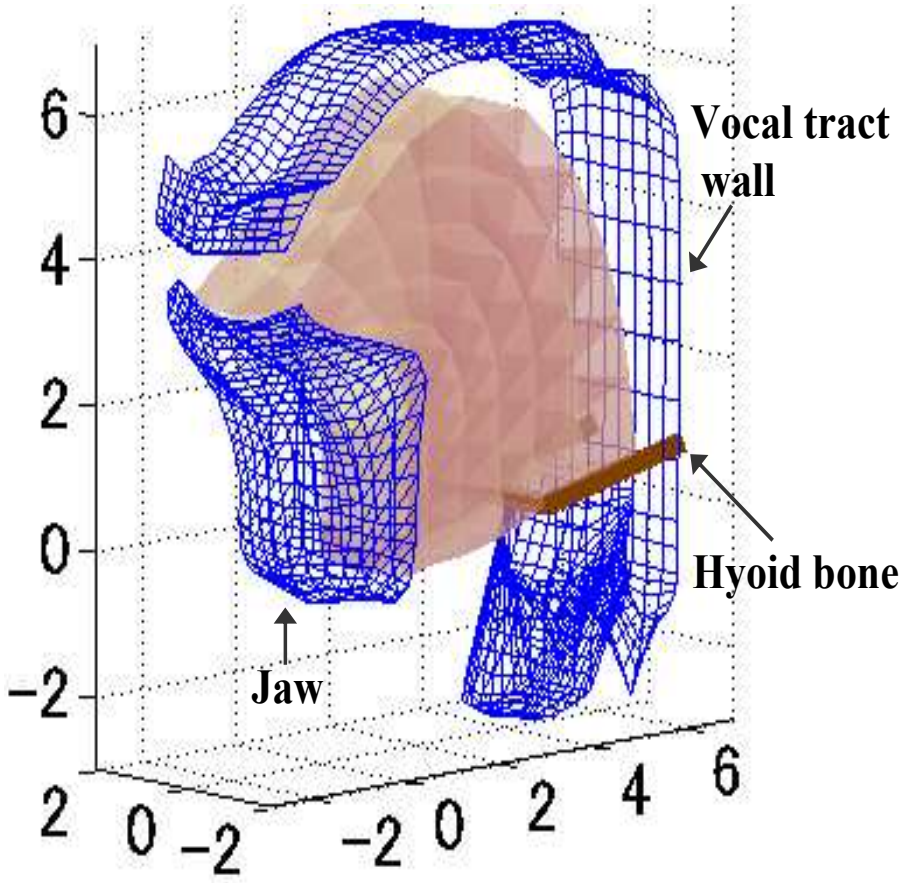

図 2.2: 部分 3 次元生理学的発話機構モデル

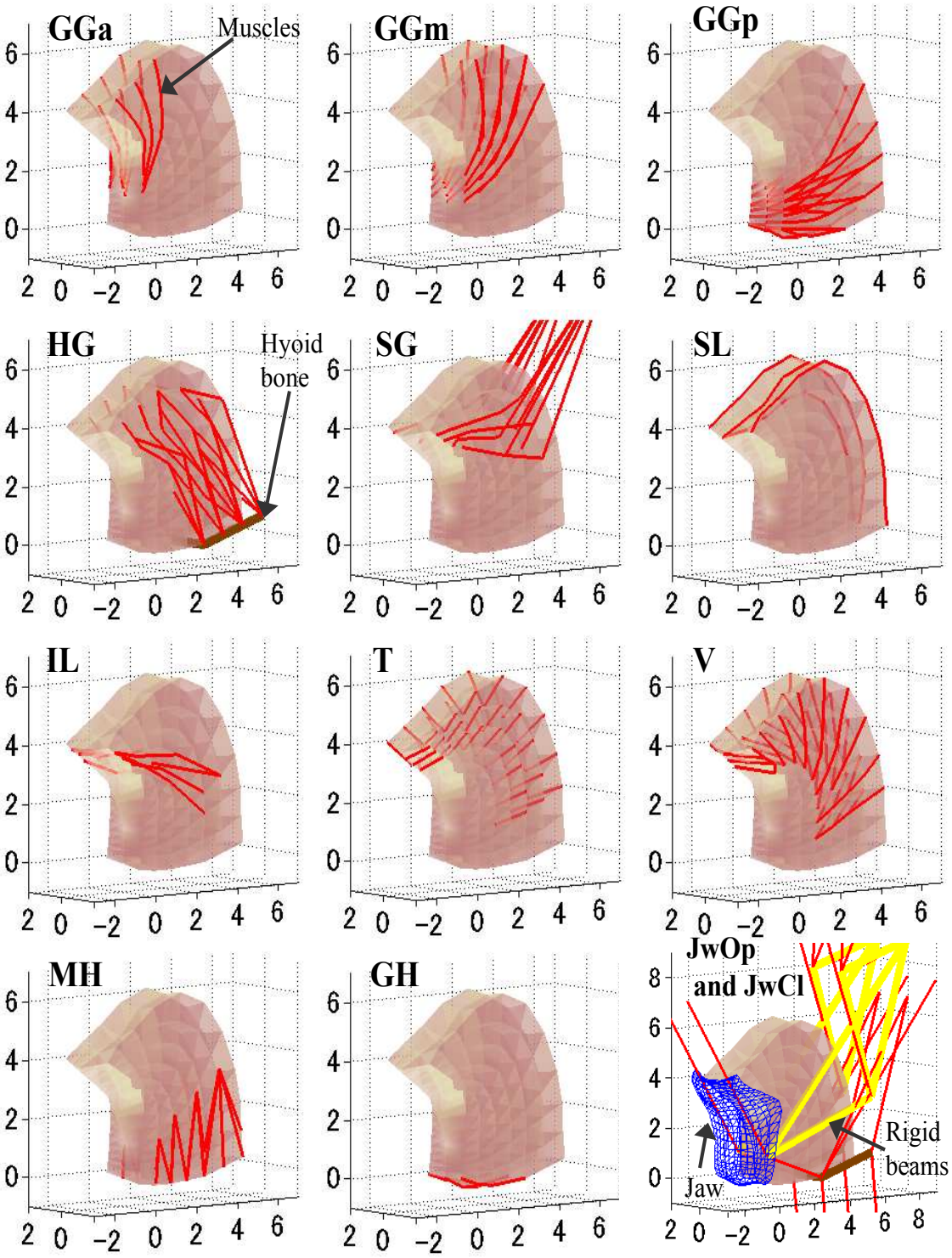

図 2.3: 舌及び下顎の筋構造

（顎舌骨筋（Mylohyoid: MH），オトガイ舌骨筋（Geniohyoid: GH））が含まれている。GG は部位によって異なる働きを行うことから，前部，中部，後部それぞれを GG anterior（GGa），GG middle（GGm），GG posterior（GGp）とする三つの部位に分けられている。また，下顎に関しては大まかに 2 種類の筋群，下顎を下げるための筋群（JwOp）及び上げるための筋群（JwCl）が含まれる。

2.4 調音状態の系統的生成及び音声合成

調音状態は，舌と下顎に関する筋に収縮力を 400ms 間加え発話機構モデルを駆動することにより生成する。また，生成した各調音状態に基づき音声合成を行う。なお，人間が生成可能な多様な調音状態を生成するために，各調音状態に基づく合成音声の音韻性を考慮せず，舌と下顎の筋収縮の組み合わせのみを考慮して調音状態を生成する。また， 2.2 節で述べた正中矢状断面上の発話器官の調音運動と比較するために，生成された声道形状も正中矢状断面の舌表面上の点と下顎の点のみを調音状態として扱う。ただし，舌の詳細な形状を表すために舌表面上のノード 17 点を調音状態のパラメータに含める。よって，調音状態は正中矢状断面の舌上 17 点と下顎 1 点を合わせた計 36 次元のベクトルとし，観測データとの比較には 18 点中の LJ 及び T1~ T4 に相当する定点を用いる。

2.4.1 調音状態の生成

調音状態を生成する際に，舌に対して，2 または 3 個の筋を 1 組とする 28 種類の筋の組み合わせを用いる。筋を組み合わせる際に，GGa 及び GGm，GGp はそれぞれ一つの筋として制御する。また，全方位への移動を可能にするため SL と T を一つの筋として制御する。組み合わせの基準は Dang と Honda の検討 [16] に基づき，次のとおりとする。まず，舌尖または舌背の全方向への移動に大きく寄与する筋をそれぞれに対して選択し，選択した筋の中から外舌筋または SL & T を主動筋として，主動筋とその協同筋，または主動筋とその拮抗筋及び協同筋を組み合わせる。具体的な 28 種類の組み合わせを表 2.1 に示す。

モデルの舌の各筋に収縮力を与える際に，収縮力が 6N より大きい場合，舌の変

表 2.1: 舌筋の組み合わせ

Agonists and synergists	GGa - IL,	GGa - V,	GGm - V,
	GGm - SL & T,	GGp - SL,	GGp - V,
	GGp - SL & T,	HG - SL,	HG - IL,
	HG - SL & T,	SG - SL,	SG - IL,
	SL & T - SL,	GGm - GGp,	GGm - HG,
	GGp - SG,	GGp - MH,	HG - SG,
	SG - MH		
Agonists, antagonists and synergists	GGm - GGp - SL,	GGm - SL & T - SL,	GGp - GGa - IL,
	GGp - SL - HG,	GGp - SL - SG,	GGp - GGm - SL,
	HG - GGm - SL & T,	SG - HG - SL & T,	SG - MH - SL & T

形がほとんど見られなくなる。従って，舌に関する筋に与える収縮力は 6N を最大値として，筋ごとに 0N~ 6N の間を 7 段階に分け，舌の変位の間隔がほぼ均等になるように各段階の値を設定する。ただし，GGm，GGp，V に関しては，舌が口蓋壁と接触する際に計算が不安定になることを避けるため，それぞれ 1N，2N，2N を最大値とする。

また，下顎に対しては JwOp と JwCl の 2 種類の筋群を用いる。筋群への収縮力は，JwOp に対しては 0N~ 6N の間を 6 段階に，JwCl に対しては最大値を 3N として 0N~ 3N の間を 3 段階に分け，舌の場合と同様に下顎の変位の間隔がほぼ均等になるように各段階の値を設定する。

上記の舌筋の 28 組及び下顎の 2 種類の筋群から選択可能なすべての組み合わせに対して，次の手順で調音状態を計算する。表 2.1 の 28 組の中から一つの舌筋の組み合わせを選択し，同時に下顎の 2 種類の筋群から一つの筋群も選択する。これらの選択した舌筋の組み合わせに含まれる 2 または 3 個の筋と下顎の筋群に対してのみ，各段階に収縮力を変化させることで調音状態を計算する。この時，他の筋及び筋群に収縮力は与えない。

本研究で用いる生理学的発話機構モデルは，人間が発話する際の発話器官の形状や生理学的運動が考慮されている。また，Sanguineti らは舌や下顎に関連する

各筋が発揮可能な最大収縮力を検討しており，SL に対する 14.3N が筋が発揮可能な最大収縮力の中で最も小さな値となっている [20]。この値は，調音状態を生成する際に用いた収縮力の最大値 6N の 2 倍を超える。従って，収縮力が 6N 以下の範囲で生成される調音状態は，生理学的に可能な状態と考えられる。

2.4.2 調音状態に基づく音声合成

音声合成は，表 2.1 に示した舌筋 28 組と下顎の 2 種類の筋群をすべて組合せた結果得られる舌と下顎の調音運動に基づき行う。音声合成を行う際には，調音状態に含まれていない口唇と喉頭も考慮する。口唇は，長さと直径をパラメータとする音響管として近似し，声道断面積関数の出力端として扱う。なお，口唇の変形の影響は，変動範囲の異なる 2 種類（通常状態と円唇化状態）のパラメータセットを用いることにより取り入れられる。各セットの変動範囲は，観測データの口唇の調音運動に基づき定められ，通常状態の場合，長さは 0.02cm 間隔で 0.81cm ~ 1.09cm，直径は 0.04cm 間隔で 1.05cm~ 1.69cm とする。これに対し，円唇化状態の場合，間隔は通常状態と同様とし，長さは 1.10cm~ 1.38cm，直径は 0.45cm~ 1.09cm とする。また，喉頭は声道断面積関数の入力端からの 3 区間として扱う。ただし，発話機構モデルの目標話者が 5 母音を発話した際の MRI 画像から求めた声道断面積関数において，喉頭部分は母音間で違いがほとんど見られなかった。従って，音声合成の際に喉頭部分の 3 区間には，母音/e/の MRI 画像に基づき求められた声道断面積関数の値を固定値として用いる。

具体的な合成手順は次のとおりとする。まず，400ms 間中の安定した 100ms 間の調音運動に基づき求めた正中矢状断面の声道の幅に 2 種類の口唇パラメータを加え，それらに改良 α-β モデル [11] を適用することにより 2 種類の声道断面積関数を得る。さらに，それぞれの声道断面積関数に基づき音響等価回路モデルを求め，音源信号を入力した結果の出力として合成音声を得る。音源信号は，Fant が提案した声門体積流モデルを声門開口面積に適用し求めた声門開口面積波形 [11] を用いる。声門開口面積波形を用いる際に，最大開口面積は 0.3cm^2 とする。また，基本周波数は観測データの収録音声に基づき 120Hz とし，音質が収録音声と等しくなるように，音源信号の Opening quotient と Closing quotient を調整する。

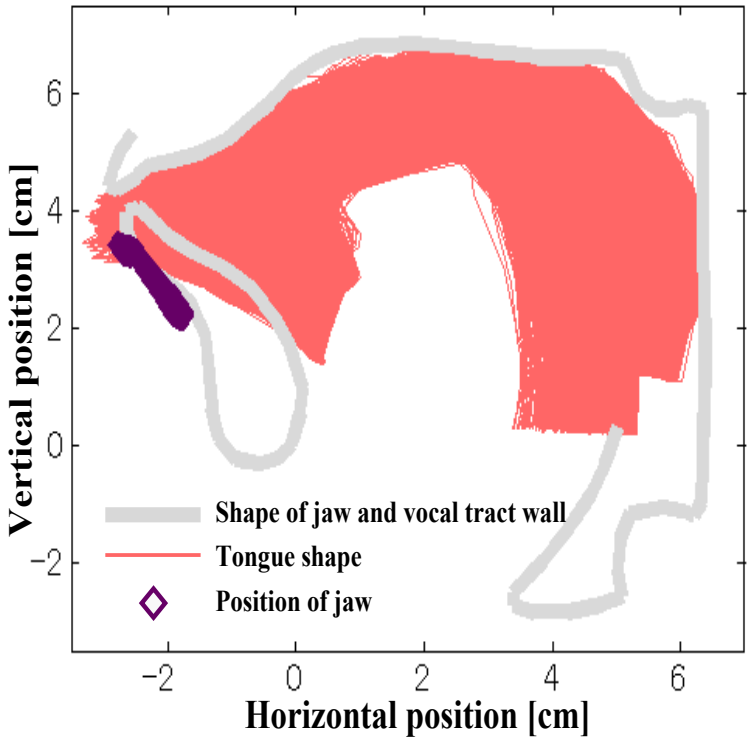

図 2.4: 生成された調音状態の形状

2.4.3 結果と考察

2.4.1 項の調音状態の生成及び， 2.4.2 項の音声合成の結果，64587 組の調音状態と合成音声の対を得た。発話機構モデルを用いて生成した調音状態を図 2.4 に示す。図中の細い線は，調音状態に含まれる舌表面上の 17 点を線形補間した舌の形状を表し，菱形の記号は，調音状態に含まれる下顎の位置を表す。また，太い線は正中矢状断面上の下顎の形状及び声道形状を表す。図 2.4 は，生成された舌の形状が発話機構モデルの口腔内をほぼ網羅していることを示している。

なお，発話機構モデルは舌と口蓋との接触が考慮されているため，舌の両側が硬口蓋と接触することにより舌の前部にくぼみが自然に生じる。従って，正中矢状断面の舌の状態には舌前部のくぼみの影響が含まれている。さらに，音声を合成する際に喉頭部分は固定値を用いているが，声道長は口唇パラメータの変化及び舌の形状の変形により 14.4cm~ 18.4cm の範囲で変化することが確認できている。

2.5　音響分析に基づく日本語5母音の調音状態の選定

2.4 節で生成された調音状態に基づき合成された音声は音韻性が考慮されていないため，日本語 5 母音に聞こえない音声も多数含まれている。つまり，生成された調音状態には日本語 5 母音のカテゴリーに含まれない音声を生成する状態も多数含まれているため，日本語 5 母音のカテゴリーに含まれる音声を生成可能な調音状態を選定する必要がある。従って， 2.2 節で述べた観測データの音声信号の音響特徴量から求めた規準範囲に基づき日本語 5 母音のカテゴリーに含まれる合成音声を抽出し，抽出された合成音声に対応する調音状態を得る。

2.5.1　音響分析

音響特徴量は，前処理として高域強調された音声信号から求めた 12 次元の MFCC と第 1 及び第 2 ホルマント周波数とする。各特徴量を求める条件は，サンプリング周波数 16kHz，窓関数はハミング窓を用い，時間長は 30ms，シフト長は 10ms とする。ホルマント周波数は，分析次数を 18 次とする線形予測分析により得られる全極型フィルタの分母多項式の根から求める。また，MFCC は 4kHz のローパスフィルタを通した後，24 個のフィルタバンク出力の離散コサイン変換から求める。なお，音源特性のパワー成分の影響を低減させるため，MFCC の最初の係数 C0 を除き，低次の係数 C1~ C12 のみを用いる。

2.5.2　合成音声の抽出

合成音声のデータ数が約 65000 と多いため， 2.2 節で述べた観測データの音声信号から MFCC とホルマント周波数を抽出し，それぞれの特徴量に対して母音ごとに求めた規準範囲を合成音声に適用することにより，日本語 5 母音のカテゴリーに含まれる合成音声を自動的に抽出する。

MFCC に対する規準範囲は，母音ごとの MFCC から求めた信頼度 68%の信頼楕円体 [21] とする。この信頼楕円体は分布の標準偏差の範囲に相当する。また，ホルマント周波数に対する規準範囲は，各母音の第 1 及び第 2 ホルマント周波数それぞれの平均± 10%（ホルマント周波数の弁別閾値に相当 [22]）を軸とする楕円

とする。ただし，ホルマント周波数は音韻性と密接に関連する特徴量 [23] であるが，周波数の一部のみしか考慮されず，また精度良く推定することは難しい。一方，MFCC はスペクトルの形状全体が考慮され，求められる特徴量の精度は高いが，音韻性との直接の関連性は明確ではない。よって，MFCC に対する規準範囲に含まれる合成音声に対して，さらにホルマント周波数に対する規準範囲を適用し，両方の特徴量の規準範囲に含まれる合成音声を抽出する。この手順により，2 種類の特徴量それぞれの短所を補い合成音声を抽出することができると考えられる。

2.5.3 結果と考察

2 種類の特徴量両方の規準範囲に含まれる特徴量を持つ合成音声を抽出した。その結果，5 母音合わせて 8229 個の合成音声が抽出された。抽出された合成音声とすべての合成音声のホルマント周波数を図 2.5 に示す。図 2.5 より，抽出された合成音声の分布は母音ごとに密集し，母音間では分離していることが示されている。ホルマント周波数の規準範囲と弁別閾値が等しいことを考慮すると，抽出された合成音声は各母音のカテゴリーに含まれる音声と考えられる。従って，抽出された合成音声に対応する調音状態は，日本語 5 母音の音声を生成可能な調音状態と考えられる。なお，5 母音に含まれないデータの中に，第 1 及び第 2 ホルマント周波数が共に高い領域に分布しているデータが見られる。一般的な音声のホルマント周波数はこのような領域には分布しない。これは，調音状態を生成する際に舌と下顎の筋収縮の組み合わせのみを考慮し音韻性が考慮されていないためと考えられる。

さらに，抽出された合成音声に対応する調音状態の分布を示す。調音状態は 36 次元であり，分布を直接示すことは難しいため，調音状態の主成分分析（Principal component analysis: PCA）[24] を行った。36 次元の要素を持つ調音状態 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i36})^T$，$i = 1, \ldots, N(= 8229)$ の共分散行列を $\boldsymbol{\Sigma}$ とすると，式 2.1 に示される共分散行列に対する固有値問題を解くことにより，調音状態 $\boldsymbol{x}_i$ の主成分 $\acute{\boldsymbol{x}}_i = (\acute{x}_{i1}, \ldots, \acute{x}_{i36})^T$ を得ることができる。ただし，主成分を求める際には $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i = 0$ となるように各調音状態から平均を引いておく。

$$\boldsymbol{\Sigma U} \;=\; \boldsymbol{\Lambda U} \tag{2.1}$$

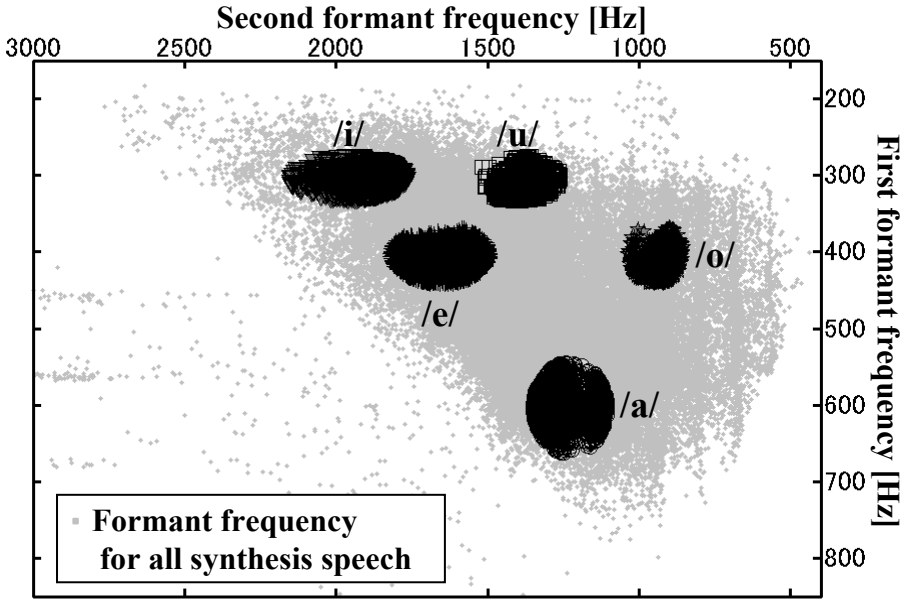

図 2.5: 抽出された 5 母音の合成音声とすべての合成音声の第 1 及び第 2 ホルマント周波数

なお，$\mathbf{\Lambda}$ は式 2.1 の固有値問題を解いた結果得られる固有値 $\lambda_1, \ldots, \lambda_{36}$ を対角要素とする対角行列，$\boldsymbol{U}$ は固有値ベクトル $\boldsymbol{\alpha}_l = (\alpha_{l1}, \ldots, \alpha_{l36})^T$，$l = 1, \ldots, 36$ を列とする行列を表す。得られた固有値を降順に並べ替えた際の最大固有値 λ_1 に対応する固有ベクトル $\boldsymbol{\alpha}_1$ を第 1 主成分ベクトルとし，それ以降降順の i 番目の固有値に対する固有ベクトルを第 i 主成分ベクトル $\boldsymbol{\alpha}_i$ とすることにより，調音状態 $\boldsymbol{x}_i$ の主成分 $\acute{\boldsymbol{x}}_i$ は下記の式から求められる。

$$\acute{\boldsymbol{x}}_i \quad = \quad \boldsymbol{U}^T \boldsymbol{x}_i \tag{2.2}$$

第 1 主成分ベクトル $\boldsymbol{\alpha}_1$ 及び第 2 主成分ベクトル $\boldsymbol{\alpha}_2$ の要素の分析により，第 1 主成分は主に舌全体の水平方向の変位を，第 2 主成分は主に舌尖の垂直方向の変位を表すことが示された。また，第 2 主成分までに対応する固有値の和をすべての固有値の和で割った累積寄与率は 77%となった。PCA により得られた 5 母音の調音状態の第 1 及び第 2 主成分を図 2.6 に示す。図 2.6 から 5 母音の相対的な位置関係はホルマント周波数空間と一致しているが，分布の重なりが大きいことが分かる。なお，通常 5 母音の調音状態の PCA において，舌全体または舌背の変位や口唇を伴う下顎の変位が主要な主成分となる [25]。しかしながら，本実験で得ら

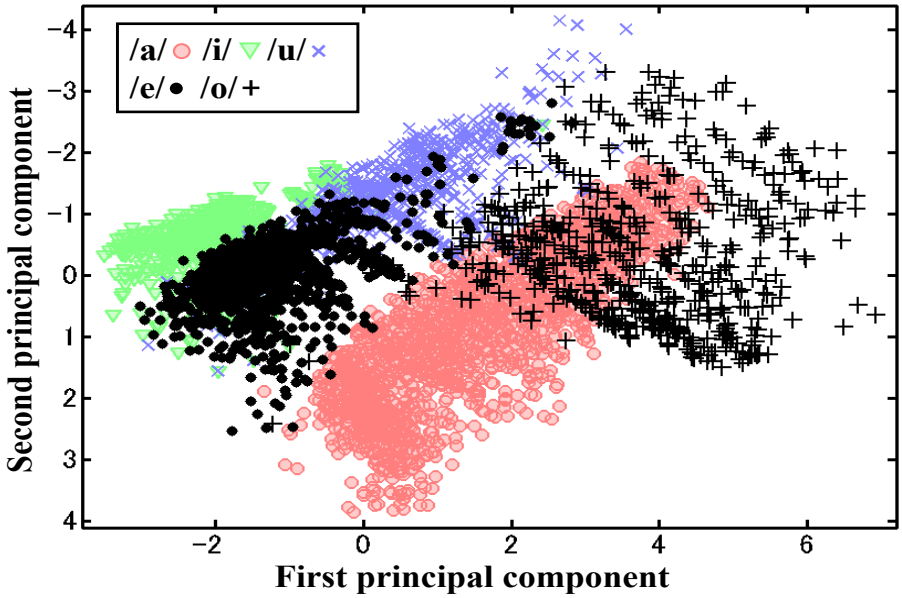

図 2.6: 抽出された 5 母音の合成音声に対応する調音状態の第 1 及び第 2 主成分

れた結果は舌尖の垂直方向の変位が第 2 主成分となっており，通常の結果と異なる。この原因として，不自然調音状態が含まれる影響で舌尖の分布の分散が大きくなっていることが考えられる。

2.6 まとめ

本章では，X-ray microbeam system により観測された発話器官の調音運動と音声信号に対して母音発話中の定常部を表す母音区間を定め，母音区間に含まれる調音運動のサンプリングデータと音声信号のフレームデータを得た。また，発話機構モデルを用いて系統的に調音状態を生成し，生成した調音状態に基づき音声を合成することで，65487 組の調音状態と合成音声の対を得て，生成された調音状態が発話機構モデルの口腔内をほぼ網羅することを示した。さらに，母音区間に含まれる観測データの音声信号の音響特徴量に基づく規準を用いて，日本語 5 母音の音響特徴量を持つ合成音声を抽出することにより，生成されたすべての調音状態の中から日本語 5 母音合わせて 8229 個の調音状態を選定した。

第 3章

音声信号と一対多の関係にある調音状態の分析

3.1　はじめに

不自然調音状態を含む音声信号と一対多の関係にある調音状態の分布の全体像を把握することにより，音声信号から調音状態の逆推定の推定候補に含まれる不自然調音状態の適切な除去が可能になると考えられる。しかしながら，自然調音状態と不自然調音状態の分布の関係や不自然調音状態の性質は明らかではないため，音声信号と一対多の関係にある調音状態の分布構造を明らかにし，自然調音状態と不自然調音状態との分布間の重なり度合い（以降，重複度とする）や，不自然調音状態の傾向を把握する必要がある。そのために，まず 2.5 節で選定された日本語 5 母音の調音状態を，自然調音状態と不自然調音状態に分類する。さらに，異なる調音状態の分布間の重複度を減少させる非線形空間に調音状態を射影し，非線形特徴量のクラスタ構造を保ったまま次元圧縮する。この分析により音声信号と一対多の関係にある調音状態の分布構造を可視化することで，不自然調音状態を含む分布構造を明らかにする。また分布構造を分析することにより，自然調音状態と不自然調音状態の位置関係を定量化し，さらに日本語 5 母音として取り得る不自然調音状態の形状の傾向を明らかにする。

3.2　自然調音状態と不自然調音状態の分類

2.5 節で選定された 5 母音の調音状態は，音声信号と一対多の関係にある調音状態を表しているため，自然調音状態と不自然調音状態が含まれていると考えられる。よって，調音状態の分布構造を明らかにするためには，まず選定された調音状態を自然調音状態と不自然調音状態に分類する必要がある。そのために， 2.2 節で述べた観測データの発話器官の調音運動に基づき分類規準を定め，選定された 5 母音の調音状態を自然調音状態と不自然調音状態に分類する。

3.2.1　分類基準

分類規準を 2.2 節で述べた下顎及び舌上 4 点のペレット位置（LJ 及び T1~ T4）の調音運動に基づき定める。LJ 及び T1~ T4 と実際に分類に用いる調音状態の定点との対応関係は次のとおりとする。LJ に対応する定点は，発話機構モデルの LJ

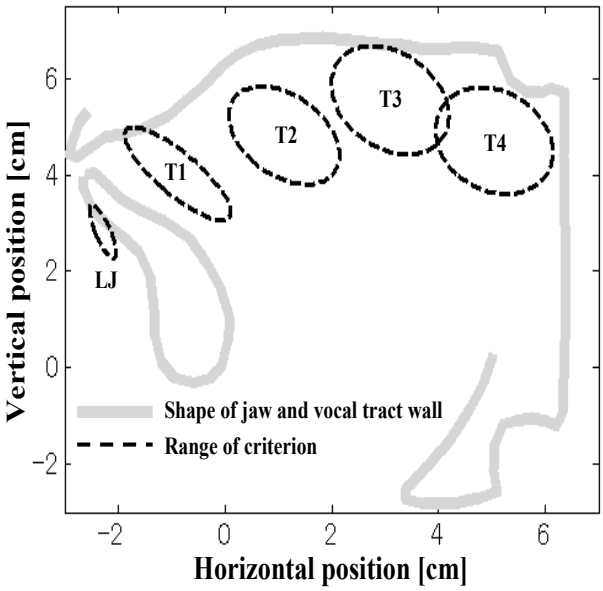

図 3.1: 各ペレットの信頼楕円（母音/a/）

と同じ箇所とする。T1~ T4 に対応する定点は，モデルの初期状態における調音状態の舌のパラメータ（舌上 17 点）を線形補間した形状と，母音/e/の平均の T1~ T4 を比較し，平均との誤差が最小となる位置とする。なお，モデルの目標話者が母音/e/を発話した際の調音状態がモデルの初期状態となっているため，初期状態の形状と母音/e/の平均の T1~ T4 を比較することにより，観測された発話器官のペレット位置に対応する調音状態の定点を求めている。

具体的な分類規準は，観測データの各発話器官に関する調音運動に含まれるサンプリングデータから求めた信頼度 99%の信頼楕円（標準偏差の 3 倍の範囲に相当）とする。母音/a/の場合の各ペレットの信頼楕円を図 3.1 に示す。5 箇所のペレット位置がすべて信頼楕円に含まれる調音運動のサンプリングデータは 96%となる。この規準を用いて，5 箇所の定点がすべて規準範囲に含まれる調音状態を自然調音状態，1 箇所でも含まれない場合は不自然調音状態として分類する。

3.2.2 結果と考察

規準に基づき自然調音状態と不自然調音状態に分類した結果を表 3.1 に示す。表 3.1 よりすべての母音で不自然調音状態のデータ数のほうが多く，自然調音状態は 5 母音合わせて 1580（全体の約 20%）となった。この結果は，調音モデルを用いる場合，不自然調音状態に基づき自然な範囲に含まれる音響特徴量を持つ合成音声が多数生じる可能性を示唆する。

表 3.1: 自然調音状態と不自然調音状態の数

Vowel	/a/	/i/	/u/	/e/	/o/
Number of natural articulations	160	94	64	1188	74
Number of unnatural articulations	2447	1235	888	1461	618

また，分類された日本語 5 母音の自然調音状態の形状を図 3.2 に，不自然調音状態の形状を図 3.3 に示す。図中の細い実線は舌上 17 点の舌のパラメータを線形補間した舌の形状を，太い実線は声道形状と下顎の形状を示す。図 3.2 と図 3.3 を比較すると，母音の発話に重要な舌背の位置だけでなく，下顎や舌尖の位置も，自然調音状態と不自然調音状態では大きく異なることが示されている。この結果は，自然調音状態と不自然調音状態を識別する際に，特定の発話部位に着目することが有用であることを示唆している。

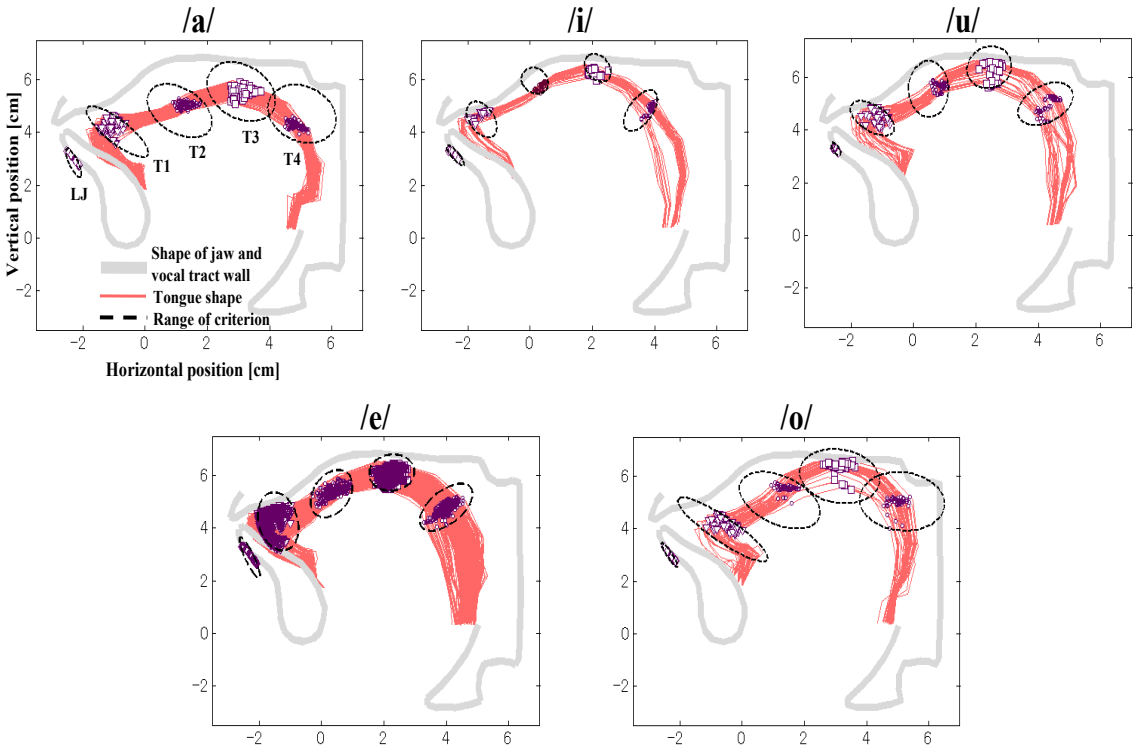

図 3.2: 自然調音状態の調音形状

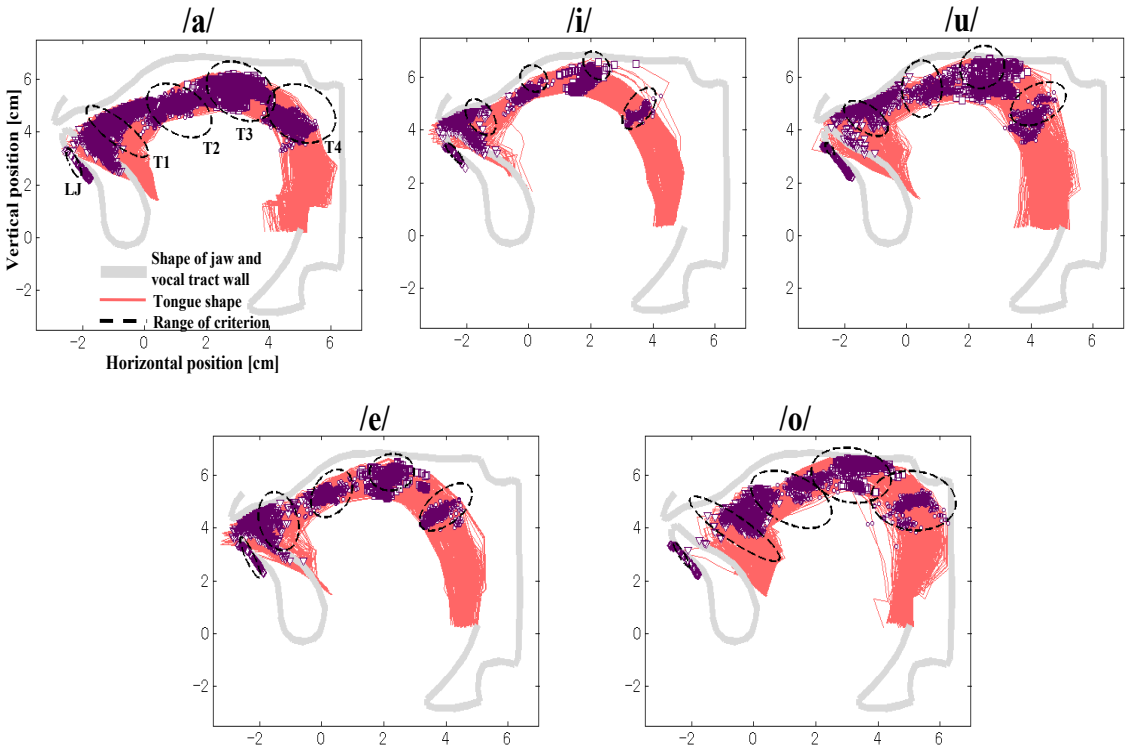

図 3.3: 不自然調音状態の調音形状

3.3 異なる調音状態間の重複度の検討

調音状態の分布の可視化によりその構造を把握するためには，自然調音状態と不自然調音状態の分布間の重複度を小さくする必要がある。3.2 節の結果から，調音状態のパラメータ空間よりも，下顎や舌尖，舌背の差異を強調した空間の方が，自然調音状態と不自然調音状態の分布間の重複度が減少する可能性が考えられる。そこで，カーネル関数を用いることにより特徴量の特性を反映させた非線形空間への射影を可能とするカーネル主成分分析（Kernel Principal Component Analysis: KPCA）[26] を用いて，自然調音状態と不自然調音状態の分布間の重複度について検討する。

3.3.1 カーネル主成分分析による調音状態の非線形射影

KPCA は，非線形空間上に射影された特徴量間の内積を表すカーネル関数を用いることにより，元の特徴量よりはるかに高次元の空間へ特徴量を非線形射影し，その高次元空間で PCA を行う非線形多変量解析の一手法である。具体的には，調音状態 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i36})^T$，$i = 1, \ldots, N$ の非線形空間への射影を $\boldsymbol{X}_i$ とすると，カーネル関数は調音状態 $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ それぞれの非線形射影 $\boldsymbol{X}_i$ と $\boldsymbol{X}_j$ の内積として下記の式により定義される。

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{X}_i \cdot \boldsymbol{X}_j) \tag{3.1}$$

可能なすべての調音状態の組み合わせに対して式 3.1 を計算し，その結果の値を要素とするグラム行列 $\boldsymbol{K}$ $(K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j))$，$i, j = 1, \ldots, N$ に対する固有値問題（式 3.2 ）を解くことにより，調音状態 $\boldsymbol{x}_l$，$l = 1, \ldots, N$ の非線形主成分 $\acute{\boldsymbol{X}}_l = (\acute{X}_{l1}, \ldots, \acute{X}_{lN})^T$ が得られる。

$$\boldsymbol{H}\boldsymbol{K}\boldsymbol{U}_K = \boldsymbol{\Lambda}_K \boldsymbol{U}_K, \quad \boldsymbol{H} = \boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^T \tag{3.2}$$

ここで，$\boldsymbol{\Lambda}_K$ は式 3.2 を解いた結果の固有値 $\lambda_{K1}, \ldots, \lambda_{KN}$ を対角要素とする対角行列，$\boldsymbol{U}_K$ は固有値ベクトル $\boldsymbol{\alpha}_{Ki} = (\alpha_{Ki1}, \ldots, \alpha_{KiN})^T$，$i = 1, \ldots, N$ を列とする行列を表し，$\boldsymbol{I}$ は $N \times N$ の単位行列，$\boldsymbol{1}$ は要素がすべて 1 の列ベクトルを表す。なお，KPCA の場合，PCA と異なり非線形主成分ベクトルを求めることはできないが，

式 3.2 を解いた結果の固有値を降順に並べ替えた際の各固有値 λ_{Ki} に対応する固有ベクトル $\boldsymbol{\alpha}_{Ki}$ を用いることにより，調音状態 $\boldsymbol{x}_l$ の非線形主成分 $\acute{\boldsymbol{X}}_l$ を下記の式から求めることができる。

$$\acute{X}_{lj} = \sum_{i=1}^{N} \alpha_{Kji} K(\boldsymbol{x}_i, \boldsymbol{x}_l),\ j = 1, \ldots, N \tag{3.3}$$

式 3.3 から，射影前の調音状態の次元数を超える，最大 N 次元の非線形主成分を得ることが可能となる。

なお，KPCA は高次元空間で PCA を行うことから，主な用途として特徴量の次元圧縮による低次元特徴量の抽出に用いられている [27]。しかしながら，本研究では，KPCA の利点の一つである特徴量の特性をカーネル関数に反映可能な点に着目し，発話部位の差異が強調された非線形空間の特徴量を抽出するために KPCA を用いて分析を行う。よって，これ以降非線形主成分 $\acute{\boldsymbol{X}}$ を調音状態を非線形空間に射影した非線形特徴量として扱う。

3.3.2 カーネル関数の設計

3.2 節の結果から，舌尖，舌背と下顎が自然調音状態と不自然調音状態を分離するための重要な発話部位と考えられる。よって，それらの発話部位の差異を強調するカーネル関数を用いることにより，KPCA により得られた非線形空間上で，自然調音状態と不自然調音状態の分布の重複度の減少が期待できる。ただし，KPCA に用いるカーネル関数は正定値性を満たす必要がある。ここで，KPCA に用いられる一般的なカーネル関数であるガウスカーネルは正定値性を満たすことが知られており，正定値性を満たすカーネル関数同士の和や積の結果のカーネル関数も正定値性を満たすことが示されている [28]。従って，ガウスカーネル及び，ガウスカーネルを組合せることにより新たに設計した 2 種類のカーネル関数の計 3 種類のカーネル関数を，自然調音状態と不自然調音状態の分布の重複度の検討に用いる。3 種類のカーネル関数を下記の式に示す。式 3.4 はガウスカーネルを，式 3.5 は下顎の差異を強調するカーネル関数を，式 3.6 は舌尖と舌背，下顎の差異を強調するカーネル関数を表している。

$$K_1(\boldsymbol{x}_i, \boldsymbol{x}_j) \ = \ \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \tag{3.4}$$

$$
\begin{aligned}
K_2(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \exp\left\{-\left(\frac{\|\boldsymbol{x}_{Ti} - \boldsymbol{x}_{Tj}\|^2}{2\sigma_T^2} + \frac{\|\boldsymbol{x}_{Ji} - \boldsymbol{x}_{Jj}\|^2}{2\sigma_J^2}\right)\right\} && (3.5)\\
K_3(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \{\exp(D_{TI}) + \exp(D_{TD}) + \exp(D_{OT})\}\exp(D_J) && (3.6)\\
D_{TI} &= -\frac{\|\boldsymbol{x}_{TIi} - \boldsymbol{x}_{TIj}\|^2}{2\sigma_{TI}^2}\\
D_{TD} &= -\frac{\|\boldsymbol{x}_{TDi} - \boldsymbol{x}_{TDj}\|^2}{2\sigma_{TD}^2}\\
D_{OT} &= -\frac{\|\boldsymbol{x}_{OTi} - \boldsymbol{x}_{OTj}\|^2}{2\sigma_{OT}^2}\\
D_J &= -\frac{\|\boldsymbol{x}_{Ji} - \boldsymbol{x}_{Jj}\|^2}{2\sigma_J^2}
\end{aligned}
$$

ここで，$\boldsymbol{x}_{Ti}$ 及び，$\boldsymbol{x}_{Ji}$ は調音状態 $\boldsymbol{x}_i$ のパラメータに含まれる舌の要素と下顎の要素をそれぞれ表す。また，$\boldsymbol{x}_{TIi}$ 及び，$\boldsymbol{x}_{TDi}$，$\boldsymbol{x}_{OTi}$ は，調音状態 $\boldsymbol{x}_i$ のパラメータに含まれる舌尖の要素，舌背の要素，舌尖と舌背を除く舌の要素をそれぞれ表す。σ^2 及び，σ_T^2，σ_J^2，σ_{TI}^2，σ_{TD}^2，σ_{OT}^2 は，カーネル関数のパラメータを表す。

3.3.3 自然調音状態と不自然調音状態の重複度の検討

自然調音状態と不自然調音状態の分布の重複度を定量化し，3.3.2 項で示した 3 種類のカーネル関数を用いて KPCA により調音状態を射影した非線形特徴量及び元の調音状態それぞれに対して，異なる状態間の重複度を検討する。ここで，ベイズ誤り確率の上限は分布の重なり度合いとして解釈することができ，値が小さいほど分布の重なりも小さくなる [29]。また，ベイズ誤り確率の上限は，有限のデータ群を評価データと判別パラメータの学習用データに分類し判別誤差の評価を行う交差確認法により求めた誤判別率として得ることができる。よって，自然調音状態と不自然調音状態の分布間の重複度を誤判別率により定量化し，調音状態及び調音状態を射影した非線形特徴量それぞれの重複度を比較する。

ここでは，異なる調音状態間の重複度を定量化し比較することに主眼をおくため，取り扱い易い解析的特性を持つ一次線形判別手法の Fisher の線形判別分析 [30] を用いて判別を行う。具体的には，判別分析の結果得られた 1 次元の軸上に判別対象の特徴量を射影し，境界値との大小関係を比較することで判別を行う。境界値は，1 次元の軸上に射影した自然調音状態と不自然調音状態それぞれの分布を正規分布で近似した場合に下記の式から求められる的中率 P を最大にする C とする

[31]。なお，これ以降1次元軸上に射影した自然調音状態が不自然調音状態より負の方向側に分布する場合を想定し数式が導かれており，自然調音状態が不自然調音状態より正の方向側に分布する場合は，式 3.7 ~ 式 3.10 の添え字 N と U が入れ替わる。

$$\begin{aligned} P &= P_N \int_{-\infty}^{C} \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left(-\frac{(\xi - m_N)^2}{2\sigma_N^2}\right) d\xi \\ &\quad + P_U \int_{C}^{\infty} \frac{1}{\sqrt{2\pi\sigma_U^2}} \exp\left(-\frac{(\xi - m_U)^2}{2\sigma_U^2}\right) d\xi \end{aligned} \tag{3.7}$$

ここで，P_N と P_U は，それぞれ自然調音状態と不自然調音状態のデータ数の割合を表す。また，m_N と m_U はそれぞれの平均を，σ_N^2 と σ_U^2 はそれぞれの分散を表す。式 3.7 の極値条件 $\frac{\partial P}{\partial C}=0$ より下記の式が得られ，

$$\frac{P_N}{\sqrt{2\pi\sigma_N^2}} \exp\left(-\frac{(C - m_N)^2}{2\sigma_N^2}\right) = \frac{P_U}{\sqrt{2\pi\sigma_U^2}} \exp\left(-\frac{(C - m_U)^2}{2\sigma_U^2}\right) \tag{3.8}$$

式 3.8 を整理した下記の式を解くことにより境界値 C が得られる。

$$\begin{aligned} &\left(\frac{1}{\sigma_N^2} - \frac{1}{\sigma_U^2}\right) C^2 + \left(\frac{m_U}{\sigma_U^2} - \frac{m_N}{\sigma_N^2}\right) C + \left(\frac{m_N^2}{\sigma_N^2} - \frac{m_U^2}{\sigma_U^2}\right) - 2L = 0 \\ &L = \log\left(\frac{P_N}{P_U} - \sqrt{\frac{\sigma_U^2}{\sigma_N^2}}\right) \end{aligned} \tag{3.9}$$

境界値 C を用いた自然調音状態と不自然調音状態の判別式は下記のとおり。

$$\begin{aligned} &\begin{cases} d(\boldsymbol{a}) < 0, & \boldsymbol{a} \in \omega_N \\ d(\boldsymbol{a}) > 0, & \boldsymbol{a} \in \omega_U \end{cases} \\ &d(\boldsymbol{a}) = \boldsymbol{\alpha}_D^T \boldsymbol{a} - C \end{aligned} \tag{3.10}$$

ここで，$\boldsymbol{a}$ は判別対象の調音状態 $\boldsymbol{x}$ または非線形特徴量 $\acute{\boldsymbol{X}}$ を表し，$\boldsymbol{\alpha}_D$ は1次元の軸への射影ベクトルを表す。なお，$\boldsymbol{\alpha}_D$ は，自然調音状態と不自然調音状態それぞれの分布の分布内散布行列の和の逆行列と，それぞれの分布の平均の差との積から得られる。式 3.10 から，判別対象の特徴量 $\boldsymbol{a}$ が ω_N に含まれる場合は自然調音状態，ω_U に含まれる場合は不自然調音状態とする。

式 3.10 から誤判別率を求めるため，交差確認法の一種である Leave-one-out 法 [29] を用いる。Leave-one-out 法は，データ群の中から一つのデータを選択し，選

択したデータを評価データ，残りを境界値Cを求めるための学習データとして判別を行う。この手順をすべてのデータに対して行うことにより，誤って判別された回数を全データ数で割った値として誤判別率を求めることができる。

判別に用いるデータは， 3.2 節で自然調音状態と不自然調音状態に分類された日本語 5 母音の調音状態及び，日本語 5 母音の調音状態を式 3.4 ~ 式 3.6 のカーネル関数を用いて射影した非線形特徴量とする。また，データの次元数は，調音状態，非線形特徴量共に 36 次元とする。なお，各カーネル関数のパラメータは，境界値Cを式 3.7 に代入し的中率を求める予備検討の結果から，式 3.4 に対して$\sigma^2 = 1$，式 3.5 に対して$\sigma_T^2 = 1.2$，$\sigma_J^2 = 0.08$，式 3.6 に対して$\sigma_{TI}^2 = 0.9$，$\sigma_{TD}^2 = 1$，$\sigma_{OT}^2 = 1.2$，$\sigma_J^2 = 0.4$とする。

3.3.4 結果と考察

判別の結果を図 3.4 に示す。調音状態の結果は，/u/と/e/で 5%以上の誤判別率となっている。式 3.4 を用いて射影した非線形特徴量の結果は，調音状態の結果に比べ/a/と/i/の誤判別率は改善されているが，逆に/e/と/o/が大きく劣化している。式 3.5 を用いて射影した非線形特徴量の結果は，式 3.4 を用いた結果に比べ，/u/，/e/，/o/の誤判別率が大きく改善されているが，/e/や/o/の誤判別率はまだ調音状態の結果より大きい。式 3.6 を用いて射影した非線形特徴量の結果は，調音状態の結果に比べ，5 母音すべてにおいて誤判別率は小さくなっており，5 母音平均で 37%改善されている。この結果は，自然調音状態と不自然調音状態の誤判別率が異なる調音状態間の分布の重なり度合いを表すことから，式 3.6 のカーネル関数を用いた KPCA により得られる非線形特徴量は，異なる調音状態間の重複度が元の調音状態に比べ 5 母音すべてにおいて減少することを示している。従って，式 3.6 のカーネル関数を用いた KPCA により得られる非線形特徴量を用いて調音状態の分布構造の可視化を行う。

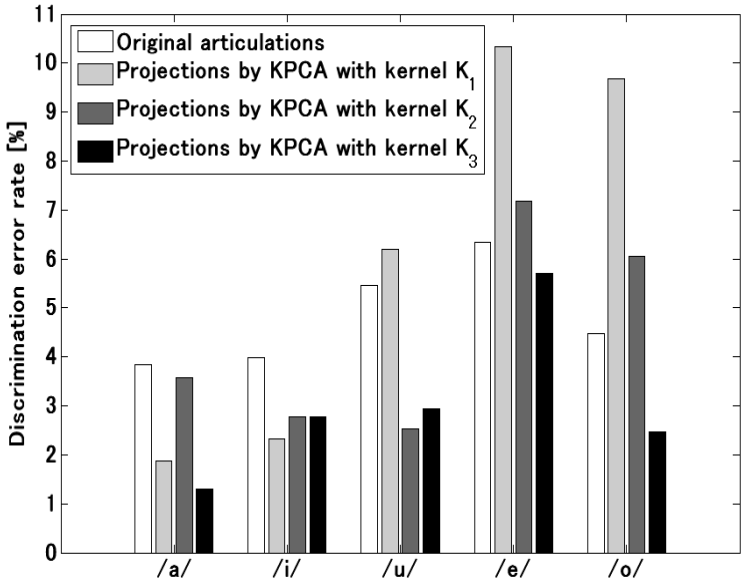

図 3.4: 5 母音の調音状態及び非線形特徴量に対する誤判別率

3.4 非線形特徴量の次元圧縮による分布構造の可視化

調音状態の分布の目視により，分布の全体像の直感的な把握が可能となる。また，自然調音状態と不自然調音状態のそれぞれの分布を類似した特徴量ごとのクラスタに分けることで，不自然調音状態を除去するための自然調音状態と不自然調音状態の識別が容易になると考えられる。ここで，クラスタリングされた高次元の特徴量に線形判別分析を適用することで，高次元空間における特徴量間の類似性を保持したまま次元圧縮することにより，高次元空間における特徴量の分布の構造を可視化する枠組み（クラスタ判別法）が提案されている [32]。従って，調音状態の KPCA により得られた非線形特徴量をクラスタ判別法を用いて次元圧縮することにより，調音状態の分布構造の可視化を行う。なお，クラスタ判別法は，使用するクラスタリングや線形判別分析の手法の選択に自由度があり，本研究ではクラスタリングにはスペクトラルクラスタリング [33] を，線形判別分析には重判別分析法 [34] を用いる。

3.4.1 特徴量のクラスタリング

調音状態の KPCA により得られた特徴量は，非線形空間上で複雑に分布しているため，複雑な分布のクラスタリングに適したスペクトラルクラスタリング [33] を用いてクラスタリングを行う。スペクトラルクラスタリングは，まず KPCA により得られた調音状態の非線形特徴量 $\acute{\boldsymbol{X}}$ 一つ一つをグラフ構造のノードとして捉え，すべてのノード間の類似度を要素とする隣接行列 $\boldsymbol{A}$(式 3.11) を求める。

$$A_{ij} = \begin{cases} 0 & if\ i = j \\ \exp\left(\dfrac{-\|\acute{\boldsymbol{X}}_i - \acute{\boldsymbol{X}}_j\|^2}{2\sigma_S^2}\right) & otherwise \end{cases} \tag{3.11}$$

なお，σ_S^2 はノード間の類似度のパラメータを表し，$\sigma_S^2 = 1$ とする。求めた隣接行列 $\boldsymbol{A}$ から下記の式により正規化グラフラプラシアン $\boldsymbol{L}$[33] を求める。

$$\boldsymbol{L} = \boldsymbol{B}^{-\frac{1}{2}}\,\boldsymbol{A}\,\boldsymbol{B}^{-\frac{1}{2}} \tag{3.12}$$

$\boldsymbol{B}$ は $B_{ii} = \sum_{j=1} \boldsymbol{A}_{ij}$ を対角要素とする対角行列を表す。正規化グラフラプラシアン $\boldsymbol{L}$ に対する固有値問題（式 3.13 ）を解いた結果得られる固有ベクトルの要素は，クラスタごとに異なる傾向の値をとる [33]。

$$\boldsymbol{L}\boldsymbol{U}_S = \boldsymbol{\Lambda}_S\boldsymbol{U}_S \tag{3.13}$$

ここで，$\boldsymbol{\Lambda}_S$ は $\boldsymbol{L}$ に対する固有値問題を解いた結果の固有値 $\lambda_{S1}, \ldots, \lambda_{SN}$ を対角要素とする対角行列を，$\boldsymbol{U}_S$ は固有ベクトル $\boldsymbol{\alpha}_{Si} = (\alpha_{Si1}, \ldots, \alpha_{SiN})^T$，$i = 1, \ldots, N$ を列とする行列を表し，N は非線形特徴量のデータ数を表す。

クラスタごとに異なる傾向の値を取る固有ベクトルに対してクラスタリングを行うことにより，精度の高いクラスタリングが可能となる。よって，式 3.13 から得られる固有値を降順に並べ替え，最大値から上位 M 個の固有値に対応する固有ベクトルを正規化した列ベクトル $\boldsymbol{\beta}_i$，$i = 1, \ldots, M$ に基づき，クラスタリングのための新たな表現を下記の式から得る。

$$\begin{aligned} &\boldsymbol{Y} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M) \\ &\beta_{ij} = \frac{\alpha_{Sij}}{\sqrt{\sum_j^M \alpha_{Sij}^2}},\ i = 1, \ldots, N \end{aligned} \tag{3.14}$$

式 3.14 の $\boldsymbol{Y}$ の行ベクトル $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iM})$，$i = 1, \ldots, N$ をクラスタリングのための非線形特徴量の新たな表現とし，$\boldsymbol{Y}_i$ に対してクラスタリングを行うことにより，非線形特徴量の適切なクラスタリングが可能となる。クラスタリングには，一般的なクラスタリング手法である k-means 法 [34] を用いる。

なお，クラスタリングを行なう際に，特徴量に最適なクラスタ数が問題となる。スペクトラルクラスタリングを用いる場合は，連続する固有値間の差の絶対値を表す Eigengap が，最適なクラスタ数を定める際の有用な指標の一つとなっている [33]。Eigengap $g(i)$ を求める式を下記に示す。

$$g(i) = |\lambda_{si+1} - \lambda_{si}|, \quad i = 1, \ldots, M-1 \tag{3.15}$$

ここで，λ_{si} は正規化グラフラプラシアンの固有値を降順に並べ替えた際の最大値から i 番目の固有値を表す。また，Eigengap $g(i)$ の極大値を取る i が最適なクラスタ数となる結果が報告されている [35]。従って，Eigengap $g(i)$ の極大値を取る i を最適なクラスタ数とし，具体的なクラスタ数は 3.4.3 項で検討する。

なお，予備検討により自然調音状態はほぼ一つのクラスタと見なせ，不自然調音状態は複数のクラスタとなった。よって，自然調音状態は母音ごとに一つのクラスタとし，不自然調音状態に対してのみスペクトラルクラスタリングを行う。

3.4.2 特徴量の線形判別分析

5 母音合わせた非線形特徴量の分布には多数のクラスタが含まれていると考えられ，クラスタ間の分離度を最大にすることは，自然調音状態と不自然調音状態の識別に有用である。従って，複数のクラスタ間の分離度を最大にする重判別分析法 [34] を線形判別分析に用いる。重判別分析法により，クラスタ内散布行列に対するクラスタ間散布行列の比を最大にする部分空間を得ることができる。ここで，クラスタ G_i，$i = 1, \ldots, M_C$ に対するクラスタ内散布行列 $\boldsymbol{S}_W$ 及びクラスタ間散布

行列 $\boldsymbol{S}_B$ は下記の式により表される。

$$
\begin{aligned}
\boldsymbol{S}_W &= \sum_{i=1}^{M_C} \boldsymbol{S}_i \\
\boldsymbol{S}_i &= \sum_{\acute{\boldsymbol{X}} \in G_i} (\acute{\boldsymbol{X}} - \boldsymbol{m}_i)(\acute{\boldsymbol{X}} - \boldsymbol{m}_i)^T \\
\boldsymbol{m}_i &= \frac{1}{n_i} \sum_{\acute{\boldsymbol{X}} \in G_i} \acute{\boldsymbol{X}} \\
\boldsymbol{S}_B &= \sum_{i=1}^{M_C} n_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T \\
\boldsymbol{m} &= \frac{1}{N} \sum_{i=1}^{N} \acute{\boldsymbol{X}}_i \\
N &= \sum_{i=1}^{M_C} n_i
\end{aligned}
\tag{3.16}
$$

式 3.16 のクラスタ内散布行列とクラスタ間散布行列に対する一般化固有値問題（式 3.17 ）を解くことにより，クラスタ内散布に対するクラスタ間散布の比を最大にする部分空間を求めることができる。

$$
\boldsymbol{S}_B \boldsymbol{U}_D = \boldsymbol{\Lambda}_D \boldsymbol{S}_W \boldsymbol{U}_D \tag{3.17}
$$

ここで，$\boldsymbol{\Lambda}_D$ は一般化固有値問題を解いた結果得られる固有値 $\lambda_{D1}, \ldots, \lambda_{Dd}$ を対角要素とする対角行列を，$\boldsymbol{U}_D$ は固有ベクトル $\boldsymbol{\alpha}_{Di} = (\alpha_{Di1}, \ldots, \alpha_{Did})^T$，$i = 1, \ldots, M_C$–1 を列とする行列を表し，$d$ は非線形特徴量の次元数を表す。なお，$\boldsymbol{S}_B$ のランクは $M_C - 1$ のため，非ゼロの固有値の数が $M_C - 1$ 個となることから，重判別分析により得られる部分空間の最大次元数はクラスタ数 -1 となる [34]。また，重判別分析を行う際に分析対象の特徴量の次元数をクラスタ数以上にする必要がある。よって， 3.4.3 項において最適な特徴量の次元数を最適なクラスタ数と合わせて検討する。

3.4.3 最適なクラスタ数と特徴量の次元数の検討

重判別分析を行う際に分析対象の特徴量の次元数をクラスタ数以上にするため，自然調音状態を合わせた総クラスタ数以上となる最適な次元数を検討する。なお，KPCA により，調音状態の次元数（36 次元）よりも高次元の非線形特徴量を得ることが可能なため，36 次元を超える次元数に対しても検討を行う。

検討方法は，調音状態の次元数と同じ 36 次元の非線形特徴量に対してクラスタリングを行い，まず不自然調音状態の最適なクラスタ数を求める。求めた不自然調音状態の最適なクラスタ数と自然調音状態のクラスタ数を合わせた総クラスタ数を新たな次元数として，再度非線形特徴量に対してクラスタリングを行い不自然調音状態の最適なクラスタ数を求める。この手順を総クラスタ数以上の最小次元数が得られるまで繰り返す。この検討で最適なクラスタ数を求める際に，式 3.15 の $M = 18$ として極大値を求めた。なお，極大値が複数存在する場合は，各クラスタに含まれる非線形特徴量の元の調音状態のバラつきを調べ，バラつきが小さい極大値に対する i を最適なクラスタ数とする。

非線形特徴量の最適な次元数及びクラスタ数の検討により，最適な次元数は 42 となり，総クラスタ数は 41 となった。Eigengap に基づいて得られた不自然調音状態の各母音の最適なクラスタ数を表 3.2 に示す。表 3.2 から/o/を除く母音では，クラスタ数は不自然調音状態の数に比例する傾向を示している。また，クラスタに含まれる非線形特徴量に対応する調音状態の形状のばらつきは小さく，さらにクラスタに含まれるデータ数が一桁のクラスタは存在しないことから，表 3.2 に示されている母音ごとの最適なクラスタ数は表 3.1 に示されている母音ごとのデータ数に対して妥当であると考えられる。

表 3.2: 自然調音状態と不自然調音状態のクラスタ数

Vowel	/a/	/i/	/u/	/e/	/o/
Number of clusters for natural articulations	1	1	1	1	1
Number of clusters for unnatural articulations	8	6	6	7	9

3.4.4 調音状態の分布構造の分析

これ以降，クラスタ判別法により得られた部分空間上の非線形特徴量のクラスタ構造を”調音状態の分布構造”，分布構造を含む部分空間を”分布構造空間”と呼ぶこととする。まず，可視化された調音状態の分布構造として，3 次元の分布構造空間における非線形特徴量の分布を図 3.5 に示す。図中の楕円体は，各クラ

スタから求めた分布の標準偏差の範囲を示している。楕円体中の文字は，最初の文字は母音を表し，2 番目の文字は調音状態によって変わり，自然調音状態の場合は N が示され，不自然調音状態の場合は各クラスタのラベル番号を表す数字が示されている。図 3.5 から非線形特徴量は多様体上に母音/a/，/i/，/o/を各頂点とする三角形状に分布している。これは，音声と一対多の関係にある調音状態は調音空間上の特定の領域に分布することを示唆している。なお，Lu と Dang は，磁気センサシステムにより観測した連続音声中の日本語 5 母音のペレット位置（口唇，下顎，舌）を 3 次元空間に非線形射影することにより，5 母音の多様体上の構造を示している [36]。その構造も/a/，/i/，/o/を頂点とする三角形を示しており，図 3.5 の結果と一致している。

得られた調音状態の分布構造に含まれる自然調音状態と不自然調音状態との位置関係を定量的に示すため，クラスタ判別法により得られる最大次元数（40 次元）の分布構造空間におけるクラスタ間の距離を求める。各クラスタの分散は各次元で異なり，クラスタ間でも分散は異なることから，クラスタ間の距離として各クラスタを確率分布と見なした際の確率分布間の差を用いることが適切と考えられる。ただし，各クラスタは非線形空間に分布しており，ガウス分布などの確率分布モデルを仮定することは適切ではない。従って，確率分布モデルを仮定せずに求めることが可能な確率分布間の距離を表す統計量である Cauchy-Schwarz（CS）divergence[37] を用いる。CS divergence は，確率分布モデルを仮定することなく分布のデータから直接密度関数を推定することにより確率分布間の差を求めることができる。分布構造のクラスタ G_i，G_j，$i, j = 1, \ldots, 41$ に含まれる非線形特徴量をそれぞれ $\acute{\boldsymbol{X}}_{G_i}$，$\acute{\boldsymbol{X}}_{G_j}$ とすると，CS divergence は下記の式から求められる。

$$
\begin{aligned}
D_{CS}(G_i, G_j) &= \frac{1}{2}\log\{V_2(G_i)\cdot V_2(G_j)\} - \log\{C_r(G_i, G_j)\} \\
V_2(G_i) &= \frac{1}{L^2 h_{CS}^4}\sum_{l=1}^{L}\sum_{s=1}^{L} V(\acute{\boldsymbol{X}}_{G_i}^{(s)} - \acute{\boldsymbol{X}}_{G_i}^{(l)}, h_{CS}) \\
C_r(G_i, G_j) &= \frac{1}{L\cdot S\cdot h_{CS}^4}\sum_{l=1}^{L}\sum_{s=1}^{L} V(\acute{\boldsymbol{X}}_{G_i}^{(s)} - \acute{\boldsymbol{X}}_{G_j}^{(l)}, h_{CS}) \\
V(\acute{\boldsymbol{X}}_{G_i} - \acute{\boldsymbol{X}}_{G_j}, h_{CS}) &= \frac{1}{(2\pi h_{CS}^2)^{\frac{d}{2}}}\exp\left(-\frac{\|\acute{\boldsymbol{X}}_{G_i} - \acute{\boldsymbol{X}}_{G_j}\|^2}{2h_{CS}^2}\right)
\end{aligned}
\tag{3.18}
$$

ここで，L はクラスタに含まれる非線形特徴量の数を表し，$\acute{\boldsymbol{X}}_{G_i}^{(s)}$ はクラスタ G_i に含まれる s 番目の非線形特徴量を表す。また，d は非線形特徴量の次元数を，h_{CS}

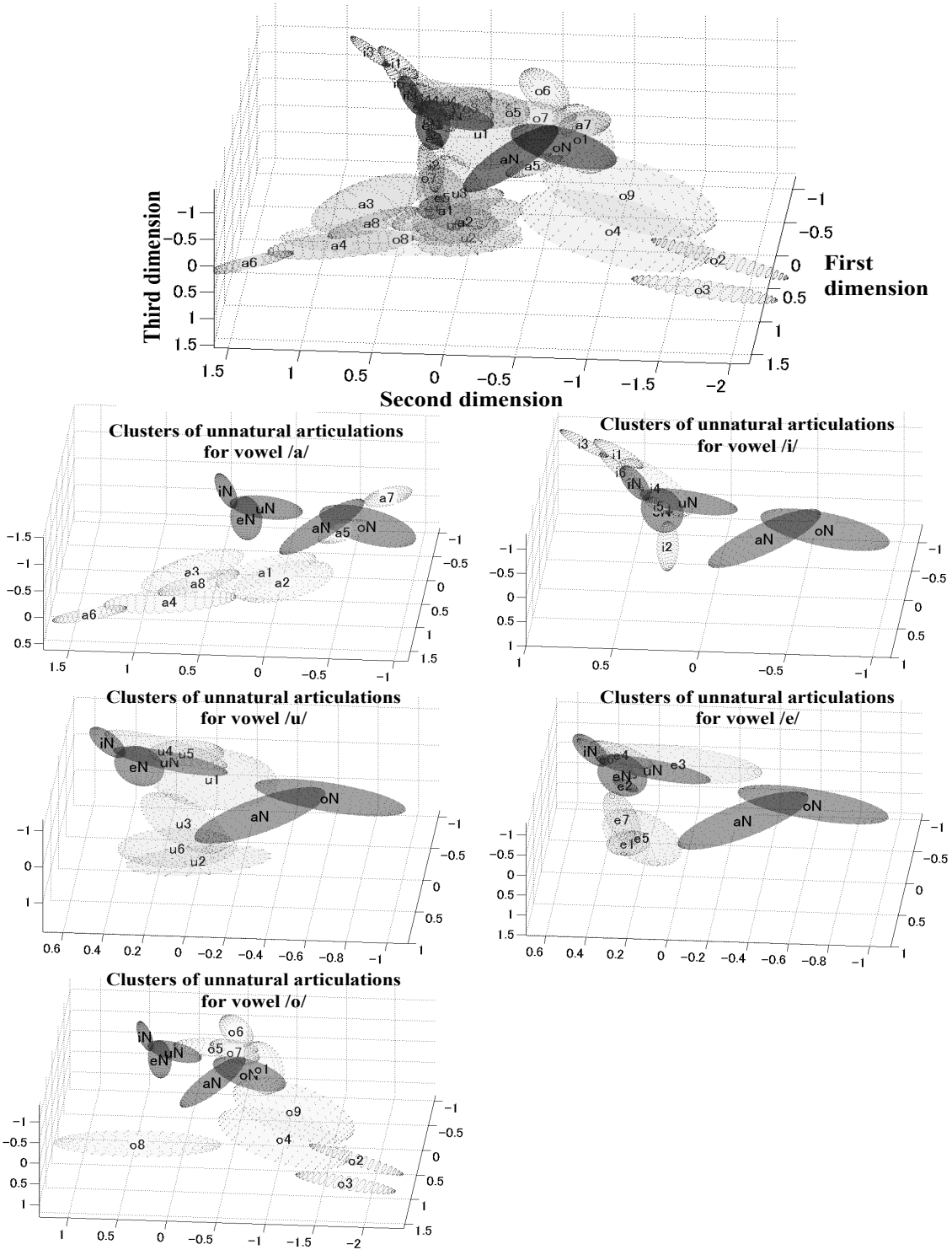

図 3.5: 3 次元分布構造空間の非線形特徴量の分布。最上段は 5 母音の自然調音状態と不自然調音状態すべてを表示。二段目以下は，各母音の不自然調音状態と 5 母音の自然調音状態のみを拡大して表示。

は推定された密度関数の滑らかさに関するパラメータを表し，$h_{CS}=1$ として CS divergence を計算する。5 母音の不自然調音状態の各クラスタと各母音の自推然調音状態との距離を図 3.6 に示す。なお，すべてのクラスタ間距離の最大値が 1 になるように距離は正規化されている。図 3.6 より，不自然調音状態の各クラスタと自然調音状態との距離は，自然調音状態が同じ母音の場合，母音ごとの平均距離は 0.14~ 0.32 の間の値をとる。一方，自然調音状態が異なる母音の場合，母音ごとの平均距離は 0.25~ 0.37 の間の値をとり，5 母音すべてにおいて後者のほうがより大きな値となった。この結果は，不自然調音状態が他の母音の自然調音状態よりも同じ母音の自然調音状態の近くに分布することを示している。

さらに，自然調音状態と不自然調音状態とのクラスタ間の分布の重複度を調べる。 3.3 節の検討に用いたベイズ誤りの上限は，分布の重複度を定量的に表すことは出来るが，各分布の一定の範囲における重複の有無について知ることは出来ない。従って，今回は，分布構造空間の非線形特徴量に対して，次元数を変えて自然調音状態と不自然調音状態とのクラスタ対ごとにクラスタ間の分離度を最大にする 1 次元空間に線形射影し，射影空間上の分布の重なりを調べた。その結果，27 次元以上では，信頼度 95%の信頼楕円体の範囲（標準偏差の 2 倍の範囲に相当）において，自然調音状態と不自然調音状態とのすべてのクラスタ対の間で重なりが見られなかった。

また，日本語 5 母音として取り得る不自然調音状態の形状を具体的に示すため，図 3.7 に母音ごとに各クラスタに含まれる非線形特徴量に対応する調音状態の形状を示す。なお，各形状の中心に示されている文字の意味は図 3.5 と同じとする。不自然調音状態の形状の大まかな傾向をみると，/a/の場合，ほとんどのクラスタでは下顎が規準より下方に位置している。しかしながら，舌尖の位置はクラスタにより異なり，規準より後方または前方下方に位置する場合と規準付近に位置する場合に分けられる。/i/の場合，一部のクラスタを除き舌尖が規準より前方に位置しており，さらに舌全体が下方に位置している。/e/の形状も/i/と同様の傾向を示している。/u/の場合，ほとんどのクラスタで舌尖が硬口蓋の付近に位置し声道中の狭めを形成している。また，下顎の位置は大きく上方に位置するクラスタと下方に位置するクラスタに分けられる。/o/の場合，ほぼすべてのクラスタにおいて舌全体が後方に位置しているが，舌尖の位置は後方に位置するクラスタと後方

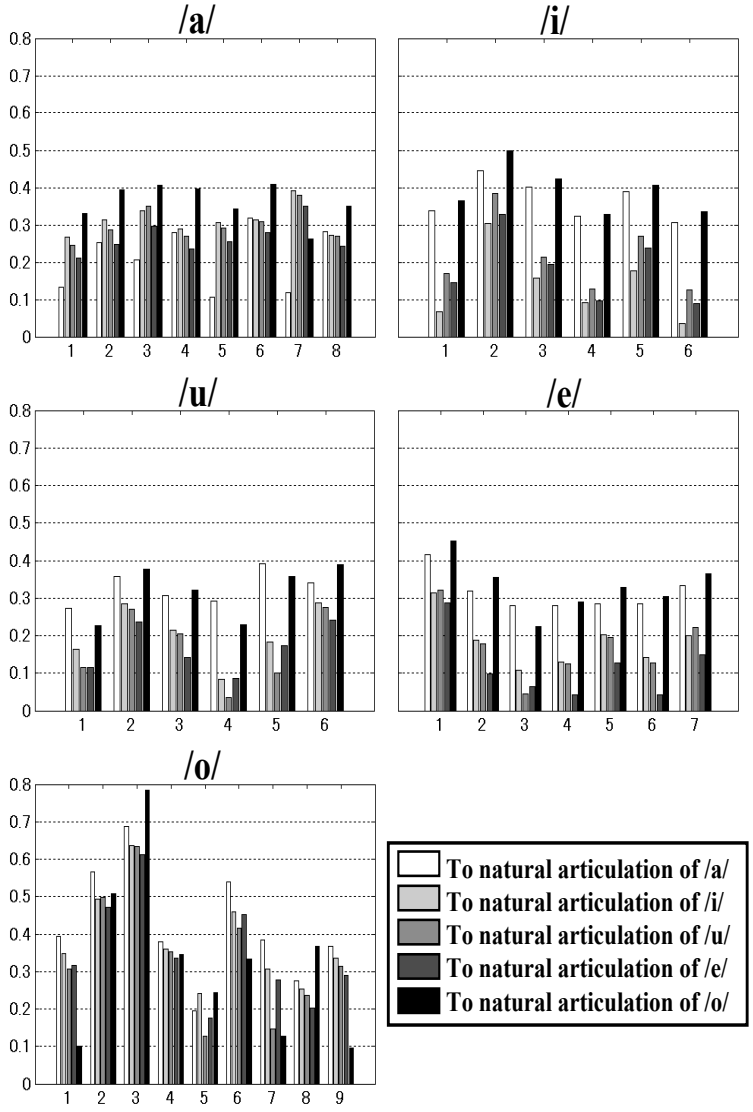

図 3.6: 不自然調音状態の各クラスタと自然調音状態との距離（横軸：不自然調音状態のクラスタラベル，縦軸：不自然調音状態の各クラスタと自然調音状態との正規化距離）

上方に位置するクラスタに分けられる。また，下顎の位置は/u/と同様，上方に位置するクラスタと下方に位置するクラスタに分けられる。

3.4.5 考察

自然調音状態と不自然調音状態を識別する場合，二つの分布間の重複度が小さいほど，高い精度の識別が可能となる。 3.4.4 項における分布構造の分析の結果から，分布構造に含まれる自然調音状態のクラスタと不自然調音状態のクラスタの重なりは小さいと考えられる。これは，調音状態の分布構造に基づくことで，自然調音状態と不自然調音状態が高い精度で識別できる可能性を示唆する。従って，次節において分布構造に基づき自然調音状態と不自然調音状態を識別する手法を提案する。

また，分類規準に対する不自然調音状態の母音ごとの形状の傾向をみると，/a/と/o/の場合，舌背より後方は分類の規準範囲内に含まれるが，下顎と舌尖が規準範囲外になる傾向が見られる。/e/の場合も下顎と舌尖が規準範囲外になる傾向が見られるが，舌背が範囲外となる割合が/a/や/o/より多い。一方，/i/の場合，下顎は規準範囲内であるが，舌尖から舌背にかけて範囲外になる傾向が見られる。/u/の場合は，下顎が範囲外となる割合が一番大きいが，他の母音と比べて，舌全体として範囲外になる割合が大きい。このように狭母音の不自然調音状態は，自然調音状態の母音の調音における声道中の狭めの形成に寄与する舌背の位置が自然調音状態と比べて大きく異なる場合が見られる。しかしながら，母音全体をとおしてみると，不自然調音状態は舌背の位置が自然調音状態と同じであるが，下顎や舌尖は大きく異なることが示されている。自然調音状態と不自然調音状態との分類規準は，単独発話と連続音声中の発話の両方を考慮し定められている。従って，不自然調音状態の形状の傾向から，連続音声の定常部に対する調音運動の目標が，単純な声道中の狭めの位置や大きさだけでは人間が自然調音状態を獲得することは難しいことが示唆される。

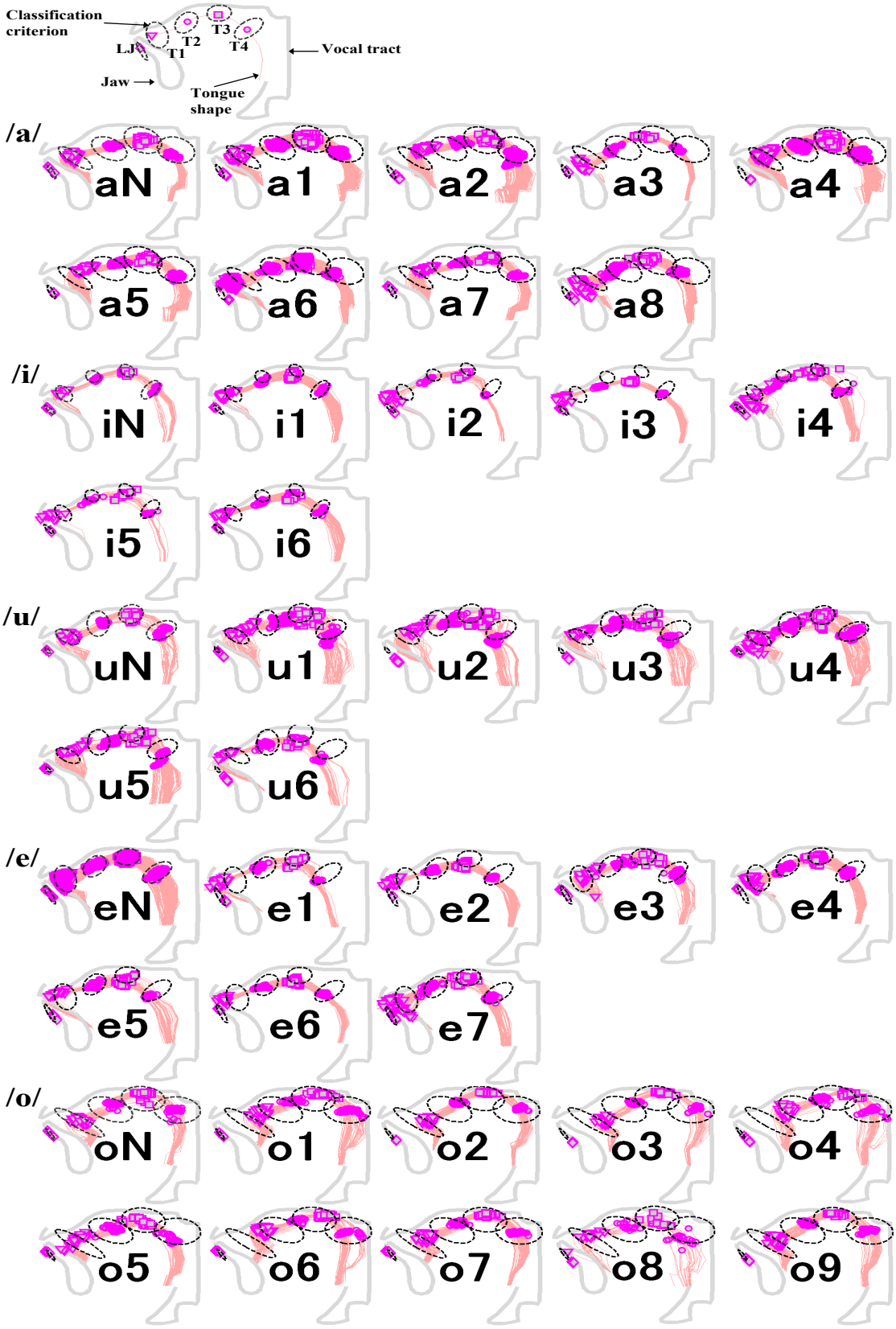

図 3.7: 調音状態の分布構造に含まれるクラスタごとの調音形状

3.5　まとめ

本章では，生成された日本語 5 母音の調音状態を観測データの発話器官の調音運動に基づく規準を用いて，自然調音状態と不自然調音状態に分類した。また，異なる調音状態の分布間の重複度を減少させるため，分類した自然調音状態と不自然調音状態を KPCA により非線形空間に射影し，調音状態と射影後の非線形特徴量それぞれに対して重複度を定量的に求め比較した。その結果，新たに提案した下顎，舌尖，舌背の差異を強調するカーネル関数を用いた KPCA により得られた非線形特徴量に対する異なる調音状態の分布間の重複度は，調音状態に対する重複度より 37%減少することが示された。さらに，KPCA により得られた非線形特徴量に対してクラスタ判別法を用いて特徴量間のクラスタ構造を考慮し次元圧縮することにより，調音状態の分布構造を可視化し，日本語 5 母音として取り得る不自然調音状態の形状の傾向を示した。

第 4章

調音状態の分布構造に基づく自然調音状態と不自然調音状態の識別手法の検討

4.1 はじめに

音声信号から調音状態の逆推定の推定候補に含まれる不自然調音状態を取り除くためには，自然調音状態と不自然調音状態を精度良く識別する必要がある。前章で示した音声信号と一対多の関係にある調音状態の分布構造空間では，自然調音状態と不自然調音状態の分布は，27 次元以上で標準偏差の 2 倍の範囲の重なりがなくなる結果が示された。つまり，調音状態の分布構造空間で識別を行うことで，精度の高い自然調音状態と不自然調音状態の識別が期待できる。従って，調音状態の分布構造に基づき自然調音状態と不自然調音状態を識別する手法を提案し，識別実験により提案手法を評価する。

4.2 調音状態の分布構造に基づく調音状態の識別手法

調音状態の分布構造空間における自然調音状態と不自然調音状態の分布間の重なりは小さいながらも存在するため，調音状態の識別には確率的手法を用いることが適切と考えられる。従って，調音状態の分布構造に基づく調音状態の識別は事後確率最大則に従い，分布構造に含まれる各クラスタに対して事後確率を求め，事後確率が最大のクラスタが自然調音状態のクラスタの場合は自然調音状態，それ以外は不自然調音状態とする。事後確率最大則に従い識別を行うことにより，識別の際に求めた事後確率を調音状態の逆推定における有用な情報として利用することも可能となる。

クラスタごとの事後確率 $p(G_i|\boldsymbol{\mathcal{X}})$ は，調音状態 $\boldsymbol{x}$ の分布構造空間への射影を $\boldsymbol{\mathcal{X}}$，クラスタ G_i，$i = 1, \ldots, 41$ の事前確率を $p(G_i)$，$\boldsymbol{\mathcal{X}}$ のクラスタ G_i に対する尤度を $p(\boldsymbol{\mathcal{X}}|G_i)$ とすると，ベイズの公式 [34] より下記の式から求めることができる。

$$
\begin{aligned}
p(G_i|\boldsymbol{\mathcal{X}}) &= \frac{p(\boldsymbol{\mathcal{X}}|G_i)p(G_i)}{\sum_{i=1}^{M} p(\boldsymbol{\mathcal{X}}|G_i)p(G_i)} \\
p(G_i) &= \frac{N_{G_i}}{N}
\end{aligned}
\tag{4.1}
$$

なお，M はクラスタ数を表し，N 及び N_{G_i} はそれぞれ，全データ数とクラスタ G_i に含まれるデータ数を表す。ここで，分布構造に含まれるクラスタは非線形空間上で定義されており分布が複雑なため，ガウス分布などの確率分布モデルを仮定するパラメトリックな手法により尤度を求めることは適切ではない。従って，式 4.1 のク

ラスタごとの尤度を求めるために，代表的なノンパラメトリック手法であるParzen窓推定法 [38] を用いる。Parzen窓推定法を用いることにより，確率分布モデルを仮定せずに分布に含まれるデータから直接尤度を計算することが可能となる。窓関数としてガウシアン関数を用いたParzen窓推定法により尤度を求める式を下記に示す。

$$p(\boldsymbol{\mathcal{X}}|G_i) \;=\; \frac{1}{N_{G_i}} \sum_{\boldsymbol{\chi}_j \in G_i} \frac{1}{(2\pi h^2)^{D/2}} \exp\left\{ -\frac{\|\boldsymbol{\mathcal{X}} - \boldsymbol{\chi}_j\|^2}{2h^2} \right\} \tag{4.2}$$

ここで，$\boldsymbol{\chi}_j$ はクラスタ G_i に含まれる分布構造空間の非線形特徴量を表す。また，D は分布構造空間の次元数を，h はクラスタに含まれるデータから確率密度関数を推定する際の平滑化パラメータを表す。この h の値により推定される確率密度関数が変わるため，h の変化は尤度に大きく影響する。

式 4.2 から求められる尤度を用いて式 4.1 により計算した事後確率に基づき自然調音状態と不自然調音状態を識別する関数を下記のように定める。

$$\begin{aligned} F(\boldsymbol{\mathcal{X}},\, c) &= \delta(c,\, \arg\max_i\, p(G_i|\boldsymbol{\mathcal{X}})) \\ where \quad \delta(j,\, k) &= \begin{cases} 1 & if\ j = k \\ 0 & otherwise \end{cases} \end{aligned} \tag{4.3}$$

なお，c は $\boldsymbol{\mathcal{X}}$ に対する自然調音状態のクラスタを示すラベル番号を表す。式4.3 より，調音状態 $\boldsymbol{x}$ は，$F(\boldsymbol{\mathcal{X}}, c) = 1$ の時は自然調音状態として，$F(\boldsymbol{\mathcal{X}}, c) = 0$ の時は不自然調音状態として識別される。

4.3 識別実験

前節で提案した調音状態の識別関数を評価するため，3 種類の識別実験を行う。なお，識別精度は識別関数の平滑化パラメータ h に大きく影響されるため，この識別実験で平滑化パラメータ h の最適値についても検討する。また，Parzen窓推定法により尤度を求める場合，式 4.2 より識別対象とクラスタに含まれるすべての非線形特徴量との差分ベクトルの内積計算が必要となり，低い次元数で高い識別精度を得られることが望ましい。従って，識別誤差と分布構造空間の次元数 D とのトレードオフについても検討する。

4.3.1 実験方法

3 章の表 3.1 に示されている自然調音状態と不自然調音状態を含む日本語 5 母音の生成された調音状態を実験データとし，3 種類の実験すべてに同じ実験データを用いる。また，実験手順は 3 種類共通で次の通りとする。実験データを評価データと各クラスタの尤度を求めるための学習データに分け，評価データに対して式 4.3 の識別関数を用いて調音状態の識別を行い，識別誤差を求める。

実験 1 : 識別誤差に対して最適な識別関数の平滑化パラメータ h の値を求める。実験条件として，平滑化パラメータ h は調音状態の分布構造に含まれるすべてのクラスタに対して同じ値とし，0.01 から 0.2 まで変化させる。また，分布構造空間の次元数 D は 27 次元とする。

実験 2 : 調音状態の分布構造に含まれる自然調音状態の近傍に位置する不自然調音状態のクラスタの，識別誤差への影響を調べる。ここで，自然調音状態の近傍に位置する不自然調音状態のクラスタを ”クリティカルクラスタ ” と呼ぶこととする。クリティカルクラスタは自然調音状態の近傍に位置するため，クリティカルクラスタに対する平滑化パラメータの変化は，それ以外のクラスタに対する変化よりも識別誤差に大きく影響することが予想される。従って，実験条件として，平滑化パラメータ h はクリティカルクラスタに対する h のみ 0.01 から 0.2 まで変化させ，その他のクラスタに対しては実験 1 で得られた最適値を固定値として用いる。また，比較のため不自然調音状態のすべてのクラスタに対する h を 0.01 から 0.2 まで変化させ，自然調音状態のクラスタに対しては最適値を固定値として用いる条件でも識別を行う。分布構造空間の次元数 D は，実験 1 と同様 27 次元とする。なお，実験 1 において自然調音状態が識別対象データの場合に誤識別された不自然調音状態のクラスタをクリティカルクラスタとして識別を行う。

実験 3 : 識別誤差と分布構造空間の次元数 D とのトレードオフについて検討する。調音状態の識別手法を音声信号から調音状態の逆推定に適用する場合，逆推定では多くの場合反復処理を必要とする（例えば [11]）ため，逆推定の計算負荷を大きく増加させる可能性がある。よって，識別誤差と次元数 D の関係を調べる。実験条件として，次元数 D を 3 から 27 次元まで変化させる。平滑化パラメータ h は，すべてのクラスタに対して実験 1 で得られた最適値を固定値として用いる。

4.3.2 識別誤差

有限のデータ群を評価データと学習データに分け識別手法を評価する際に，評価データが少ない場合は，高い識別精度が得られたとしても識別手法の汎化能力については保証されない。一方，学習データが少ない場合は，汎化能力は妥当であっても，逆に高い識別精度を得ることが難しくなる。従って，有限のデータ群に対して評価データと学習データ双方のデータ数の確保に適した評価方法である，3.3.3 項で述べた Leave-one-out 法をここでも用いる。

通常，Leave-one-out 法は，誤識別した回数を全データ数で割った値を識別誤差とする。よって，今回の識別実験では，自然調音状態のデータに対して誤識別された回数と不自然調音状態のデータに対して誤識別された回数の和を全データ数で割った値を識別誤差とすることが考えられる。しかしながら，今回用いる実験データは自然調音状態と不自然調音状態のデータ数の割合に大きな偏りがあり，単純に自然調音状態と不自然調音状態それぞれの誤識別の回数を足すことは適切ではない。従って，それぞれのデータ数の割合を考慮した下記の式により識別誤差を求める。

$$Error\,[\%] = \frac{w_N \times N_{N_e} + w_U \times N_{U_e}}{N} \times 100 \tag{4.4}$$

ここで，N_{N_e} と N_{U_e} はそれぞれ，自然調音状態に対する誤識別の回数と，不自然調音状態に対する誤識別の回数を表し，N は自然調音状態と不自然調音状態を合わせたデータ数を表す。また，w_N と w_U はそれぞれ，自然調音状態の誤識別の回数に対する重み係数と，不自然調音状態の誤識別の回数に対する重み係数を表す。識別誤差として両方の誤識別を平等に扱うため，表 3.1 の両方のデータ数の比を正規化した値，$w_N = 0.8$，$w_U = 0.2$ とする。

4.3.3 結果と考察

実験 1 の結果を図 4.1 に示す。識別誤差 $Error$ が最小となる平滑化パラメータ h の値は 0.06 となり，その時の全体の誤差は 2.7%，自然調音状態の誤差は 3.1%，不自然調音状態の誤差は 0.8% となった。ここで，分布構造空間では，自然調音状態の分布と不自然調音状態の分布は標準偏差の 2 倍の範囲で重なりが無い。つま

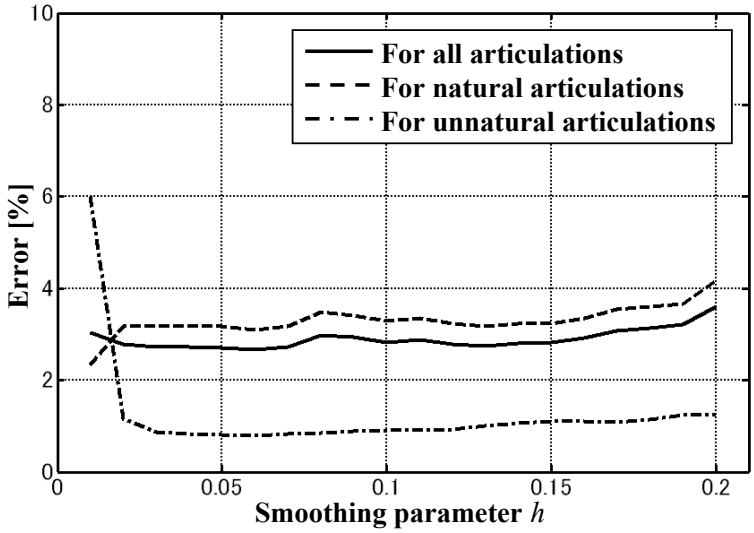

図 4.1: 平滑化パラメータの変化に対する識別誤差。次元数は $D = 27$。

り，それぞれの分布の 95%は，尤度を求める際の密度関数が適切に求められた場合正しく識別することができ，残りの 5%のデータに対して識別誤差が生じる可能性があると考えられる。前述のように h に適切な値を設定することにより誤差が 5%未満に収まることが示されていることから，密度関数は適切に求められており，分布構造に基づくことが誤差を 5%未満の小さな誤差に抑えられた要因と考えられる。また，平滑化パラメータの値が 0.02 未満の範囲では，不自然調音状態の識別誤差は指数関数的に増大し，自然調音状態の識別誤差は減少している。この結果は，不自然調音状態の識別誤差は $h < 0.02$ の平滑化パラメータの変化に敏感であることを示している。なお，これ以降平滑化パラメータの最適値として $h = 0.06$ を用いる。また，実験 1 の結果，自然調音状態が誤って識別された不自然調音状態のクラスタ数は 17 となった。これらのクラスタは，実験 2 においてクリティカルクラスタとして扱われた。

実験 2 の結果を図 4.2 に示す。図 4.2 の横軸は不自然調音状態に対する識別誤差を，縦軸は自然調音状態に対する識別誤差を示し，特定のクラスタの平滑化パラメータ h を変化せた場合の自然調音状態と不自然調音状態それぞれの識別誤差の変化を表している。不自然調音状態のすべてのクラスタに対する h を変化させた場合の結果と比較し，クリティカルクラスタに対する h のみ変化させた場合は，識別誤差が同等または減少している。この結果は，クリティカルクラスタが識別誤

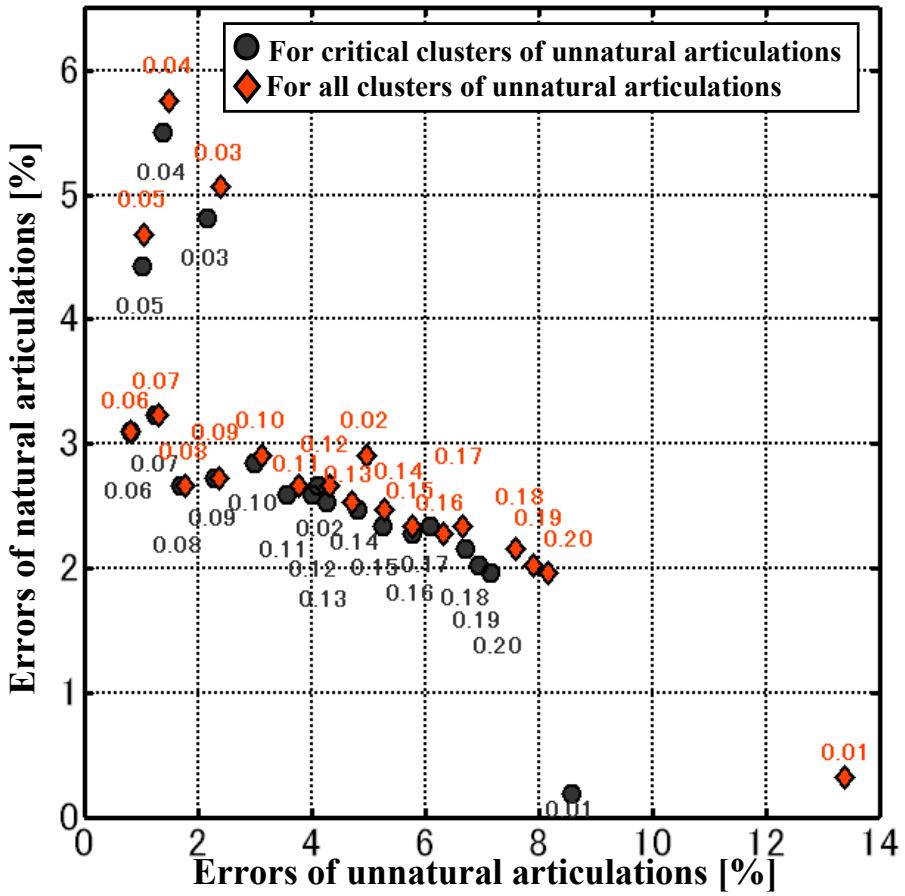

図 4.2: クリティカルクラスタの平滑化パラメータの変化に対する識別誤差。記号の上または下の数値は平滑化パラメータの値を示す。次元数は $D = 27$。

差に大きく影響していることを示している。また，h の値が $h > 0.08$ 及び $h < 0.02$ の範囲では，クリティカルクラスタのみ変化させた場合の方が，自然調音状態の識別誤差が減少する割合はより大きく，不自然調音状態の識別誤差が増加する割合はより小さくなっている。この結果は，クリティカルクラスタの平滑化パラメータを調整することで，自然調音状態と不自然調音状態の識別誤差の割合を制御できる可能性を示唆している。

実験 3 の結果を図 4.3 に示す。次元数 D の増加に伴い，識別誤差 $Error$ は単調減少しているが，16 次元以上では，変化はほとんどみられなくなっている。次元数が 27 次元を超える場合についても調べてみたが，27 次元における識別誤差と同等の値となった。この結果は，27 次元と同程度の識別精度を保った状態で 16 次元

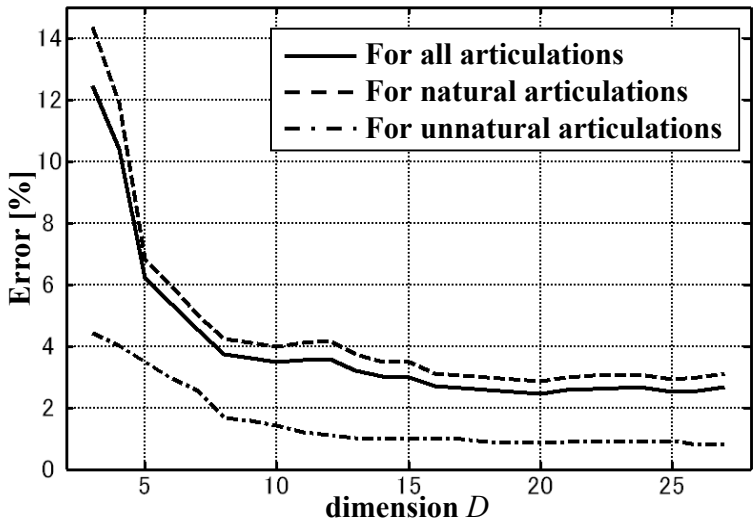

図 4.3: 次元数の変化に対する識別誤差。平滑化パラメータはすべてクラスタに対して $h = 0.06$

まで次元数を削減できる可能性を示唆している。

実験 1~ 3 の結果から，自然調音状態と不自然調音状態の識別手法を音声信号から調音状態の逆推定に適用した場合の効果について考察する。調音状態の識別において，2 種類の識別誤差が存在する。一つは，不自然調音状態が自然調音状態として誤識別された場合の誤差であり，もう一つは，自然調音状態が不自然調音状態として誤識別された場合の誤差である。識別手法を逆推定に適用する場合，前者の識別誤差は不自然調音状態の除去の程度に影響し，誤差が小さいほど不自然調音状態は正確に除去される。一方，後者の識別誤差は逆推定の推定精度に直接影響し，誤差が小さいほど識別誤差が推定精度に影響を与えるリスクを抑えることができる。実験 1 の結果から，平滑化パラメータが最適値の場合，前者の識別誤差は 0.8% であり，後者の識別誤差は 3.1% となっている。これは，提案した識別手法を調音状態の逆推定に適用することにより，従来の逆推定の手法では取り除くことができなかった不自然調音状態を 99%以上の精度で除去し，その際に識別誤差が推定精度に影響を与えるリスクを数%に収められる可能性を示唆する。また，提案手法を逆推定に適用した際に，推定精度に与えるリスクをさらに抑える必要がある場合は，クリティカルクラスタの平滑化パラメータを調整することで対応可能と考えられる。

4.4 まとめ

調音状態の分布構造に基づき，自然調音状態と不自然調音状態を確率的に識別する手法を提案した。提案した識別手法を評価するため，3.2.1 節で分類した5母音の調音状態を評価データ及び識別関数のパラメータ学習データとして，3種類の実験を行った。

実験 1 の結果から，識別誤差を最小とする識別関数の平滑化パラメータ h の最適値（$h = 0.06$）が得られ，その際に不自然調音状態に対しては 99.2%，自然調音状態に対しては 96.9%の識別精度が示された。

実験 2 の結果から，調音状態の分布構造において自然調音状態近傍に位置する不自然調音状態のクリティカルクラスタの平滑化パラメータを調整することで，自然調音状態と不自然調音状態の識別誤差の割合を制御できる可能性が示唆された。

実験 3 の結果から，27 次元と同程度の識別精度を保ったまま 16 次元まで分布構造空間の次元数を削減できる可能性が示唆された。

第 5章

音声信号から調音状態の逆推定における不自然調音状態の除去

5.1 はじめに

音声信号から調音状態の逆推定における一対多の問題を解決するためには，従来の研究では考慮されていなかった不自然調音状態を逆推定の推定候補から除去する必要がある。そのために， 4 章で提案した自然調音状態と不自然調音状態を識別する手法を適用した新たな逆推定システムを構築し，推定候補に含まれる不自然調音状態の除去を試みる。新たなシステムのベースとなる逆推定処理には，部分 3 次元生理学的発話機構モデルを用いて音声信号から調音状態を逆推定する手法 [11] を用いる。さらに，構築したシステムに対して，観測データの発話器官の調音運動と音声信号を実験データとして逆推定を行い，実際の逆推定における不自然調音状態の除去の精度について検証する。また，不自然調音状態を取り除いた場合の効果についても検討する。

5.2 識別手法の逆推定への適用による不自然調音状態の除去

自然調音状態と不自然調音状態の識別手法を音声信号から調音状態の逆推定に適用し，推定候補に含まれる不自然調音状態を取り除くためには，識別結果に基づき次の二つの処理を行う必要がある。一つは，識別結果が自然調音状態の場合，その後不自然調音状態への収束を避ける処理であり，もう一つは，識別結果が不自然調音状態の場合，その後自然調音状態へ収束させる処理である。

本研究では，前者の処理を行うために，推定された調音状態の分布構造空間への射影と自然調音状態の平均とのマハラノビス平方距離 [34] を，逆推定の目標となる調音ターゲットを更新する際の評価関数の新たな項として加える。この分布構造空間上のマハラノビス平方距離は自然調音状態のクラスタから離れるほど大きな値をとることから，この項に重み係数を掛け，更新した調音ターゲットが自然調音状態になるように重み係数を調整することにより，不自然調音状態への収束の回避が期待できる。なお，自然調音状態の平均は目標音声の母音に対する自然調音状態のクラスタ平均とする。

また，後者の処理を行うために，推定結果が不自然調音状態だった場合は，推

定結果が自然調音状態のクラスタに含まれることを保証する調音ターゲットに代替する。代替する手順は，不自然調音状態を分布構造空間へ射影し，その射影と自然調音状態の平均とを結ぶ直線と自然調音状態の分布に対する信頼度 68%の信頼楕円体との交点を求め，その交点の非線形特徴量に対応する調音ターゲットを新たな調音ターゲットとする。この処理により，推定結果が不自然調音状態だった場合，次の段階で推定された調音状態は自然調音状態となることが保証される。

上記の処理の詳細については，次節で述べる。

5.3 新たな逆推定システムの構成

Dang と Honda により，部分 3 次元生理学的発話機構モデルを用いて音声信号から調音状態を逆推定する手法が提案されている [11]。この手法は，従来の制約条件を内包する発話機構モデル [14] を用いることにより，他の逆推定の手法に比べより効率的に制約条件を推定結果に反映できると考えられる。さらに推定された調音状態の可視化も容易なことから，本研究では Dang と Honda の手法をベースに識別手法を適用した新たな逆推定システムを構築する。ただし，Dang と Honda の手法で用いられた発話機構モデルはその後筋収縮力の推定方法が改善され，より正確な人間の発話運動の再現が可能になっている [16]。また，Dang と Honda の手法では，音響特徴量として LPC cepstrum が用いられているが，高次の周波数より低次の周波数に大きな重みが掛けられている MFCC を用いた方が，音韻性との関連性が高い低次のホルマント周波数の影響がより大きく反映されることから，逆推定に適していると考えられる。よって，新たな逆推定システムでは，最新の部分 3 次元生理学的発話機構モデル [16] を用い，また音響特徴量には LPC cepstrum の代わりに MFCC を用いる。

新たに構築した逆推定システムの処理の流れを図 5.1 に示す。構築した逆推定システムは，大まかに二つの部分から成る。一つは，音声信号に基づき発話機構モデルの制御パラメータを調音ターゲットに近づけることにより調音状態を逆推定する部分であり，これを逆推定部とする。もう一つは，推定された調音状態を識別し，識別結果に基づき評価関数を最小化する調音ターゲットの更新または調音ターゲットの代替を行う部分であり，これをターゲット更新部とする。ここで，

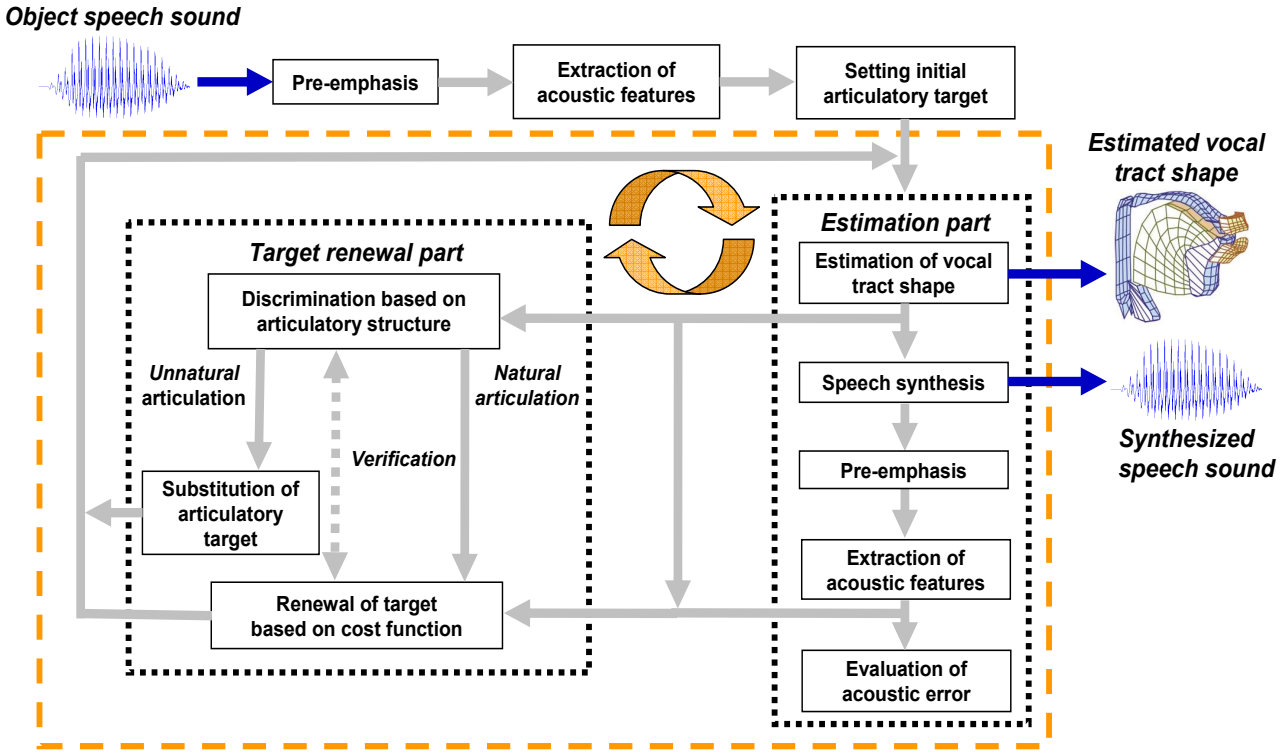

図 5.1: 音声信号から調音状態の逆推定処理の流れ

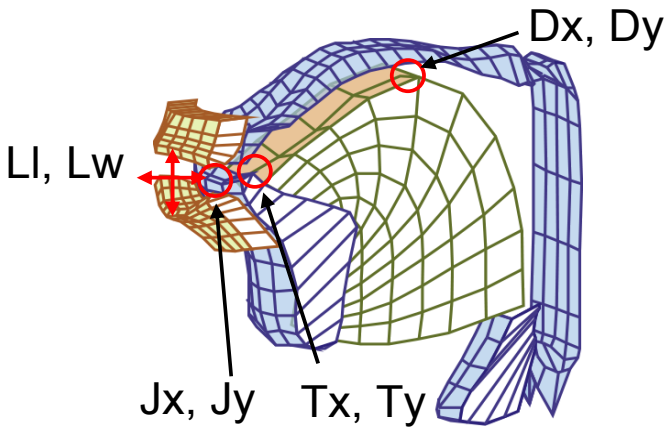

図 5.2: 制御パラメータ

制御パラメータは，調音状態 $\boldsymbol{x}$ に含まれる正中矢状断面上の舌尖 (Tx，Ty) と舌背 (Dx，Dy) 及び下顎 (Jx，Jy) のパラメータに加え，口唇を近似する音響管の長さ (Ll) と直径 (Lw) を含む 8 次元のベクトル $\boldsymbol{z} = (z_1, \ldots, z_8)$ とする。制御パラメータを図 5.2 に示す。調音ターゲットも制御パラメータと同じ要素を持つ 8 次元のベクトルとする。

主な逆推定処理の流れは次の通りとする。まず入力された目標音声信号に対して前処理を行った後に，音響特徴量（MFCC の 0 次~ 12 次の係数及び第 1，第 2 ホルマント周波数）を抽出する。次に抽出したホルマント周波数に基づき初期調音ターゲットを設定し，逆推定部において初期調音ターゲットに基づき調音状態の逆推定及び音声合成を行う。合成された音声に対しても目標音声と同様の手順で音響特徴量を抽出し，目標音声の音響特徴量との誤差を評価する。誤差が基準値より大きい場合は，ターゲット更新部に移り，推定された調音状態の識別を行う。識別結果が自然調音状態の場合は，評価関数の値が最小且つ推定結果が自然調音状態となるように調音ターゲットを更新し，不自然調音状態の場合は，推定結果が自然調音状態となる調音ターゲットに代替する。その後再び逆推定部に移り，更新または代替した調音ターゲットを用いて逆推定を行う。この逆推定部とターゲット更新部の反復処理を，目標音声と合成音声の音響特徴量の誤差が基準値未満になるまで行い，誤差が基準値未満となる推定された調音状態と合成音声を最終結果として出力する。

次項以降で逆推定システムの各処理の詳細について述べる。

5.3.1 音響特徴量の抽出

音声信号に対する前処理及び，音響特徴量（MFCC とホルマント周波数）の抽出処理とその条件は， 2.5.1 項と同様とする。また，推定された調音状態に基づき音声合成を行なう方法は， 2.4.2 項の音声合成の手順と同様とする。ただし，合成音声の母音区間は，音声信号のフレームごとに求めた MFCC の低次の係数に対する回帰係数に基づき定める。n 番目のフレームの MFCC の i 次の係数を $b_i(n)$ とすると，t 番目のフレームの回帰係数 $r(t)$ は下記の式から求められる。

$$
\begin{aligned}
r(t) &= \frac{1}{6}\sum_{i=0}^{5} r_i(t)^2 \\
r_i(t) &= \frac{\left(\sum_{n=-n_0}^{n_0} b_i(n)\cdot n\right)}{\left(\sum_{n=-n_0}^{n_0} n^2\right)}
\end{aligned}
\tag{5.1}
$$

ここで，n_0 は係数ごとに回帰係数を計算する際に考慮するフレームの範囲を表し，$n_0 = 2$ とする。式 5.1 から得られた各フレームの回帰係数に基づき，$r(t)$ が最小となるフレーム及びその前後合わせた 6 フレームを合成音声の母音区間とし，音声信号の母音区間に含まれる各フレームの平均を音響特徴量として用いる。なお，これ以降合成音声の音響特徴量は，推定された調音状態に含まれる制御パラメータ $\boldsymbol{z}$ の関数ベクトル $\boldsymbol{f}(\boldsymbol{z})$ として表す。

5.3.2 初期調音ターゲットの設定

初期推定用の調音ターゲットを母音ごとに用意し，目標音声の第 1 及び第 2 ホルマント周波数から選択された母音の調音ターゲットを初期調音ターゲットとして用いる。これは，母音ごとに第 1 及び第 2 ホルマント周波数の分布する範囲がおおまかに定まるためである [39, 40]。ただし，初期値が母音に依存して固定されることを避けるため，目標音声の第 1 及び第 2 ホルマント周波数と，発話機構モデルを作成する際の被験者の第 1 及び第 2 ホルマント周波数との差を調べ，差が最小の母音及び 2 番目に小さい母音を選択し，それぞれの差の割合を重みとする加

重加算された調音ターゲットを初期調音ターゲットとする。初期調音ターゲットを求める際に用いられる母音ごとの調音ターゲットは， 2.2 節で述べた観測データの発話器官の調音運動に対する各母音の平均とする。

5.3.3 調音ターゲットに対する調音状態の推定

調音状態の推定は Dang と Honda の方法 [14, 16] に従う。具体的には，発話機構モデルの制御パラメータ $\boldsymbol{z}$ に含まれる舌尖，舌背及び下顎の制御点が，設定された調音ターゲットの舌尖，舌背及び下顎に一致するように，舌と下顎に関する筋の収縮力を求めることにより発話機構モデルを駆動し，調音状態を推定する。

筋の収縮力を求める方法は，まず設定された調音ターゲットの位置から Equilibrium position map (EP map) に基づき初期収縮力を求める。EP map は，舌に関する筋の主動筋と協調筋，または主動筋と協調筋及び拮抗筋の複数の組み合わせに対して，組み合わせに含まれる各筋に 8 段階の収縮力を設定した場合の舌尖と舌背の制御点の軌跡により構成される。この EP map を事前に求めておくことにより，EP map 上の調音ターゲットの位置から，舌筋と下顎の筋への初期収縮力を求める。

筋への初期収縮力を求めた後に，現在の制御点の位置から調音ターゲットの位置へのベクトルを筋ワークスペース上に射影することにより，制御点が調音ターゲットへ近づくための筋収縮力を求める。筋ワークスペースは，制御点の変位に対応する主要な筋の収縮ベクトルにより構成された空間であり，制御点から調音ターゲットへの位置のベクトルを筋ワークスペースに射影し各筋の収縮ベクトルに分解することで，分解された各筋収縮ベクトルの大きさとして収縮力を求める。この処理が，計算ステップごとに行なわれることにより，計算ステップごとに制御点が調音ターゲットに近づくことで，調音状態の推定が行なわれる。

5.3.4 音響誤差の評価

調音状態の逆推定処理は，目標音声と推定結果に基づく合成音声との音響特徴量の誤差 Err_C が基準値未満の場合に終了する。Err_C は，MFCC の 1 次から 12 次

までの係数を用いて下記の式から求められる。

$$Err_C = \left(\sum_{i=1}^{12} \{Q_i(f_{Oi} - f_i(\boldsymbol{z}))\}^{\eta} \right)^{\frac{1}{\eta}} \tag{5.2}$$

ここで，f_{Oi} は目標音声に対する MFCC の i 次の係数を，$f_i(\boldsymbol{z})$ は合成音声に対する MFCC の i 次の係数を，Q_i は i 次の係数に対する重み係数を表す。なお，ヒューリスティックな検討から $\eta = 10$ とし，重み係数 Q_i は全ての係数に対して 0.083 とする。また，基準値は 0.01 とする。

5.3.5 推定された調音状態の識別

音響誤差の評価において誤差が基準値を超える場合，ターゲット更新部に移り，推定した調音状態に対して自然調音状態と不自然調音状態の識別を行う。識別には，4 章で提案した手法を用いる。識別は分布構造空間で行なうため推定した調音状態を分布構造空間へ射影し，推定結果の射影に対する事後確率が最大となるクラスタが，目標音声の自然調音状態のクラスタの場合は自然調音状態として，そうでなければ不自然調音状態として識別する。なお，これ以降推定した調音状態 $\boldsymbol{x}$ の分布構造空間への射影を，制御パラメータ $\boldsymbol{z}$ を引数とする非線形関数ベクトル $\boldsymbol{\Phi}(\boldsymbol{z})$ として表す。

識別処理の際に，下記の式で表される目標音声の自然調音状態のクラスタ平均 $\boldsymbol{\Phi}_N$ と推定結果の射影 $\boldsymbol{\Phi}(\boldsymbol{z})$ との差 $\boldsymbol{D}_{\Phi}(\boldsymbol{z})$ を求めておく。

$$\boldsymbol{D}_{\Phi}(\boldsymbol{z}) = \boldsymbol{\Phi}_N - \boldsymbol{\Phi}(\boldsymbol{z}) \tag{5.3}$$

式 5.3 により求めた $\boldsymbol{D}_{\Phi}(\boldsymbol{z})$ は，調音ターゲットを更新する際の評価関数に用いる。

5.3.6 評価関数に基づく調音ターゲットの更新

識別結果が自然調音状態だった場合，推定結果の制御パラメータ $\boldsymbol{z}$ に対する評価関数の最小化問題を解くことにより新たな調音ターゲットを求める。Dang と Honda により音響特徴量の重み付き二乗誤差の和及び，調音状態の形態的制約と調音状態の連続性に関する動的制約を考慮した評価関数が提案されている [11]。し

かしながら，提案されている評価関数には，推定結果に含まれる不自然調音状態の影響は考慮されていないため，推定結果の分布構造空間への射影と自然調音状態の平均とのマハラノビス平方距離を新たな項として評価関数に加える。この新たな項の追加により，不自然調音状態から自然調音状態に近づくように調音ターゲットが更新されるため，不自然調音状態への収束を避ける効果が更新された調音ターゲットに直接反映される。

推定結果の制御パラメータ $\boldsymbol{z}$，合成音声の音響特徴量 $\boldsymbol{f}(\boldsymbol{z})$ 及び，式 5.3 から求められる目標音声の自然調音状態のクラスタ平均と推定結果の分布構造空間への射影との差 $\boldsymbol{D}_{\Phi}(\boldsymbol{z})$ に基づく新たな評価関数 $J(\boldsymbol{z})$ を式 5.4 に示す。式 5.4 の右辺の第 1 項は音響特徴量の重み付き二乗誤差を，第 2 項は調音状態の異常な形状を避けるための形態学的制約を，第 3 項は連続音声における調音状態の変化の滑らかさを保証するための動的制約を，そして第 4 項は調音ターゲットの更新による推定結果の不自然調音状態への収束を避けるための調音状態の分布構造的制約を意味する。

$$J(\boldsymbol{z}) = \|\boldsymbol{f}_{O} - \boldsymbol{f}(\boldsymbol{z})\|^2_{\boldsymbol{W}_Q} + \|\boldsymbol{z} - \boldsymbol{z}_0\|^2_{\boldsymbol{W}_R} + \|\boldsymbol{z} - \boldsymbol{z}_P\|^2_{\boldsymbol{W}_P} + \mu\|\boldsymbol{D}_{\Phi}(\boldsymbol{z})\|^2_{\boldsymbol{W}_{\Phi}} \qquad (5.4)$$

ここで，$\boldsymbol{z}_0$ は発話機構モデルの初期制御パラメータを，$\boldsymbol{z}_P$ は前フレームの制御パラメータを表す。また，$\boldsymbol{W}_Q$，$\boldsymbol{W}_R$，$\boldsymbol{W}_P$，$\boldsymbol{W}_{\Phi}$ は右辺の各項に対する重み行列を表す。具体的には，$\boldsymbol{W}_Q$ は 5.3.4 項で示した重み係数 $\boldsymbol{Q}$ を対角要素とする対角行列，$\boldsymbol{W}_R$ は対角要素がすべて $R_k = (R_0 - R)\gamma^k + R$ である対角行列とする。k は調音ターゲットの更新回数を表し，R_k は k の増加に伴い徐々に R の値に漸近するため，初期推定結果の精度が悪い場合でも推定値の発散を防ぐことができる。R_k を求める際の定数は，それぞれ $R_0 = 0.1$，$R = 0.05$，$\gamma = 0.5$ とする。第 3 項の重み行列 $\boldsymbol{W}_P$ は対角要素がすべて 0.1 である対角行列とし，第 4 項の重み行列 $\boldsymbol{W}_{\Phi}$ は，目標音声の自然調音状態のクラスタに含まれる分布構造空間上の非線形特徴量から求めた共分散行列の逆行列とする。つまり，式 5.4 の右辺の第 4 項は，分布構造空間における目標音声の自然調音状態のクラスタ平均と推定結果の射影とのマハラノビス平方距離を表す。分布構造空間では 27 次元以上の場合，自然調音状態のクラスタと不自然調音状態のクラスタ間で標準偏差の 2 倍の範囲，つまり信頼度 95%の信頼楕円体に重なりは見られず，信頼楕円体はクラスタ平均からの等マハラノビス平方距離を表す。よって，第 4 項は推定結果が自然調音状態の場

合よりも不自然調音状態の場合により大きな値となることから，評価関数の最小化の結果得られる新たな調音ターゲットに対する調音状態が不自然調音状態となることを回避できると考えられる。なお，第 4 項の μ は重み係数を表し，下記の式により求められる。

$$\mu = \mu_0 \psi^l \ (l = 1 \sim 10) \tag{5.5}$$

更新された調音ターゲットに基づき推定された調音状態の識別結果が不自然調音状態だった場合は，式 5.5 の l を増加させることにより，識別結果が自然調音状態になるように第 4 項の制約の影響を大きくし調音ターゲットを更新しなおす。なお，l の増加に伴い μ の値を最大で 10 倍まで大きくするように $\psi = 1.25$ とする。この処理により，更新後の調音ターゲットに基づく逆推定の結果が自然調音状態となることを保証する。ただし，逆推定における第 4 項の影響は不明なため，μ_0 の値は次節で検討する。

評価関数 $J(\boldsymbol{z})$ を最小とする推定結果の制御パラメータ $\hat{\boldsymbol{z}}$ を得るために，$\hat{\boldsymbol{z}}$ の第 k 次近似解を $\hat{\boldsymbol{z}}_k$ として，まず式 5.4 の $\boldsymbol{f}(\boldsymbol{z})$ 及び式 5.3 の $\boldsymbol{\Phi}(\boldsymbol{z})$ を $\hat{\boldsymbol{z}}_k$ 周辺で線形化（式 5.6 及び式 5.7 ）する。

$$\boldsymbol{f}(\boldsymbol{z}) \ \cong \ \boldsymbol{f}(\hat{\boldsymbol{z}}_k) + \left.\frac{\partial \boldsymbol{f}(\boldsymbol{z})}{\partial \boldsymbol{z}}\right|_{\boldsymbol{z}=\hat{\boldsymbol{z}}_k} \bullet (\boldsymbol{z} - \hat{\boldsymbol{z}}_k) \tag{5.6}$$

$$\boldsymbol{\Phi}(\boldsymbol{z}) \ \cong \ \boldsymbol{\Phi}(\hat{\boldsymbol{z}}_k) + \left.\frac{\partial \boldsymbol{\Phi}(\boldsymbol{z})}{\partial \boldsymbol{z}}\right|_{\boldsymbol{z}=\hat{\boldsymbol{z}}_k} \bullet (\boldsymbol{z} - \hat{\boldsymbol{z}}_k) \tag{5.7}$$

次に式 5.6 と，式 5.3 に式 5.7 を代入した結果を式 5.4 に代入し，制御パラメータ $\boldsymbol{z}$ で偏微分するとこにより下記の式を得る。

$$\begin{aligned}\frac{\partial J(\boldsymbol{z})}{\partial \boldsymbol{z}} &= \nabla J(\boldsymbol{z}) \\ &= -2\Big\{ \boldsymbol{A}_f^T \boldsymbol{W}_Q \Big(\boldsymbol{f}_O - \boldsymbol{f}(\hat{\boldsymbol{z}}_k) - \boldsymbol{A}_f(\boldsymbol{z} - \hat{\boldsymbol{z}}_k) \Big) + \boldsymbol{W}_R(\boldsymbol{z}_0 - \boldsymbol{z}) \\ &\quad + \boldsymbol{W}_P(\boldsymbol{z}_P - \boldsymbol{z}) + \boldsymbol{A}_\Phi^T \boldsymbol{W}_\Phi \Big(\boldsymbol{D}_\Phi(\hat{\boldsymbol{z}}_k) - \boldsymbol{A}_\Phi(\boldsymbol{z} - \hat{\boldsymbol{z}}_k) \Big) \Big\}\end{aligned} \tag{5.8}$$

ただし，$\boldsymbol{D}_\Phi(\hat{\boldsymbol{z}}_k) = \boldsymbol{\Phi}_N - \boldsymbol{\Phi}(\hat{\boldsymbol{z}}_k)$，$\boldsymbol{A}_f = \frac{\partial \boldsymbol{f}(\hat{\boldsymbol{z}}_k)}{\partial \boldsymbol{z}}$，$\boldsymbol{A}_\Phi = \frac{\partial \boldsymbol{\Phi}(\hat{\boldsymbol{z}}_k)}{\partial \boldsymbol{z}}$ とする。

$\nabla J(\boldsymbol{z})$ を 0 と置き，第 k 次近似解 $\hat{\boldsymbol{z}}_k$ 周辺で線形化すると，

$$\begin{aligned}&\nabla J(\boldsymbol{z}) \cong \nabla J(\hat{\boldsymbol{z}}_k) + \nabla^2 J(\hat{\boldsymbol{z}}_k)(\boldsymbol{z} - \hat{\boldsymbol{z}}_k) = 0 \\ &\nabla^2 J(\hat{\boldsymbol{z}}_k) = -\left(\boldsymbol{A}_f^T \boldsymbol{W}_Q \boldsymbol{A}_f + \boldsymbol{W}_R + \boldsymbol{W}_P + \boldsymbol{A}_\Phi^T \boldsymbol{W}_\Phi \boldsymbol{A}_\Phi \right)\end{aligned} \tag{5.9}$$

式 5.9 を解くと，

$$\begin{aligned}\boldsymbol{z} &= \hat{\boldsymbol{z}}_k + \Big\{\boldsymbol{A}_f^T\boldsymbol{W}_Q\boldsymbol{A}_f + \boldsymbol{W}_R + \boldsymbol{W}_P + \mu\boldsymbol{A}_\Phi^T\boldsymbol{W}_\Phi\boldsymbol{A}_\Phi\Big\}^{-1} \\ &\quad \Big\{\boldsymbol{A}_f^T\boldsymbol{W}_Q(\boldsymbol{f}_O - \boldsymbol{f}(\hat{\boldsymbol{z}}_k)) + \boldsymbol{W}_R(\boldsymbol{z}_0 - \hat{\boldsymbol{z}}_k) + \boldsymbol{W}_P(\boldsymbol{z}_P - \hat{\boldsymbol{z}}_k) \\ &\quad +\mu\boldsymbol{A}_\Phi^T\boldsymbol{W}_\Phi\boldsymbol{D}_\Phi(\hat{\boldsymbol{z}}_k)\Big\} \qquad (5.10)\end{aligned}$$

ここで，$\boldsymbol{f}(\hat{\boldsymbol{z}}_{k+1}) \leq \boldsymbol{f}(\hat{\boldsymbol{z}}_k)$ を満たすための係数 ζ を導入することにより，第 $k+1$ 次近似解 $\hat{\boldsymbol{z}}_{k+1}$ を下記の式から得る。

$$\begin{aligned}\hat{\boldsymbol{z}}_{k+1} &= \hat{\boldsymbol{z}}_k + \zeta\Big\{\boldsymbol{A}_f^T\boldsymbol{W}_Q\boldsymbol{A}_f + \boldsymbol{W}_R + \boldsymbol{W}_P + \mu\boldsymbol{A}_\Phi^T\boldsymbol{W}_\Phi\boldsymbol{A}_\Phi\Big\}^{-1} \\ &\quad \Big\{\boldsymbol{A}_f^T\boldsymbol{W}_Q(\boldsymbol{f}_O - \boldsymbol{f}(\hat{\boldsymbol{z}}_k)) + \boldsymbol{W}_R(\boldsymbol{z}_0 - \hat{\boldsymbol{z}}_k) + \boldsymbol{W}_P(\boldsymbol{z}_P - \hat{\boldsymbol{z}}_k) \\ &\quad +\mu\boldsymbol{A}_\Phi^T\boldsymbol{W}_\Phi\boldsymbol{D}_\Phi(\hat{\boldsymbol{z}}_k)\Big\} \qquad (5.11)\end{aligned}$$

式 5.11 から求めた近似解 $\hat{\boldsymbol{z}}_{k+1}$ を新たな調音ターゲットとして，逆推定部において調音状態の逆推定を行う。ただし，偏導関数 $\boldsymbol{A}_f$ 及び $\boldsymbol{A}_\Phi$ を解析的に求めることは，発話機構モデルの構造上及び分布構造空間の非線形性により困難である。そのため，白井と誉田の手法 [10] に従い，$\hat{\boldsymbol{z}}_k$ に微小変動を与え $\boldsymbol{f}(\hat{\boldsymbol{z}}_k)$ 及び $\boldsymbol{\Phi}(\hat{\boldsymbol{z}}_k)$ を計算することにより，偏導関数 $\boldsymbol{A}_f$ 及び $\boldsymbol{A}_\Phi$ を求める。

5.3.7 調音ターゲットの代替

初期推定結果が自然調音状態の場合は，前項で述べた調音ターゲットの更新処理により，その後の反復処理を経て自然調音状態への収束が期待できる。しかしながら，初期推定結果が不自然調音状態の場合は，更新した調音ターゲットに基づき逆推定した結果が必ず自然調音状態となる保証はない。そのため，初期推定結果が不自然調音状態の場合は，通常の調音ターゲットの更新処理を行う代わりに，推定結果が自然調音状態となることが保証された代替の調音ターゲットを求める。

この代替の調音ターゲットを求める手順は次の通りとする。識別の結果が不自然調音状態だった場合，不自然調音状態の分布構造空間への射影と目標音声の自然調音状態のクラスタ平均を結ぶ直線と，自然調音状態のクラスタに対する信頼度 68%の信頼楕円体との交点を求める。この交点上の非線形特徴量は，不自然調

音状態の射影が自然調音状態方向に遷移し確実に自然調音状態のクラスタに含まれる。そのため，対応する調音ターゲットによる推定結果が自然調音状態となることが保証される。従って，交点上の非線形特徴量に対応する調音ターゲットを代替のターゲットとして用いる。ただし，高次元空間上の楕円体と直線を代数的に求めることは難しい。ここで，信頼楕円体は等マハラノビス平方距離を表すことから，クラスタ平均と不自然調音状態の射影を結ぶ直線上の交点までのマハラノビス平方距離は既知となる。よって，既知のマハラノビス平方距離を用いて交点を二分法 [41] により数値解析的に求める。また，KPCA の場合，非線形主成分から元の特徴量を求めることは原理的に困難なため [42]，分布構造を構築する際に用いた非線形特徴量の中から求めた交点に最も近い非線形特徴量を，交点上の非線形特徴量として用いる。

5.4 逆推定実験

新たに構築した音声信号から調音状態の逆推定システムを用いて逆推定実験を行うことにより，不自然調音状態の除去の精度を検証する。逆推定実験の実験データには，観測された発話器官の調音運動と音声信号を用いる。なお，逆推定の推定候補から不自然調音状態を取り除いた場合の効果についても検討する。

5.4.1 実験方法

実験データの音声信号を逆推定システムに入力し，反復処理の後に出力された調音状態と合成音声を推定結果とし，反復過程及び出力結果に対する不自然調音状態の除去の精度と効果について調べる。実験データには， 2.2 節で述べた観測データの発話器官の調音運動と音声信号を用いる。音声資料は，9 個の母音連鎖（/iu/，/ua/，/au/，/ai/, /ui/，/ae/，/ou/，/ie/，/ei/）とする。

実験の逆推定処理に関する条件として，評価関数に含まれる第 4 項の重み係数 μ のパラメータ μ_0 の値を，0，0.01，0.05，0.1，0.5，1，5 の 7 段階に変化させる。ここで，$\mu_0 = 0$ の場合は調音状態の識別を行わない従来の逆推定処理を意味し，調音ターゲットの代替処理も行わない。なお，逆推定の際に目標音声に含まれる音

素の母音の種類及び，音素の先頭と終端の時間は既知とする。

識別処理に関する条件として，識別関数のパラメータである分布構造空間の次元数 D は 27 次元とする。これは，27 次元の分布構造空間では，自然調音状態と不自然調音状態のクラスタ間の重なりが信頼度 95%の信頼楕円体の範囲で無くなるためである。識別関数の平滑化パラメータ h は，逆推定処理中に現れる調音状態に対して自然調音状態の識別誤差が最小になるように，各母音のクリティカルクラスタと他のクラスタそれぞれに対して調整した値（表 5.1）を用いる。これは，予備検討として，すべてのクラスタに対する h の値に 4.3 節で得られた最適値 0.06 を用いて逆推定処理中に現れる調音状態を識別した結果， 4.3 節で示された識別精度よりも大きく劣化したためである。この精度の劣化の原因として，分布構造の構築に用いた調音状態の生成時の舌筋の組み合わせより，逆推定で用いられる舌筋の組み合わせのほうがより複雑なことが考えられる。

表 5.1: 識別関数の平滑化パラメータ h の最適値

	/a/	/i/	/u/	/e/	/o/
h for critical clusters	0.02	0.1	0.2	0.18	0.04
h for other clusters	0.015	0.04	0.3	0.1	0.01

5.4.2 不自然調音状態の除去の精度と効果

不自然調音状態の除去の精度を調べるために，不自然調音状態の削減率及び推定結果に対する自然調音状態と不自然調音状態の割合を求める。不自然調音状態の削減率は，従来法（$\mu_0 = 0$）における不自然調音状態に収束した音素数からの減少率とする。これは，逆推定の最終結果に対する不自然調音状態の除去の精度を表す。一方，推定結果に対する自然調音状態と不自然調音状態の割合は，初期推定及び更新された調音ターゲットに基づく推定結果の調音状態に対する自然調音状態と不自然調音状態の割合とする。これは，逆推定の途中過程に現れる不自然調音状態も含めた除去の精度を表す。なお，識別における識別誤差の影響を避けるため，不自然調音状態の削減率及び推定結果に対する自然調音状態と不自然調

音状態の割合を求める際は，推定された調音状態に 3.2.1 項で定めた規準を当てはめて，自然調音状態か不自然調音状態かの判断を行う。

不自然調音状態を除去した場合の効果を調べるために，目標音声の調音状態の形状と推定結果の調音状態の形状を比較し，また調音状態と音響特徴量それぞれの誤差を求める。目標音声の調音状態は，実験データの各音素に対して，発話器官の調音運動に対する母音区間の平均を調音ターゲットとして発話機構モデルを駆動することにより求める。調音状態の誤差は，制御パラメータの下顎及び舌尖，舌背を対象とし，発話器官の調音運動に対する母音区間の平均と推定結果の調音状態の制御パラメータとの差の絶対値の平均とする。音響特徴量の誤差は，第 1 及び第 2 ホルマント周波数を対象とし，目標音声と合成音声それぞれの母音区間の差の絶対値を目標音声のホルマント周波数で割った値のパーセンテージの平均とする。

5.4.3 結果と考察

従来法（$\mu_0 = 0$）における逆推定の結果，自然調音状態に収束した音素数が 9 個，不自然調音状態に収束した音素数が 9 個となり，5 割が不自然調音状態に収束した。この従来法の結果に対する，$\mu_0 = 0.01 \sim 5$ における不自然調音状態の削減率を図 5.3 に示す。図 5.3 から，μ_0 のすべての場合に不自然調音状態に収束する音素は削減されており，$\mu_0 = 1$ の場合には 9 割削減可能なことが示された。

また，表 5.2 に各 μ_0 の推定における自然調音状態と不自然調音状態の割合を示す。表 5.2 から，反復処理中に現れる不自然調音状態の割合は $\mu_0 \geq 0.01$ のすべての場合に $\mu_0 = 0$ の場合よりも減少しており，$\mu_0 = 1$ の時に 3%となることが示された。$\mu_0 = 0$ に対する $\mu_0 = 1$ の不自然調音状態の割合の減少率は 90%となっており，最終結果だけでなく逆推定の過程も含めて不自然調音状態を 9 割除去可能なことが示された。なお，$\mu_0 = 0.01 \sim 5$ において推定された調音状態が不自然調音状態だった場合でも，$\mu_0 = 0.05$ の 1 回を除くほぼすべての推定結果に対して識別結果は自然調音状態となっていた。この結果は，識別誤差の影響により除去されていない不自然調音状態が存在しているが，識別手法の適用による不自然調音状態の除去の処理自体は正しく機能していることを示している。つまり，μ_0 の値を適切

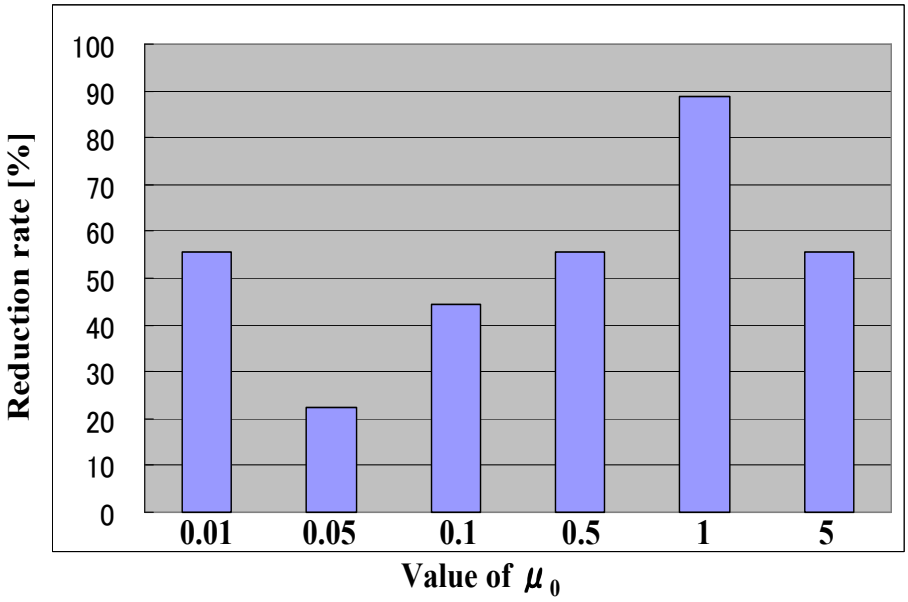

図 5.3: 不自然調音状態の削減率（従来法（$\mu_0 = 0$）において不自然調音状態に収束した音素数に対する減少率）

に設定することにより，識別手法の逆推定への適用により不自然調音状態を除去できる可能性が示された。これ以降，不自然調音状態の削減率が一番大きい $\mu_0 = 1$ の結果を提案法の結果とする。

表 5.2: 推定結果に対する自然調音状態と不自然調音状態の割合

μ_0	0	0.01	0.05	0.1	0.5	1	5
Total number of times for estimation	59	40	45	48	45	37	40
Rate of natural articulations [%]	64	88	71	83	84	97	83
Rate of unnatural articulations [%]	36	12	29	17	16	3	17

次に，逆推定における不自然調音状態を除去した場合の効果を検討するために，推定結果として出力された調音状態の正中矢状断面の形状を比較する。比較に用いた母音連鎖は/ua/とする。/ua/に含まれる母音は共に，従来法（$\mu_0 = 0$）では不自然調音状態に収束したが，提案法（$\mu_0 = 1$）では自然調音状態に収束するように改善された。図 5.4 に，/ua/の/u/の調音状態の形状を，図 5.5 に，/ua/の/a/の

調音状態の形状を示す。図 5.4 ，図 5.5 共に，左から目標音声の調音状態，従来法による推定結果の調音状態，提案法による推定結果の調音状態を表す。形状を比較した結果，/u/と/a/共に従来法よりも提案法の方が目標音声の調音状態により近い舌の形状となっており，/a/に関しては下顎の開きぐらいも提案法の方がより近い状態となっている。また舌の状態は，概形だけでなく舌内部のノードの状態も提案法でより近い状態となっている。この結果は，不自然調音状態を除去することにより，特定の発話部位の位置ではなく，発話器官全体の状態として正しく調音状態を推定できる可能性を示唆する。

なお，従来法において自然調音状態に収束した母音と不自然調音状態に収束した母音に分けて，調音誤差と音響誤差を従来法と提案法で比較した。その結果，不自然調音状態を除去した場合，調音状態の誤差は 0.01 ～ 0.1cm 減少し，音響誤差は 0.8 ～ 3.7%増加した。音響誤差が増加した原因として，本実験では，評価関数の第 4 項のパラメータ μ_0 のみを変化させ，第 1~ 3 項の重みを含む他のパラメータを考慮していないため，音響誤差の項よりも分布構造に関する項の影響がより大きかったことが考えられる。従って，不自然調音状態の除去を調音状態の逆推定の新たな制約条件として逆推定の誤差の精度を高めるためには，不自然調音状態の除去に関するパラメータだけでなく，評価関数における他のパラメータも考慮したパラメータの最適化を行う必要があると考えられる。

5.5 まとめ

調音状態の逆推定における一対多の問題を解決するために，提案した調音状態の分布構造に基づき自然調音状態と不自然調音状態を識別する手法を，最新の部分 3 次元生理学的発話機構モデルを用いた調音状態の逆推定システムに適用し，逆推定の推定候補に含まれる不自然調音状態の除去を試みた。その結果，推定候補に含まれる不自然調音状態の 9 割を除去可能なことが示された。さらに，従来法及び提案法により推定された調音状態の形状の比較から，不自然調音状態を除去することにより，発話器官全体の状態として正しい調音状態が推定される可能性が示唆された。

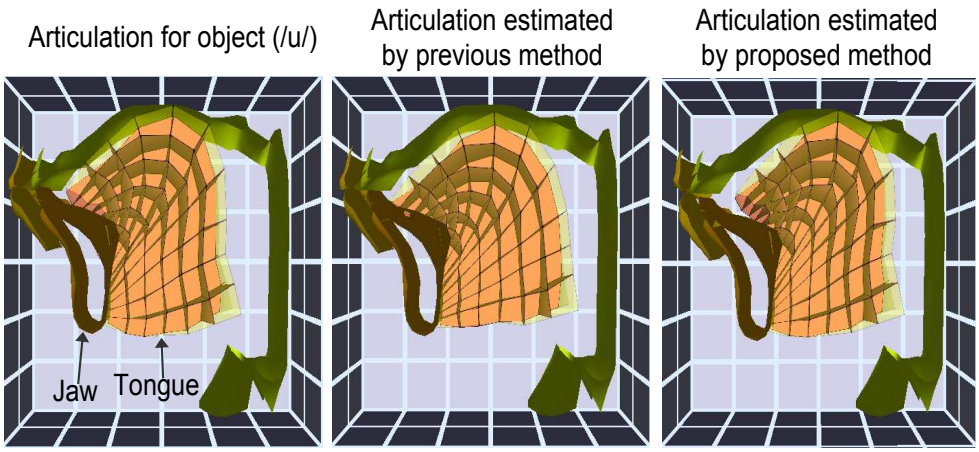

図 5.4: 目標音声（/ua/の/u/）の調音状態及び推定された調音状態の形状。左から目標音声の調音状態，従来法により推定された調音状態，提案法により推定された調音状態。

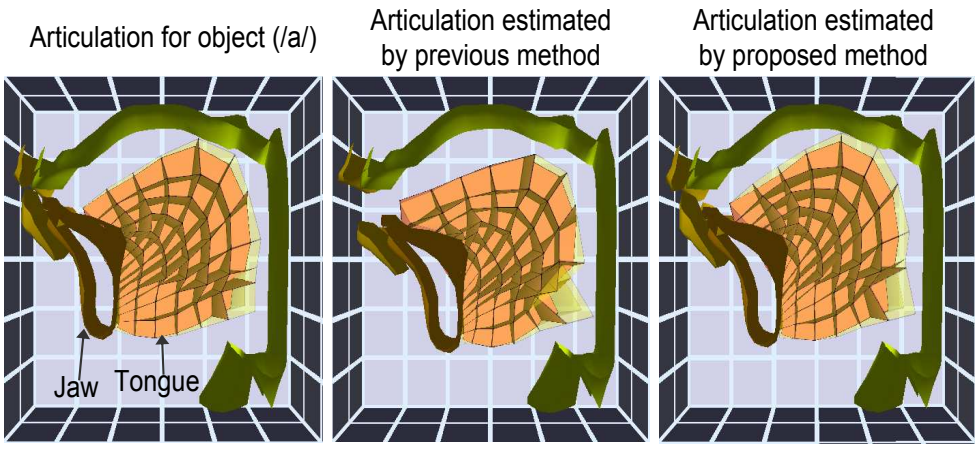

図 5.5: 目標音声（/ua/の/a/）の調音状態及び推定された調音状態の形状。左から目標音声の調音状態，従来法により推定された調音状態，提案法により推定された調音状態。

第 6章

全体に対する考察

本研究は，音声信号から調音状態を逆推定する際の一対多の問題を解決するために，推定候補に含まれる不自然調音状態を取り除くことを目的として，生理学的発話機構モデルを用いて生成した調音状態を分析することにより，音声信号と一対多の関係にある調音状態の分布の全体像を明らかにすることを試みた。さらに，明らかにした調音状態の分布の全体像に基づき自然調音状態と不自然調音状態を識別する手法を提案し，提案手法を調音状態の逆推定システムに適用し逆推定の推定候補に含まれる不自然調音状態を除去することを試みた。その結果得られた成果について考察する。

3 章では，生理学的発話機構モデルを用いて系統的に生成した調音状態を分析し，自然調音状態及び不自然調音状態に対する複数のクラスタから構成される非線形空間上のクラスタ構造を可視化することにより，音声信号と一対多の関係にある調音状態の分布の全体像を示した。このような発話の計算モデルを用いて調音状態と音声信号を生成する他の研究として，Perrier らは 2 次元生体力学モデルを用いて，多数の舌の形状を生成し，その形状に基づき音声合成を行っている [43]。しかしながら，彼らが生成した舌の形状は人間が生成可能であるという保証はなく，また舌の形状に対する分析も PCA が行われているのみで，詳細な調音状態の分布の構造までは示されていない。従って，本研究で示した調音状態の分布構造は新たな知見であり，音声生成の研究に大きく貢献できると考えられる。また，調音モデルを用いたパラメトリックな音声合成 [1] において，分布構造に基づきパラメータを制御することで，人間が生成可能な範囲でパラメータを制御することができ，音声合成の研究にも寄与できると考えられる。

さらに，調音状態の分布構造を分析することにより，日本語 5 母音として取り得る不自然調音状態の具体的な形状とその傾向を明らかにした。本研究で明らかにした不自然調音状態の形状と，腹話術や調音の補償動作における形状を比較することにより，腹話術や調音の補償動作における舌や下顎の筋肉の動きを発話機構モデルを用いて分析することが可能となり，人間の音声生成機構の解明に寄与できると考えられる。

4 章では，調音状態の分布構造に基づき自然調音状態と不自然調音状態を非線形空間で確率的に識別する手法を提案し，提案手法を用いた識別実験の結果，不自然調音状態に対して 99.2%，自然調音状態に対して 96.9%の精度で識別可能な

ことを示した。この結果は，従来の手法では識別できなかった不自然調音状態を99%以上の精度で識別できる可能性を示したことから，音声信号から調音状態の逆推定に対して貢献できると考えられる。なお，実際には逆推定システムの性能や推定対象のデータなどにより，識別手法に必要とされる自然調音状態と不自然調音状態それぞれに対する識別精度は異なると考えられられる。実際 5.4 節では逆推定システムに対して識別関数の平滑化パラメータを調整しており，その結果自然調音状態の識別精度は 4.3 節の結果と同等の精度が得られている。このように，提案した調音状態の識別手法はクリティカルクラスタ及びその他のクラスタの二種類の平滑化パラメータを調整することで自然調音状態と不自然調音状態の識別精度を変更できる柔軟性を持っており， 5 章で示した逆推定システム以外のシステムにも適用可能と考えられる。

5 章では，提案した自然調音状態と不自然調音状態を識別する手法を，音声信号から調音状態の逆推定に適用することにより，逆推定の候補に含まれる不自然調音状態を 9 割除去可能なことを示した。さらに，推定候補に含まれる不自然調音状態を取り除くことにより，発話器官全体の状態として正しい調音状態が推定される可能性が示唆された。逆推定における不自然調音状態を人間が言語を獲得する過程において淘汰された調音状態として捉えると，推定候補に含まれる不自然調音状態の除去を伴う調音状態の逆推定は，人間が言語を獲得する過程における調音状態の取捨選択と捉えることができる。従って，この成果は調音状態の逆推定の研究に大きく貢献するだけでなく，人間の音声生成機構の解明にも寄与できると考えられる。

なお，本研究で行った逆推定の推定候補に含まれる不自然調音状態の除去は，逆推定の新たな制約条件として捉えられる。この新たな制約条件により抑えられる逆推定の解の多様性は，図 1.1 に示されているように，音声信号と一対多の関係にある調音状態に含まれる人間が発話可能な調音状態の中の不自然調音状態に対してのみである。従って，従来の制約条件に本研究で提案した新たな制約条件を加えても，人間が発話可能な調音状態に含まれる自然調音状態に関してまだ解決されていない一対多の問題が存在し，その要因として連続音声中の前後の音素の調音結合 [44] の影響などが考えられる。

第 7章

結論

7.1　本論文で得られた成果の要約

音声信号から調音状態を逆推定する際の一対多の問題を解決するために逆推定の推定候補に含まれる不自然調音状態を除去することを目的として，音声信号と一対多の関係にある人間が発話可能な調音状態の分布の全体像を明らかにすることを試みた。さらに，調音状態の分布に基づき自然調音状態と不自然調音状態の識別手法を提案し，提案手法を調音状態の逆推定に適用することにより，推定候補に含まれる不自然調音状態の除去を試みた。その結果，以下成果が得られた。

まず，今まで明らかにされていなかった音声信号と一対多の関係にある人間が発話可能な調音状態の分布の全体像を，生理学的発話機構モデルを用いて系統的に生成した調音状態を分析することにより，自然調音状態と不自然調音状態を含む複数のクラスタにより構成される非線形空間上のクラスタ構造として示した。

また，調音状態のクラスタ構造に基づき確率的に調音状態を識別する手法を提案し，提案手法を音声信号から調音状態の逆推定に適用した新たな調音状態の逆推定システムを構築した。

さらに，構築した逆推定システムによる逆推定実験から，従来の手法では取り除くことが出来なかった推定候補に含まれる不自然調音状態を 9 割除去可能なことを示した。

本研究で得られた上記の成果は，音声信号から調音状態の逆推定に大きく寄与するだけでなく，聴覚障害者や語学の学習者のための理想的な発話訓練システムの実現に貢献できると考えられる。また，本研究で得られた知見は，人間の音声生成機構の解明や，音声合成に関する研究に大きく寄与できると考えられる。

7.2　今後の課題

本研究で得られた成果を，理想的な発話訓練システムの実現や，音声生成機構の解明などに応用するための克服すべき課題として，下記の点が挙げられる。

1. 自然調音状態と不自然調音状態を識別する手法の音声信号から調音状態の逆推定への適用方法の改善
2. 識別手法を適用した音声信号から調音状態の逆推定システムの汎用化

3. 自然調音状態と不自然調音状態の詳細な比較

課題 1 に対して，提案手法を逆推定に適用する際に， 5.4.3 項で考察したように式 5.4 で表される逆推定の評価関数の第 4 項の重み係数は，評価関数に含まれる他のパラメータも含めて最適化を行う必要がある。また，自然調音状態の誤差を小さく抑え且つ不自然調音状態の誤差をできるだけ小さくする識別関数のパラメータの検討も必要と考えられる。

課題 2 に対して，識別手法を適用した逆推定システムを汎用化するためには，口唇の対応，子音の対応，不特定話者への対応が必要となる。口唇に関して，調音状態の系統的生成において音声を合成する際に観測信号に基づく範囲の口唇の変形が考慮されており，自然な発話を行う際に観測され得る口唇の変形の影響は分布構造に暗に含まれている。しかしながら，調音状態の分布構造を求める際に不自然な口唇の状態は考慮されていないため，観測され得ない口唇の変形の分布構造への影響は不明である。従って，観測される範囲を超えた口唇のパラメータの値を用いて発話機構モデルにより調音状態を生成及び音声合成を行い，今回生成した調音状態と合わせて 5 母音の調音状態を分析し調音状態の分布構造を求めることにより，口唇の不自然な状態も考慮した分布構造が得られると考えられる。

また子音に関しては，本研究で行われた母音に対する分析過程と同じ過程を子音に対しても行い子音も含めた調音状態の分布構造を構築することにより，自然調音状態と不自然調音状態を識別することが可能になると考えられる。ただし，子音の場合は，生成された調音状態から各子音の調音状態を選定する用いる音響特徴量としてホルマント周波数は適切ではない。そのため，子音の各様式に関連性のある音響特徴量（例えば，閉鎖子音に対する有声開始時間 (Voice onset time: VOT)[23] など）と MFCC を用いて選定を行う必要がある。なお，母音の場合は定常部の区間を対象としたが，子音の場合は遷移部の区間を対象とする必要がある。

さらに，本研究で用いた観測データ及び発話機構モデルは同じ 1 名の成人男性を被験者としており，他の話者に対して音声信号から調音状態を逆推定する場合，声道形状には個人差があるため，同等の精度が得られる保証はない。よって，他の話者に対しても精度良く逆推定を行うために，調音状態の分布構造及び発話機構モデルの話者適応が必要になると考えられる。

課題 3 に対して，本研究では，自然調音状態と不自然調音状態それぞれの状態の声道断面積関数が得られている。従って，自然調音状態と不自然調音状態の声道断面積関数を比較し音響理論的分析を行うことにより，大きく異なる調音状態が同じ音素カテゴリーに含まれる音響特徴量を生成可能な要因の検討が可能となる。また，自然調音状態と不自然調音状態それぞれを生成するための舌及び下顎に関する筋の情報も得られており，それぞれの筋の種類や収縮力の傾向を比較することにより，自然調音状態と不自然調音状態の運動指令の差に関する検討も可能となる。これらの自然調音状態と不自然調音状態の詳細な比較に基づき人間が発話を行う際の調音運動の目標に関する検討を行うことで，音声生成機構の解明へのさらなる貢献が期待できる。

謝辞

本研究を行なうに当たり，終始御指導を賜わった党建武 教授に深謝致します。

本論文を執筆するにあたり，草稿の段階から貴重な御助言及び御指導を賜りました北陸先端科学技術大学院大学 情報科学研究科 赤木正人 教授，徳田功 准教授，小谷一孔 准教授，甲南大学 知能情報学部 知能情報学科 北村達也 准教授に心より感謝致します。

また，日頃から有益な御助言をいただき，多面に渡って励ましていただいた北陸先端科学技術大学院大学 情報科学研究科 鵜木祐史 准教授，末光厚夫 助教，川本真一 助教，宮内良太 助教に感謝致します。

最後に，本研究を進めるにあたり，日頃から御協力いただいた北陸先端科学技術大学院大学 情報科学研究科 生体情報処理分野及び，音情報処理分野の皆様，また諸先輩方に厚く御礼申し上げます。

参考文献

[1] 鏑木 時彦, “音声生成の計算モデルと可視化,” コロナ社, 東京, pp.192–194 (2010).

[2] 川人 光男, “脳の計算理論,” 産業図書, 東京 (1996).

[3] J. H. Ryalls and J. Ryalls, “A basic introdution to speech perception,” Singular Pub Group, Sun Diego (1996).

[4] 児島 宏明, 大村 浩, 奥村 真知, 小張 敬之, “調音状態の評価と音響分析に基づく英語発音矯正システムの開発,” 音講論, 3-10-8, pp.475–478 (2008).

[5] M. R. Schroeder, “Determination of the geometry of the human vocal tract by acoustic measurements,” *J. Acoust. Soc. Am.*, vol.41, pp.1002–1010 (1967).

[6] S. Atal, J. Chang, J. Mathews and W. Tukey, “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique,” *J. Acoust. Soc. Am.*, vol.63, pp.1535–1555 (1978).

[7] J. Schroeter and M. M. Sondhi, “Techniques for estimating vocal-tract shapes from the speech signal,” *IEEE Trans. Speech Audio Process.*, vol.2, pp.135–150 (1994).

[8] 鈴木 紳, 岡留 剛, 誉田 雅彰, “音響調音対コードブックを用いた音声からの調音運動の逆推定,” 信学論 A, J85-A, pp.840–846 (2002).

[9] S. Hiroya and M. Honda, “Estimation of articulatory movements from speech acoustics using an HMM-based speech production model,” *IEEE Trans. Speech Audio Process.*, vol.12, pp.175–185 (2004).

[10] 白井 克彦, 誉田 雅彰, “音声波からの調音パラメータの推定,” 信学論 A, J61-A, pp.409–416 (1978).

[11] J. Dang and K. Honda, “Estimation of vocal tract shapes from speech sounds with a physiological articulatory model,” *J. Phonet.*, vol.30, pp.511–532 (2002).

[12] 伊福部 達, “九官鳥, インコ, そして超腹話術,” 音響学会誌, vol.56, pp.6570–662 (2000).

[13] B. Lindblom, J. Lubker and T. Gay, “Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation,” *J. Phonet.*, vol.7, pp.147–161 (1979).

[14] J. Dang and K. Honda, “Speech production of vowel sequences using a physiological articulatory model,” *Proc. ICSLP*, vol.5, pp.1767–1770 (1998).

[15] C. -H. Jo, T. Kawahara, S. Doshita and M. Dantsuji, “Automatic pronunciatoin error detection and guidance for foreign language learning,” *Proc. ICSLP*, pp.2639–2642 (1998).

[16] J. Dang and K. Honda, “Construction and control of a physiological articulatory model,” *J. Acoust. Soc. Am.*, vol.115, pp.853–870 (2004).

[17] J. Dang and K. Honda, “Investigation of the acoustic characteristics of the velum for vowels,” *Proc. ICSLP*, pp.603–606 (1994).

[18] T. Okadome and M. Honda, “Generation of articulatory movements by using a kinematic triphone model,” *J. Acoust. Soc. Am.*, vol.110, pp.453–463 (2001).

[19] 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, 山本 幹雄, “音声認識システム,” オーム社, 東京, pp.13–15 (2001).

[20] V. Sanguineti, R. Laboissière and D. J. Ostry, “A dynamic biomechanical model for neural control of speech production,” *J. Acoust. Soc. Am.*, vol.103, pp.1615–1627 (1998).

[21] T. W. Anderson, “An introduction to multivariate statistical analysis third edition,” Wiley, New York, pp.91–101 (2003).

[22] T. Nakagawa, S. Saito and T. Yoshino, “Tonal difference limens for second formant frequencies of synthesized Japanese vowels,” *Ann. Bull. RILP*, vol.16, pp.81–88 (1982).

[23] R. D. Kent and C. Read, “The acoustic analysis of speech 2nd ed.,” Singular, New York (2002).

[24] H. Hotelling, “Analysis of complex statistical variables into principal components, ” *J. Educ. Psychol.*, vol.24, pp.417–441 (1933).

[25] P. Mokhtari, T. Kitamura, H. Takemoto and K. Honda, “Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients,” *J. Phonet.*, vol.35, pp.20–39 (2007).

[26] B. Schölkopf, A. Smola and K. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Comput.*, vol.10, pp.1299–1319 (1998).

[27] A. Lima, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda and T. Kitamura, “On the use of kernel PCA for feature extraction in speech recognition,” *IEICE Trans. Inf. & Syst.*, vol.E87-D, pp.2802–2811 (2004).

[28] 赤穂 昭太郎, “カーネル多変量解析 –非線形データ解析の新しい展開–,” 岩波書店, 東京, pp.124–127 (2008).

[29] 石井 健一郎, 上田 修功, 前田 英作, 村瀬 洋, “わかりやすいパターン認識,” オーム社, 東京, pp.76–97 (1998).

[30] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Ann. Eugenics.*, vol.7, pp.179–188 (1936).

[31] 水野 欽司, “多変量データ解析講義,” 朝倉書店, 東京, pp.92–93 (1996).

[32] 末永 高志, 佐藤 新, 坂野 鋭, “クラスタ構造に着目した特徴空間の可視化–クラスタ判別法–,” 信学論 D-II, J85-D-II, pp.785–795 (2002).

[33] A. Y. Ng, M. I. Jordan and Y. Weiss, “On spectral clustering: analysis and an algorithm,” *NIPS*, vol.14, pp.849–856 (2002).

[34] R. O. Duda, P.E. Hart and D. G. Stork, “Pattern Cassification 2nd ed.,” Wiley, New York (2001).

[35] D. Cai, X. He, Z. Li, W. Ma and J. Wen, “Hierarchical clustering of WWW image search results using visual, textual and link information,” *Proc. 12th ACM Int. Conf. Multimedia*, pp.952–959 (2004).

[36] X. Lu and J. Dang, “Vowel production manifold: intrinsic factor analysis of vowel articulation,” *IEEE Trans. Audio Speech Language Process.*, pp.37–40 (2008).

[37] Th. Villmann, B. Hammer, F. -M. Schleif, T. Geweniger, T. Fischer and M. Cottrell, “Prototype based classification using information theoretic learning,” LNCS: Neural Information Processing, 4233, Springer, Belin, pp.40–49 (2006).

[38] E. Parzen, “On estimation of a probability density function and mode,” *Ann. Math. Statist.*, vol.33, pp.1065–1076 (1962).

[39] 粕谷 英樹, 鈴木 久喜, 城戸 健一, “年齢, 性別による日本語 5 母音のピッチ周波数とホルマント周波数の変化,” 音響学会誌, vol.24, pp.355–364 (1968).

[40] 佐藤 大和, “男女声の声質情報を決める要素,” 研究実用化報告 (NTT), vol.24, pp.977–993 (1975).

[41] 戸川 隼人, “数値計算,” 岩波書店, 東京, pp.20–24 (1991).

[42] C. M. Bishop, “パターン認識と機械学習　下,” シュプリンガー・ジャパン, 東京, pp.307–308 (2008).

[43] P. Perrier, J. Perkell, Y. Payan, M. Zandipour, F. Guenther and A. Khalighi, “Degrees of freedom of tongue movements in speech may be constrained by biomechanics,” *Proc. ICSLP*, vol.2, pp.162–165 (2000).

[44] J. Wei, X. Lu and J. Dang, "A model-based learning process for modeling coarticulation of human speech," *IEICE Trans. Inf. & Syst.*, vol.E90-D, pp.1582–1591 (2007).

本研究に関する発表

学術論文

1. 錦戸 信和, 党 建武, “発話機構モデルに基づく音声と調音状態との一対多の関係に関する考察,” 音響学会論文誌, vol.67, pp.1–12 (2011).

2. Q. Fang, A. Nishikido and J. Dang, “Feedforward control of a 3-D physiological articulatory model for vowel production,” *Tsinghua Science & Technology*, vol.14, pp.617–622 (2009).

学会発表（査読付き）

1. A. Nishikido, S. Kawamoto and J. Dang, “Discrimination between natural and unnatural articulations based on articulatory structure,” *Proc. The 7th ISCSLP*, pp.50–54 (2010).

2. Q. Fang, A. Nishikido, J. Dang and A. Li, “Feedforward control of a 3D physiological articulatory model for vowel production,” *Proc. Interspeech2009*, pp.52–55 (2009).

3. Q. Fang, A. Nishikido, J. Dang and T. B. Ho, “Feedforward control of a 3D physiological articulatory model for the investigation of speech production,” *Proc. International Workshop on Nonlinear Circuit and Signal Processing*, pp.169-172 (2009).

4. A. Nishikido and J. Dang, “Analysis of normal and infrequent articulation based on comparison of simulation and observation,” *Proc. International Seminars on Speech Production*, pp.201–208 (2006).

学会発表

1. 錦戸 信和, 党 建武, “音声と一対多の関係にある調音状態の分布構造：発話機構モデルに基づく考察,” 電子情報通信学会技術研究報告, SP2009-157, pp.51–56 (2010).

2. 錦戸 信和, 党 建武, “音声に対する多意性を考慮した自然発話状態の判別,” 音講論, pp.367–370 (2009.9).

3. Q. Fang, A. Nishikido and J. Dang, “Feedforward control of a 3D physiological articulatory model for the investigation of speech production,” *Proc. International Symposium on Biomechanical and Physiological Modeling and Speech Science*, pp.72-77 (2009).

4. 錦戸 信和, 党 建武, “自然発話状態と不自然発話状態との分離に適した調音特徴量の検討,” 音講論, pp.449–540 (2009.3).

5. 錦戸 信和, 党 建武, “母音発話状態の特異調音についての考察,” 音講論, pp.371–372 (2008.9).

6. A. Nishikido and J. Dang, “Investigation of usual and unusual articulation based on simulations and observations,” *Asian Workshop on Speech and Technology*, *IEICE teqhnical report*, pp.23–28 (2008).

7. 方 強, 錦戸 信和, 藤田 覚, 盧 緒剛, 党 建武, “モデル制御に対する 3D 舌形状の解析,” 音講論, pp.327–328 (2008.3).

8. 錦戸 信和, 党 建武, “GMM を用いた通常発話状態と特異発話状態の弁別,” 音講論, pp.443–444 (2007.9).

9. 錦戸 信和, 党 建武, “通常発話状態と特異発話状態との判別規準の検討,” 電子情報通信学会技術研究報告, SP2007-24, pp.1–6 (2007).

10. 金野 武司, 錦戸 信和, 党 建武, “乳幼児の音声模倣能力の獲得過程における調音ジェスチャーの役割,” 電子情報通信学会技術研究報告, SP2007-31, pp.43–48 (2007).

11. 錦戸 信和, 党 建武, “調音モデルを用いた特異発話状態の調査,” 音講論, pp.319–320 (2007.3).

12. 錦戸 信和, 党 建武, “日本語5母音の調音・音響的観測とモデルシミュレーションとの比較,” 電子情報通信学会技術研究報告, SP2006-21, pp.5–10 (2006).

13. 錦戸 信和, 党 建武, “モデルを用いた模擬に基づく発話状態の多意性の分析,” 音講論, pp.261–262 (2006.3).

14. 錦戸 信和, 党 建武, “シミュレーションによる日本語5母音の音響特性と発話状態の関連性についての検討,” 音講論, pp.311–312 (2005.9).