JAIST Repository
https://dspace.jaist.ac.jp/

Title	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
Author(s)	錦戸，信和；党，建武
Citation	日本音響学会誌，67(1)：3-14
Issue Date	2011
Type	Journal Article
Text version	publisher
URL	http://hdl.handle.net/10119/9601
Rights	Copyright (C)2011 日本音響学会，錦戸信和，党建武，日本音響学会誌，67(1)，2010，3-14.
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

論　　文 43.70.−h,Jt

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*

錦戸信和*1　党　建武*1,*2

[要旨] 音声から調音状態への逆問題における一対多の関係は古くから知られている。しかし，同一の範疇に含まれる音声に対する人間が調音可能なすべての状態を得ることは難しいため，その実体についてまだ十分に調査されていない。本論文は，音声と一対多の関係にある調音状態の全体像を明らかにすることを目的として，発話機構モデルを用いて日本語 5 母音を生成可能な調音状態を作成した。更に，作成した日本語 5 母音の調音状態を，自然調音状態（連続音声に含まれる定常部の音響特性に基づく規準を満たし，かつ自然な発話を行う際に観測され得る状態）と，不自然調音状態（音響特性の規準を満たすが，自然な発話を行う際に観測され得ない状態）に分類し，分類した調音状態を非線形空間に射影した。射影した調音状態を次元圧縮することにより，音声と一対多の関係にある調音状態の分布構造を可視化した。また，分布構造に基づき，異なる調音状態間の位置関係を定量化し，母音ごとの不自然調音状態の傾向を明らかにした。分布構造から得られた知見は，音声から調音状態を逆推定する際の新たな制約条件への利用が期待できる。

キーワード　一対多の関係，生理学的発話機構モデル，母音生成，不自然調音状態，逆推定
One-to-many relationship, Physiological articulatory model, Vowel production, Unnatural articulation, Inverse estimation

1. はじめに

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，逆に同一の範疇に含まれる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 [1]。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 [2]。

音声信号と調音状態との一対多の関係は，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 [3]。また，Lindblom らは，バイトブロックにより下顎が不自然な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 [4]。このように同一の範疇に含まれる音声と一対多の関係にある調音状態には，2 種類の調音状態が含まれると考えられる。一つは，連続音声に含まれる定常部の音響特性に基づく規準を満たし，かつ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，音響特性の規準を満たし，生理学的に発話可能な状態だ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。なお，本論文では母音を対象とし，母音定常部の音響特性を満たす調音状態は，観測された音声信号の音響特徴量と観測された発話器官の位置に基づき定める。

また，音声信号と調音状態との一対多の関係は，音声信号から調音状態を逆推定する場合に大きな問題となる。音声信号から調音状態を逆推定する場合，入力音声に対して多数の調音状態が推定候補となるため，一対多の関係は推定精度を劣化させる。このため，調音状態の逆推定に関する研究では，制約条件を導入し一対多の関係性を抑えることに焦点が当てられている。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 [2]。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 [5]。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，更に調音運動の軌跡を最適化することによ

---

* Model-based investigation on one-to-many relationship between speech sound and articulation, by Akikazu Nishikido and Jianwu Dang.

*1 北陸先端科学技術大学院大学情報科学研究科
*2 天津大学計算機科学技術学院
(問合先：錦戸信和　e-mail: a-nishi@jaist.ac.jp)

(2010 年 1 月 6 日受付，2010 年 8 月 4 日採録決定)

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，構築したコードブックを用いて調音運動の逆推定を行った [6]。観測されたデータに基づきコードブックを構築することにより，導入された形態学的制約及び動的制約には実際の調音形状や調音運動が反映されている。一方，白井と誉田は，幾何学的調音モデルの調音パラメータを直接推定することで，調音運動の逆推定を行っている [7]。形態学的制約が，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で考慮されている。更に，動的制約として，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が含まれている。また，Dang と Honda は部分 3 次元生理学的発話機構モデル [8] を構築し，構築したモデルの調音パラメータの逆推定を行った [9]。その際，生理学的発話機構モデルを用いることにより空間的，動的及び生理学的制約が有機的に結合され取り入れられている。

このように音声と調音状態との一対多の関係が問題となる場合，従来の研究では一対多の関係性を抑えることに焦点があてられ推定結果の正誤にしか着目されておらず，音声に対して一対多の関係にある調音状態空間に対する詳細な分析は行われていない。これは，同一の範疇に含まれる音声に対する人間が調音可能なすべての状態を観測することが困難なためと考えられる。しかし，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，調音状態の全体像，すなわち調音状態の分布構造を把握することが必要となる。前述の不自然調音状態は人間が調音可能であり，かつ滑らかな連続音声も生成可能なことから，調音状態の逆推定における従来の空間的（形態学的），動的及び生理学的制約条件を満たす。また，これまで不自然調音状態に関する詳細な分析は行われていないため，調音モデルにより形成された状態が自然調音状態か不自然調音状態かを自動的に判断することはできない。そのため，自然調音状態により生成された音声を入力として調音モデルを用いて調音状態の逆推定を行う場合，推定候補に含まれる不自然調音状態は取り除くことができず，推定精度を劣化させる。もし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の分布の重なり具合や不自然調音状態の傾向を把握することができれば，その知見は自然調音状態と不自然調音状態との識別関数，すなわち調音状態の逆推定における新たな制約条件に利用できると考えられる。更に，調音状態の分布構造から得られる知見は，人間の音声生成機構の解明にも寄与すると考えられる。

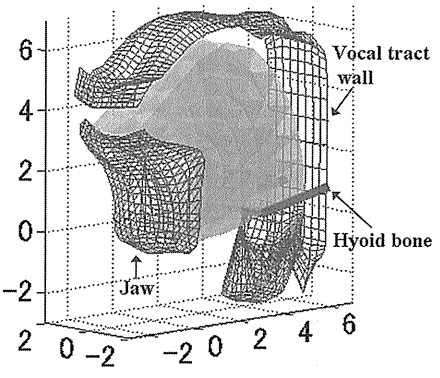
よって，本論文は音声に対して一対多の関係にある調音状態の分布構造を明らかにすることを目的とし，2 章以降は次のように構成される。まず，2 章では生理学的発話機構モデルを用いた日本語 5 母音の調音状態の作成方法とその結果を示す。3 章では，作成した 5 母音の調音状態の分析方法を述べ，可視化された音声と一対多の関係にある調音状態の分布構造を示す。4 章で，分布構造に基づく自然調音状態と不自然調音状態の識別，及び不自然調音状態の傾向について考察し，最後に 5 章で結論を述べる。

2. 日本語 5 母音を生成可能な調音状態の作成

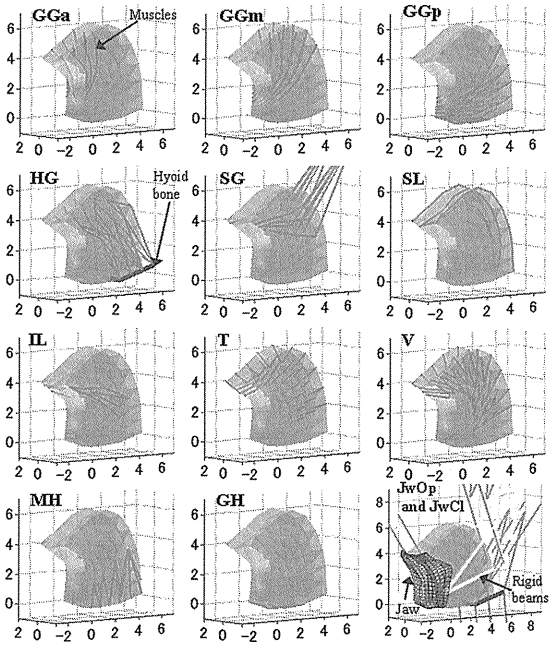
日本語 5 母音に対して，人間が取り得るすべての調音可能な状態を観測することは困難である。従って，まず部分 3 次元生理学的発話機構モデルを用いて調音状態を系統的に生成し，各状態に基づき音声を合成する。更に，観測された音声信号に基づき日本語 5 母音の範疇に含まれる合成音声を抽出することにより，抽出された合成音声に対応する調音状態を選定する。選定された調音状態は，人間が日本語 5 母音の音声を生成可能な調音状態と言える。なお，合成した音声と収録音声を比較することで，より実際の状況を反映した 5 母音の調音状態の選定が可能となる。ただし，合成音声には音源特性の影響が含まれるため，音質が収録音声に等しくなるように合成音声の音源信号を調整し，音響特徴量を求める際に音源特性のパワー成分の影響を低減させる。

2.1 生理学的発話機構モデル

本研究では，日本人成人男性 1 名の MRI 画像に基づき構築された部分 3 次元生理学的発話機構モデル [8] を用いる。このモデルは，舌，下顎，舌骨及び声道壁により構成されており（図–1），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）。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 2 cm 幅のみとなっている。



図–1 部分 3 次元生理学的発話機構モデル



図–2 舌及び下顎の筋構造

舌の筋構造は，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（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。

このモデルの精度に関して，Dang と Honda は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を比較した [8]。その結果，下顎に関してわずかな誤差があるが，舌尖と舌背の最大速度はほぼ一致することが示された。従って，このモデルは人間が発話する際の状態を精度良く再現可能であると考えられる。

2.2 日本語 5 母音の発話器官の観測信号と音声信号

生成される調音状態の日本語 5 母音の選定，及び自然調音状態と不自然調音状態の分類には，母音区間の観測信号と音声信号を用いる。よって，発話器官の位置の観測信号と音声信号，及びそれぞれの母音区間について述べる。

発話器官の位置の観測信号は，モデルの目標話者を被験者とする X 線マイクロビームシステムによるペレット位置の観測信号 [10] とし，本論文では，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 箇所の観測信号を用いる。具体的な各ペレット位置は，LJ の場合は，下側歯列の歯と歯茎の境とする。また，T1～T4 の場合は，舌の先端から 1 cm 程度後方を T1，装着可能な最後方を T4 とし，T1 と T4 の間を等間隔に分ける 2 箇所を T2 及び T3 とする。なお，観測信号のサンプリング周波数は 146 Hz 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。

収録された音声信号は，EMU-4545 マイクロホンを用いて，X 線マイクロビームシステム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 [10] とし，サンプリング周波数は 16 kHz とする。

ペレット位置の観測信号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（75.3 ms）とする。母音の中心位置は，Okadome と Honda の基準 [11] 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。

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（34 ms）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）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いた。この結果，5 母音合わせて 5,892 個の観測信号と，2,946 フレームの音声信号を得た。

2.3 調音状態の系統的生成及び音声合成

調音状態は，舌と下顎の筋に収縮力を 400 ms 間加え生理学的発話機構モデルを駆動することにより生成する。なお，調音状態を生成する際に，各調音状態に基づく合成音声の音韻性は考慮せず，舌と下顎の筋収縮の組み合わせのみを考慮する。

舌に対して，2 又は 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，

表–1 舌筋の組み合わせ

主動筋と共同筋	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
主動筋と拮抗筋及び共同筋	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

GGa 及び GGm，GGp はそれぞれ一つの筋として扱う。また，全方位への移動を可能にするため SL と T を一つの筋として扱う。組み合わせの基準は Dang と Honda の検討 [8] に基づき，次のとおりとする。まず，舌尖又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又は SL+T を主動筋として，主動筋とその共同筋，又は主動筋とその拮抗筋及び共同筋を組み合わせる。具体的な 28 種類の組み合わせを表–1 に示す。

筋に与える収縮力は，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–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を選択する。これらの選択した舌筋の組み合わせに含まれる 2 又は 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。このとき，他の筋及び筋群に収縮力は与えない。

本論文で用いる生理学的発話機構モデルは，人間の調音における形状的及び生理学的要素が考慮されている。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，最も小さな値として SL の場合の 14.3N が示されている [12]。この値は，収縮力の範囲の最大値 6N の 2 倍を超える値である。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。なお，6N より大きい収縮力を用いた場合，舌の変形がほとんど見られなくなることから，収縮力の最大値を 6N としている。

一方，音声の合成は，上記の舌筋 28 組と下顎の 2 種類の筋群に対するすべての組合せの結果得られる舌と下顎の調音運動に基づき行う。更に，音声を合成する際には，調音状態に含まれていない口唇と喉頭を考慮する。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口唇のペレット位置の観測信号に基づき定められ，通常状態の場合，長さは 0.02 cm 間隔で 0.81 cm～1.09 cm，直径は 0.04 cm 間隔で 1.05 cm～1.69 cm 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 1.10 cm～1.38 cm，直径は 0.45 cm～1.09 cm 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，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，400 ms 間中の安定した 100 ms 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 [9] を適用することにより 2 種類の声道断面積関数を得る。更に，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 [9] 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 0.3 cm^2 とする。また，基本周波数は収録音声に基づき 120 Hz 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。

上記の調音状態の生成及び音声合成の結果，64,587 組の調音状態と合成音声の対を得た。なお，本論文で用いた生理学的発話機構モデルは部分 3 次元モデルであるため，用いる情報はペレット位置の観測信号と正確な比較が行える正中矢状断面の情報のみとする。よって，生成された調音状態は，正中矢状断面の舌表面上

の 17 点と下顎 1 点の位置，計 36 次元をパラメータとする特徴量ベクトルとして扱う。また，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舌前部のくぼみの影響は含まれている。更に，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 14.4 cm～18.4 cm の範囲で変化することが確認できている。

2.4 音響分析に基づく調音状態の選定

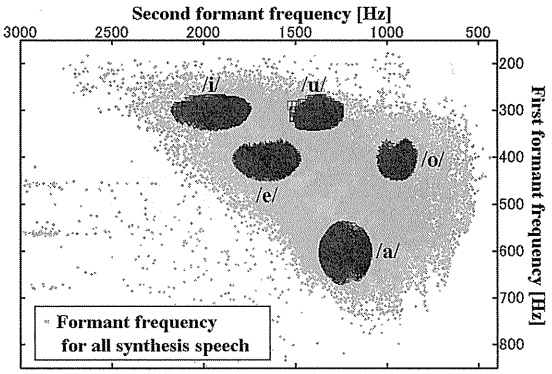
発話機構モデルを用いて生成された調音状態に基づき合成された音声には，子音や音声に聞こえない無意味な音も含まれている。そのため，生成されたすべての調音状態の中から，日本語 5 母音の範疇に含まれる音声を生成可能な調音状態を選定する必要がある。従って，2.2 節で述べた母音区間の音声信号の音響特徴量から求めた規準範囲に基づき合成音声を抽出し，抽出された合成音声に対応する調音状態を得る。

音響特徴量は，母音区間の音声信号から求めた 12 次元の MFCC と第 1 及び第 2 ホルマント周波数を用いる。各特徴量を求める条件は，サンプリング周波数 16 kHz，窓関数は時間長 30 ms のハミング窓を用い，シフト長は 10 ms 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 4 kHz 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。

MFCC 空間の規準範囲は，母音ごとの MFCC から求めた信頼度 0.68 の信頼楕円 [13] とする。この信頼楕円は分布の標準偏差の範囲に相当する。また，ホルマント周波数空間の規準範囲は，第 1 及び第 2 ホルマント周波数それぞれの各母音の平均 ±10%（ホルマント周波数の弁別閾値に相当 [14]）を軸とする楕円とする。

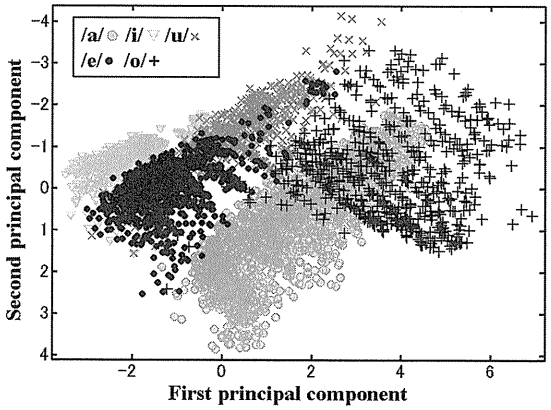
ただし，ホルマント周波数は音韻性と密接に関連する特徴量だが，周波数の一部のみしか考慮されず，また精度良く推定することは難しい。一方，MFCC はスペクトルの形状全体が考慮され，求められる特徴量の精度は高いが，音韻性との直接の関連性は明確ではない。従って，ホルマント周波数と MFCC の両空間の規準範囲に特徴量が含まれる合成音声を抽出する。この結果，2 種類の特徴量それぞれの短所を補い合成音声を抽出することができると考えられる。

MFCC 空間の規準範囲に含まれる特徴量を持つ合成音声に対して更にホルマント周波数空間の規準範囲を適用し，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 8,229 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–3 に示す。図–3 より，MFCC とホルマント周波数両方を用いて抽出された 5 母音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範疇に含まれる音声と考えることができる。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は，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。



図–3 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

更に，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，分布を直接把握することは難しいため，調音状態の主成分分析（Principal component analysis：PCA）[15] を行った。その結果，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し，第 2 主成分までの累積寄与率は 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–4 に示す。図–4 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，舌全体又は舌背の変位が主要な主成分となる。しかし，本論文の結果では舌尖の垂直方向の変位が第 2 主成分となっており，通常と異なる結果となっている。この原因として，不自然調音状態



図–4 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

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。

3. 音声と一対多の関係にある調音状態の分析

音声から調音状態を逆推定する際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，音声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の重なり及び不自然調音状態の傾向を把握する必要がある。従って，まず 2 章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。更に，異なる調音状態の分布間の重なりが減少する非線形空間に調音状態を射影し，非線形特徴量間の類似性の構造を保ったまま次元圧縮する。この分析により分布構造を可視化することで，分布構造の把握が容易になる。また分布構造に基づき，自然調音状態と不自然調音状態の位置関係を定量化し，更に不自然調音状態の傾向を示す。

3.1 自然調音状態と不自然調音状態の分類

音声に対して一対多の関係にある調音状態には自然調音状態と不自然調音状態が含まれるため，選定された 5 母音の調音状態にも両方の調音状態が含まれると考えられる。従って，調音状態の分布構造を明らかにするためには，選定された調音状態を自然調音状態と不自然調音状態とに分類する必要がある。よって，まず分類規準を 2.2 節で述べた母音区間の下顎及び舌上 4 点のペレット位置（LJ 及び T1～T4）の観測信号に基づき定める。LJ 及び T1～T4 と実際に分類に用いる調音状態の固定点との対応関係は次のとおりとする。LJ に対応する固定点は，生理学的発話機構モデルの LJ と同じ箇所とする。T1～T4 に対応する固定点は，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。

表–2 5 母音の自然調音状態と不自然調音状態の数

母音	/a/	/i/	/u/	/e/	/o/
自然調音状態	160	94	64	1188	74
不自然調音状態	2447	1235	888	1461	618

分類規準は，5 箇所のペレット位置ごとに求めた信頼度 0.997 の信頼楕円（標準偏差の 3 倍の範囲に相当）とする。5 箇所のペレット位置がすべて信頼楕円に含まれる観測信号は 96%となる。この規準を用いて，5 箇所の固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。

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–2 に示す。表–2 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 1,580 と全体の約 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。

3.2 非線形空間における調音状態の分析

調音状態の分布構造を目視により捉えるため，高次元空間上の分布を次元圧縮し，調音状態の分布構造を可視化する。ただし，自然調音状態と不自然調音状態の分布の重なりが大きい場合，分布構造の把握は難しい。従って，まず異なる調音状態間の分布の重なりが減少する非線形空間に調音状態を射影し，非線形特徴量を次元圧縮することにより分布構造を可視化する。なお，類似している特徴量は密集し，異なる特徴量は離れて分布することにより分布構造が明確になることから，次元圧縮には，特徴量間の類似性の構造を考慮し次元圧縮を行うクラスタ判別法 [16] を用いる。

3.2.1 調音状態の非線形空間への射影

分布構造を目視により把握するためには，自然調音状態と不自然調音状態の分布の重なりを減少させる必要がある。先行研究 [17] では，母音の調音特徴の類似性を強調するカーネル関数を用いたカーネル主成分分析（Kernel Principal Component Analysis：KPCA）[18] により，36 次元の調音状態が非線形空間に射影された。射影空間では，自然調音状態と不自然調音状態との推定におけるベイズ誤り確率の上限が，元の調音状態空間と比べて減少した。ベイズ誤り確率は分布の重なりの度合いと解釈でき，値が小さいほど分布の重なりも

小さくなる。従って，異なる調音状態の分布の重なりを減少させるため，KPCA を用いて調音状態を非線形空間に射影する。KPCA は，射影空間上の特徴量間の内積を表すカーネル関数を用いて，元の次元数よりも遥かに高い次元の非線形空間上に射影された特徴量の PCA を行う。非線形空間上の PCA により得られる射影ベクトルを用いることで，調音状態を非線形空間に射影した非線形特徴量が得られる。KPCA に用いる母音の調音特徴を強調したカーネル関数を次式に示す。

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \{\exp(D_{\mathrm{TI}}) + \exp(D_{\mathrm{TD}}) + \exp(D_{\mathrm{OT}})\} \times \exp(D_{\mathrm{J}}) \quad (1)$$

$$D_{\mathrm{TI}} = -\|\boldsymbol{x}_{\mathrm{TI}\,i} - \boldsymbol{x}_{\mathrm{TI}\,j}\|^2 / 2\sigma_{\mathrm{TI}}^2 \quad (2)$$

$$D_{\mathrm{TD}} = -\|\boldsymbol{x}_{\mathrm{TD}\,i} - \boldsymbol{x}_{\mathrm{TD}\,j}\|^2 / 2\sigma_{\mathrm{TD}}^2 \quad (3)$$

$$D_{\mathrm{OT}} = -\|\boldsymbol{x}_{\mathrm{OT}\,i} - \boldsymbol{x}_{\mathrm{OT}\,j}\|^2 / 2\sigma_{\mathrm{OT}}^2 \quad (4)$$

$$D_{\mathrm{J}} = -\|\boldsymbol{x}_{\mathrm{J}\,i} - \boldsymbol{x}_{\mathrm{J}\,j}\|^2 / 2\sigma_{\mathrm{J}}^2 \quad (5)$$

ここで，$\boldsymbol{x}_i$ $(i = 1, \cdots, M)$ は調音状態の各データ，M はデータ数を表す。$\boldsymbol{x}_{\mathrm{TI}}$，$\boldsymbol{x}_{\mathrm{TD}}$，$\boldsymbol{x}_{\mathrm{OT}}$，$\boldsymbol{x}_{\mathrm{J}}$ は各データのパラメータの舌尖要素，舌背要素，舌尖と舌背を除く舌要素及び下顎要素を表す。また，σ_{TI}，σ_{TD}，σ_{OT}，σ_{J} はカーネル関数のパラメータを表し，それぞれの値は 0.9，1，1.2，0.4 とする [17]。

3.2.2 類似性の構造を考慮した非線形特徴量の次元圧縮

調音状態の分布構造の把握を容易にするため，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により次元圧縮する。クラスタ判別法による次元圧縮の手順は，まず非線形特徴量をクラスタリングし，クラスタリングされた特徴量の線形判別分析を行う。クラスタリングにはスペクトラルクラスタリング [19] を用い，線形判別分析には重判別分析法 [20] を用いる。なお，予備検討により自然調音状態は一つのクラスタとなり，不自然調音状態は複数のクラスタとなったため，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用いる。

スペクトラルクラスタリングは，分布の各データをグラフ構造のノードとして捉え，分布から求めたノード間の重み行列に対してグラフラプラシアンの固有値分解を行う。固有値分解により得られる固有ベクトルの成分はクラスタごとに異なる傾向の値をとるため，固有ベクトルの成分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。

一方，重判別分析法は，Fisher の線形判別分析法を多クラスタに拡張して，クラスタ間散布行列とクラスタ内散布行列の逆行列との積の固有値分解を行う。この固有値分解により，クラスタ内散布対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，全クラスタに対する分離が最も良い線形射影分布が得られる。

ただし，重判別分析を行う際に，分析対象の特徴量の次元数がクラスタ数以上である必要がある。KPCA により，調音状態の次元数（36 次元）よりも高い次元への射影を可能とする射影ベクトルが得られている。従って，非線形特徴量の次元数を変化させ不自然調音状態のクラスタリングを行い，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。この検討では，最適なクラスタ数も同時に検討する必要がある。スペクトラルクラスタリングにおいて，連続する固有値間の差の絶対値を表す Eigengap が有用な指標の一つとなっている [19] ことから，この指標を用いて最適なクラスタ数を検討する。Eigengap は次式から求められる。

$$g(k) = |\lambda_k - \lambda_{k+1}| \quad (6)$$

ここで，λ_k は固有値を表し，k は固有値の最大値からの降順の順位を表す。また，Eigengap $g(k)$ を k に関する関数とみた場合，関数の極大値が最適なクラスタ数となる結果が報告されている [21]。従って，固有値の最大値から降順に最大値を除いた上位 18 個に対する Eigengap $g(k)$ を求め，$g(k)$ を k に関する関数とみた場合の極大値の k を最適なクラスタ数とする。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調音状態のバラつきが少ない極大値の k を最適なクラスタ数とする。

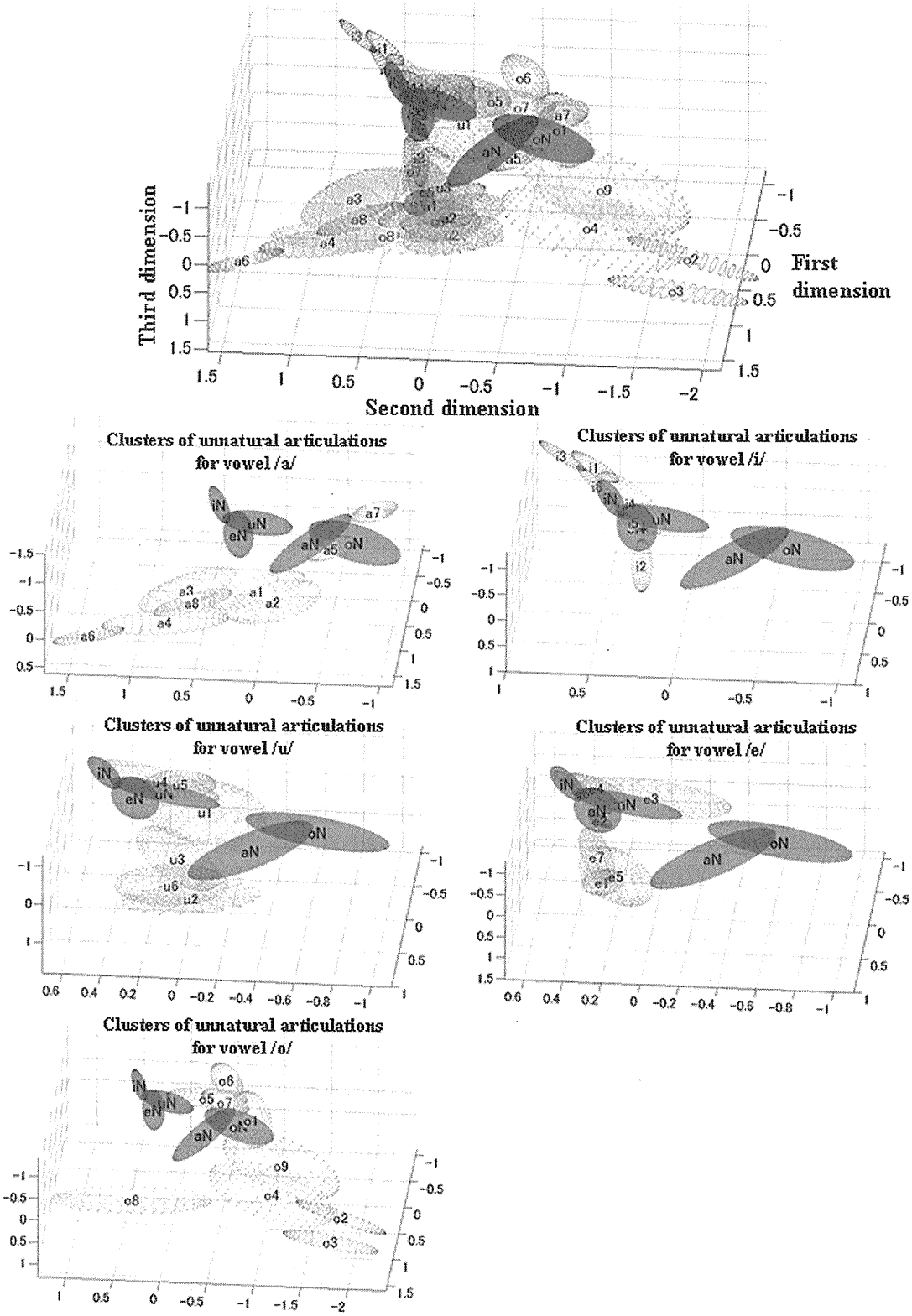
非線形特徴量の最適な次元数及びクラスタ数の検討の結果，最適な次元数は 42，総クラスタ数は 41 となった。Eigengap に基づいて得られた不自然調音状態の最適なクラスタ数を**表–3** に示す。表–3 から/o/を除く母音では，クラスタ数は不自然調音状態の数に比例する傾向を示している。

表–3 5 母音の不自然調音状態の最適なクラスタ数

母音	/a/	/i/	/u/	/e/	/o/
最適なクラスタ数	8	6	6	7	9

3.3 調音状態の分布構造

クラスタ判別法により次元圧縮された，3 次元空間上の非線形特徴量の分布を**図–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，最初の文字が母音を表



図–5　3 次元空間上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。2 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。

している。2番目の文字は調音状態によって異なり，自然調音状態の場合はNとなり，不自然調音状態の場合は，クラスタ番号を表す数字となっている。図–5から特徴量は多様体上に母音/a/，/i/，/o/を各頂点とする3角形に分布している。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，Dangらは，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[22]。その構造も/a/，/i/，/o/を頂点とする3角形を示しており，図–5の結果と一致している。

また，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，クラスタ判別分析により得られる部分空間の最大次元数（40次元）の空間上におけるクラスタ間の距離を求める。クラスタ間の距離として，確率分布間の距離を表す統計量であるCauchy-Schwarz (CS) divergence[23]を用いる。CS divergenceは，確率分布モデルを仮定せずに分布の密度関数を推定することにより，分布のデータから直接求められる。CS divergenceの式を下記に示す。

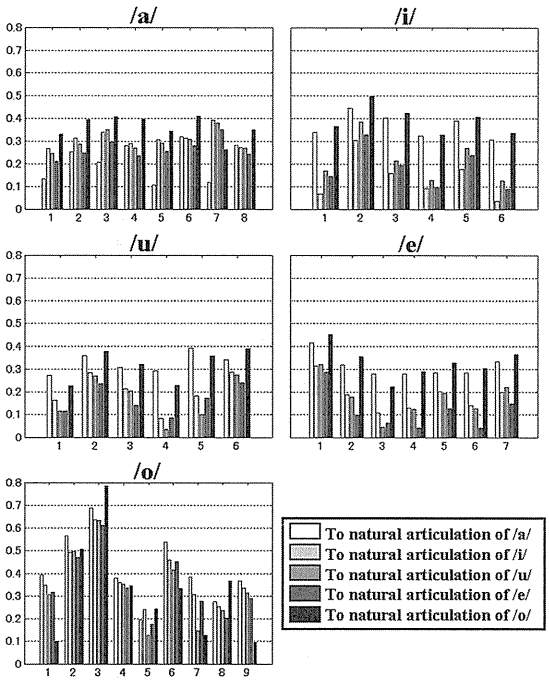
$$D_{\mathrm{CS}}(\boldsymbol{V},\boldsymbol{Y}) = \frac{1}{2}\log\{V_2(\boldsymbol{V})\cdot V_2(\boldsymbol{Y})\} - \log\{C_r(\boldsymbol{V},\boldsymbol{Y})\} \tag{7}$$

$$V_2(\boldsymbol{V}) = \frac{1}{L^2h^4}\sum_{l=1}^{L}\sum_{s=1}^{L} G(\boldsymbol{v}_s - \boldsymbol{v}_l, h) \tag{8}$$

$$C_{\mathrm{r}}(\boldsymbol{V},\boldsymbol{Y}) = \frac{1}{L\cdot S\cdot h^4}\sum_{l=1}^{L}\sum_{s=1}^{S} G(\boldsymbol{v}_s - \boldsymbol{y}_l, h) \tag{9}$$

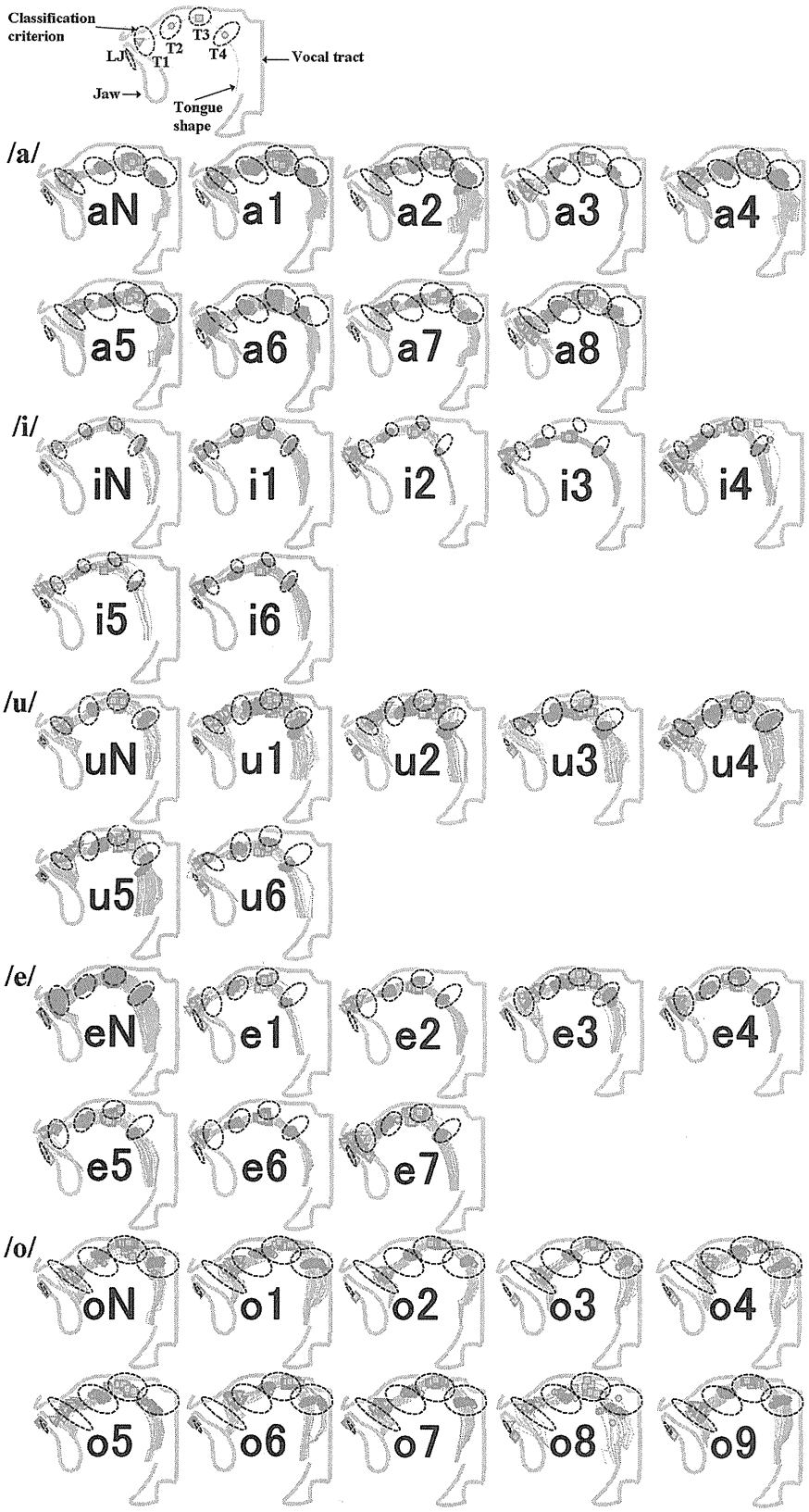
$$G(\boldsymbol{v}-\boldsymbol{y},h) = \frac{1}{(2\pi h^2)^{d/2}}\exp\left(-\|\boldsymbol{v}-\boldsymbol{y}\|^2/2h^2\right) \tag{10}$$

ここで，$\boldsymbol{V}$と$\boldsymbol{Y}$はそれぞれ一つのクラスタを，$\boldsymbol{v}$と$\boldsymbol{y}$はクラスタ$\boldsymbol{V}$と$\boldsymbol{Y}$それぞれに含まれる非線形特徴量を表し，LとSはクラスタに含まれる非線形特徴量の数を表す。また，dは非線形特徴量の次元数を，h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，$h=1$とする。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–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–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。



図–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番号，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）

更に，各クラスタの調音形状を具体的に示すため，図–7に各クラスタに含まれる非線形特徴量の調音状態の正中矢状断面の形状を示す。なお，各調音形状の中心に示されている文字の意味は図–5と同じである。不自然調音状態の調音形状を見ると，/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，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，更に舌全体が下方に位置している。/e/の調音形状も/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。



図–7　クラスタごとの調音形状

4. 考　　察

分布構造に基づき，自然調音状態と不自然調音状態の識別，及び不自然調音状態の傾向について考察する。

4.1 自然調音状態と不自然調音状態の識別

自然調音状態と不自然調音状態を識別する場合，二つの分布間の重なりが小さいほど，高い精度での識別が可能となる。よって，分布構造における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なりを調べる。具体的には，非線形空間における 40 次元の部分空間上の調音状態に対して，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る。その結果，標準偏差の 2 倍の範囲において，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は見られなかった。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，分布構造に基づき自然調音状態と不自然調音状態の識別関数を作成する場合，その識別関数は音声から調音状態の逆推定における新たな制約条件としての利用が期待できる。

なお，分布構造を求めた際の調音状態に口唇は含まれていないが，自然調音状態と不自然調音状態の識別を調音状態の逆推定に適用する場合，口唇も含めて考える必要がある。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，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り，その検討は今後の課題である。

4.2 不自然調音状態の傾向

分類規準に対する不自然調音状態の母音ごとの傾向として，/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，下顎は規準範囲内だが，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，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だが，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，不自然調音状態の傾向から，連続音声に含まれる定常部に対する調音の運動目標が，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。

5. ま　と　め

部分 3 次元生理学的発話機構モデルを用いて生成した調音状態の音響分析により，日本語 5 母音の範疇に含まれる音声と一対多の関係にある調音状態を得た。更に，得られた 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し，非線形空間上に射影した特徴量を次元圧縮することで，日本語 5 母音の範疇に含まれる音声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た。また，分布構造に基づき，自然調音状態と不自然調音状態の位置関係が定量化され，今まで詳細が明らかになっていなかった不自然調音状態の傾向が母音ごとに示された。

今後，調音状態の分布構造から得られた知見を調音状態の逆推定に適用するため，分布構造に基づき自然調音状態と不自然調音状態との識別関数を検討する。更に，識別関数を新たな制約条件として，音声から調音状態を逆推定するシステムを構築する予定である。

謝　　辞

本研究の遂行にあたり，有益な助言をいただいた北陸先端科学技術大学院大学徳田功准教授，末光厚夫助教並びに，本論文に対し有益なコメントをいただいた甲南大学北村達也准教授，株式会社エーアイ藤田覚氏に深く感謝いたします。なお，本研究の一部は，基盤研究（2250150）によりサポートされている。

文　　献

[1] M.R. Schroeder, “Determination of the geometry of the human vocal tract by acoustic measurements,” *J. Acoust. Soc. Am.*, 41, 1002–1010 (1967).
[2] B.S. Atal, J.J. Chang, M.V. Mathews and J.W. Tukey, “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique,” *J. Acoust. Soc. Am.*, 63, 1535–1555 (1978).
[3] 伊福部達, “九官鳥，インコ，そして超腹話術—その声の謎解き—,” 音響学会誌, 56, 657–662 (2000).
[4] 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.*, 7, 147–161 (1979).
[5] J. Schroeter and M.M. Sondhi, “Techniques for estimating vocal-tract shapes from the speech signal,” *IEEE Trans. Speech Audio Process.*, 2, 133–150 (1994).
[6] 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**, 840–846 (2002).
[7] 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**, 409–416 (1978).
[8] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *J. Acoust. Soc. Am.*, 115, 853–870 (2004).
[9] J. Dang and K. Honda, "Estimation of vocal tract shapes from speech sounds with a physiological articulatory model," *J. Phonet.*, 30, 511–532 (2002).
[10] J. Dang and K. Honda, "Investigation of the acoustic characteristics of the velum for vowels," *Proc. ICSLP*, pp. 603–606 (1994).
[11] T. Okadome and M. Honda, "Generation of articulatory movements by using a kinematic triphone model," *J. Acoust. Soc. Am.*, 110, 453–463 (2001).
[12] V. Sanguineti, R. Laboissière and D.J. Ostry, "A dynamic biomechanical model for neural control of speech production," *J. Acoust. Soc. Am.*, 103, 1615–1627 (1998).
[13] T.W. Anderson, *An introduction to multivariate statistical analysis third edition* (Wiley, New York, 2003), pp. 91–101.
[14] T. Nakagawa, S. Saito and T. Yoshino, "Tonal difference limens for second formant frequencies of synthesized Japanese vowels," *Ann. Bull. RILP*, 16, 81–88 (1982).
[15] H. Hotelling, "Analysis of complex statistical variables into principal components," *J. Educ. Psychol.*, 24, 417–441 (1933).
[16] 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**, 785–795 (2002).
[17] 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," 音講論集, pp. 367–370 (2009.9).
[18] B. Schölkopf, A. Smola and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, 10, 1299–1319 (1998).
[19] A.Y. Ng, M.I. Jordan and Y. Weiss, "On spectral clustering: analysis and an algorithm," *NIPS*, 14, 849–856 (2002).
[20] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern classification second edition* (Wiley, New York, 2001), pp. 121–124.
[21] D. Cai, X. He, Z. Li, W.-Y. Ma and J.-R. Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information," *Proc. 12th ACM Int. Conf. Multimedia*, pp.952-959 (2004).
[22] J. Dang, X. Lu, M. Tiede and K. Honda, "Inherent vowel structures in speech production and perception spaces," *Proc. 8th ISSP*, pp. 37–40 (2008).
[23] 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, Berlin/Heidelberg, 2006), pp. 40–49.