博士論文

音響情報・言語情報に基づく調音運動の復元

2016 年 12月 1日

指導教員　峯松 信明 教授

東京大学大学院　工学系研究科
電気系工学専攻

37-147054　内田 秀継

あらまし

音声の生成過程における舌や口唇の運動を調音運動と呼ぶ。調音運動によって音声の音響特性が変化し、それが音韻性として知覚される。つまり、調音運動は、音声に言語的情報を付加する役割を持つ。調音運動に関する情報は、音声生成過程における音韻性の表現として、音声認識や声質変換、発音トレーニングなど様々な分野での応用が期待されている。

調音運動の情報を利用する場合、何らかの方法によって発話中の調音運動を測定する必要がある。しかし、調音運動の大部分は口腔内でおこなわれるため、ビデオカメラなどの簡便な機材では、測定することが難しいという問題がある。調音観測システムと呼ばれる特殊な機材を用いれば、発話中の話者の調音運動を測定することが可能だが、観測器は高価であり測定も大掛かりなものになるため、総じて高コストな方法と言える。したがって、多くのユーザーを想定する音声認識や声質変換、発音トレーニングなどのシステムに、高コストな調音観測システムを組み込むことは難しい。そこで、音声から調音運動の情報を復元（推定）する音声–調音マッピングが検討されている。音声–調音マッピングは調音観測システムに頼らずに調音運動情報を取得する技術として注目を集めている。

音声–調音マッピングは、音声と調音運動の同時測定データ（音声–調音パラレルデータ）を用いて、音響空間と調音運動空間の対応（変換）関係を機械学習によってモデル化する。このとき、特定の話者（モデル話者）の音声–調音パラレルデータを用いて構築した変換モデルは、その話者の専用モデルとなり、モデル話者以外の音声を入力した場合、推定精度は著しく悪化する。つまり、従来の音声–調音マッピングは、入力としてモデル話者の音声（音響情報）が与えられた場合にしか適用できないという制約がある。この適用範囲の狭さが音声–調音マッピングを実用的な応用から遠ざけている一因となっている。

そこで、本研究では、音声–調音マッピングの適用範囲を拡張するために、様々な情報から調音運動を推定することを試みる。まず、モデル話者以外の音声、つまり、任意の話者の音声から調音運動を推定する手法として、話者正規化音声–調音マッピングを提案する。話者正規化音声–調音マッピングでは、声質変換を用いて任意話者の音声をモデル話者の音声に変換することで、入力音声と専用モデルの間での音響的なミスマッチを軽減し、高い精度で調音運動を推定することが可能になる。次に、音響情報以外の情報から調音運動を推定する手法を提案する。本研究では、音声の構造的表象によって抽出される言語的特徴量に注目し、従来の音声–調音マッピングでは困難であった、音響情報が存在しない音声に対する調音運動の推定を試みた。

目次

図目次

表目次

第1章

序論

1.1　本論文の背景

我々は、日常的に他者と音声によるコミュニケーションをおこなっている。このコミュニケーションでは、音声を媒体として、言語的な情報に加えて、発話者の意図や感情など、様々な情報がやり取りされている。音声は、物理現象として見れば単なる空気の振動現象（音波）である。しかし、発話者は様々な情報を空気の振動に埋め込むことができ、聴取者はその空気の振動から様々な情報を抽出することができる。そして、その音声の生成過程・聴取過程には複雑なプロセスが数多く存在するにも関わらず、我々はそれらを意識することなく、自然に音声コミュケーションをおこなうことができる。ここで、普段は意識することのない、発話中の口に意識を向けると、口唇や舌が様々な動きをしていることがわかる。これらの運動は調音運動と呼ばれ、音声の生成過程において非常に重要な役割を担うものとなっている。

音声の生成過程において、声帯は肺からの呼気流に振動を与え、音源波を生成する。この音源波には、言語的情報（音韻性）はなく、我々が通常聞く音声でいうところの声の高さ（韻律）に相当する情報が含まれている。この音源波が声道を通過することで、音韻性を伴った音声となる。このとき、声道はある種のフィルタとして働き、音源波に様々な音色を付加する。その音色が音韻性となって知覚される。声道のフィルタの特性は、声道の形状によって変化するのだが、その形状を変化させているのが、調音運動である。したがって、調音運動は、音声の生成過程において、音声の音韻性を付加する役割を担っている。したがって、調音運動は生成過程における音韻性の表現とも言える。

調音運動が持つ音韻性に関する情報は、音声認識や音声合成、声質変換などの音声工学的分野、発音トレーニングなどの語学学習に関する分野や発声障害などに関する医学的応用などで有益な情報として注目されており、調音運動情報を利用した様々な応用システムが検討されている [1][2][3][4][5]。これらの応用例については、次章で詳しく紹介するが、多くのシステムに共通するのが、音声の音韻性に注目した制御をおこなう際に、音響特性をそのもの操作するのではなく、調音運動を介して操作するという点である。同じ音韻性を有する音声であっても、話者やコンテキスト、感情といった様々な要因によって、音響特性は複雑に変化する。音響特性そのものでは複雑で取り扱いづらい、もしくは理解が難しい変化も、生成過程の調音運動まで立ち返ることで、合理的に、または直感的に扱うことができる。

調音運動を音声認識や発音トレーニングなどに応用する場合、発話中の話者の調音運動の情報を何らかの方法で取得する必要がある。これまでに、磁気センサシステム（Electro-Magnetic Articulography, EMA）[6] や核磁気共鳴画像法（Magnetic Resonance Imaging, MRI）[7] といった特殊な機材（調音観測システム）を用いて調音運動を測定する手法が検討されてきた。しかし、一般に調音観測システムは、機材が高価であり、測定も簡便でなく、非常に高コストな測定となるため、音声認識や発音トレーニングなどの多くのユーザーを想定するシステムに組み込むことは難しい。一方で、音声（音響情報）から調音運動を復元する技術が検討されている。この技術は音声の生成過程の逆過程を辿るため、逆推定と呼ばれている。逆推定を用いることで、調音観測システムに頼らずに、発話者から調音

運動情報を取得することが可能になる。

これまでに検討されている逆推定法は、音響理論に基づく手法とコーパスに基づく手法の二つに分けることができる。音響理論に基づく手法では、調音運動によって定まる声道の形状を音響管で近似し、その管の共鳴特性と、実際に観測された音声の音響特性が一致するように、声道形状を推定する [8]。この手法の利点として、後に述べるコーパスに基づく方法とは異なり、調音運動の事前収録データを必要としないという点が挙げられる。一方で、フォルマント周波数が安定しない子音に対して適用するのが難しく、様々な音素が連続して現れる連続発話の推定には向かないという問題点がある。コーパスに基づく推定法では、あらかじめ調音観測システムで測定された音声と調音運動の同時測定データ（音声–調音パラレルデータ）を用いて、音声と調音運動の変換関係をモデル化し、それに基づき音声から調音運動を推定する [9]。この手法は、音声–調音パラレルデータという形で調音運動の事前収録データを必要とする代わりに、音響理論に基づく手法では難しかった子音に対しても、高い精度で推定できるという利点がある。

近年では、多量で高品質な音声–調音パラレルデータが利用可能になったこと、及び、統計的メディア変換の高度化を背景にコーパスベースの逆推定法は様々な発展を見せている。Hiroya ら（2004）は、音声の生成過程を隠れマルコフモデル（Hidden Markov Model, HMM）を用いてモデル化した HMM 音声生成モデルを提案し、統計的モデルを用いることで音声から調音運動を高い精度で推定できることを示した [10]。Toda ら（2007）は、混合ガウス分布を用いた変換モデルを提案し、音声から調音運動への変換だけではなく、調音運動から音声への変換でも高い変換精度が得られることを示した [11]。さらに、近年では、深層学習を用いた手法が注目を集めている [12]。

コーパスベースの逆推定法は、高い精度で音声から調音運動を推定することが可能だが、限定的な状況にしか適用できない。すなわち、推定対象となる音声の発話者が、モデル構築に用いた音声–調音パラレルデータの話者と一致する場合である。音声の音響特性や調音運動は、話者によって異なる性質を持つため、ある特定の話者の音声–調音パラレルデータから構築された変換モデルは、その話者の音声と調音運動の性質に特化した専用モデルとなる。したがって、変換対象として他話者の音声が与えられた場合、その音声はモデル化された音声と異なる性質を持つため、正常に調音運動が推定できなくなる。この適用範囲の狭さが、コーパスベースの逆推定法を実用的な応用から遠ざけている要因となっている。

1.2　本論文の目的

特定の話者、即ち音声–調音パラレルデータの存在する話者の音声にしか適用できないコーパスベースの逆推定法に対して、その適用範囲の拡張を試みる。まず、音声–調音パラレルデータの存在しない任意の話者の音声を推定対象とする変換法を提案する。これは、任意の話者の音響情報から調音運動を推定する試みである。さらに、音響情報以外の情報に注目し、言語情報から調音運動を推定する方法を検討する。これにより、音響情報が存在しない音声に対する調音運動の推定を試みる。つまり、本研究では、コーパスベースの逆推定の適用範囲を、1) 特定話者の音響情報から任意の話者の音響情報へ、2) 音響情報が

存在する音声から音響情報が存在しない音声へ拡張する。

1.3　本論文の構成

本論文では、第２章で調音運動と逆推定問題について説明する。まず、音声生成過程における調音運動の役割と特性について詳しく述べ、その後、逆推定問題の重要性と従来の逆推定法に述べる。最後に逆推定法の課題について改めて議論する。第３章では、逆推定法の適用範囲を、特定話者の音響情報から任意の話者の音響情報へ拡張するために、話者変換の技術を応用した話者正規化音声–調音マッピングを提案する。第４章では、逆推定法の適用範囲を音響情報が存在音声から音響情報が存在しない音声へ拡張するために、音声の構造的表象で表される言語的特徴量を用いた推定法を提案し、話者の音声が存在しないという特殊な条件における逆推定問題に取り組む。そして、第５章で本論文をまとめる。なお、本論文の内容とは直接関係はないが、研究の一環として構築した日中二カ国語話者の音声–調音パラレルデータコーパスについて付録に記した。

第2章

音声の生成過程と逆推定問題

2.1　はじめに

本章では、まず音声生成過程の詳細を述べ、その中における調音運動の役割、および観測法について説明する。次に、本研究の主題となる逆推定法について、その重要性について応用技術を紹介しながら述べる。さらに、従来の逆推定法の概要と問題点を説明し、本論文の目的を詳しく述べる。

2.2　音声の生成過程と調音運動

2.2.1　音声生成過程の概要

音声を物理的な音波として捉えた場合、その始まりは、肺で生成される呼気流である。一方で、音声に含まれるの言語情報始まりは、脳である。発話者の脳内で、音声に乗せて伝えたい概念や意図が、言語に変換される。そして、それに基づき発話指令が形成される (図 2.1)。発話指令は、所望の音声を発話するために必要な身体的な運動が記述された、言わば音声のレシピのようなものである。発話指令に基づき、身体的な運動が開始され物理的な音声が生成される [13]。

先に述べたように、物理的な音声の始まりは、肺で生成される呼気流である。呼気流は、喉頭を通過する際に、声帯を振動させ、周期的な振る舞いを持った音源波となる。音源波は、高周波域になるほどエネルギーがなだらかに減衰する調波構造を持つ。音源波には、音韻的な特徴はなく、声の低さ・高さの印象を与える。一般に男性の声が低く女性の声が高いのは、男性の声帯の方が女性よりも大きて重く振動の周期が低くなるためである。また、音源波の調波構造における基本周波数の時間に沿った変動は、音声の韻律として知覚される。

音源波は、声道を通過し口唇から放射され、音声として空気中を伝番していく。声道を通過する際に、音源波の特定の周波数が強調され、幾つかのピークを伴った音波となる（図 2.2)。ここで、声道は音源波に対してある種のフィルタとして作用する。この声道のフィルタの特性を声道スペクトルと呼び、声道スペクトルに存在するピーク周波数をフォルマント周波数と呼ぶ。声道スペクトルには、複数のフォルマント周波数が存在しており、周波数が低い順に第一フォルマント、第二フォルマントと呼ばれる。音声の音韻性は、声道スペクトルによって特徴付けられる。そして、その声道スペクトルは、声道の形状によって定まる。発話者は、発声中に顎や舌、口唇を複雑に動かし声道の形状を柔軟に変化させることによって、声道スペクトルを連続的に変化させ様々な音韻性を持つ音声を生成している。顎や舌、口唇などの発話中の声道形状に寄与する器官を調音器官といい、その運動を調音運動という。

2.2.2　調音器官とその役割

図 2.3 に主な調音器官を示す。顎の運動は、舌や下口唇といった他の調音器官を巻き込む運動であり、声道形状を大きく変形させる。顎が下がることによって、声道内にスペース

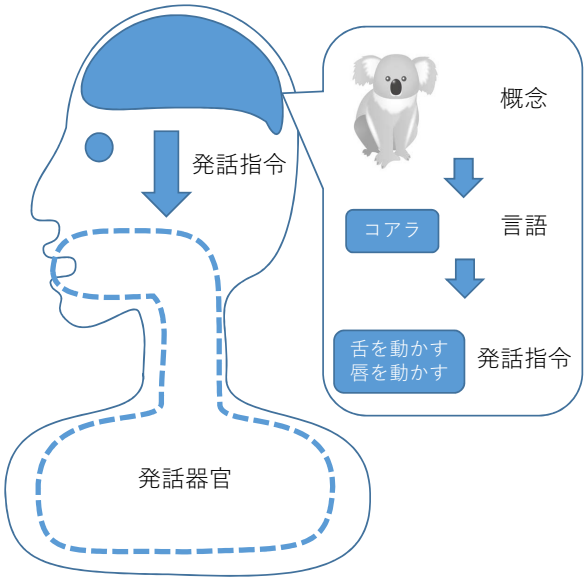

図 2.1: 音声生成過程・発話指令の生成

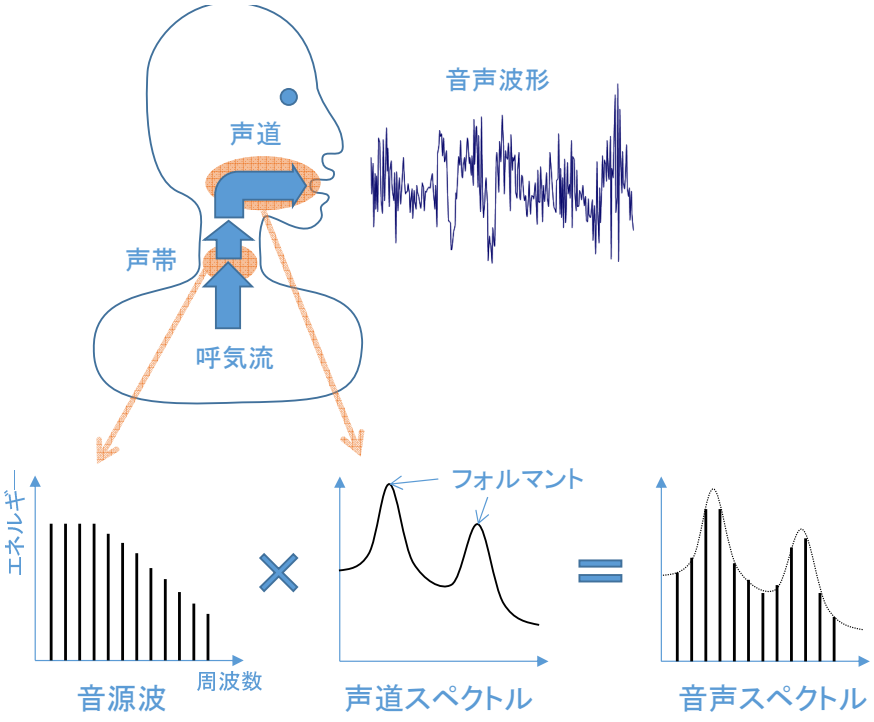

図 2.2: 音声生成過程・音源波と声道スペクトル

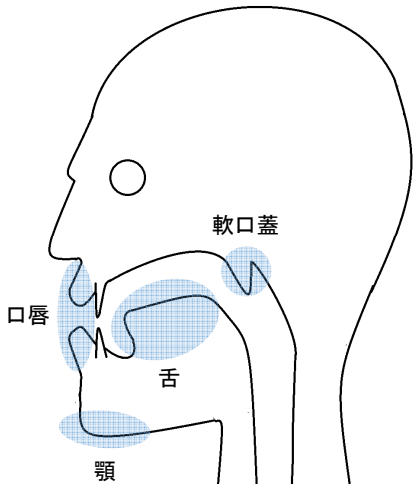

図 2.3: 主な調音器官

が確保され舌が自由に運動することができるようになる。舌は調音器官の中で最も複雑な運動をおこなう器官であり、声道スペクトルに様々な変化を与えることができる。口唇の形状は、声道内を通過した音波が体外に放射される際の放射特性をコントロールする。一般的に、口唇の放射特定は、音声の高域を強調する働きがあるといわれている [17]。軟口蓋は、声道と鼻腔の接続部に存在し、その運動によって音波の鼻腔への分岐をコントロールすることができる。鼻腔から放射された音波は、声道から放射された音波と干渉し、音声のスペクトルにディップ（谷）が形成され、鼻音特有の響きが生じる。

2.2.3　調音運動と音声の個人性

図 2.4 は、男性話者と女性話者の母音/a/,/i/の発話時の頭部 MRI 画像である。画像を比較すると、話者によって声道の長さや太さ、調音器官の形状やサイズが異なっていることがわかる。さらに、同じ音韻性を発音する際の舌の形状も話者によって異なっている。話者間における、調音器官の形態的な違いを含めた調音運動の差を調音運動の個人性という。調音運動の個人性は音声にも反映され、同じ音韻性を持つ音声でも話者によって異なる音響特性を持つことになる（音声の個人性）[29]。

話者一人ひとりが異なった調音運動で異なった音響特性の音声を発声するという、調音運動と音声の個人性は後に述べる逆推定問題において重要な話題となる。

2.2.4　音声生成・知覚における符号器としての調音運動

音声の生成過程において調音運動は音声に言語情報を付加するいう重要な役割を果たしている。その役割は、言い換えれば言語情報を音声へ埋め込む符号器とも言える。一方で、音声知覚における運動説は、調音運動の復号器としての役割を示唆している。運動説というのは、人が音声に含まれる言語情報を知覚する際に、調音運動を参照しているという説

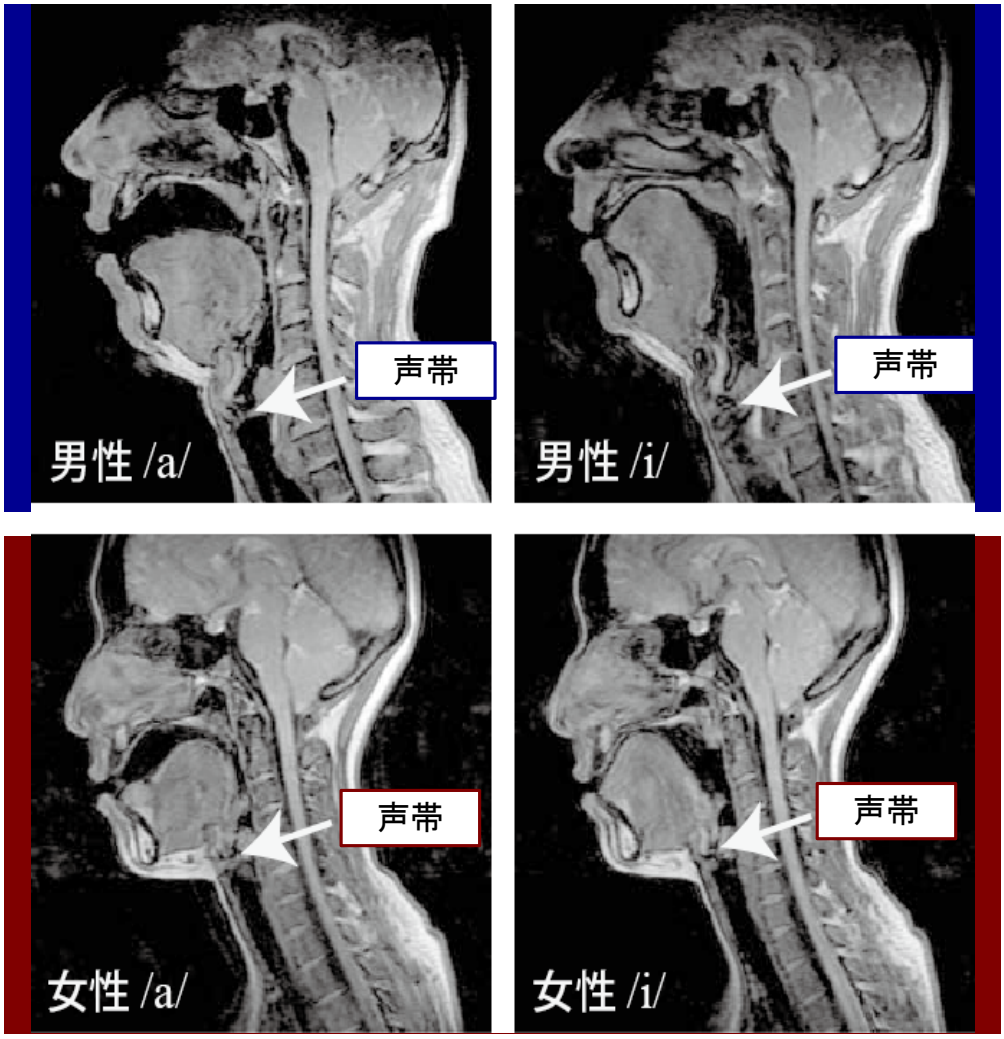

図 2.4: 男性話者と女性話者の MRI 画像

である [15][16]。例えば、発話者が/a/という母音を発声したとする。それを聞いた聴取者は、その音を生成するためにどのような調音運動が必要なのかを考える。そして、必要とされる調音運動が/a/を生成するための運動であるという情報をもとに/a/を知覚する、というのが音声における運動説である。この過程で、調音運動は音声から言語情報を復元する役割を果たしており、生成過程における符号器とは逆の役割、つまり復号をおこなっていると言える（図 2.5)。

2.3　調音運動の観測

前節で述べたように調音運動は音声生成過程において音韻性の付加という重要な役割を果たす。したがって、調音運動を観測し実態を明らかにすることは、音声生成過程を解明する上で重要な課題となる。しかし、舌や軟口蓋の運動は、口腔内でおこなわれるためビ

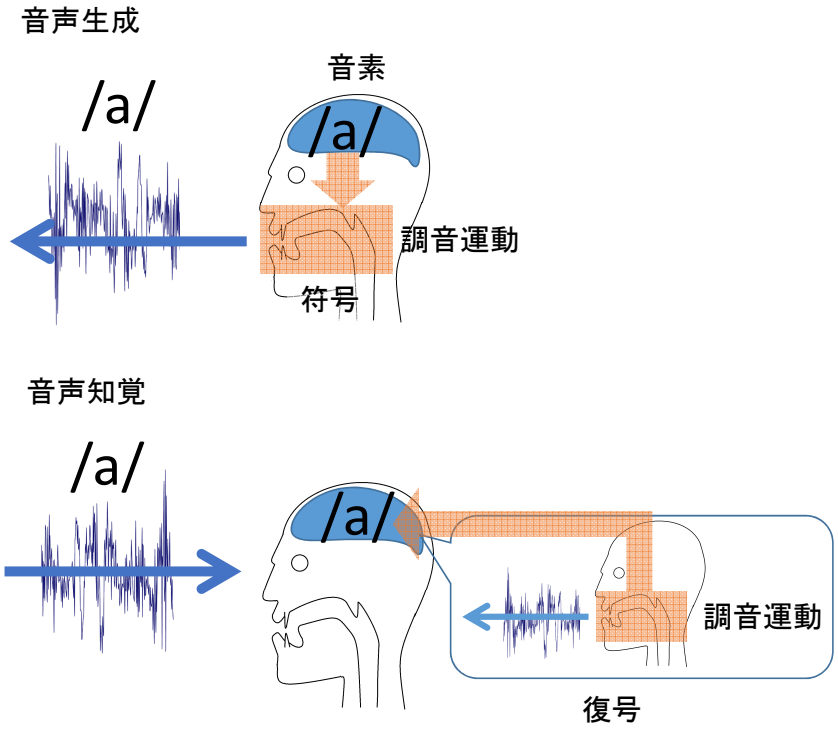

図 2.5: 音声知覚の運動説と調音運動

デオカメラなどを用いた外部からの観測が難しい。さらに、調音運動を十分に観測するためには、以下のような様々な条件を満たす必要がある [17]。

測定範囲　全ての調音器官が測定範囲に含まれる。

時間・空間分解能　舌の運動の場合は、時間分解能として 100Hz、空間分解能として 2mm 以上が望ましい。

自然発話　測定によって自然な発話を阻害しない。自然な音声を同時に測定可能。

非侵襲　人体に対して無害な測定でなくてはならない。

これまでに検討されている調音観測システムを、上記の条件を踏まえて紹介する。

2.3.1　MRI

核磁気共鳴画像法（Magnetic Resonance Imaging, MRI）は医学分野で広く使用されている測定システムであるが、調音観測システムとしても使用することができる [7]。調音観測システムとしての MRI は、最大の利点として、その広い測定範囲が挙げられる。MRI は、全ての調音器官を含む頭部全域を撮像することができ、声帯から口唇までの声道全体の形

状が画像データとして得られる。MRI の画像データから得られる情報は多く、そのデータから精工な 3 次元声道形状を復元することもできる。さらに、それを用いた声道内での音波伝搬シミュレーションによって、声道形状と声道スペクトルの関係性についての知見を得ることができる [18]。しかし、MRI を用いた調音運動観測には時間分解能・自然発話に問題がある。通常の MRI の時間分解能は極めて低く、調音運動を時間に沿って測定することは困難である。被験者に発声を繰り返しおこなってもらい、その発声ごとにタイミングをずらして撮像した画像を再構成することで、調音運動の時間的な変化を観測する手法 (発話同期撮像法) が検討されているが、この手法は被験者の負担が大きいものとなっている [19]。また、MRI の撮像中は、被験者は仰向けの姿勢をとらなけらばならず、さらに非常に大きなノイズ音にさらされているため、自然な発話が難しい。

2.3.2　超音波測定器

超音波測定器を用いた調音観測システムでは、舌表面の形状を高い時間分解能で測定することができる [20]。超音波測定器の最大の利点としては、測定機材がコンパクトであり扱いやすく、測定コストが他の手法よりも比較的に低いことがあげられる。その特性を活かした Silent Speech Interface(SSI) の開発がおこなわれている [21]。しかし、超音波測定器は、測定範囲に問題を抱えている。超音波測定器によって測定できる範囲は、舌表面と口蓋の一部のみと狭い。さらに、取得した画像から舌の表面形状を抽出するのには、人手または複雑な画像処理を必要とする。

2.3.3　磁気センサシステム

磁気センサシステム（ElectroMagnetic Articulography, EMA）は、これまでに紹介した 2 つのシステムとは異なり調音運動を観測するためだけに開発されたシステムである。調音観測システムとしては、現在最も一般的なシステムとなっている。

EMA は、送信コイルと受信コイルの 2 種類のコイルから構成されている（図 2.6）。送信コイルは、交流磁界の発生器として被験者の頭部を囲むように設置される。一方、受信コイルは、被験者の調音器官に取り付けられ[1]、観測用のマーカー（センサー）として用いられる。交流磁界の中に置かれた受信コイルには、誘導起電力が発生し、それが受信信号として観測される。受信信号の信号パターンは受信コイルの位置によって変化するため、受信コイルの位置、つまり受信コイルが取り付けられた調音器官の運動を信号パターンから推定することができる。

初期の EMA は、測定範囲が特定の平面に限定されていた（二次元磁気センサシステム,2D-EMA）[22][24]。このタイプの EMA には、受信コイルが測定面から逸脱すると推定精度が悪化するという問題がある。しがって、測定中は、被験者に取り付けられた受信コイルが測定面から逸脱しないように、被験者の頭部をヘルメット状の装置で固定しなければならない。この測定法は、被験者に対する負担が大きく、長時間の測定を難しくしていた。後に、測定範囲を 3 次元空間に拡張した三次元磁気センサシステム（3D-EMA）[6] が

[1]受信コイルは、接着剤で調音器官に取り付けられ、測定中は取り外すことはできない。


図 2.6: EMA（磁気センサシステム）。左図:システム全体と被験者, 右図:調音器官に取り付けれた受信コイル。

登場した。これにより、受信コイルの配置に制限がなくなり、被験者の頭部の運動も許されるようになった。さらに、2D-EMA では困難であった舌の側面の運動の観測や、口唇の左右方向の運動の観測が可能になった。近年では、持ち運びが可能なポータブルタイプの EMA も登場している [25]。

EMA の時間分解能と空間分解能は非常に高く、連続発話中の調音運動を精度よく測定することができる。しかし、EMA の測定範囲は、受信コイルが取り付けられた点に限定される。そして、同時に測定できる受信コイルの数は多くても 8 個程度である。つまり、先に述べた 2 つの観測システムが調音器官の形状を画像データとして取得するのに対して、EMA から得られるデータは数点の測定点の座標データである（図 2.7）。したがって、EMA のデータは、空間的な情報が非常にスパースなデータとなっている。通常、EMA の測定では、舌尖などの調音運動において重要となる箇所に受信コイルを取りけるなどして、少ない測定点で十分な情報を得られるように工夫をしている。

EMA の測定の問題点として、受信コイルが自然発話を阻害するという点があげられる。受信コイルは有線であるため、被験者は口からそのコードを垂らした状態で発話をしなけらばない。通常、口腔内の測定点は 6 点ほどだが、数本のコードが口の中に存在しているというのは自然発話を阻害するには十分である。それを考慮して、被験者が受信コイルを装着した状態での発話に十分に慣れた状態で測定をおこなうが、そこで得られるデータはあくまでも受信コイルの装着感に慣れた発話であり通常の自然発話でないということに気を付けなければならない。

2.4　逆推定問題

音声生成の過程では、調音運動によって声道の形状が変化し、それに伴って声道スペクトルが変化する。そして、様々な音韻性を持つ音声が生成される。この生成過程を順過程としたとき、その逆の過程を辿る、すなわち音声の音響情報から調音運動を復元すること

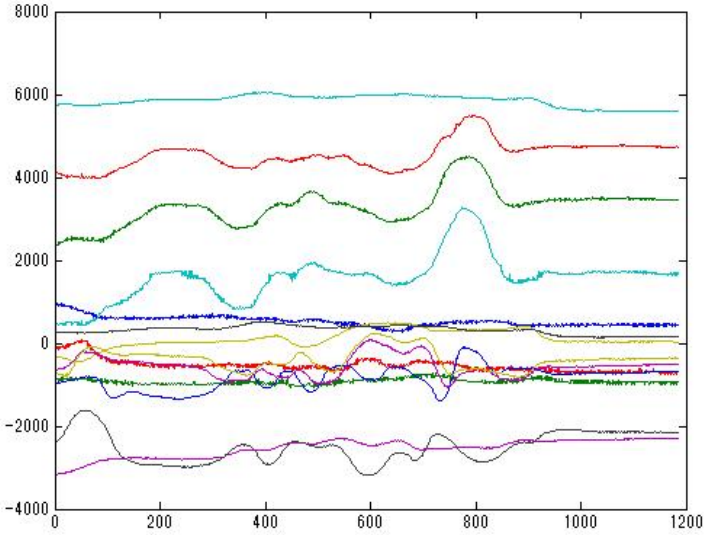

図 2.7: EMA の測定データの例。7 点の測定点の時系列に沿った座標データである。縦軸が座標、横軸が時間。各測定点が二元座標を持つため 14 次元の時系列データとなる。

を調音運動の逆推定問題をいう。この逆推定問題は、音声科学的な側面と音声工学的な側面のどちらにおいても非常に重要な研究テーマとなっている。

音声から調音運動を推定する逆推定問題は、音声の音響特性と調音運動の関係をモデル化する問題と考えることができる。しかし、音響特性と調音運動の関係には、単純なモデルでは表現できない強い非線形性がある。例えば、声道内において強い狭窄が生じている箇所は、わずかな調音運動によって音響特性が急激に変化する。その一方で、音響特性にほとんど関与しない調音運動もある。このような複雑な関係をどのようにモデル化するかが逆推定問題の重要な課題となる。

本節では、まず各分野における逆推定問題の役割・重要性に述べ、その後、これまでに検討されてきた逆推定法を紹介する。

2.4.1　音声生成分野における逆推定問題

2.2.4 節で述べたように、音声知覚の運動説に従えば、聴取者は音声から調音運動を復元し、その結果を知覚に役立てていることなる。つまり、我々は音声知覚の過程において逆推定問題を解いていることを仮定している。しかし、音声知覚の運動説が実際に音声知覚において有効に働いているか、また、どのように働いているのかは今でも議論がなされている [26]。

逆推定の過程がモデル化できれば、聴取者がどのような音響特性に注目して調音運動を復元しているのか、復元された調音運動が音声知覚において有効な情報になりうるのかといったことが明らかになる。そして、それの知見に基づき、音声知覚の運動説について更

なる議論が可能となる。

また、調音運動の観測データは音声生成過程の解明において重要な資料となるが、2.3 節で述べたように、調音運動は基本的に観測コストが高い。もし、逆推定問題を解くことによって、音声から調音運動が正確に推定できれば、本来ならば測定コストが非常に高い調音運動情報を、測定コストが低い音声から取り出すことが可能になる。これによって、利用できる調音運動の観測データが増加し、音声生成過程の解明に近づくことが期待される。

2.4.2　音声工学分野における調音運動と逆推定問題

i)　音声合成

音声工学的な分野では、しばしば音声生成過程の情報を援用し、既存の技術の高度化を目指した研究がおこなわれている。音声合成の分野では、調音運動をコントロールパラメータとして合成音声の性質を変化させる手法が検討されている [1]。Zhen ら (2013) では、HMM 音声合成において、音声と調音運動の同時測定データを用いて、音声の音響特性が調音運動に基づくパラメータ（調音運動特徴量）の重回帰によって表現された音声合成用の隠れマルコフモデル（Hidden Markov Model, HMM）のモデル学習をおこなった。この HMM から合成される音響特性は、学習された重回帰を通して調音運動特徴量によってコントロールされる。音響特性を調音運動に基づいてコントロールする利点として、調音音声学的、または生成的な知見に基づいた操作ができるという点がある。例えば、音素/I/の音声を/æ/の音声に変化させることを考える。音響特性における/I/から/æ/へ変化は複雑なものとなるが、調音運動における/I/から/æ/へ変化は舌の高さを下げることに対応する。これは、音声生成における音素と舌の高さの関係に関する知見に基づくものである。このように、音響的には複雑な操作であっても、調音運動を介することで簡単な操作によって実現することができる。

同じように調音運動によって合成音声を制御する研究として、発音の訛りや誤りを矯正する手法が検討さている [3][27]。Aryal ら (2015) では、EMA を用いて取得した母語話者の調音運動を非母語話者の調音運動の空間に写像し、その調音運動から音声を合成することで、母語話者の発音をもった非母語話者の音声を合成する手法が検討された。また、Lumban ら (2015) では、発音の誤りを含む音声から調音運動を推定し、その推定された調音運動の発音を誤りを修正し、修正後の調音運動から改めて音声を合成することで、正しい発音の音声を合成する手法が検討された。これらの手法は、Zhen ら (2013) の手法と同様に、音声の音韻性を操作する際に、音響特性を直接操作するのではなく、調音運動を介して合理的におこなっている。そして、これらの技術を中核をなすのが調音運動から音声を合成する技術と音声から調音運動を推定する技術、すなわち順推定問題と逆推定問題である。順推定問題は、調音運動から音声の音響的特性を推定するもので、音声の生成過程をそのまま辿るものである。これらの技術において、逆推定問題は音声から調音運動という生成過程の情報を抽出する重要な役割を果たしている。

ii)　音声認識

音声認識の分野でも逆推定問題を扱った研究が検討されている [2]。Mitra ら (2011) では、ニューラルネットワーク (Artificial neural network, ANN) を用いて認識対象の音声から調音運動を推定し、その推定された調音運動と元々の音声の音響特性を合わせたものを特徴量として、ノイズ環境下での音声認識をおこなった。音声認識に調音運動の情報を用いる一つの利点として、調音運動の連続性をモデルに組み込むことで、音声のコンテキスト依存の多様性を合理的にモデル化できるという点が挙げられる。

音声生成過程には、調音結合という現象が存在する。調音結合は、ある音素の調音運動が前後の音素の調音運動の影響を受けて変化するというものである。その結果、音響特性も前後の音素の影響を受けて変化する。これにより音声には、コンテキスト依存の多様性が生じる。音声認識では、音響モデルという各音素の音響特性をモデル化したものを用いて認識を行うが、調音結合を考慮するために、当該音素の前後 1 音素や 2 音素の関係に依存したモデル化をおこなう場合がある。そのようなモデル化は、モデルの規模が膨大になり有限のデータからでは十分に学習できない恐れがある。調音結合によって生じる音響的多様性は複雑な現象であるが、調音運動から眺めれば見通しが良くなる。調音結合の簡単な説明の一つとして、調音運動のエネルギー最小化がある [23]。調音運動は、ある音素毎に決めれた調音動作を連続的に達成していく運動である。このとき、運動のエネルギーを最小化するように各調音動作が最適化され、単独で発声する場合とは異なる運動となり、それが調音結合として観測される。例えば、/aka/と/aki/という二つ VCV（母音–子音–母音）音節の子音/k/の調音運動を比較すると、後続の音素が後舌母音/a/である/aka/よりも、後続の音素が前舌母音/i/である/aki/のほうが、/k/を調音位置（舌背と口蓋の接触位置）が前方に移動したものになる。これは、あらかじめ/k/の調音位置を前方にずらしたほうが、後続の母音の調音運動まで考慮した場合のエネルギーが少なくなるためである。このように、調音結合は、調音運動のエネルギーという観点で合理的に考えることができる。音声認識の音響モデルに調音運動を組み込むことで、調音結合による音響的多様性が、調音運動によって合理的に説明されることが期待できる。音声認識に調音運動を組み込む際にも、やはり音声から調音運動を推定する必要があるため、逆推定問題が重要な課題となる。

iii)　発音トレーニング

近年、新しい発音トレーニングとして調音運動情報を利用したトレーニングが検討されている。一般的な語学学習における発音トレーニングでは、学習者は教師のお手本の発音を聞いて自ら発音し、教師がその発音の修正点を示す、といった手順を繰り返すことで進められる。このトレーニングは、学習者は自らの耳に頼って発音を直す、いわば「聞いて真似る」トレーニングである (図 2.8)。発音を直すためには、発音を作り出している調音運動を直さなければならない。つまり、「聞いて真似る」トレーニングでは、調音運動における修正箇所を、教師または学習者が学習者の音声（音情報）から探し出さなけらばならず、直感的なトレーニングとは言えない。

一方で、調音運動情報を利用したトレーニングは、「見てまねる」トレーニングを目指し

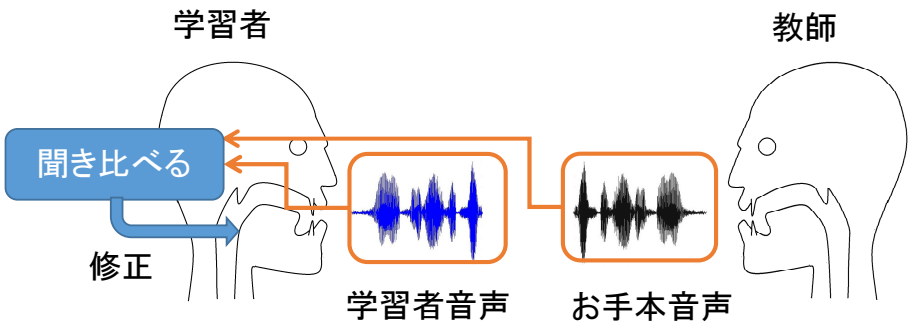

図 2.8: 従来の発音トレーニング

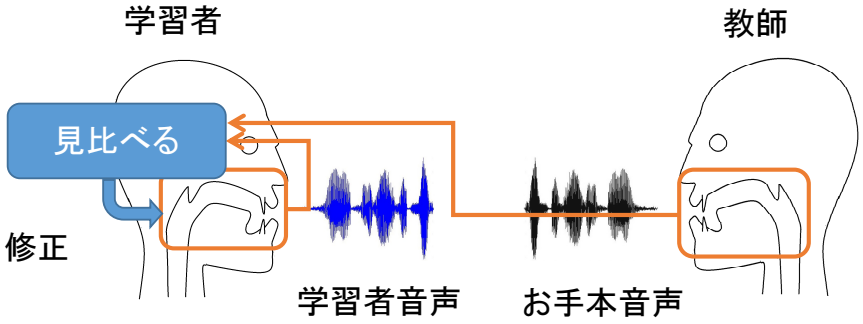

図 2.9: 調音情報を利用した発音トレーニング

たものである。このトレーニングでは、学習者の発音の誤りを含む調音運動と学習者が本来おこなうべき正しい調音運動をアニメーションによって視覚的に学習者に提示する（調音フィードバック）。この視覚的な調音フィードバックによって学習者は正しい発音と誤った自分の発音の調音運動を「見て」比べることができ、直感的に発音を修正することができる (図 2.9)。実際に、Suemitsu ら (2013) は、EMA を使った調音フィードバックシステムを用いて、調音運動情報を利用したトレーニングと音情報のみを利用したトレーニングの学習効果を比較した。その結果、調音フィードバックを用いることで音情報のみを用いた場合よりも高い学習効果が得られることが明らかになった [28]。

調音フィードバックは、発声中の学習者から何らかの方法を用いて調音運動の情報を取得する必要がある。これまで、電気的パラトグラフ（ Electro Palategraphy, EPG) や EMA などの調音観測システムして調音運動を直接的に観測する手法が検討されてきたが、実際の語学学習の現場への普及を考えると、高コストな調音観測システムを用いるのは望ましくない。そこで、調音運動を音声から逆推定することで調音観測システムを用いずに調音

フィードバックを作成する手法が検討されている [4]。Badin ら (2010) は、統計的手法を用いて音声から調音運動を逆推定し、その推定結果をもとにヴァーチャルヘッドを駆動し学習者に提示する調音フィードバックシステムを検討した。

逆推定された調音運動をフィードバックに用いた場合、学習者はその推定された調音運動を自らの調音運動と信じて発音を修正することになる。もし仮に、学習者の実際の発音が誤っているにもかかわらず、推定された調音運動が正しい発音の調音運動になってしまった場合、もしくはその逆に実際は正しいにもかかわらず、誤った発音の調音運動になってしまった場合、調音フィードバックが学習を妨げる恐れがある。そういった点を考慮すると、発音トレーニングにおける逆推定問題では、推定精度が他の応用例よりも重視されると言える。

2.4.3　様々な逆推定法

様々な分野で重要な課題となっている調音運動の逆推定問題に対して、これまでにも数多くの手法が検討されてきた。それらの手法は、音声における音響特性と調音運動の関係をどのようなモデルで表現するのかという点で異なっている。そして、それらの手法は音響理論に基づく手法とコーパスに基づく手法の二つに分けることができる。その二つの手法について以下にまとめる。

i)　音響理論に基づく手法

音響理論に基づく手法では、音声の音響的特徴と調音運動の関係を音響管を用いてモデル化する。ここで、音響管とは幾つかの円筒管が接続したものである。この手法では、まず調音運動によって変化する声道内の空間の形状を、音響管よって近似する。この音響管の周波数特性は、音響理論に沿って求めることができる。その周波数特性におけるピーク周波数と逆推定の対象となる音声から抽出したフォルマント周波数を比較し、それらが一致するように、音響管の形状を定め、その音響管の形状に基づき声道形状を求める (図 2.10)[8][29]。音声の音響的特徴と調音運動の間には複雑な関係があるため、音響的特性から声道形状を直接計算することは難しい。しかし、声道形状から音響的特性を求める順過程に関しては、音響管という近似を導入することによって、音響理論から求めることができる。

音響理論に基づく手法の利点としては、後述するコーパスに基づく手法と異なり、原理的に調音運動に関する事前収録データが必要ないという点が挙げられる。事前収録データが必要ないということは、調音運動に関してデータ依存性がないということである。つまり、逆推定の対象の音声が誰の発話であっても、フォルマント周波数が抽出できれば、声道形状を推定できる。一方で、この手法では最終的に推定される結果は声道形状であり個々の調音器官に関する情報（例えば、舌の形状など）を得ることができない。先に紹介した発音トレーニングのようなシステムでは、調音運動における個々の調音器官の状態が重要となるため、この手法を用いる場合、声道形状から個々の調音器官の運動を改めて推定する必要がある。また、音響管の周波数特性と音声のフォルマント周波数を比較するため、逆推定の精度がフォルマント周波数の抽出精度に大きく依存する。子音の多くは、フォル

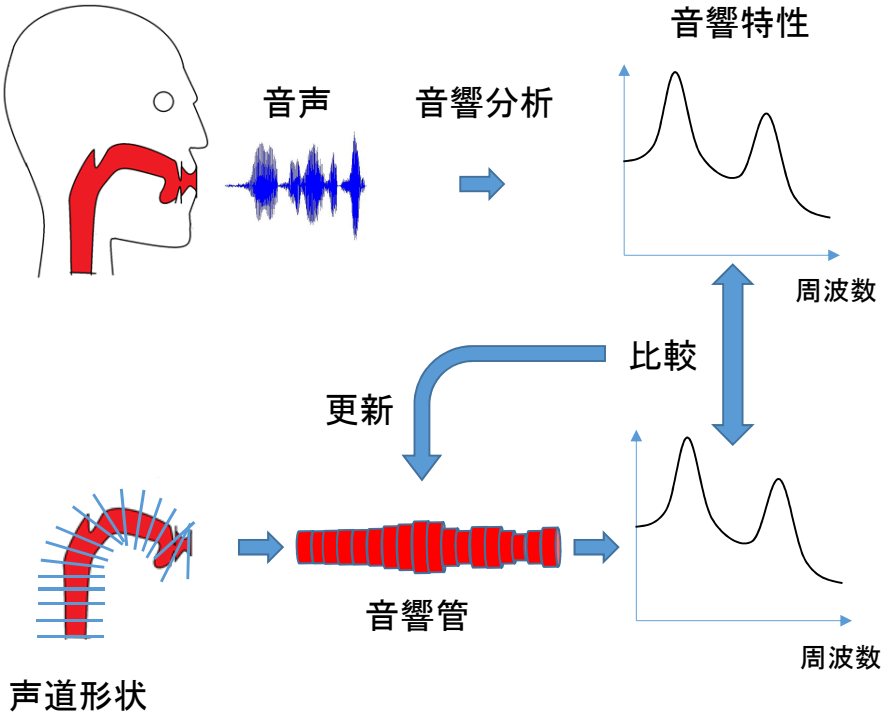

図 2.10: 音響理論に基づく手法

マント周波数を抽出することが難しいため、安定した推定をおこなうことが難しいという問題がある。

ii)　コーパスに基づく方法

コーパスに基づく手法では、音声と調音運動の同時観測データ（音声–調音パラレルデータ）を用いて音声の音響特性と調音運動の関係をモデル化する。そして、そのモデルを用いて音声から調音運動を推定する。音響理論に基づく方法では、音響特性と調音運動の複雑な関係を、音波の物理的な性質に基づいてモデル化していたのに対して、この手法におけるモデル化は観測データに対する機械学習によっておこなわれる。このとき、調音運動によって生じる声道形状の変化と音声の音響的変化の物理的な意味合いは考慮せずに、調音運動と音響特性の観測データ上での対応係性がモデル化される。つまり、パラレルデータ内の音声の音響特性によって構成される特徴量空間（音響空間）と調音運動によって構成される特徴量空間（調音空間）の対応関係（変換関係）をモデル化していると言える（図 2.11）。このとき、音声の音響特性に関する特徴量（音響特徴量）として、音韻性の情報が含まれている声道スペクトルとよい対応を示す特徴量を用いるのが一般的である。本論文では、逆推定法の中でもコーパスに基づく手法を音声–調音マッピングと呼ぶ。また、音声–調音マッピングは音声を入力として調音運動を出力する変換過程（逆過程）であるが、

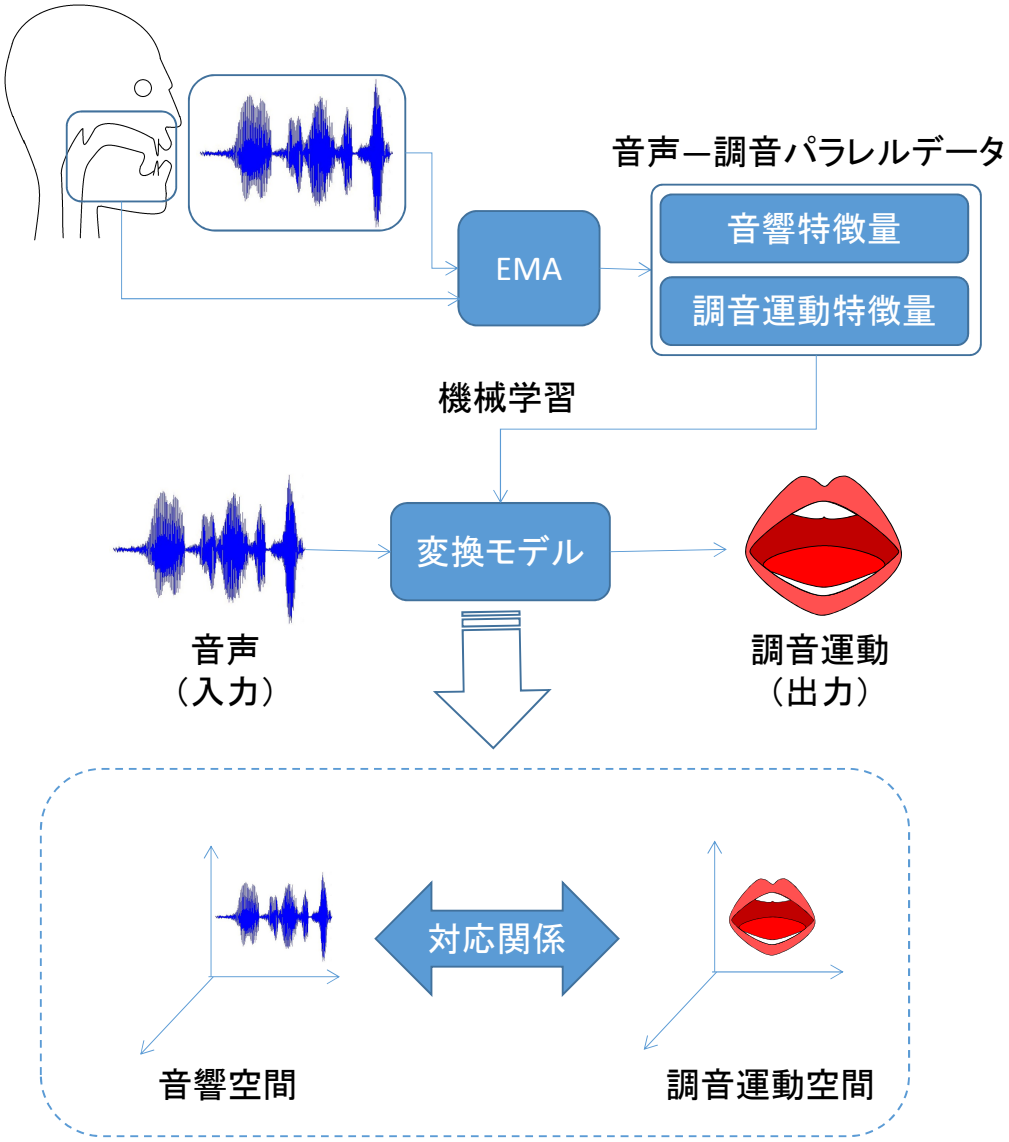

図 2.11: コーパスに基づく手法

調音運動を入力として音声を出力する変換過程（順過程）は調音–音声マッピングと呼ばれる。そして、これらの変換過程に用いるモデルを、単純に変換モデル、またはマッピングモデルと呼ぶ。

音声–調音マッピングの利点として、個別の調音器官の状態を推定できるという点があげられる。音声–調音マッピングでは、変換元（入力）と変換先（出力）の観測データから変換関係を学習するため、変換先の観測データとして EMA データなどの各調音器官の運動が明示的に表現されているデータを採用すれば、入力の音声から各調音器官の運動状態を出力するような変換モデルが学習できる。もう一つの利点として、音響特徴量の空間を構

成した上で変換をおこなうため、子音などの音響特徴量の抽出精度が低いような音素に関しても、その抽出された特徴量に対する空間が多数の観測データから構成されるので、推定精度に影響が表れにくい。したがって、音声–調音マッピングは、様々な音素を含む連続発話から各調音器官の運動を推定することが求められる場合に有効な手法と言える。

音声–調音マッピングは、近年の機械学習の発展と高品質な音声–調音パラレルデータの登場を背景に様々な検討がおこなわれており、その性能は十分に高いものとなっている。さらに、音声認識や音声合成といった工学的分野は、音声–調音マッピングと同じように機械学習による手法が中心となっているため、それらの分野において逆推定問題を扱う場合、音声–調音マッピングは親和性が高く、自然な形で既存のモデルと統合できる。2.4.2 節で紹介した、Zhen ら (2013) や Mitra ら (2011) の研究がその例として挙げられる。したがって、発音トレーニングや音声工学的な分野への応用という観点では、逆推定問題を音声–調音マッピングで解くことは重要な課題となり、大きな注目を集めている。以降では、これまでに検討されてきた代表的な音声–調音マッピング手法を紹介する。

2.4.4　調音 HMM を用いた音声–調音マッピング

隠れマルコフモデル（Hidden Morkov Model, HMM) を用いた音声–調音マッピング法が検討されている [10]。HMM とは系列データのモデリング手法の一つである。HMM では、系列データについて、特定の状態（例えば音素）ごとに出力確率と遷移確率をパラメータとするモデルを作成する。出力確率とは、特徴量の確率分布であり、その状態においてどのような特徴量が生成されるかを表現したものである。遷移確率は、状態が遷移していく様子を表現したものである (図 2.12)。

Hiroya ら (2004) は、音素を状態として調音運動を出力する HMM(調音 HMM) と、状態ごとに定められた調音運動と音声の変換関係を組み合わせた HMM 音声生成モデル (図 2.13) を提案し、それを用いた音声–調音マッピング法を検討した。調音運動特徴量系列を $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T)$, 音響特徴量系列を $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_T)$ とすると、HMM 音声生成モデルにおける音響特徴量系列の出力確率は、

$$p(\boldsymbol{Y}; \lambda) = \sum_{\boldsymbol{q}} \int p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{q}; \lambda) p(\boldsymbol{X}|\boldsymbol{q}; \lambda) p(\boldsymbol{q}; \lambda) d\boldsymbol{X} \tag{2.1}$$

となる。ここで、λ は HMM のモデルパラメータ、$\boldsymbol{q} = (q_1, q_2, \cdots, q_L)$ は HMM の音素状態系列、T は観測系列の長さである。また、$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{q}; \lambda)$ は音素状態と調音運動特徴量が与えられたときの音響特徴量の出力確率、 $p(\boldsymbol{X}|\boldsymbol{q}; \lambda)$ は音素状態が与えれたときの調音運動特徴量の出力確率である。このとき、時刻 t における音響特徴量は、調音運動特徴量からの線形変換として以下のようにモデル化される。

$$\boldsymbol{y}_t = \boldsymbol{A}_{q_t} \boldsymbol{x}_t + \boldsymbol{b}_{q_t} + \boldsymbol{w}_{q,t} \tag{2.2}$$

ここで、$\boldsymbol{A}_q, \boldsymbol{b}_q$ は音素状態 q に依存した線形変換のパラメータ、w_t は線形変換の誤差である。

今、$p(\boldsymbol{x}|\boldsymbol{q}; \lambda)$ について

$$p(\boldsymbol{x}|\boldsymbol{q}; \lambda) = \mathcal{N}(\boldsymbol{\mu}_q^{(x)}, \boldsymbol{\Sigma}_q^{(x)}) \tag{2.3}$$

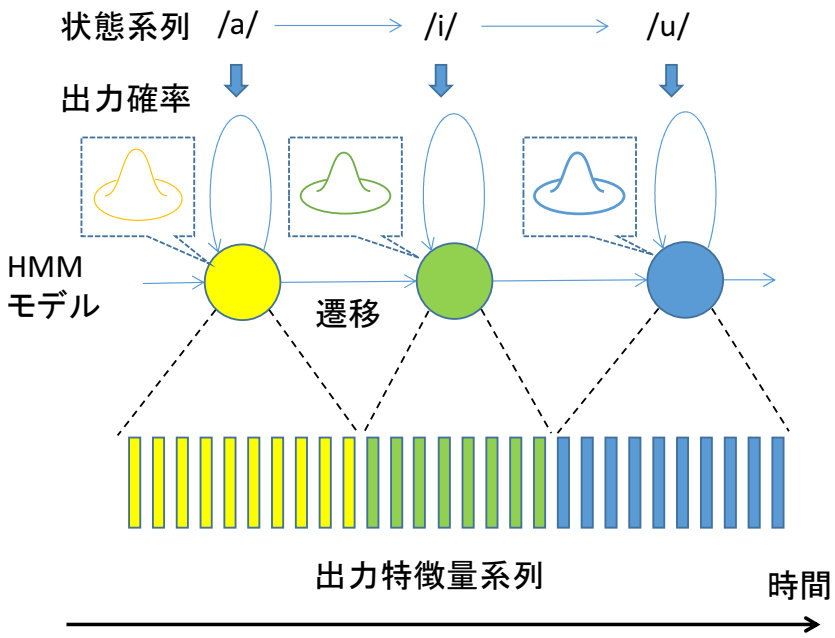

図 2.12: 音素を状態とした HMM。音素毎に出力確率が定義されている。出力確率に基づき特徴量（例えば音響特徴量）が生成される。

を仮定する。ここで、$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル μ, 分散共分散行列 $\boldsymbol{\Sigma}$ のガウス分布であり、$\boldsymbol{\mu}_q^{(x)}, \boldsymbol{\Sigma}_q^{(x)}$ はそれぞれ音素状態 q における調音運動特徴量の平均ベクトルと分散共分散行列である。このとき、$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{q}; \lambda)$ は以下の分布になる。

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{q}; \lambda) = \mathcal{N}(\boldsymbol{A}_q \boldsymbol{x} + \boldsymbol{b}_q, \boldsymbol{\Sigma}_q^{(w)}) \tag{2.4}$$

ここで、 $\boldsymbol{\Sigma}_q^{(w)}$ は $w_{q,t}$ の分散共分散行列である。

このモデルは、音素ごとに調音運動が決まり、そして音声が生成されるという音声の生成過程を HMM で表現したものになっている。HMM のモデルパラメータと式 (2.2) の線形変換のパラメータの学習は、音声–調音パラレルデータを用いて音響特徴量の出力確率が最大になるようにおこなう。

観測された音声から調音運動を逆推定する方法について述べる。まず Viterbi アルゴリズムを用いて HMM の最適な音素状態系列を推定する。そして、最適な音素状態系列と観測された音響特徴量系列に対して、事後確率が最大となる調音運動特徴量を求める。

式 (2.4) および式 (2.3) より、与えれた音響特徴量と音素状態に対して事後確率が最大になる調音特徴量は、以下のコスト関数を最小化することで求まる。

$$J = (\boldsymbol{y} - \boldsymbol{A}_q \boldsymbol{x} - \boldsymbol{b}_q)^\top {\boldsymbol{\Sigma}_q^{(w)}}^{-1} (\boldsymbol{y} - \boldsymbol{A}_q \boldsymbol{x} - \boldsymbol{b}_q) + (\boldsymbol{x} - \boldsymbol{\mu}_q^{(x)})^\top {\boldsymbol{\Sigma}_q^{(x)}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_q^{(x)}) \tag{2.5}$$

コスト関数 J を最小化する調音特徴量 $\hat{\boldsymbol{x}}$ は、

$$\hat{\boldsymbol{x}} = ({\boldsymbol{\Sigma}_q^{(x)}}^{-1} + \boldsymbol{A}_q^\top {\boldsymbol{\Sigma}_q^{(w)}}^{-1} \boldsymbol{A}_q)^{-1} ({\boldsymbol{\Sigma}_q^{(x)}}^{-1} \boldsymbol{\mu}_q^{(x)} + \boldsymbol{A}_q^\top {\boldsymbol{\Sigma}_q^{(w)}}^{-1} (\boldsymbol{y} - \boldsymbol{b}_q)) \tag{2.6}$$

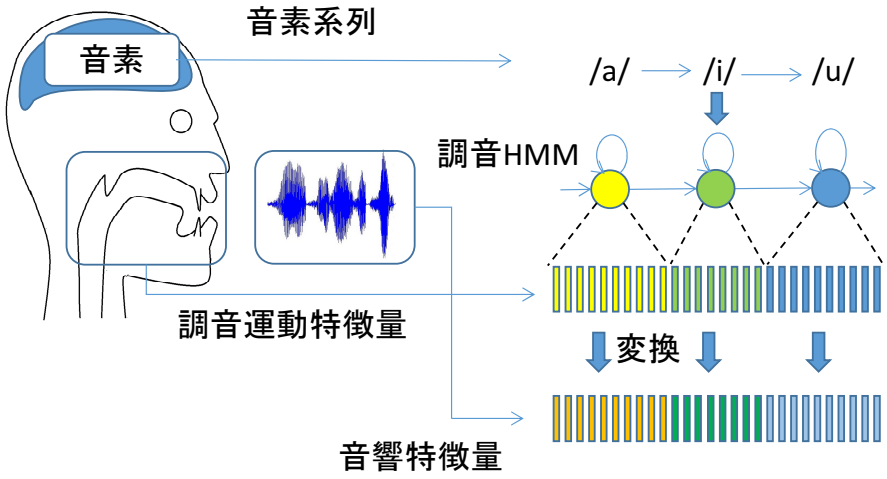

図 2.13: HMM 音声生成モデル

となる。ここで、

$$\boldsymbol{\Sigma}_q = (\boldsymbol{\Sigma}_q^{(x)^{-1}} + \boldsymbol{A}_q^\top \boldsymbol{\Sigma}_q^{(w)^{-1}} \boldsymbol{A}_q)^{-1} \tag{2.7}$$

を導入すると、

$$\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}} + \boldsymbol{\Sigma}_q \boldsymbol{A}_q^\top \boldsymbol{\Sigma}_q^{(w)^{-1}} (\boldsymbol{y} - \boldsymbol{A}_q \boldsymbol{x} - \boldsymbol{b}_q) \tag{2.8}$$

となる。これを式 (2.1) に代入すると、

$$p(\boldsymbol{Y}; \lambda) = \sum_{\boldsymbol{q}} p(\boldsymbol{q}; \lambda) \tilde{p}(\boldsymbol{Y}|\boldsymbol{q}) \int p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{q}; \lambda) d\boldsymbol{X} \tag{2.9}$$

となる。ここで、

$$\tilde{p}(\boldsymbol{y}|q; \lambda) = \mathcal{N}(\boldsymbol{A}_q \boldsymbol{\mu}_q^{(x)} + \boldsymbol{b}_q, \boldsymbol{\Sigma}_q^{(w)} + \boldsymbol{A}_q \boldsymbol{\Sigma}_q^{(x)} \boldsymbol{A}_q^\top) \tag{2.10}$$

$$p(\boldsymbol{x}|\boldsymbol{y}, q; \lambda) = \mathcal{N}(\boldsymbol{\mu}_q^{(x)}, \boldsymbol{\Sigma}_q) \tag{2.11}$$

である。式 (2.9) の積分項は 1 になるので、

$$p(\boldsymbol{Y}; \lambda) = \sum_{\boldsymbol{q}} p(\boldsymbol{q}; \lambda) \tilde{p}(\boldsymbol{Y}|\boldsymbol{q}) \tag{2.12}$$

となる。さらに、出力確率を最大化する状態系列のみで近似することで、

$$p(\boldsymbol{Y}; \lambda) \approx \max_{\boldsymbol{q}} p(\boldsymbol{q}; \lambda) \tilde{p}(\boldsymbol{Y}|\boldsymbol{q}) \tag{2.13}$$

となる。上式に対して Viterbi アルゴリズムを用いることで最適な音声状態系列を求めることができる。そして、その状態系列を用いて式 (2.6) を計算することで、調音運動を逆推定することができる。先に述べたように、音声–調音マッピングでは、音響空間と調音運動空間の複雑な非線形の対応関係をどのように表現するかが最大の課題となる。この調音 HMM を用いた音声–調音マッピングでは、音響空間と調音運動空間を音素という共通の要素で分割し、その部分空間では音響特徴量と調音運動特徴量の間に単純な線形変換の関係があると仮定している。つまり、複雑な非線形の対応関係を、音素を単位とした区分線形変換で表現している。

2.4.5　GMM に基づく音声–調音マッピング

次に GMM に基づく音声–調音マッピングについて述べる [11]。GMM とは、Gaussian Mixture Model(混合ガウス分布モデル) の略称で、その名のとおり複数のガウス分布を足し合わせたものである。今、混合ガウス分布から確率変数 $\boldsymbol{x}$ が生成されているとすると、その確率分布は、

$$p(\boldsymbol{x};\lambda) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{2.14}$$

と表される。ここで、M はガウス分布の混合数、$\alpha_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m$ は、それぞれ m 番目の混合成分における混合重み、ガウス分布の平均ベクトルと分散共分散行列である。

GMM を用いて音声–調音マッピングを行う場合、音響特徴量と調音運動特徴量の結合確率分布を考える。音響特徴量系列を $\boldsymbol{x}_t = (\boldsymbol{x}_t(1), \boldsymbol{x}_t(2), \cdots, \boldsymbol{x}_t(d_x))^\top$, 調音運動特徴量系列を $\boldsymbol{y}_t = (\boldsymbol{y}_t(1), \boldsymbol{y}_t(2), \cdots, \boldsymbol{y}_t(d_y))^\top$ とし、その二つの特徴量の結合ベクトル、$\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{x}_t^\top]^\top$, を定義する。ここで、$t$ は時間インデックス、d_x, d_y はそれぞれ音響特徴量と調音運動特徴量の次元数である。結合ベクトルの確率分布は、GMM を用いて、

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

と表される。さらに、この分布のパラメータは、

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix} \tag{2.16}$$

$$\boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(x,x)} & \boldsymbol{\Sigma}_m^{(x,y)} \\ \boldsymbol{\Sigma}_m^{(y,x)} & \boldsymbol{\Sigma}_m^{(y,y)} \end{bmatrix} \tag{2.17}$$

と表される。ここで、$\boldsymbol{\mu}_m^{(x)}, \boldsymbol{\mu}_m^{(y)}$ はそれぞれ音響特徴量と調音運動特徴量の平均ベクトル、$\boldsymbol{\Sigma}_m^{(x,x)}, \boldsymbol{\Sigma}_m^{(y,y)}$ はそれぞれ音響特徴量と調音運動特徴量の分散共分散行列、$\boldsymbol{\Sigma}_m^{(x,y)}, \boldsymbol{\Sigma}_m^{(y,x)}$ は音響特徴量と調音運動特徴量の間の相互共分散行列である。この相互共分散行列に特徴量間の関係性が記述されている (図 2.14)。この GMM のパラメータは、EM アルゴリズムを用いて推定される。

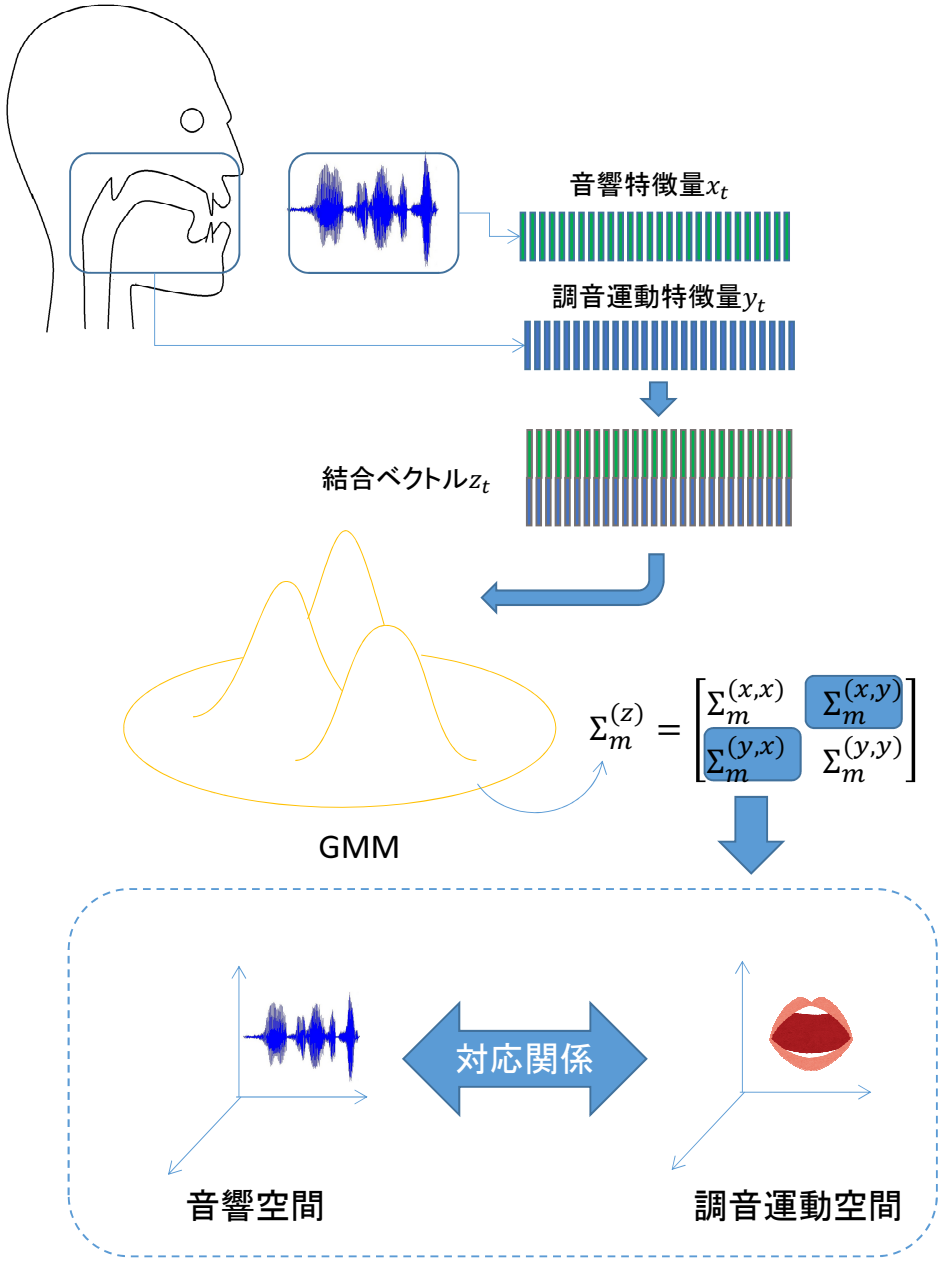

図 2.14: GMM による音声–調音マッピング

音響特徴量から調音運動特徴量への変換には、入力となる音響特徴量 $\boldsymbol{x}_t$ が与えれた場合の調音運動特徴量 $\boldsymbol{y}_t$ の事後分布 $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ を用いる。$p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ は、結合ベクトルの GMM を

用いて以下のように表される。

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t;\lambda^{(z)}) = \sum_{m=1}^{M} p(m|\boldsymbol{x}_t;\lambda) p(\boldsymbol{y}_t|\boldsymbol{x}_t, m;\lambda^{(z)}) \tag{2.18}$$

$$\tag{2.19}$$

ここで、

$$p(m|\boldsymbol{x}_t;\lambda^{(z)}) = \frac{\alpha_m \mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_m^{(x)},\boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M}\alpha_n \mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_n^{(x)},\boldsymbol{\Sigma}_n^{(xx)})} \tag{2.20}$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, m;\lambda^{(z)}) = \mathcal{N}(\boldsymbol{y}_t;\boldsymbol{E}_{m,t}^{(y|x)},\boldsymbol{D}_{m,t}^{(y|x)}) \tag{2.21}$$

$$\boldsymbol{E}_{m,t}^{(y|x)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(y,x)}{\boldsymbol{\Sigma}_m^{(x,x)}}^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_m^{(x)}) \tag{2.22}$$

$$\boldsymbol{D}_{m,t}^{(y|x)} = \boldsymbol{\Sigma}_m^{(y,y)} - \boldsymbol{\Sigma}_m^{(y,x)}{\boldsymbol{\Sigma}_m^{(x,x)}}^{-1}\boldsymbol{\Sigma}_m^{(x,y)} \tag{2.23}$$

である。$\boldsymbol{E}_{m,t}^{(y|x)}, \boldsymbol{D}_{m,t}^{(y|x)}$ は、$\boldsymbol{x}_t$ が与えたときの $\boldsymbol{y}_t$ の条件付き分布の m 番目の混合成分の平均ベクトルと分散共分散行列である。変換式は、異なる基準に基づく 2 種類がある。

i)　最小二乗誤差基準 (Minimum Mean Square Error, MMSE)

MMSE 基準による変換では、入力 $\boldsymbol{x}_t$ が与えたときの出力 $\hat{\boldsymbol{y}}_t$ が以下の式で表される。

$$\hat{\boldsymbol{y}}_t = E[\boldsymbol{y}_t|\boldsymbol{x}_t] \tag{2.24}$$

ここで、$E[\cdot]$ は期待値を表す。変換式は、以下で与えられる。

$$\hat{\boldsymbol{y}}_t = E[\boldsymbol{y}_t|\boldsymbol{x}_t] \tag{2.25}$$

$$= \int p(\boldsymbol{y}_t|\boldsymbol{x};\lambda^{(z)})\boldsymbol{y}_t d\boldsymbol{y}_t \tag{2.26}$$

$$= \int \sum_{m=1}^{M} p(m|\boldsymbol{x}_t;\lambda^{(z)}) p(\boldsymbol{y}_t|\boldsymbol{x}_t, m;\lambda^{(z)})\boldsymbol{y}_t d\boldsymbol{y}_t \tag{2.27}$$

$$= \sum_{m=1}^{M} p(m|\boldsymbol{x}_t,\lambda^{(z)})\boldsymbol{E}_{m,t}^{(y|x)} \tag{2.28}$$

この変換式において出力は、入力 $\boldsymbol{x}_t$ が与えられた場合の $\boldsymbol{y}_t$ の条件付き分布の各混合成分の平均値を、混合成分の事後確率 $p(m|\boldsymbol{x}_t;\lambda^{(z)})$ で重み付けし足し合わせたものとなっている。

ii)　最尤基準 (Maximum Likelihood Estimation, MLE)

MLE 基準による変換では、入力 $\boldsymbol{x}_t$ が与えたときの出力 $\hat{\boldsymbol{y}}_t$ が以下の式で表される。

$$\hat{\boldsymbol{y}}_t = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}_t|\boldsymbol{x}_t;\lambda^{(z)}) \tag{2.29}$$

尤度関数 $p(\boldsymbol{y}_t|\boldsymbol{x}_t;\lambda^z)$ を最大化するために、EM アルゴリズムを用いる。EM アルゴリズムでは、以下の補助関数（Q 関数）を $\hat{\boldsymbol{y}}_t$ について最大化する。

$$Q(\boldsymbol{y}_t,\hat{\boldsymbol{y}}_t)=\sum_{m=1}^{M}p(m|\boldsymbol{x}_t,\boldsymbol{y}_t;\lambda^{(z)})\log p(\hat{\boldsymbol{y}}_t,m|\boldsymbol{x}_t;\lambda^{(z)}) \tag{2.30}$$

$$=-\frac{1}{2}\hat{\boldsymbol{y}}_t^{\top}\overline{\boldsymbol{D}_t^{(y)^{-1}}}\hat{\boldsymbol{y}}_t+\hat{\boldsymbol{y}}_t^{\top}\overline{\boldsymbol{D}_t^{(y)^{-1}}\boldsymbol{E}_t^{(y)}}+K_t \tag{2.31}$$

ここで、K_t は $\hat{\boldsymbol{y}}_t$ と無関係な項である。また、

$$\overline{\boldsymbol{D}_t^{(y)^{-1}}}=\sum_{m=1}^{M}\gamma_{m,t}^{(z)}\boldsymbol{D}_m^{(y|x)^{-1}} \tag{2.32}$$

$$\overline{\boldsymbol{D}_t^{(y)^{-1}}\boldsymbol{E}_t^{(y)}}=\sum_{m=1}^{M}\gamma_{m,t}^{(z)}\boldsymbol{D}_m^{(y|x)^{-1}}\boldsymbol{E}_{m,t}^{(y|x)} \tag{2.33}$$

$$\gamma_{m,t}^{(z)}=p(m|\boldsymbol{x}_t,\boldsymbol{y}_t;\lambda^{(z)}) \tag{2.34}$$

である。Q 関数を最大化する $\hat{\boldsymbol{y}}_t$、即ち推定結果は、

$$\hat{\boldsymbol{y}}_t=(\overline{\boldsymbol{D}_t^{(y)^{-1}}})^{-1}\overline{\boldsymbol{D}_t^{(y)^{-1}}\boldsymbol{E}_t^{(y)}} \tag{2.35}$$

となる。

二つの変換式を改めて示す。

MMSE

$$\hat{\boldsymbol{y}}_t=\sum_{m=1}^{M}p(m|\boldsymbol{x}_t,\lambda)\boldsymbol{E}_{m,t}^{(y|x)} \tag{2.36}$$

MLE

$$\hat{\boldsymbol{y}}_t=\left(\sum_{m=1}^{M}\gamma_{m,t}^{(z)}\boldsymbol{D}_m^{(y|x)^{-1}}\right)^{-1}\sum_{m=1}^{M}\gamma_{m,t}^{(z)}\boldsymbol{D}_m^{(y|x)^{-1}}\boldsymbol{E}_{m,t}^{(y|x)} \tag{2.37}$$

これらを比較すると、どちらも MLE 基準では、MMSE 基準では考慮されていなかった $D_m^{(y|x)}$ が考慮されていることがわかる。この条件付き分布の分散共分散行列は、各混合の平均値の重み付け和を計算する際に、各混合に対する更なる重みとして作用する。ここでの重みは、混合成分の事後確率の重み付けとは異なり、特徴量の各次元に対して作用する。

iii)　動的特徴量を考慮した最尤系列マッピング

MLE 基準の変換式を発展させ、特徴量の時系列特性を考慮した変換法が検討されている。音響特徴量と調音特徴量はどちらも時間方向の相関が強く、時系列特性を考慮することで変換性能が向上する。

今、出力特徴量の静的成分と動的成分を考慮した特徴量を

$$\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta^{(1)}\boldsymbol{y}_t^\top, \cdots, \Delta^{(n)}\boldsymbol{y}_t^\top, \cdots, \Delta^{(N)}\boldsymbol{y}_t^\top]^\top \tag{2.38}$$

とする。ここで、n 番目の動的特徴量 $\Delta^{(n)}\boldsymbol{y}_t^\top$ は、以下で表される。

$$\Delta^{(n)}\boldsymbol{y}_t^\top = \sum_{\tau=-L_-^{(n)}}^{\tau=-L_+^{(n)}} w^{(n)}(\tau)\boldsymbol{y}_{t+\tau} \tag{2.39}$$

静的成分 $\boldsymbol{y}_t$ の時系列を、

$$\boldsymbol{y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top \tag{2.40}$$

とすると、静的成分の時系列と $\boldsymbol{Y}_t$ の時系列 $\boldsymbol{Y}$ の関係は、

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{2.41}$$

となる。W は以下の行列で表される行列である。

$$\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_t, \cdots, \boldsymbol{W}_T]^\top \otimes \boldsymbol{I}_{D_x \times D_y} \tag{2.42}$$

$$\boldsymbol{W}_t = [\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \cdots, \boldsymbol{w}_t^{(n)}, \cdots, \boldsymbol{w}_t^{(N)}] \tag{2.43}$$

$$w_t^{(n)} = [\overbrace{0, \cdots, 0}^{1\text{st}}, \overbrace{w^{(n)}(-L_-^{(n)})}^{(t-L_-^{(n)})-\text{th}}, \cdots, \overbrace{w^{(n)}(0)}^{(t)-\text{th}}, \cdots, \overbrace{w^{(n)}(L_+^{(n)})}^{(t+L_+^{(n)})-\text{th}}, 0, \cdots, \overbrace{0}^{T-\text{th}}] \tag{2.44}$$

この変換行列 $\boldsymbol{W}$ を用いて、特徴量の時系列特性を考慮した変換を求める。

動的成分を考慮した入力特徴量（音響特徴量）$\boldsymbol{X}_t$ と出力特徴量（調音運動特徴量）$\boldsymbol{Y}_t$ の結合ベクトル $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ を定義する。この $\boldsymbol{Z}_t$ に対して求めた GMM のパラメータセットを $\lambda^{(Z)}$ とする。入力特徴量系列 $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ が与えられた場合の出力特徴量の静的成分の時系列は、

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p(\boldsymbol{Y}|\boldsymbol{X}; \lambda^{(Z)}) \tag{2.45}$$

となる。ここで、

$$p(\boldsymbol{Y}|\boldsymbol{X}; \lambda^{(Z)}) = \sum_{\text{all}\boldsymbol{m}} p(\boldsymbol{m}|\boldsymbol{X}, \lambda^{(Z)}) p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}; \lambda^{(Z)}) \tag{2.46}$$

であり、$\boldsymbol{m} = [m_1, m_2, \cdots, m_T]$ である。時刻 t における $\boldsymbol{Y}_t$ の条件付き分布は、

$$p(\boldsymbol{Y}_t|\boldsymbol{X}_t, \lambda^{(Z)}) = \sum_{m=1}^{M} p(\boldsymbol{m}|\boldsymbol{X}_t; \lambda^{(Z)}) p(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{m}_t; \lambda^{(Z)}) \tag{2.47}$$

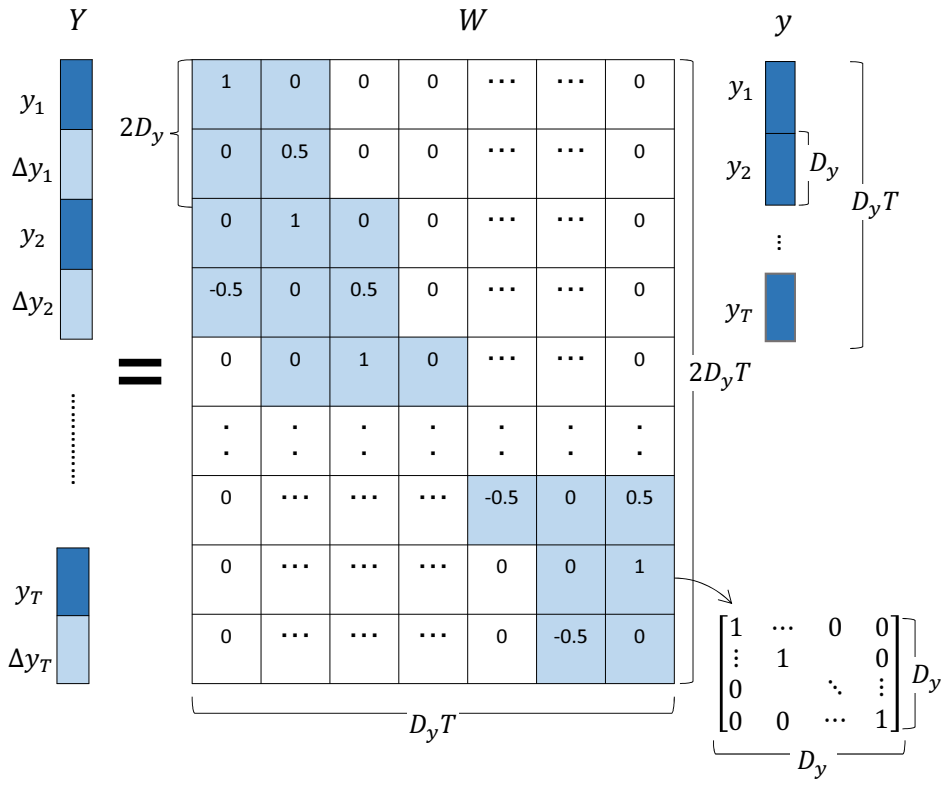

図 2.15: 静的成分と動的成分に関する変換行列の例

である。ここで、

$$p(m|\boldsymbol{X}_t;\lambda^{(Z)}) = \frac{\alpha_m \mathcal{N}(\boldsymbol{X}_t;\boldsymbol{\mu}_m^{(X)},\boldsymbol{\Sigma}_m^{(X,X)})}{\sum_{n=1}^{M}\alpha_n \mathcal{N}(\boldsymbol{X}_t;\boldsymbol{\mu}_n^{(X)},\boldsymbol{\Sigma}_n^{(X,X)})} \tag{2.48}$$

$$p(\boldsymbol{Y}_t|\boldsymbol{X}_t,m,\lambda^{(Z)}) = \mathcal{N}(\boldsymbol{Y}_t;\boldsymbol{E}_{m,t}^{(Y|X)},\boldsymbol{D}_{m,t}^{(Y|X)}) \tag{2.49}$$

$$\boldsymbol{E}_{m,t}^{(Y|X)} = \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(Y,X)}\boldsymbol{\Sigma}_m^{(X,X)^{-1}}(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}) \tag{2.50}$$

$$\boldsymbol{D}_{m,t}^{(Y|X)} = \boldsymbol{\Sigma}_m^{(Y,Y)} - \boldsymbol{\Sigma}_m^{(Y,X)}\boldsymbol{\Sigma}_m^{(X,X)^{-1}}\boldsymbol{\Sigma}_m^{(X,Y)} \tag{2.51}$$

である。ここで、式 (2.52) と同様に以下の補助関数を考える。

$$Q(\boldsymbol{Y},\hat{\boldsymbol{Y}}) = \sum_{\text{all}\boldsymbol{m}} p(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{Y};\lambda^{(Z)}) \log p(\hat{\boldsymbol{Y}},\boldsymbol{m}|\boldsymbol{X};\lambda^{(Z)}) \tag{2.52}$$

ここで、式 (2.41) を用いると、

$$Q(\boldsymbol{Y},\hat{\boldsymbol{Y}}) = -\frac{1}{2}\hat{\boldsymbol{y}}^\top \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}} \boldsymbol{W}\hat{\boldsymbol{y}} + \hat{\boldsymbol{y}}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} + K' \tag{2.53}$$

となる。ここで、

$$\overline{\boldsymbol{D}^{(Y)^{-1}}} = \mathrm{diag}\left[\overline{\boldsymbol{D}_1^{(Y)^{-1}}}, \overline{\boldsymbol{D}_2^{(Y)^{-1}}}, \cdots, \overline{\boldsymbol{D}_t^{(Y)^{-1}}}, \cdots, \overline{\boldsymbol{D}_T^{(Y)^{-1}}}\right] \tag{2.54}$$

$$\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} = \left[\overline{\boldsymbol{D}_1^{(Y)^{-1}}\boldsymbol{E}_1^{(Y)}}, \overline{\boldsymbol{D}_2^{(Y)^{-1}}\boldsymbol{E}_2^{(Y)}}, \cdots, \overline{\boldsymbol{D}_t^{(Y)^{-1}}\boldsymbol{E}_t^{(Y)}}, \cdots, \overline{\boldsymbol{D}_T^{(Y)^{-1}}\boldsymbol{E}_T^{(Y)}}\right] \tag{2.55}$$

$$\overline{\boldsymbol{D}_t^{(Y)^{-1}}} = \sum_{m=1}^{M} \gamma_{m,t}^{(Z)} \boldsymbol{D}_m^{(Y|X)^{-1}} \tag{2.56}$$

$$\overline{\boldsymbol{D}_t^{(Y)^{-1}}\boldsymbol{E}_t^{(Y)}} = \sum_{m=1}^{M} \gamma_{m,t}^{(Z)} \boldsymbol{D}_m^{(Y|X)^{-1}} \boldsymbol{E}_{m,t}^{(Y|X)} \tag{2.57}$$

$$\gamma_{m,t}^{(Z)} = p(m|\boldsymbol{X}_t, \boldsymbol{Y}_t; \lambda^{(Z)}) \tag{2.58}$$

である。$Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$ を最大化する出力特徴量系列、即ち推定結果は、

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} \tag{2.59}$$

で与えられる。この変換式は、入力特徴量系列に対して、特徴量の時系列特性を考慮した出力特徴量系列を出力する。$\lambda^{(Z)}$ で表される GMM には、動的成分を含む入力特徴量と出力特徴量の結合分布が学習されている。変換の際には、式 (2.35) と同様に混合成分の平均を混合成分の事後確率と分散行列の逆行列で重み付けし足し合わせるのだが、変換行列 $\boldsymbol{W}$ を用いることで、全ての時刻において動的成分から静的成分への変換を考慮した上で、足し合わせることになる。結果的に、時系列全体の動的特性を考慮した出力系列を得ることができる。

GMM を用いた音声–調音マッピングは、先に紹介した調音 HMM を用いた音声–調音マッピングと類似点がある。調音 HMM でのマッピングでは、音響空間と調音運動空間の複雑な変換関係を、音素を単位とした HMM の状態ごとの区分線形変換で近似することで、音声から調音運動を推定する。一方で、GMM における変換式である式 (2.36) と式 (2.37) に注目すると、どちらも混合成分における調音運動の条件付き分布の平均ベクトル $\boldsymbol{E}_{m,t}^{(y|x)}$ の重み付け足し合わせになっていることがわかる。そして、式 (2.28) からわかるように $\boldsymbol{E}_{m,t}^{(y|x)}$ は入力 $\boldsymbol{x}_t$ からの線形変換によって表現されている。つまり、GMM を用いた変換においても混合成分を単位とした線形変換が行われることになる。調音 HMM と異なる点は、調音 HMM における線形変換の区分は音素という単位で明示的に与えられるのに対して、GMM での線形変換の区分である混合成分は観測データから学習される。さらに、調音 HMM では入力特徴量から各フレームにおける音素状態が一つ定まり、その状態で規定される線形変換が適応されてるのに対して、GMM での変換では各フレームにおける入力特徴量に対する混合成分の事後確率を重みとして、線形変換の結果が足し合わされる。つまり、調音 HMM がハードな区分線形変換であるのに対して GMM はソフトな区分線形変換であると言える。

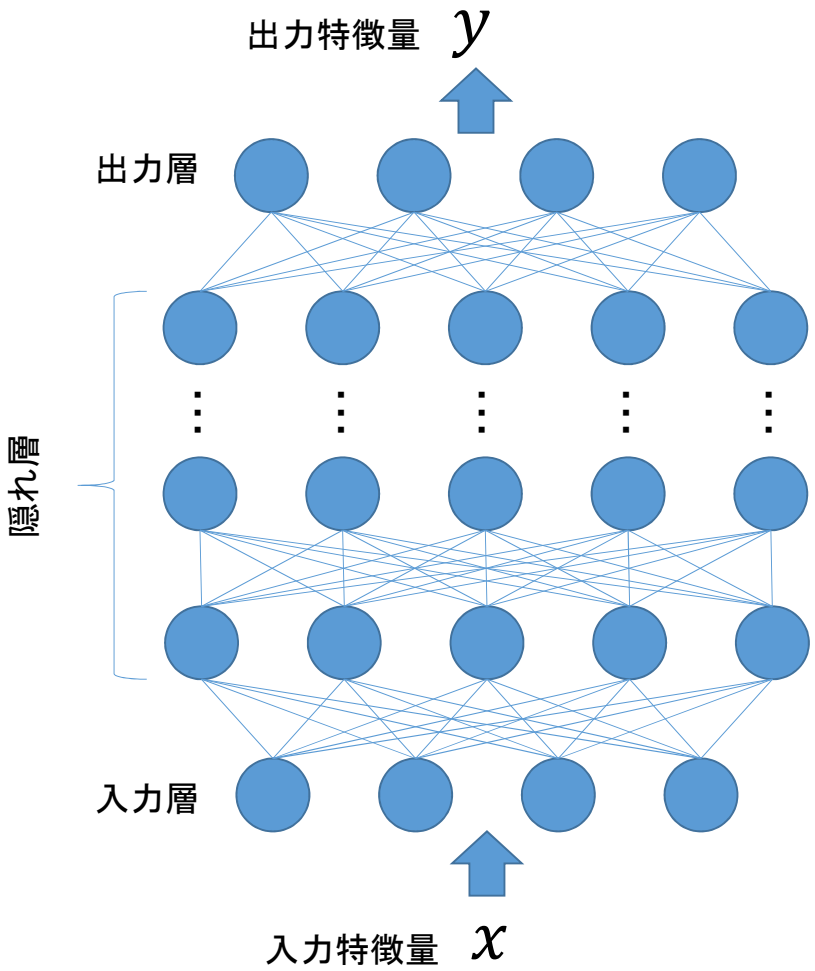

図 2.16: フィードフォワード型ニューラルネットワーク

2.4.6　ニューラルネットワークを用いた音声–調音マッピング

最後にニューラルネットを用いた音声–調音マッピングについて述べる。ニューラルネットワークは、多数のノードと、その結合によって構成される。特にフィードフォワード型のニューラルネットワークでは、幾つかのノードによって構成される層の間を情報（特徴量）が一方向へ伝搬していく (図 2.16)。このフィードフォワード型のネットワークは音声工学分野において、音声認識の音響モデルや、声質変換における特徴量の変換器など様々な応用が検討されている。音声–調音マッピングにおいても、声質変換と同様にフィードフォワード型のネットワークが特徴量変換器として用いられている [9]。フィードフォワード型のネットワークは、入力特徴量を受け入れる入力層、変換結果を出力する出力層、そして、入力層と出力層の間に設けられる隠れ層（中間層）によって構成される (図 2.16)。以下、フィードフォワード型のネットワークを用いた特徴量変換について説明する。

今、入力特徴量 $\boldsymbol{x} = [x_1, x_2, \cdots, x_i, \cdots, x_{D_x}]$ が与えられたとする。一層目の隠れ層の j 番目のノードへの入力は、

$$a_j^{(1)} = \sum_{i=1}^{D_x} w_{i,j}^{(1)} x_i + w_{j,0}^{(1)} \tag{2.60}$$

となる。ここで、(1) は隠れ層の 1 層目のパラメータということを表している。$w_{i,j}^{(1)}$ は、入力特徴量の i 次元目が j 番目のノードへ伝達される際の重みパラメータである。また、$w_{j,0}^{(1)}$ はバイアスパラメータである。ここで、$x_0 = 1$ を付加するとこで、

$$a_j^{(1)} = \sum_{i=0}^{D_x} w_{i,j}^{(1)} x_i \tag{2.61}$$

となり、バイアスパラメータを重みパラメータ集合に組み込むことができる。j 番目のノードへの入力 $a_j^{(1)}$ は、非線形活性化関数 $h(\cdot)$ によって変換され、

$$z_j^{(2)} = h(a_j) \tag{2.62}$$

となる。$z_j^{(1)}$ は、隠れ層の 1 層目の j 番目のノードの出力、即ち隠れ層の 2 層目へ伝搬される情報となる。ここで、非線形活性化関数には一般に、ロジスティクシグモイド関数や tanh 関数のような関数が用いられる。ロジスティックシグモイド関数は、

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \tag{2.63}$$

で表される関数である。

隠れ層が N 層の場合、出力層の k 番目のノードへの入力は、

$$a_k^{(N+1)} = \sum_{j=0}^{M^{(N)}} w_{k,j}^{(N+1)} z_j^{(N+1)} \tag{2.64}$$

となる。ここで、$k = 1, \cdots, D_y$ であり、D_y は出力特徴量の次元数である。また、$M^{(N)}$ は隠れ層の N 層目のノード数である。声質変換や音声–調音マッピングなどの変換タスクでは、出力特徴量は、出力層の各ノードの入力の恒等写像によって求めることが一般的である。したがって、出力特徴量 $\boldsymbol{y} = [y_1, \cdots, y_k, \cdots, y_{D_y}]$ は、

$$y_k = a_k^{(N+1)} \tag{2.65}$$

となる。このネットワークのパラメータは重みパラメータの集合である。重みパラメータは、入力特徴量と出力特徴量のパラレルデータから学習される。このとき用いられる手法が誤差逆伝搬法である。誤差逆伝搬法は、学習データ内の入力特徴量から得られる出力と、それに対応する出力特徴量から計算される誤差関数の勾配に基づいて、重みパラメータを出力層側から順に更新していく手法である。ネットワークは、隠れ層の総数が増えるほ

ど、その表現力が増し複雑な変換が可能になるが、そのような層が深いネットワークは誤差逆伝搬による重みパラメータの学習が難しくなるという問題があった。その問題を解決したのが、ネットワークの事前学習である。事前学習は、ネットワークに対して。誤差逆伝搬に適した重みパラメータの初期値を与える手法である。これによって、層が深いネットワーク (Deep Nueral Network, DNN) の学習が可能になり、複雑な変換にも対応できるようになった。音声–調音マッピングにおいても、DNN を用いた変換が検討され、高い変換精度が示された [12]。さらに近年では、特徴量の時系列特性を考慮したネットワークである BLSTM-RNN(Bidirectional Long Short-Term Memory based Recurrent Neural Network) を用いた変換なども検討されている [30]。

2.4.7　音声–調音マッピング法の比較

ニューラルネットワークによる音声–調音変換は、先に紹介した HMM や GMM を用いた変換とは大きく異なる点がある。HMM や GMM を用いた変換では、音響空間と調音空間の複雑な対応関係を、音響特徴量と調音特徴量で共通する要素（例えば音素）を単位とした区分線形変換で表現していたのに対して、ニューラルネットでは、多数の非線形活性化関数の積み重ねで表現している。この違いは、2.4.1 節で述べた音声知覚における運動説の解明という点において意味を持つ。HMM や GMM を用いた変換モデルでは、そのモデルパラメータが明確な意味を持つため、逆推定における音声と調音運動の関係の理解を助ける。例えば、HMM の各状態における線形変換の係数や、GMM の各混合における共分散行列は、調音運動に対する音響特徴量の感度を表現していると捉えることができる。ここでの感度とは、調音運動が変化した際にどのような音響特徴量がどの程度変化するかを捉えたものである。この感度を調べることによって、逆推定に有効な音響特徴量が明らかになる可能性がある。一方で、非線形活性化関数の積み重ねで表現されたニューラルネットワークからそのような知見を得ることは難しい。したがって、音声生成分野的な側面を考慮すると、HMM や GMM を用いた変換モデルを検討することは大きな意味がある。

2.4.8　音声–調音パラレルコーパス

これまで紹介した音声–調音マッピングは、いずれの変換モデルも、音声と調音運動の多量の同時測定データ (音声–調音パラレルデータ) を用いて構築（学習）される。逆に言えば、変換モデルが学習できるほどの量の音声–調音パラレルデータが登場したことが、音声–調音マッピングの始まりだと言える。そして、このパラレルデータの質が変換モデルの変換精度を左右する要因の一つにもなる [31]。

2.3 節で述べたように、調音運動を観測するためには、特殊な調音観測システムが必要となる。さらに、大規模な音声–調音パラレルコーパスを収録するためには、調音観測システムの中でも、自然発声に近い音声データとの同時測定、しかも長時間の測定が可能なシステムが必要となる。それらの条件を（不完全ながら）満たす調音観測システムとして EMA(磁気センサシステム) が注目され、EMA を用いた音声–調音パラレルデータコーパスが登場した。現在、音声–調音マッピングの研究において音声–調音パラレルデータとして

主に用いられているデータコーパス、MOCHA-TIMIT と mngu0 について紹介する。

i)　MOCHA-TIMIT

MOCHA-TIMIT は、Carstens 社の 2D-EMA(AG200) よって計測された音声–調音パラレルデータコーパスである [33]。イギリス英語話者の男女各 1 名のデータが収録されている。発話内容は、TIMIT[34] の音素バランス文 460 文である。EMA の測定点（受信コイル, センサー）の位置は、上下の口唇、下の前歯、舌上の三点、軟口蓋の計 7 点である (図 2.17)。それらは全て正中矢状面に沿って取り付けられている。各センサーの運動が 2 次元座標系における時系列データであるため、調音運動データは計 14 次元の時系列データとなっている。この調音運動データのサンプリング周波数は、200Hz である。データコーパスには、各発話の音声データ、調音運動データに加え、読み上げ文のテキストファイルと、各発話の音素情報を記述したファイル（音素ラベルファイル）が同梱されている。音素ラベルファイルには、音素の種類と、その時間情報が記されている。全 460 文の発話時間の合計は、無音区間を除けば、およそ 16 分間ほどで、音声–調音マッピングの変換モデルの学習データとしては、十分な量となっている。

MOCHA-TIMIT の利点として、2 名分のデータが収録されているため、話者の違いに注目した研究がおこなえることが挙げられる。2.2.3 節で述べたように、音声の個人性は、調音運動と関わりが深く、複数話者の音声–調音パラレルデータは、それらに関する検討をおこなう際に貴重な資料となる。また、データベースの自体がよく整理されており、音素ラベルなどの補助情報も充実していて、扱いやすいコーパスとなっている。

その一方で、MOCHA-TIMIT は、測定データの質に問題があることが指摘されている [31]。指摘されている問題点は、計測中のセンサーの脱着と、頭部の運動の影響の 2 点である。まず、センサーの脱着だが、女性話者のデータの測定において、測定点のセンサーが外れてしまい、改めて付け直されていることが報告されている。全 460 発話のうち、125 発話目の発話で軟口蓋のセンサーが、284 発話目の発話で舌上のセンサーの一つが、それぞれ付け直されている。EMA の受信センサーは手作業で各調音器官（測定点）に取り付けられるため、センサーが外れた場合、元々取り付けられていた位置と正確に一致する箇所に付け直すのは難しい。したがって、センサーの付け直しがおこなわれた発話の前後で、そのセンサーが計測している点は異なっている。つまり、全発話データを通して見たときに、調音運動データの特定のパラメータが意味しているもの（具体的には計測点）に一貫性が無いという問題点がある。

もう一つの問題は、MOCHA-TIMIT で使用された磁気センサシステム特有の問題点である。MOCHA-TIMIT で使用された EMA は、測定領域が 2 次元に限定された 2D-EMA である。2.3 節で述べたように、2D-EMA では、受信センサーが測定面から逸脱しないように、被験者の頭部の動きを固定しながらおこなう。このとき、ヘルメット型の器具を使い被験者の頭部を固定するのだが完全に固定するということは難しく、測定中に少しずつ頭部が動いてしまう。その結果、センサーの位置推定の結果に時間とともに変化するバイアスが混入することになる。 図 2.18 は、女性話者の舌尖のセンサの水平軸方向の座標を各発話で平均した値をプロットしたものである。横軸が発話番号（1〜460)、縦軸が座標の

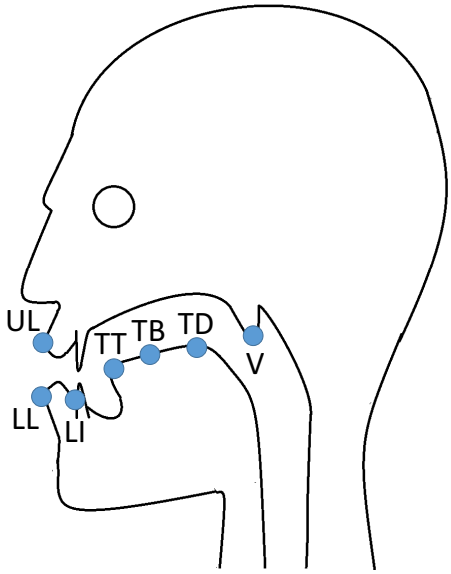

図 2.17: MOCHA-TIMIT の調音運動データの測定点

値 (0.01mm) である。発話が進むごとに、値が上昇していることがわかる。これが頭部のずれによって生じるバイアスだと考えられる。研究によっては、このバイアスを前処理によって取り除いている [32]。

データの質としては、次に紹介する mngu0 に劣るところがあるが、複数名のデータが存在するという点は、mngu0 にはない利点であり、異なる話者用いた様々な実験を設計することができる。本研究でも、次章からの実験ではこの MOCHA-TIMIT を音声–調音パラレルコーパスとして用いる。

ii)　mngu0

mngu0 は、男性話者 1 名のデータが収録された音声–調音パラレルデータコーパスである [35]。収録されている発話数は 1000 以上で、MOCHA-TIMIT よりもデータ量が豊富である。一般に、モデル学習に必要なデータ量が GMM や HMM よりも多いニューラルネット系の変換モデルも学習可能なデータ量である。このコーパスは、3 次元の測定領域を持つ 3D-EMA である Carstens 社の AG500 によって収録されており、MOCHA-TIMIT で問題となった頭部の運動によるバイアス混入の問題が解決されている。また、測定中のセンサーの脱着もなく、総じて質の高いデータを言える。実際に Richmond(2009) らの研究では、ニューラルネットワークの一種である MDN(Mixture Density Network) を MOCHA-TIMIT と mngu0 をそれぞれ用いて学習された場合、mngu0 を用いたネットワークの方が良い変換精度を示すことが報告されている。この比較は、ネットワークの学習に用いたデータ量がそれぞれのコーパスで異なる (MOCHA-TMIT:368 発話, mngu0:1137 発話) ため、簡単に

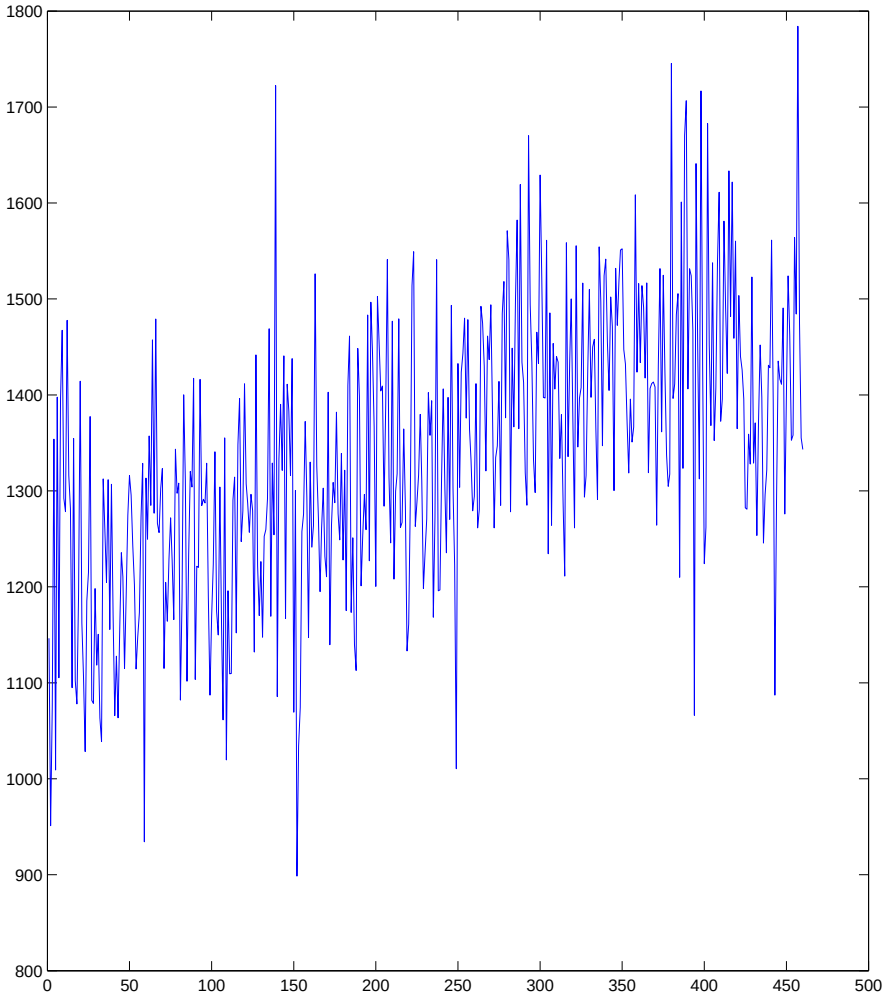

図 2.18: MOCHA-TIMIT のデータ例

データの質だけの差とは言えないものの、コーパスの違いが変換精度に影響を与えることを示した一つの結果だと言える。

mngu0 には、EMA データのほかに、MRI データも公開されおり、同時測定ではないものの一人の話者の EMA データと MRI データの組み合わせは、音声生成過程に関するモデル化をする際に有益なデータとなる。

iii)　日中二カ国語話者の音声–調音パラレルデータ

前述した二つのコーパスを含め、これまで公開されてきた音声-－調音パラレルデータコーパスは、主に発話者の母語を対象としたものであった。これは、一般的な音声の生成

過程の解明や工学的モデルの構築を目的にしているためである。一方で、語学学習の研究領域では、しばしば二カ国語話者の音声データが第二言語習得の過程の解明において重要となる。そこで、研究の一環として、EMA を用いて日中二カ国語話者を対象とした日中・音声–調音パラレルデータの収録を行った。その収録過程と統計的手法を用いた分析結果を付録 A にまとめた。測定したデータは、音声生成過程に関する研究資料として公開する予定である。

2.5　逆推定の課題と本研究の目的

本節では音響理論に基づく逆推定法とコーパスに基づく逆推定法（音声–調音マッピング）について紹介してきたが、これらの手法にはいくつかの課題が残っている。i) で述べたように、音響理論に基づく手法の課題として、子音に対する推定が挙げられる。子音の多くは、明確なフォルマント周波数を持たないため、音響管の周波数特性を利用した推定をおこなうことが難しい。したがって、子音を含んだ様々な音素が出現する連続発話に適用することが難しい。

一方、音声–調音マッピングは、音声–調音パラレルデータを使うことで、音声の音響空間と調音運動空間の対応を機械学習し、変換モデルを構築する。この手法は、音響理論に基づく方法とはことなり、声道形状と音響特性の間の物理的な関係には注目せず、パラレルデータに存在する音声データと調音データの観測上の関係のみに注目する。そのため、子音であっても、音声から音響特徴量が抽出できれば、同時観測された調音運動との対応が多数のデータから学習され変換が可能になる。さらに、音響特徴量と調音特徴量の動的な特性（時系列特性）を考慮する変換方法も検討され、音響理論に基づく手法の課題であった連続発話を対象とする逆推定でも良好な精度を得られることが示されている。これらの利点は、音声と調音運動の間に物理的な仮定を設けず、その関係を音声–調音パラレルデータに語らせることで得られたものである。しかし、それが利点だけなく、この手法の欠点も作り出している。

音声–調音マッピングを用いた変換は、音響空間と調音運動空間の対応をモデル化することでおこなわれるが、ここでモデル化される空間は、あくまでも音声–調音パラレルデータに基づく空間である。つまり、ある特定話者（以下、モデル話者）の音声–調音パラレルデータを用いて学習された変換モデルは、その話者の音響空間と調音運動空間の対応のみをモデル化したものである。このモデルに、他の話者（以下、入力話者）の音声を変換の入力として用いた場合、音声の個人性に由来する音響空間のミスマッチが生じることになる。また、仮にモデル話者と入力話者の音響空間の間でミスマッチが生じなかったとしても、出力となる調音運動についての話者性でのミスマッチが生じる。パラレルデータ内の調音運動は、モデル話者の調音器官のサイズや形状を反映したものであり、それが入力話者と一致する保証はない。このことから、音声–調音マッピングは、音響空間（入力空間）と調音運動空間（出力空間）のそれぞれ、つまり、二重で話者依存性の制限を受けていると言える。モデル話者の音声–調音パラレルデータを大量に収集し、それを用いて高度な変換モデルを構築したとしても、その変換モデルの精度は、モデル話者の音声にしか保証さ

れない。しかし、2.4.2 節で述べたような様々な応用を想定すると、この話者依存性の問題は解決しなければならない問題の一つであると言える。

音響理論に基づく逆推定と音声–調音マッピングに共通する特性として、あくまでも推定対象となる音声の音響的な側面のみに注目して推定しているということが挙げられる。これは、一見当たり前のことように思えるが、従来の逆推定法の大きな制限の一つだと考えられる。音声の中には、言語情報、話者の情報、発話意図や感情など様々な情報が含まれている。しかし、従来の逆推定で用いられてきた情報は、どれも音声の音響的な（表面上の）振る舞いに関するもので、その中に含まれる情報に注目したものは検討されていない。従って、従来の推定法は、音声には様々な情報が含まれているのも関わらず音響情報のみに頼っており、その結果、音響情報が存在する音声にしか適用できないという制限がある。一方で、2.4.2 節の iii) で述べた発話トレーニングなどでは、学習者が学習対象となる言語を正確に発音した場合の調音運動を推定する技術が必要とされている。しかし、学習者は学習対象を正確に発音することは難しいため、推定対象の音声、つまり音響情報が存在しないことになる。このような推定問題に対して、従来の音声–調音マッピングを適用することは難しい。

以上を踏まえて、逆推定法の課題についてまとめる。

音響理論に基づく手法の課題であった連続発話を対象とした逆推定は、コーパスに基づく手法（音声–調音マッピング）によって、その推定精度が向上した。しかし、音声–調音マッピングには、以下の課題が残されている。

- 特定話者のデータから構築された変換モデルは話者依存モデルとなり、多様な話者の音声を対象とした推定が難しい。
- 音声の音響情報のみに注目した推定であるため、適用できる問題が限られている。

そこで、本研究では、これらの課題について検討をおこなう。本章に続く第 3 章では、話者変換の技術を用いて入力話者とモデル話者の音響空間のミスマッチを軽減することで、多様な話者の音声を入力とすることができる変換手法を提案し、第 4 章では、音声の言語的な情報を利用することで、音響情報が存在しない音声の調音運動の推定法を提案する。

2.6　まとめ

本章では、まず音声生成過程と、その中での調音運動の役割について述べた。さらに、調音運動を観測するために必要な調音観測システムを紹介した。次に、本研究のテーマとなる逆推定問題について、主な応用先と従来手法について説明した。そして、最後に本研究で検討する課題について述べた。

第3章

話者正規化音声–調音マッピング

3.1　はじめに

前章の 2.5 節で述べたように、音声–調音マッピングには、

- 特定話者のデータから構築された変換モデルは話者依存モデルとなり、多様な話者の音声を対象とした推定が難しい。

という問題がある。この問題の原因は、2.2.3 節で述べた音声と調音運動の個人性（話者性）にある。音声–調音マッピングは、音声–調音パラレルデータに基づき、音響空間と調音運動空間の対応（変換関係）をモデル化する。このとき特定の話者（モデル話者）のデータを用いた場合、その変換モデルで対応付けられるのは、モデル話者の音響空間と調音運動空間である。そして、それらの空間はモデル話者に特有のものである。その変換モデルに対して、他の話者（入力話者）の音声を入力した場合、その音響特徴量と変換モデルが学習したモデル話者の音響空間との間で音響的なミスマッチが生じ、結果として変換後の調音運動もミスマッチの影響を受けて大きな推定誤差を含んだものになる (図 3.1)。ここで、注意しなけらばならないのが、もし仮に、音響的なミスマッチが存在しないような入力話者の音声を用いたとしても、変換先の調音運動空間がモデル話者ものであるため、モデル話者の調音運動が出力されるということである。

この問題の最も単純な解決方法は、入力話者ごとに音声–調音パラレルデータを測定し、その話者の専用モデルを構築することである。入力話者の専用モデルであれば、モデル話者が入力話者自身であるため、先に述べたような音響的なミスマッチの問題は起こらず、正常に音声-調音マッピングがおこなえる。しかし、2.4.2 節で紹介した音声認識や発音トレーニングなどの多くのユーザーを想定するシステムへの応用を考えた場合、そのユーザーごとに専用モデルを構築するのは、音声–調音パラレルデータの測定コストを考えると現実的ではない。つまり、目標とするべきは、入力話者の調音運動のデータを必要としない手法である。

ここで、入力となる音声の音響特徴量に注目する。これまでに、音声の話者性を変換する技術として話者変換が検討されている [36]。話者変換は、ある話者によって発声された音声を、その言語的内容を保持したまま、他の話者が発声した音声に変換する技術である。この技術を用いて、入力音声と変換モデルの間に生じる音響的なミスマッチを解消することを目指す。問題となる音響的なミスマッチを軽減するために、音声–調音マッピングの前処理として話者変換を導入する。ここでの話者変換は、入力話者の音声をモデル話者の音声へと変換する話者変換である。したがって、話者変換後の音声（変換音声）は、モデル話者の話者性を持った音声となる。この変換音声を、音声–調音マッピングの入力音声として用いれば、音声–調音マッピングにおける変換モデルの入力空間（モデル話者の音響空間）と入力特徴量（変換音声）の話者性が一致するため、音響的なミスマッチの問題が軽減されると考えられる。

話者変換を用いて、入力話者の音声をモデル話者の音声に変換するという前処理は、任意の話者をモデル話者にする、いわば話者の正規化と言える。正規化後の音声に対する音声–調音マッピングは、モデル話者音声が入力された場合と同様に考えられるので、変換

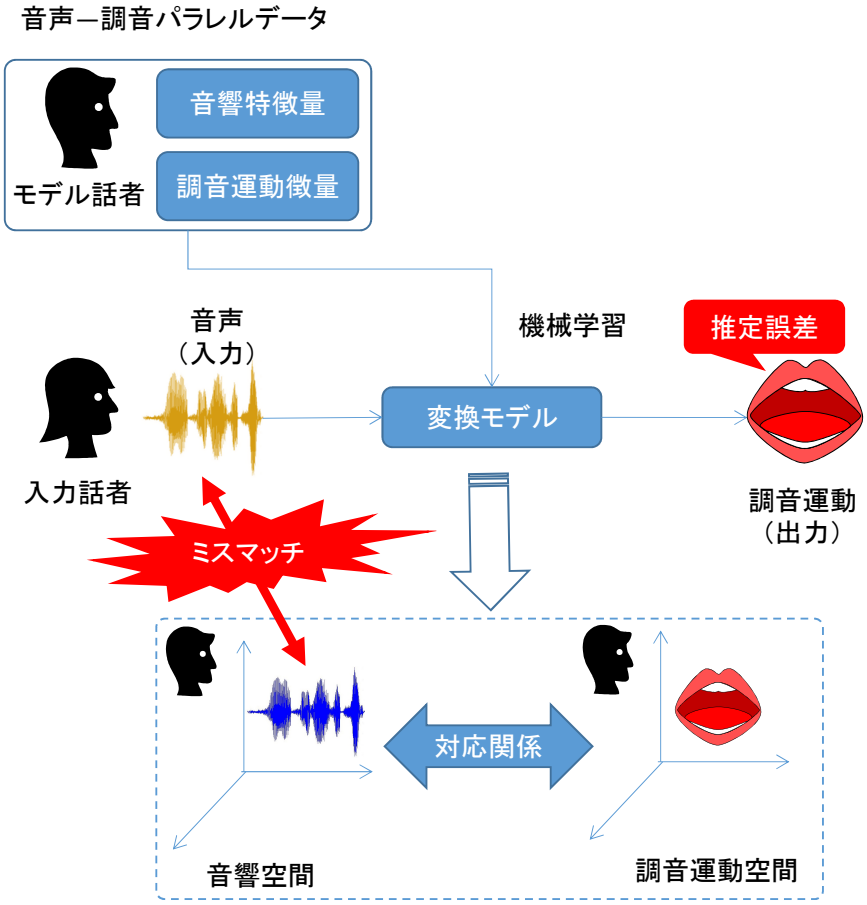

図 3.1: 音声–調音マッピングにおける話者依存性

によって得られる調音運動は、モデル話者の調音運動になる。つまり、話者変換によって、話者正規化をおこなった音声–調音マッピングは、

- 任意の話者の音声を特定の話者（モデル話者）の調音運動に変換するシステム

となる。このシステムの特徴は、入力の音声と出力の調音運動で話者性が異なっている点である。したがって、入力話者自身の調音運動が必要となる場合には、適用することが難しい。その場合は、モデル話者の調音運動から入力話者の調音運動への変換する後処理が新たに必要となる。しかし、2.4.2 節で紹介した調音運動情報を利用した音声認識や音声合成では、主に調音運動が保持する音声生成過程に関する情報に注目しており、その調音運動の話者性の再現まで厳密に求めることは多くない。つまり、音声–調音マッピングによって推定された調音運動が誰のものであるかに関わらず、音声生成過程に関する情報が十分に含まれていればよいである。したがって、「任意の話者の音声を特定の話者（モデル話者）の調音運動に変換するシステム」は、十分に意味のあるシステムと考えられる。

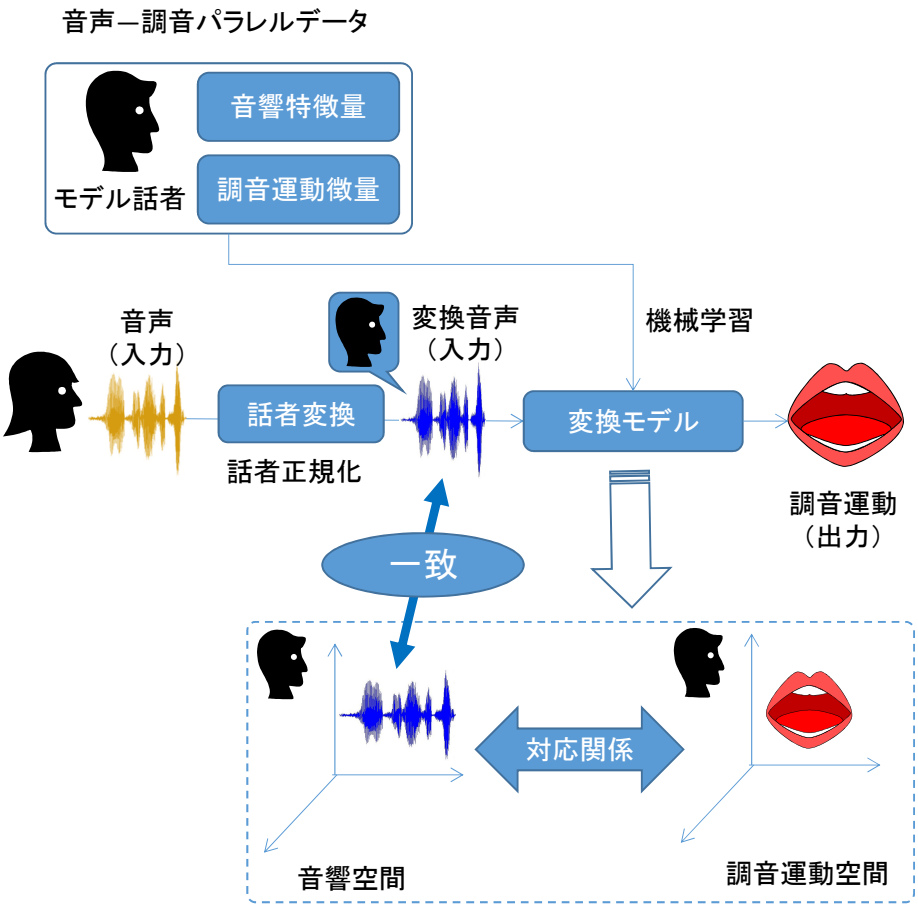

図 3.2: 話者正規化音声–調音マッピング

3.2　話者変換を利用した音声-調音変換モデルの話者正規化法

本章で検討するのは、話者変換によって入力話者を正規化した上で音声–調音マッピングをおこなう、話者正規化音声–調音マッピングである。このシステムは、話者変換と音声–調音マッピングという二つの要素技術からなる。本節では、それぞれの技術について説明したのち、それらの技術の統合し、話者正規化音声–調音マッピングを構築する方法について述べる。

3.2.1　話者変換と音声–調音マッピング

話者変換は、音声の言語的内容を保持したまま、話者性、つまり、誰の音声なのかという情報を変換する技術である。具体的には、変換対象となる話者（対象話者）の音声の音響特徴量を変換目標となる話者（目標話者）の音響特徴量へ変換する。したがって、特徴量を

変換するという点で、音声–調音マッピングと共通している。そのため、音声–調音マッピングで検討されてきた変換モデルは、話者変換においても同じように検討されている。特に、2.4.5 節で紹介した GMM による音声–調音マッピングで用いられている手法は、元々は話者変換で検討されていたものであり [37]、GMM を用いた話者変換と音声–調音マッピングは、変換する特徴量（話者変換では音響から音響、音声–調音マッピングでは音響から調音運動）が異なることを除けば、ほぼ同じ技術となっている。そこで、本章で検討する話者正規化音声–調音マッピングでは、話者変換と音声–調音マッピングの二つの特徴量変換器を GMM という共通の枠組みで考えることで、二つの変換器を合理的に統合することを目指す。

ここで、GMM を用いた話者変換を簡単に説明する。GMM を用いた話者変換では、式 (3.1) と同様に結合ベクトル $\boldsymbol{z}_t$ の GMM、

$$p(\boldsymbol{z};\lambda^{(z)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{3.1}$$

を考える。ここで、音声–調音マッピングでは結合ベクトル $\boldsymbol{z}_t$ が音響特徴量と調音運動特徴量の結合ベクトルであったのに対して、話者変換では、対象話者の音響特徴量 $\boldsymbol{x}_t = (\boldsymbol{x}_t(1), \boldsymbol{x}_t(2), \cdots, \boldsymbol{x}_t(d_x))^\top$ と目標話者の音響特徴量 $\boldsymbol{y}_t = (\boldsymbol{y}_t(1), \boldsymbol{y}_t(2), \cdots, \boldsymbol{y}_t(d_y))^\top$ の結合ベクトル $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{x}_t^\top]^\top$ となる。この GMM は、音声–調音マッピングと同様に変換の入力となる特徴量と出力となる特徴量の事前収録データ（音声–調音マッピングの場合は、音声–調音パラレルデータ）を用いて構築される。話者変換における事前収録データは、対象話者と目標話者の同一文読み上げ音声（音声–音声パラレルデータ）である。音声–調音パラレルデータは、音声と調音運動が同時に測定されるため、各特徴量における時間構造が一致している（時間対応がとれている）が、音声–音声パラレルデータの場合、発話速度が話者によって異なるため、各特徴量の時間構造が一致していない。そこで、動的時間伸縮法（Dynamic Time Warping, DTW) を用いて、 話者間の発話の時間構造の対応を求め、その時間対応に沿って、対象話者と目標話者の音響特徴量の結合ベクトルを用意する。話者間で時間構造の対応が取れた結合ベクトルを用いることで、発話内容を保持し話者性のみを変換する GMM が構築できる。話者変換式は、 音声–調音マッピングと同様に、式 (2.28) または式 (2.35) における $\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{z}_t$ を、それぞれ対象話者の音響特徴量、目標話者の音響特徴量、それら結合ベクトル、としたもので表される。

話者正規化音声–調音マッピングの大まかな変換の流れは、1)GMM 話者変換を用いて入力話者音声をモデル話者音声に変換し、その後、2)GMM 音声–調音マッピングよって変換音声をモデル話者調音運動に変換する、となる。GMM による変換では、入力特徴量の時間構造が保存されるため、本手法の変換系では、元々の入力音声の時間構造が最終的な出力である調音運動まで保存されることになる。したがって、最終的な出力は、入力音声の時間構造をもったモデル話者の調音運動となる。これは、モデル話者が入力話者の発話（入力音声）を真似した場合のモデル話者自身の調音運動と考えることができる。本研究では、この調音運動を、模擬調音運動と呼ぶことにする。前節で述べたように、模擬調音運動を推定すること、即ち、不特定話者の音声から特定の話者性を持つ調音運動を推定することは工学的に価値のあるものである。さらに、模擬調音運動の推定は、2.4.2 節の iii) で紹介

したような調音運動情報を利用した発音トレーニングにも応用できる。発音トレーニングの現場では、教師が学習者に対して、その学習者が行っている誤った発音を真似して示すことで、学習者の発音の誤りに対する気づきを促すということが、しばしばおこなわれている。ここで、本手法のモデル話者を教師、入力話者を学習者と考えれば、模擬調音運動は学習者の発音を教師が真似した場合の調音運動となる。したがって、模擬調音運動の推定を利用すれば、先に述べた発音トレーニングの例を調音運動に基づいておこなうことが可能になる。また、模擬調音運動の推定は、2.2.4 節で述べた音声知覚における運動説と関係がある。運動説において聴取者が音声から調音運動を復元する過程は、他話者の音声から聴取者自身の調音運動を復元しているので、これは聴取者による模擬調音運動の推定と言える。したがって、話者正規化音声–調音マッピングを構築することで、音声知覚における運動説に関する新たな知見が得られる可能性がある。

3.2.2　話者変換と音声ー調音変換の統合

本節では、話者変換モデルと音声–調音マッピングモデルを統合し、話者正規化音声–調音マッピングモデルを構築する方法について述べる。この変換モデルを構築する際に、a) モデル話者の音声–調音パラレルデータ及び、b) 任意の入力話者とモデル話者との音声–音声パラレルデータが存在する、というパラレルデータに関する条件を設ける。a) の音声–調音パラレルデータは、音声–調音マッピングモデルの構築に、b) の音声–音声パラレルデータは、話者変換モデルの構築にそれぞれ必要となる (図 3.3)。ここで、改めて、入力話者の調音運動データは用いないということを強調しておく。任意の入力話者に関して必要となるのは、モデル話者との音声–音声パラレルデータである。このデータの測定コストは読み上げ文と音声収録環境されあれば用意することができ、音声–音声パラレルデータに比べ圧倒的に低コストである。したがって、音声認識や発音トレーニングなど多数の入力話者を想定する技術へ応用することも十分に可能な条件と言える。

話者正規化音声–調音マッピングモデルの構築法として、1) モデル話者への話者変換と音声–調音変換を多段で適用し連結する変換手法（連結モデル）と、2) 話者変換と音声–調音マッピングを一つの変換モデルとして構築し、任意の入力話者の音声を調音運動に直接変換する手法（分布共有モデル）の 2 つについて検討する。

3.2.3　連結モデル

連結モデルでは、話者変換と音声–調音マッピングを別々の独立した変換モデルとしてそれぞれ GMM を用いて構築する。音声から調音運動へ変換する際は、入力音声に話者変換を適用し、その出力を音声–調音マッピングの入力として調音運動を求める。つまり、入力に対して二種類の変換が多段に作用することになる。

今、入力話者の音声特徴量を $\boldsymbol{x} \in \mathcal{R}^{d_x}$、モデル話者の音声特徴量及び調音運動特徴量を $\boldsymbol{y} \in \mathcal{R}^{d_y}$、$\boldsymbol{a} \in \mathcal{R}^{d_a}$ とし、音声–音声パラレルデータから構成される結合ベクトル $\boldsymbol{z}^{(xy)} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$、音声–調音パラレルデータから構成される結合ベクトルを $\boldsymbol{z}^{(ya)} = [\boldsymbol{y}^\top, \boldsymbol{a}^\top]^\top$ で

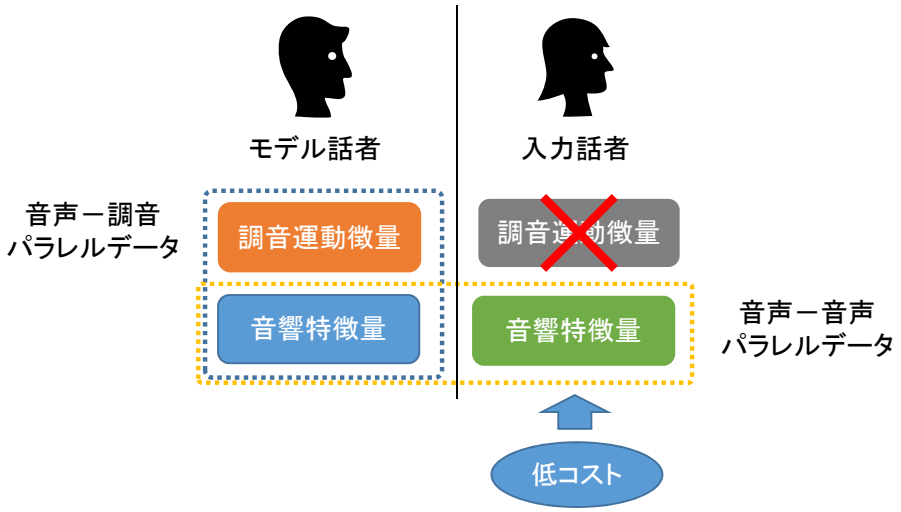

図 3.3: パラレルデータの条件

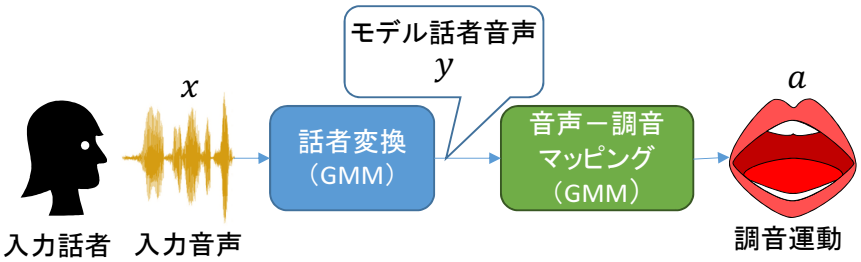

図 3.4: 連結モデル

表す。これらの結合ベクトルの確率密度を以下の二つの混合ガウス分布（GMM）で表す。

$$p(\boldsymbol{z}^{(xy)};\boldsymbol{\lambda}^{(xy)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}^{(xy)};\boldsymbol{\mu_m}^{(xy)},\boldsymbol{\Sigma}_m^{(xy)}) \tag{3.2}$$

$$p(\boldsymbol{z}^{(sa)};\boldsymbol{\lambda}^{(ya)}) = \sum_{n=1}^{N} \beta_n \mathcal{N}(\boldsymbol{z}^{(ya)};\boldsymbol{\mu_n}^{(ya)},\boldsymbol{\Sigma}_n^{(ya)}) \tag{3.3}$$

$$\boldsymbol{\mu}_m^{(xy)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix} \tag{3.4}$$

$$\boldsymbol{\Sigma}_m^{(xy)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(x,x)} & \boldsymbol{\Sigma}_m^{(x,y)} \\ \boldsymbol{\Sigma}_m^{(y,x)} & \boldsymbol{\Sigma}_m^{(y,y)} \end{bmatrix} \tag{3.5}$$

$$\boldsymbol{\mu}_n^{(ya)} = \begin{bmatrix} \boldsymbol{\mu}_n^{(y)} \\ \boldsymbol{\mu}_n^{(a)} \end{bmatrix} \tag{3.6}$$

$$\boldsymbol{\Sigma}_n^{(ya)} = \begin{bmatrix} \boldsymbol{\Sigma}_n^{(y,y)} & \boldsymbol{\Sigma}_n^{(y,a)} \\ \boldsymbol{\Sigma}_n^{(a,y)} & \boldsymbol{\Sigma}_n^{(a,a)} \end{bmatrix} \tag{3.7}$$

ここで、二つの GMM からそれぞれ求まる $\boldsymbol{y}$ の周辺確率分布 $p(\boldsymbol{y}; \lambda^{(xy)})$ および $p(\boldsymbol{y}; \lambda^{(ya)})$ は、同じ特徴量に関する分布だが、異なる分布であることに気を付けなければならない。この分布の違いは、次節で述べる分布共有モデルの構築で重要となる。

話者変換では式 (3.2) の GMM を、音声–調音マッピングでは式 (3.3) の GMM を用いて変換をおこなう。入力話者の音声を調音運動に変換する場合は、まず、入力話者の音声と話者変換モデルを用いて、以下の式に従って話者変換音声 $\hat{\boldsymbol{y}}$ を求める。

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\lambda}^{(xy)}) \tag{3.8}$$

この $\hat{\boldsymbol{y}}$ は、入力話者の音声をモデル話者の音声に変換したものである。次に、この変換音声を音声–調音変換モデルの入力として、以下の式に従って調音運動に変換する。

$$\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}} P(\boldsymbol{a}|\hat{\boldsymbol{y}}; \boldsymbol{\lambda}^{(ya)}) \tag{3.9}$$

この手法では、入力話者の音声に対して、音声変換と音声-調音運動変換という 2 つの変換が多段に作用することになる。このとき、話者変換での変換誤差を含んだ $\hat{\boldsymbol{y}}$ を音声–音声調音マッピングの入力に用いるため、最終的な出力である調音運動に話者変換の変換誤差と音声–音声調音マッピングの変換誤差が蓄積する可能性がある。

3.2.4　分布共有モデル

分布共有モデルでは、二つの変換モデルによる多段の適用が変換誤差の蓄積を招く可能性がある。そこで、話者変換と音声-調音変換を統合し、一つの変換モデルとして構築することによって、話者変換音声を経由せずに、入力話者の音声を調音運動へ一度に変換する手法を検討する。

前節の連結モデルで述べたように、式 (3.2) と式 (3.3) で表される 2 つの GMM には、部分空間としてモデル話者の音声特徴量 $\boldsymbol{y}$ の空間が存在しているが、それらの分布は 2 つの GMM で異なったものとなっている。分布共有モデルでは、まず、これらの部分空間が共通のモデルによって表されていると考える（分布共有）[38]。そして、2 つの GMM に共通

図 3.5: 分布共有モデル

するモデル話者の音声特徴量を周辺化することで、入力話者の音声を調音運動に直接変換するモデルを構築する。

式 (3.2) 及び、式 (3.3) において、モデル話者の音声特徴量 $\boldsymbol{y}$ に関する部分空間が共通のモデルで表されている場合、式 (3.2) 及び、式 (3.3) は、以下のように改められる。

$$p(\boldsymbol{z}^{(xy)};\boldsymbol{\lambda}^{(xy)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}^{(xy)};\boldsymbol{\mu_m}^{(xy)},\boldsymbol{\Sigma}_m^{(xy)}) \tag{3.10}$$

$$p(\boldsymbol{z}^{(sa)};\boldsymbol{\lambda}^{(ya)}) = \sum_{n=1}^{M} \beta_n \mathcal{N}(\boldsymbol{z}^{(ya)};\boldsymbol{\mu_m}^{(ya)},\boldsymbol{\Sigma}_m^{(ya)}) \tag{3.11}$$

この二つの GMM は、混合成分のインデックス (隠れ変数)m を共有しており、各混合におけるモデル話者の音声特徴量 $\boldsymbol{y}$ に関する部分空間が一致している。このとき、入力話者の音声から調音運動への変換式を求めるために、入力話者の音声特徴量 $\boldsymbol{x}$ が与えられたときの調音運動 $\boldsymbol{a}$ の条件付き確率分布 $p(\boldsymbol{a}|\boldsymbol{x})$ を考える。$p(\boldsymbol{a}|\boldsymbol{x})$ を、混合成分 m 及び、モデル話者の音声特徴量 $\boldsymbol{y}$ について展開すると、

$$p(\boldsymbol{a}|\boldsymbol{x}) = \sum_{m=1}^{M} \int_{\mathcal{Y}} p(\boldsymbol{a},\boldsymbol{y},m|\boldsymbol{x}) d\boldsymbol{y} \tag{3.12}$$

$$= \sum_{m=1}^{M} p(m|\boldsymbol{x}) \int_{\mathcal{Y}} p(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{x},m) p(\boldsymbol{y}|\boldsymbol{x},m) d\boldsymbol{y} \tag{3.13}$$

となる。ここで、

$$p(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{x},m) \approx p(\boldsymbol{a}|\boldsymbol{y},m) \tag{3.14}$$

の近似を導入すると、

$$p(\boldsymbol{a}|\boldsymbol{x}) \approx \sum_{m=1}^{M} p(m|\boldsymbol{x}) \int_{\mathcal{Y}} p(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{x},m) p(\boldsymbol{y}|\boldsymbol{x},m) d\boldsymbol{y} \tag{3.15}$$

となる。$\boldsymbol{y}$ の全積分をおこなうと、

$$p(\boldsymbol{a}|\boldsymbol{x}) \approx \sum_{m=1}^{M} p(m|\boldsymbol{x})\mathcal{N}(\boldsymbol{a}; \boldsymbol{E}_m^{(a|x)}, \boldsymbol{D}_m^{(a|x)}) \tag{3.16}$$

を得る。ここで、

$$\boldsymbol{E}_m^{(a|x)} = \boldsymbol{\mu}_m^{(a)} + \boldsymbol{\Sigma}'_m(\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)}) \tag{3.17}$$

$$\boldsymbol{D}_m^{(a|x)} = \boldsymbol{\Sigma}_m^{(a,a)} - \boldsymbol{\Sigma}'_m\boldsymbol{\Sigma}_m^{(x,x)}\boldsymbol{\Sigma}'_m{}^{\top} \tag{3.18}$$

$$\boldsymbol{\Sigma}'_m = \boldsymbol{\Sigma}_m^{(a,y)}\boldsymbol{\Sigma}_m^{(y,y)^{-1}}\boldsymbol{\Sigma}_m^{(y,x)} \tag{3.19}$$

である。この条件付き分布 $p(\boldsymbol{a}|\boldsymbol{x})$ を用いて変換式を求めることができる。2.4.5 節の ii) で述べた MLE 基準の変換式では、

$$\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}} p(\boldsymbol{a}|\boldsymbol{x}) \tag{3.20}$$

となり、式 (2.37) と同じ形式で表すことができる。この分布共有モデルでは、調音運動 $\boldsymbol{y}$ を、入力音声 $\boldsymbol{x}$ に対する条件付き分布 $p(\boldsymbol{a}|\boldsymbol{x})$ を用いて推定する。式 (3.17) で示す通り、この $p(\boldsymbol{a}|\boldsymbol{x})$ のパラメータは、$\boldsymbol{x}$ の関数として表されているため、変換式も $\boldsymbol{x}$ の関数となっている (式 (3.20))。この変換過程では、連結モデルと異なり、モデル話者の音声 $\boldsymbol{y}$ は明示的に経由しない。式 (3.15) における、モデル話者音声 $\boldsymbol{y}$ に関する周辺化によって、入力話者の音声から変換され得る全てのモデル話者音声が確率的に考慮されることになる。その影響は、式 (3.19) で表される共分散行列の変換によって記述されている。

3.2.5　分布共有モデルの構築法

前節で述べた分布共有モデルでは、話者変換モデルと音声–調音変換モデルの間の GMM において、共通する特徴量であるモデル話者音声の分布が、各混合成分おいて一致している必要がある。しかし、一部の（分布）パラメータが一致するように二つの GMM を個別に構築することは困難である。そこで、本手法では、話者変換モデルと音声–調音変換モデルを個別に構築するのではなく、一つの GMM によって構築する方法について検討する。今、入力話者の音声特徴量およびモデル話者の音声特徴量と調音特徴量（三種類の特徴量）の結合ベクトル $\boldsymbol{z}^{(x,y,a)} = [\boldsymbol{x}^{\top}, \boldsymbol{y}^{\top}, \boldsymbol{a}^{\top}]^{\top}$ を考える。この結合ベクトルの GMM は、

$$p(\boldsymbol{z}^{(x,y,a)}; \boldsymbol{\lambda}^{(x,y,a)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}^{(x,y,a)}; \boldsymbol{\mu_m}^{(x,y,a)}, \boldsymbol{\Sigma}_m^{(x,y,a)}) \tag{3.21}$$

で表され、各混合における平均ベクトル及び分散共分散行列は、図 3.6 のようになる。図 3.6 に示すように、この GMM では話者変換を表す部分空間と音声–調音変換を表す部分空間がモデル話者の音声特徴量の部分空間を共有した形で構成されている。つまり、$p(\boldsymbol{z}^{(x,y,a)}; \boldsymbol{\lambda}^{(x,y,a)})$ を $\boldsymbol{a}$ について周辺化すれば、$p(\boldsymbol{z}^{(x,y)}; \boldsymbol{\lambda}^{(x,y)})$ となり、これは、式 (3.10) で表される話者変

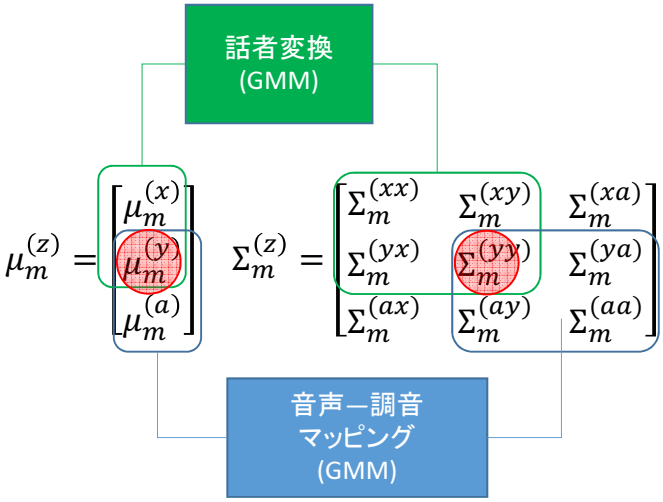

図 3.6: 結合ベクトル $\boldsymbol{z}$ から得られる GMM における、一つの混合成分の平均ベクトルと分散共分散行列。赤は、話者変換と音声–調音マッピングで共有するべきモデル話者の音声に関する分布パラメータである。

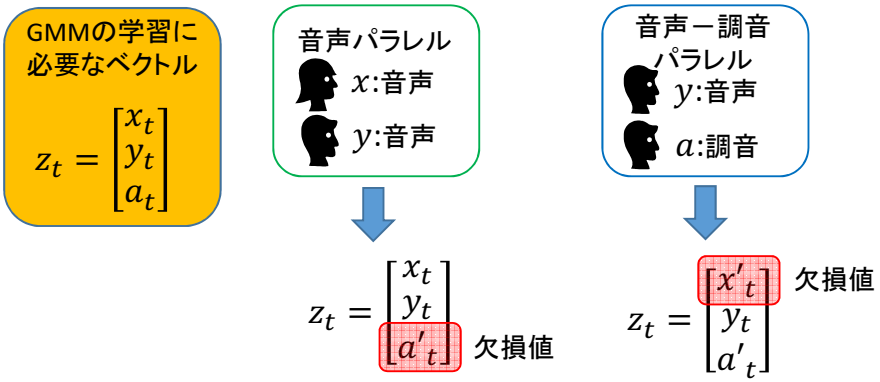

図 3.7: パラレルデータに生じる欠損値

換モデルの GMM となる。一方で、$\boldsymbol{x}$ について周辺化すれば、$p(\boldsymbol{z}^{(y,a)}; \boldsymbol{\lambda}^{(y,a)})$ となり、これは、式 (3.11) で表される音声–調音マッピングモデルの GMM となる。

式 (3.21) の GMM は、a) 入力話者の音声とモデル話者の音声–音声パラレルデータ、b) モデル話者の音声–調音パラレルデータを用いて、結合ベクトル $\boldsymbol{z}^{(x,y,a)}$ を用意することで構築できるが、そのためには、パラレルデータ a) 及び b) の発話内容（読み上げ文）が一

致している必要がある。しかし、実際的な応用を考えると、パラレルデータ a) 及び b) の間で読み上げ文を一致させることが難しい場合がある。前述したように、発音トレーニングに応用する場合、モデル話者は語学教師、入力話者は生徒（学習者）となる。このとき、パラレルデータ a) は生徒の母国語、パラレルデータ b) は学習の対象となる言語を含むことが望ましい。その場合、パラレルデータ a) 及び b) は、それぞれ異なった言語の読み上げ文から構成されることになる。その結果、これらのコーパスから用意できる結合ベクトルは、a) からは $\boldsymbol{z}^{(x,y)} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$、b) からは $\boldsymbol{z}^{(x,y)} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ となる。所望の $\boldsymbol{z}^{(x,y,a)}$ と比較すると、それぞれに欠損値が存在することがわかる。つまり、a) においては調音運動 $\boldsymbol{a}$ が、b) においては入力話者音声 $\boldsymbol{x}$ が、観測されない値となる (図 3.7)。

そこで、二つのパラレルデータから得られる結合ベクトル、$\boldsymbol{z}_t^{(x,y)} = [{\boldsymbol{x}_t}^\top, {\boldsymbol{y}_t}^\top]^\top$ および、$\boldsymbol{z}_t^{(y,a)} = [{\boldsymbol{y}_t}^\top, {\boldsymbol{a}_t}^\top]^\top$ に対して、擬似的な結合ベクトル、${\boldsymbol{z}_t}^{(xya')} = [{\boldsymbol{x}_t}^\top, {\boldsymbol{y}_t}^\top, {\boldsymbol{a'}_t}^\top]^\top (t = 1 \sim T_1)$,${\boldsymbol{z}_t}^{(x',y,a)} = [{\boldsymbol{x'}_t}^\top, {\boldsymbol{y}_t}^\top, {\boldsymbol{a}_t}^\top]^\top (t = T_2 \sim T_3)$ を考える。ここで、$\boldsymbol{a'}, \boldsymbol{x'}$ がそれぞれのコーパスの欠損値、t は時間インデックスである。この欠損値を含む二つの結合ベクトルを時間軸方向に並べたベクトル、$\boldsymbol{z}^{(x,y,a)'} = \{\boldsymbol{z}_1^{(x,y,a')}, \cdots, \boldsymbol{z}_{T_1}^{(x,y,a')}, \boldsymbol{z}_{T_2}^{(x',y,a)}, \cdots, \boldsymbol{z}_{T_3}^{(x'ya)}\}$ について、GMM を構築する。GMM の構築には、EM アルゴリズムを用いる。EM アルゴリズムでは、以下の補助関数 (Q 関数) を最大化する。

$$
\begin{aligned}
Q(\theta, \theta^{old}) = &\sum_{m=1}^{M} \int p(m, \boldsymbol{A'} | \boldsymbol{X}, \boldsymbol{Y}^{(1)}, \theta^{old}) \log p(\boldsymbol{X}, \boldsymbol{Y}^{(1)}, \boldsymbol{A'} | m, \theta) d\boldsymbol{a} \\
&+ \sum_{m=1}^{M} \int p(m, \boldsymbol{X'} | \boldsymbol{Y}^{(2)}, \boldsymbol{A}, \theta^{old}) \log p(\boldsymbol{X'}, \boldsymbol{Y}^{(2)}, \boldsymbol{A} | m, \theta) d\boldsymbol{x}
\end{aligned} \tag{3.22}
$$

ここで、

$$\boldsymbol{X} = \{{\boldsymbol{x}_1}^\top, \cdots, {\boldsymbol{x}_{T_1}}^\top\} \tag{3.23}$$
$$\boldsymbol{Y}^{(1)} = \{{\boldsymbol{y}_1}^\top, \cdots, {\boldsymbol{y}_{T_1}}^\top\} \tag{3.24}$$
$$\boldsymbol{Y}^{(2)} = \{{\boldsymbol{y}_{T_2}}^\top, \cdots, {\boldsymbol{y'}_{T_3}}^\top\} \tag{3.25}$$
$$\boldsymbol{A} = \{{\boldsymbol{a}_{T_2}}^\top, \cdots, {\boldsymbol{a}_{T_3}}^\top\} \tag{3.26}$$
$$\boldsymbol{A'} = \{{\boldsymbol{a'}_1}^\top, \cdots, {\boldsymbol{a'}_{T_1}}^\top\} \tag{3.27}$$
$$\boldsymbol{X'} = \{{\boldsymbol{x'}_{T_2}}^\top, \cdots, {\boldsymbol{x'}_{T_3}}^\top\} \tag{3.28}$$

である。また θ は、GMM のパラメータである。この補助関数は、以下の E ステップと M ステップを繰り返し計算することで最大化される。

E ステップ

$$\gamma_{m,t}^{(x,y,a')} = \frac{\pi_m \mathcal{N}(\boldsymbol{z}_t^{(x,y,a')} | {\boldsymbol{\mu}_m}^{(x,y,a)}, {\boldsymbol{\Sigma_m}}^{(x,y,a)})}{\sum_{j=1}^{M} \pi_j \mathcal{N}(\boldsymbol{z}_t^{(x,y,a')} | {\boldsymbol{\mu}_j}^{(x,y,a)}, {\boldsymbol{\Sigma_j}}^{(x,y,a)})} \tag{3.29}$$

$$\gamma_{m,t}^{(x',y,a)} = \frac{\pi_m \mathcal{N}(\boldsymbol{z}_t^{(x',y,a)} | {\boldsymbol{\mu}_m}^{(x,y,a)}, {\boldsymbol{\Sigma_m}}^{(x,y,a)})}{\sum_{j=1}^{M} \pi_j \mathcal{N}(\boldsymbol{z}_t^{(x',y,a)} | {\boldsymbol{\mu}_j}^{(x,y,a)}, {\boldsymbol{\Sigma_j}}^{(x,y,a)})} \tag{3.30}$$

M ステップ

$$\boldsymbol{\mu}_m = \frac{1}{\gamma_m}\left(\sum_{t=1}^{T_1} {\gamma_{m,t}}^{(x,y,a')}\boldsymbol{z_t}^{(x,y,a')} + \sum_{t=T_2}^{T_3} {\gamma_{m,t}}^{(x',y,a)}\boldsymbol{z_t}^{(x',y,a)}\right) \tag{3.31}$$

$$\begin{aligned}\boldsymbol{\Sigma}_m = \frac{1}{\gamma_m}\Bigg(&\sum_{t=1}^{T_1} {\gamma_{m,t}}^{(x,y,a')}\left\{(\boldsymbol{z_t}^{(x,y,a')} - \boldsymbol{\mu}_m)(\boldsymbol{z_t}^{(x,y,a')} - \boldsymbol{\mu}_m)^\top - {\hat{\boldsymbol{D}}_m}^{(x,y,a')}\right\} \\ &+ \sum_{t=T_2}^{T_3} {\gamma_{m,t}}^{(x,y,a')}\left\{(\boldsymbol{z_t}^{(x,y,a')} - \boldsymbol{\mu}_m)(\boldsymbol{z_t}^{(x,y,a')} - \boldsymbol{\mu}_m)^T - {\hat{\boldsymbol{D}}_m}^{(x,y,a')}\right\}\Bigg)\end{aligned} \tag{3.32}$$

$$\pi_m = \frac{\gamma_m}{\sum_{m=1}^{M}\gamma_m} \tag{3.33}$$

ここで、$\gamma_m = \Sigma_{t=1}^{T_1}\gamma_{m,t}^{(x,y,a')} + \Sigma_{t=1}^{T_2}\gamma_{m,t}^{(x',y,a)}$ である。E ステップにおける欠損値 $\boldsymbol{a'}_t, {\boldsymbol{x'}_t}^{(s)}$ は、$\boldsymbol{a}_t, \boldsymbol{x}_t$ の推定値である。各推定値は、測定値の結合ベクトル $\boldsymbol{z}_t^{(x,y)}, \boldsymbol{z}_t^{(y,a)}$ と GMM の部分空間を用いて以下の式によって求めることが出来る。

$$\boldsymbol{a'}_t = \sum_{m=1}^{M} p(m|\boldsymbol{z}_t^{(x,y)}; \boldsymbol{\mu}^{(x,y)}, \boldsymbol{\Sigma}^{(x,y)})E_{m,t}^{(a|x,y)} \tag{3.34}$$

$$\boldsymbol{x'}_t = \sum_{m=1}^{M} p(m|\boldsymbol{z}_t^{(y,a)}; \boldsymbol{\mu}^{(y,a)}, \boldsymbol{\Sigma}^{(y,a)})E_{m,t}^{(x|y,a)} \tag{3.35}$$

ここで、$\boldsymbol{\mu}_m^{(x,y)}, \boldsymbol{\Sigma}_m^{(x,y)}$ は、結合ベクトル $\boldsymbol{z}^{(x,y)}$ の平均ベクトルと分散共分散行列、$\boldsymbol{\mu}_m^{(y,a)}, \boldsymbol{\Sigma}_m^{(y,a)}$ は、結合ベクトル $\boldsymbol{z}^{(ya)}$ に関する平均ベクトルと分散共分散行列である。また、

$$E_{m,t}^{(a|x,y)} = \boldsymbol{\mu}_m^{(a)} + \boldsymbol{\Sigma}_m^{(a,xy)}{\boldsymbol{\Sigma}_m^{(xy,xy)}}^{-1}(\boldsymbol{z}_t^{(x,y)} - \boldsymbol{\mu}_m^{(x,y)}) \tag{3.36}$$

$$E_{m,t}^{(x|y,a)} = \boldsymbol{\mu}_m^{(x)} + \boldsymbol{\Sigma}_m^{(x,ya)}{\boldsymbol{\Sigma}_m^{(ya,ya)}}^{-1}(\boldsymbol{z}_t^{(ya)} - \boldsymbol{\mu}_m^{(ya)}) \tag{3.37}$$

である。ここで、$\Sigma_m^{(a,xy)}, \Sigma_m^{(x,ya)}$ は、それぞれ $\boldsymbol{a}$ と $\boldsymbol{z}^{(xy)}$、$\boldsymbol{y}$ と $\boldsymbol{z}^{(ya)}$ の共分散行列、$\boldsymbol{\Sigma}_m^{(xy,xy)}, \boldsymbol{\Sigma}_m^{(ya,ya)}$ は、それぞれ $\boldsymbol{z}^{(xy)}$,$\boldsymbol{z}^{(ya)}$ の分散共分散行列である。

M ステップにおける欠損値 $\boldsymbol{a'}_t, \boldsymbol{x'}_t$ は、それぞれ式 (3.36), 式 (3.37) で表される平均ベクトルで置き換えられる。また、式 (3.32) における ${\hat{\boldsymbol{D}}_m}^{(xya')}, {\hat{\boldsymbol{D}}_m}^{(x'ya)}$ は、以下の行列である。

$${\hat{\boldsymbol{D}}_m}^{(x,y,a')} = \begin{bmatrix} \boldsymbol{0}^{(d_1,d_1)} & \boldsymbol{0}^{(d_1,d_2)} \\ \boldsymbol{0}^{(d_2,d_1)} & \boldsymbol{D}_m^{(a|x,y)} \end{bmatrix} \tag{3.38}$$

$${\hat{\boldsymbol{D}}_m}^{(x',y,a)} = \begin{bmatrix} \boldsymbol{D}_m^{(x|y,a)} & \boldsymbol{0}^{(d_3,d_4)} \\ \boldsymbol{0}^{(d_4,d_3)} & \boldsymbol{0}^{(d_4,d_4)} \end{bmatrix} \tag{3.39}$$

ここで、

$$\boldsymbol{D}_m^{(a|x,y)} = \Sigma_m^{(a,a)} - \Sigma_m^{(a,xy)}{\Sigma_m^{(xy,xy)}}^{-1}\Sigma_m^{(xy,a)} \tag{3.40}$$

$$\boldsymbol{D}_m^{(x|y,a)} = \Sigma_m^{(x,x)} - \Sigma_m^{(x,ya)}{\Sigma_m^{(ya,ya)}}^{-1}\Sigma_m^{(ya,x)} \tag{3.41}$$

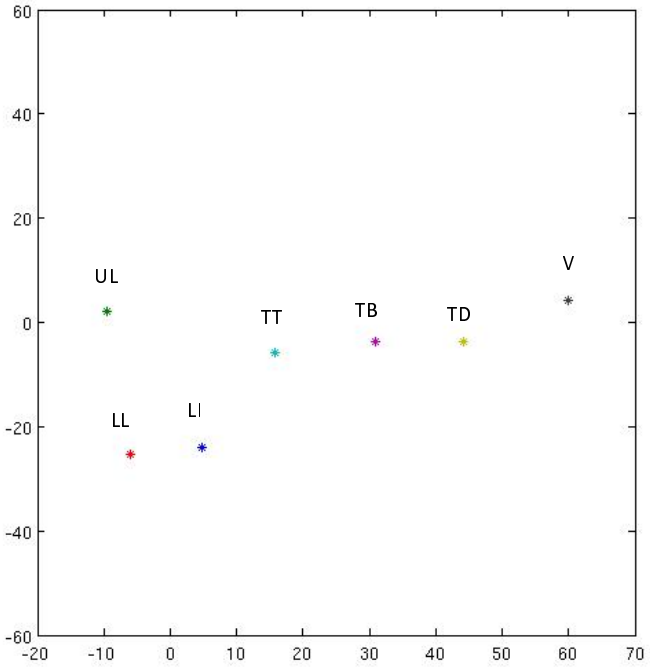

図 3.8: MOCHA-TIMIT に収録されている EMA の測定データの特定のサンプルにおける 7 個のセンサー (LL=下唇、UL・LL=上口唇・下口唇、TT・TB・TD=舌尖・舌背・舌の後方、V=軟口蓋）の位置を正中矢状面に沿ってプロットした図。横軸が水平方向、縦軸が垂直方向に対応する。各調音器官の大まかな位置は把握できるが、形状や他の器官との関係性は表現されていない。

である。また、$0^{(m,l)}$ は m 行 l 列の零行列であり、$d_1 = d_x + d_y, d_2 = d_a, d_3 = d_a, d_4 = d_y + d_a$ である。この分布共有モデルは、基本的な構造は通常の GMM を用いた特徴量変換と同様なので、2.4.5 節の iii) で述べた動的成分を考慮した最尤系列マッピングに拡張することができる。

3.3　調音モデルの構築

音声-調音マッピングにおいて、一般的に調音運動の特徴量として用いられるのは EMA(2.3.3 節) の測定データに基づく特徴量である（[11]）。EMA の測定データは、幾つかの調音器官に取り付けられたセンサーの座標である（図 2.7)。したがって、マッピングの出力として得られるのも、座標データとなる。座標データは、各調音器官の運動を表すものだが、センサー（測定点）の数は少なく（例えば、舌上のセンサーは多くて 3 点)、その数点の座標データだけでは各調音器官の形状や他の器官（口蓋や声道壁）の形状は解らず、調音運動を直感的に理解することは難しい (図 3.8)。しかし、座標データから各調音器官やその他の

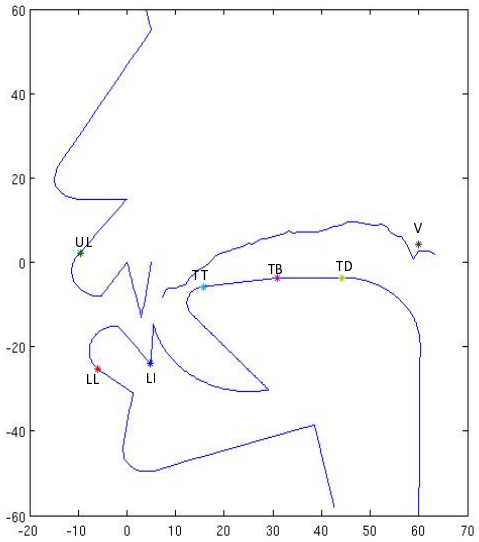

図 3.9: 調音モデル

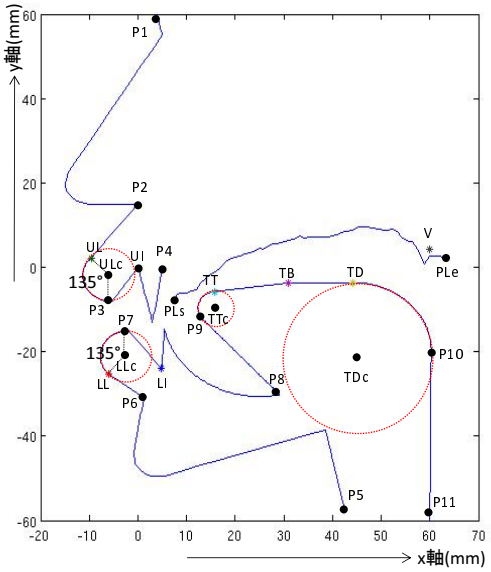

図 3.10: 調音モデルの構築

器官（口蓋や声道壁）の形状を可視化できれば、マッピング結果の妥当性を視覚的（直感的）に評価することができる。

少ないパラメータで調音器官の形状を表現する手法として、幾何的調音モデルが検討されている [39]。幾何的調音モデルでは、調音器官や声道壁の形状を単純な図形で近似し、それらの図形に対して駆動点を与えることで調音器官の運動・変形を表現する。本研究では、Birkholz ら（2006）のモデルを参考にして、図 3.9 に示すモデルを構築した。この調音モデルは、MOCHA-TIMIT の測定データから調音器官の形状を可視化することを目的としたモデルである。MOCHA-TIMIT の 7 点の測定点から軟口蓋を除いた 6 個の測定点が駆動点となって、調音器官の運動・変形を表現することができる。

提案した調音モデルは、図 3.9 に示すように円と直線によって構築されている。調音モデルの構築法を部位ごとに説明する。

舌

舌は、点 P8 から点 P11 の線によって表される。点 TT,TB,TD は EMA データの測定点である。点 TTc は点 TT と同じ x 座標に存在し、円 TTc の中心点となっている。点 P8 は、舌の付け根を表し、点 P9 は点 P8 から円 TTc の接線を引いた時の接点である。測定点 TT と測定点 TB は直線で結ばれる。点 TDc は点 TD は同じ x 軸座標に存在し、円 TDc の中心点である。点 P11 は舌根を表し、点 P10 は点 P11 から円 TDc の接線を引いた時の接点である。測定点 TB と測定点 TD は直線で結ばれる。曲線 LI-P8、点 P11 の座標、円 TTc,TDc の半径は調整パラメータであり、測定データから決定される。

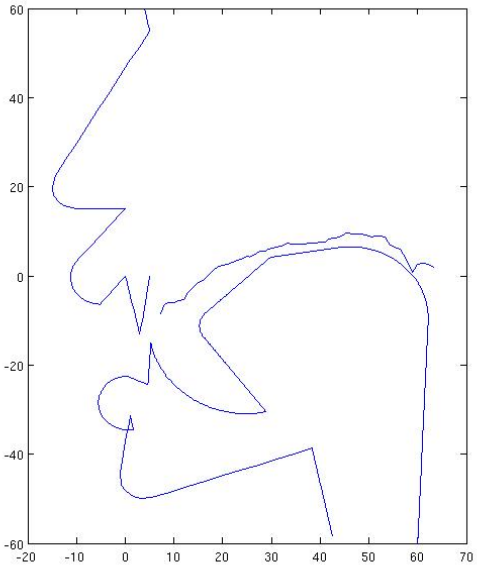

図 3.11: /i/発音時の様子

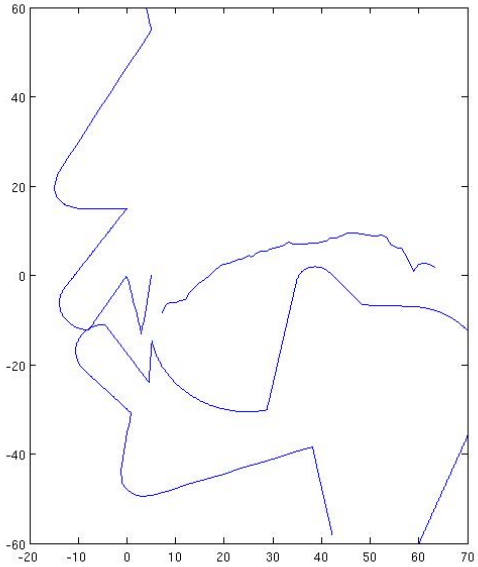

図 3.12: /b/発音時の様子

顎・下口唇・下前歯

顎、下口唇、及び、下前歯は点 P5 から点 P8 の線によって表される。測定点 LL および点 P7 は、点 LLc を中心点とする円上に存在し、点 P7,LLc,LL からなる角の優角は 135 度である。また、点 P7 と点 LLc は同じ x 座標に存在する。下口唇の付け根を表す点 P6 と測定点 LL、及び、点 P7 と測定点 LI はそれぞれ直線で結ばれる。点 P5,P6 の座標, 円 LLc の半径, 曲線 P5-P6, 曲線 LI-P8 の形状は調整パラメータである。

鼻・上口唇・上前歯

鼻、上口唇、及び、上前歯は点 P1 から点 P4 の線によって表される。測定点 UL および点 P3 は、点 ULc を中心点とする円上に存在し、点 P3,ULc,UL からなる角の優角は 135 度である。また、点 P3 と点 ULc は同じ x 座標に存在する。鼻の付け根を表す点 P2 と測定点 UL、及び、点 P3 と点 UI はそれぞれ直線で結ばれる。点 P2,P4 の座標, 円 ULc の半径, 曲線 P1-P2, 曲線 UI-P4 の形状は調整パラメータである。点 PLs から点 PLe までの曲線は口蓋の形状を表しており、その形状は調整パラメータである。また、点 UI は、上前歯の付け根を表し、座標系の原点となる。

各調整パラメータは、MOCHA-TIMIT に収録されている全測定データを考慮して、手作業で決定される（MOCHA-TIMIT の男性話者に対する調音モデルの調整パラメータを付録 B に示す)。MOCHA-TIMIT の男性話者に対する調音モデルを実際に測定データ（実測データ）を用いて駆動した例を図 3.11 と図 3.12 に示す。図 3.11 は短文中の単語 “thieve” の音素/i/発音時の測定データで調音モデルを駆動した結果である。/i/は舌背を口蓋に接近させて声道内に狭めを作ることで発音される音素であるが、図 3.11 の調音モデルでも舌背が

口蓋に接近し狭めが形成されていることがわかる。また、図 3.12 は短文中の単語 “bright” の音素/b/発音時の測定データで調音モデルを駆動した結果である。/b/は、上下の口唇で閉鎖を作り、そこで破裂音を生成することで発音される音素である。図 3.12 の調音モデルでも口唇が接触していることがわかる。さらに、舌の形状に注目すると、後続の音素である/r/を発音するために、舌の先端が口蓋に向かってせり上がっている様子が表現されている。これらの結果から、構築した調音モデルは EMA の調音運動データから口腔内の様子を、十分な妥当性をもって可視化できていると言える。

3.4　実験

本章で提案した話者正規化音声–調音マッピングの性能を確かめるために、2.4.8 節の i) で紹介した MOCHA-TIMIT を用いて、音声から調音運動への変換実験をおこなう。MOCHA-TIMIT には、話者 2 名分の音声–調音パラレルデータが収録されているため、そのいずれかの話者を入力話者、もう一方の話者をモデル話者とした実験をおこなうことができる。

3.4.1　データの準備

本実験では、音声–調音パラレルデータとして MOCHA-TIMIT を用いるが、2.4.8 節の i) で述べたように、MOCHA-TIMIT には、調音運動データの質に問題がある。そこで、実験の前にいくつかの前処理をおこなう。

まず、調音運動データに混入した時間変化するバイアス成分の除去をおこなう。図 3.13 に、調音運動データの例を示す (図 2.18 の再掲)。図 3.13 は、女性話者の舌尖のセンサーの水平軸方向の座標を各発話で平均した値をプロットしたものである。横軸が発話番号（1〜460)、縦軸が座標の値 (0.01mm) である。発話が進むごとに、値が上昇していることがわかる。この時間変化するバイアス成分を取り除くために、先行研究 [32] にしたがって、以下の前処理をおこなう。

まず最初に、i 番目の発話における調音運動データの平均値 $\bar{\boldsymbol{a}}^{(i)}$、

$$\bar{\boldsymbol{a}}^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} \boldsymbol{a}_t^{(i)} \tag{3.42}$$

を求める。このとき、T_i は i 番目の発話の長さ (サンプル数) であり、$\boldsymbol{a}_t^{(i)}$ は、i 番目の発話の t 番目のサンプルにおける調音運動データである。調音運動データは、7 つの測定点における 2 次元座標データであるため、14 次元のベクトルとなる（$\boldsymbol{a}_t^{(i)} \in \mathcal{R}^{d_a=14}$)。次に、$\bar{\boldsymbol{a}}^{(i)}$ を全発話にそって並べた系列データ $\bar{\boldsymbol{A}} = \{\bar{\boldsymbol{a}}^{(1)}, \cdots, \bar{\boldsymbol{a}}^{(460)}\}$ を考える。この系列データの時間（発話番号）方向にローパスフィルタを作用させる。本実験では、ローパスフィルタとして、以下の移動平均フィルタに基づいた処理をおこなった。

$$\mathcal{F}(\bar{\boldsymbol{a}}^{(i)}) = \frac{1}{10} \sum_{n=0}^{10} \frac{1}{2} (\bar{\boldsymbol{a}}^{(i-n+1)} + \bar{\boldsymbol{a}}^{(i+n-1)}) \tag{3.43}$$

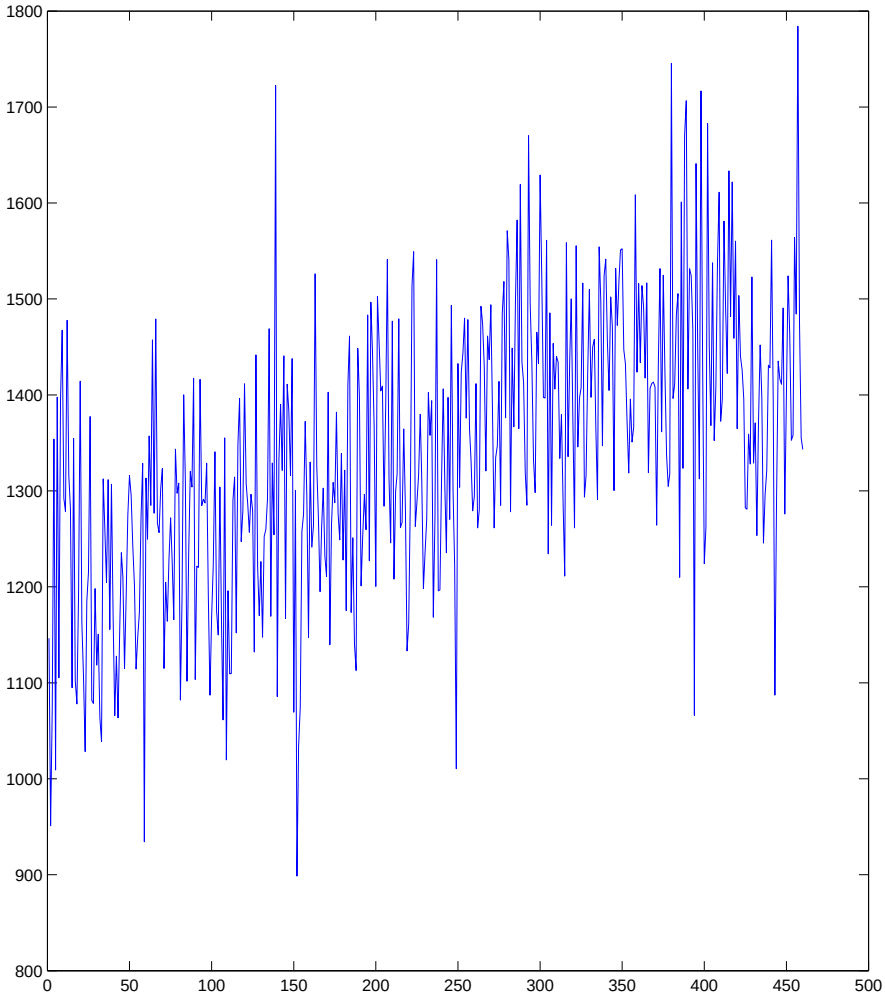

図 3.13: MOCHA-TIMIT のデータ例 (再掲)

この処理によって得られる、$\bar{\boldsymbol{A}}' = \{\mathcal{F}(\bar{\boldsymbol{a}}^{(1)}), \cdots, \mathcal{F}(\bar{\boldsymbol{a}}^{(460)})\}$ を図 3.14 に示す。図 3.14 は、図 3.13 に $\bar{\boldsymbol{A}}'$(赤) を重ねたものである。$\bar{\boldsymbol{A}}'$ が時間変化するバイアス成分を捉えていることがわかる。i 番目の発話の t 番目のサンプルにおける、バイアスを除去した調音運動データは $\bar{\boldsymbol{A}}'$ を用いて、

$$\hat{\boldsymbol{a}}_t^{(i)} = \boldsymbol{a}_t{}^{(i)} - \mathcal{F}(\bar{\boldsymbol{a}}^{(i)}) \tag{3.44}$$

となる。

さらに本研究では、調音運動データに混入するシステムノイズを除去する。図 3.15 は、女性話者の特定の発話における下口唇のセンサの水平方向の座標データである。横軸が時間軸（サンプル番号)、縦軸が座標である。図からわかるように、非常に高い周期の変動が

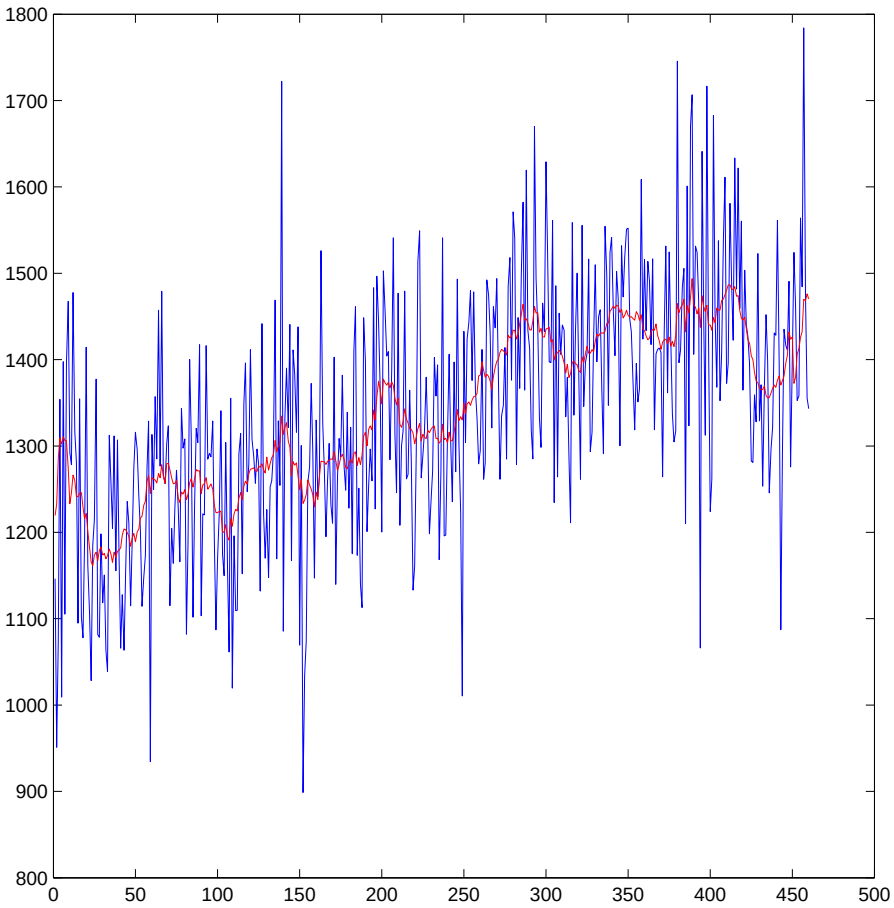

図 3.14: ローパスフィルタの結果

発生している。この変動の周期は、100Hz 以上であり、一般的な調音運動に比べ非常に高い周期となっている。つまり、この変動は、システムノイズだと考えられる。そこで、カットオフ周波数 30Hz のバタワースフィルタを用いて、全ての調音運動パラメータからシステムノイズを除去した。

3.4.2　実験条件

前節の前処理をおこなった MOCHA-TIMIT を用いて実験をおこなう。MOCHA-TIMIT のいずれかの話者を入力話者、もう一方をモデル話者として、連結モデルと分布共有モデルをそれぞれ構築する。

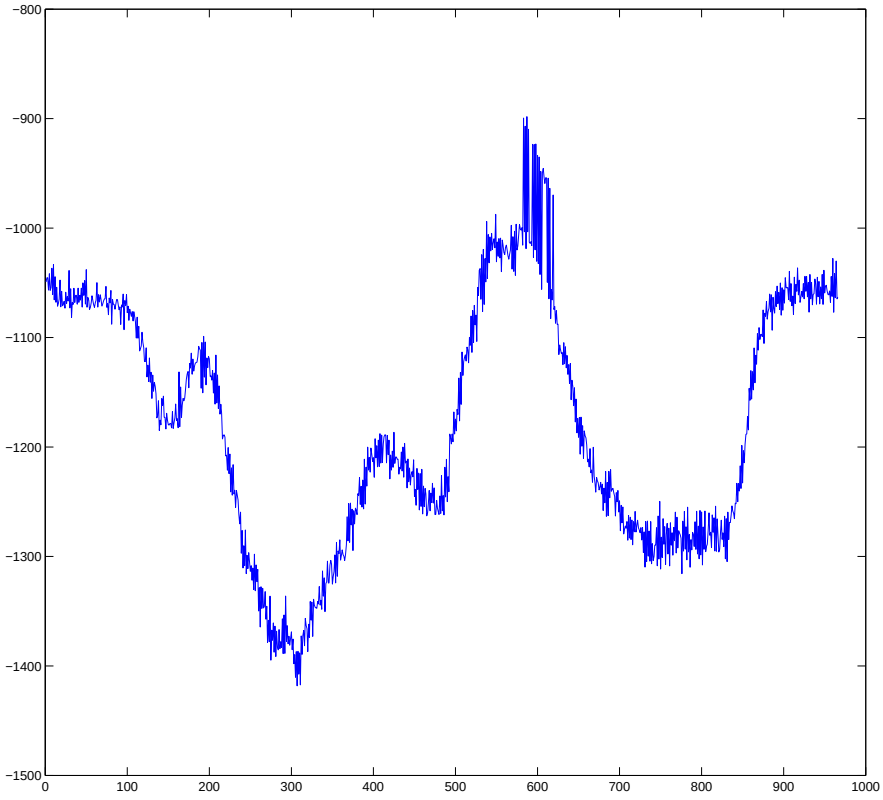

図 3.15: システムノイズの例

i)　データの分割

MOCHA-TIMIT を分割して、変換モデルの構築用データと評価用データを用意する。まず、各話者の全 460 発話の音声–調音パラレルデータを構築用 398 発話と評価用 92 発話に分ける (5-fold cross-validation)。次に、構築用データを用いて、モデル話者の音声–音声パラレルデータと入力話者とモデル話者の音声–音声パラレルデータを用意する。この二つのデータは、構築用 398 発話を 199 発話ずつに分割し、お互いに共通する読み上げ文章が存在しないように用意した (図 3.16)。

ii)　連結モデルの構築

話者変換モデルの構築では、音響特徴量として 24 次元のメルケプストラム[1][40] を用いる。モデルの構築、および特徴量変換では、2.4.5 節の iii) で述べた動的成分を考慮した最尤系列マッピングを用いる。2.4.5 節では、音声–調音マッピングの文脈で紹介したが、入力特徴量を入力話者の音声の音響特徴量、出力特徴量をモデル話者の音声の音響特徴量とす

[1]人間の聴覚特性を考慮して、音声の周波数特性を分析したもの。

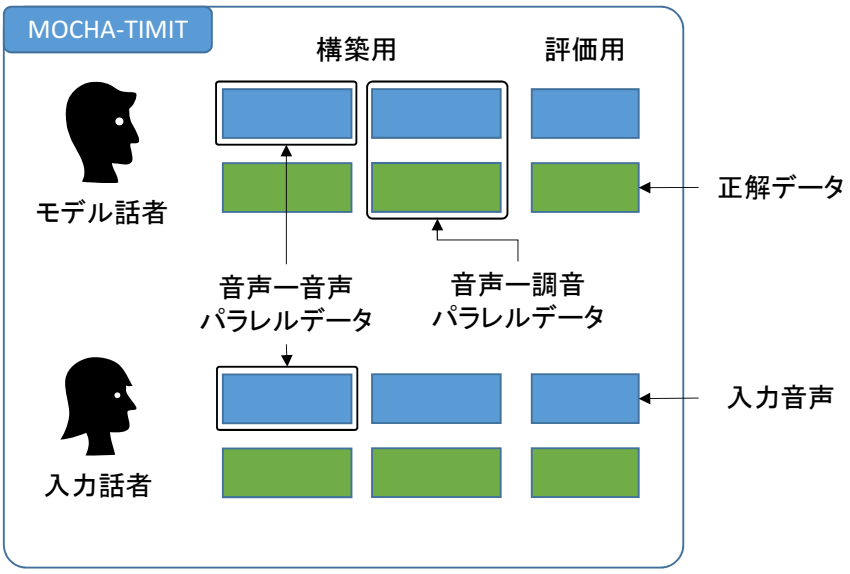

図 3.16: データの分割

ることで、話者変換としても用いることができる。ここで動的成分は、入力特徴量・出力特徴量ともに、式 (2.39) において $N = 1$, $L_{-}^{(1)} = L_{+}^{(1)}$, $w^{(1)}(-1) = -0.5$, $w^{(1)}(-1) = -0.5$, $w^{(1)}(0) = 0$, $w^{(1)}(1) = 0.5$ として求めた。

音声–調音マッピングモデルの構築では、調音運動特徴量として、14 次元の調音運動データ[2]を各次元ごとに、平均が 0 で分散が 1 の正規分布に従うように正規化したものを特徴量として用いた [11]。このマッピングモデルでも、2.4.5 節の iii) で述べた動的成分を考慮した最尤系列マッピングを用いる。このとき、調音運動特徴量の動的成分は、話者変換モデルと同様に式（2.39）において $N = 1$, $L_{-}^{(1)} = L_{+}^{(1)}$, $w^{(1)}(-1) = -0.5$, $w^{(1)}(-1) = -0.5$, $w^{(1)}(0) = 0$, $w^{(1)}(1) = 0.5$ としたものを用いたが、入力となる音響特徴量はフレーム特徴量を用いた [11]。ここで用いたフレーム特徴量とは、当該フレームの前後 10 フレームを連結したものを主成分分析し、得られた主成分の上位 75 成分から構成される特徴量である。音響特徴量および調音特徴量のフレーム幅は、10ms である。

入力音声から調音運動への変換を行う際には、話者変換モデルで出力されたメルケプストラムを、モデル話者の主成分分析の結果を用いてフレーム特徴量に変換し、音声–調音変換モデルの入力として調音運動を求める。

話者変換および、音声–調音マッピングの各変換過程では、2.4.5 節の iii) で述べた動的成分を考慮した最尤系列マッピングを用いる。

[2]MOCHA-TIMIT の EMA の調音運動データ（測定データ）は 7 点の測定点ごとに 2 次元座標を持つ時系列データである。

iii)　分布共有モデルの構築

分布共有モデルでは、入力話者とモデル話者の音響特徴量、および調音運動特徴量として、先に述べた連結モデルの音声–調音マッピングで用いたメルケプストラムのフレーム特徴量と動的成分を考慮した調音運動特徴量を用いた。また、変換は連結モデルと同様に動的成分を考慮した最尤系列マッピングを用いた。

iv)　比較対象となるモデルの構築

提案法の話者正規化音声–調音マッピングの性能を評価するために、モデル話者の音声–調音パラレルデータから構築したの音声–調音マッピングモデルを用意した（従来法）。この音声–調音マッピングは、連結モデルにおける音声–調音マッピングと同じものである。

提案法と従来法の性能を比較するために、以下の 4 つの条件で音声–調音マッピングをおこなった。

reference　モデル話者の音声–調音変換モデルに、モデル話者の音声を入力する。

baseline　モデル話者の音声–調音変換モデルに、入力の音声を入力する。

concatenated　連結モデルに、入力話者の音声を入力する。

shared　分布共有モデルに、入力話者の音声を入力する。

reference 条件は、モデル話者と入力音声の話者が一致する状態、つまり、変換モデルの話者依存性の問題が生じない理想的な状態である。一方、baseline 条件は、変換モデルにモデル話者以外の音声（入力話者音声）がそのまま入力される状態、つまり、話者依存性の問題が生じる状態である。本実験では、連結モデル (conatenated) と分布共有モデル (shared) を用いることで、変換精度が baseline に比べ、改善されるかを検証する。

v)　評価

各条件で求めた調音運動と実測の調音運動（評価用データのモデル話者の調音運動データ（正解データ））の二乗平均平方根誤差 (Root Mean Square Error, RMSE) を評価値とする。この RMSE は変換誤差を表す。reference 条件以外は、入力話者とモデル話者が異なるため、変換結果の調音運動と正解データの調音運動の時間構造が異なる。そこで、DTW を用いて、評価用データのモデル話者音声と入力の時間構造の対応付けをおこない、それに基づき RMSE を計算する。また、RMSE は調音運動データの各次元ごとに計算する。MOCHA-TIMIT に収録されている話者は男女 1 名の計 2 名であるため、モデル話者としていずれかの話者を選択することができる。したがって、各条件において、話者の組み合わせは 2 パターンずつ考えることができる。評価に用いる RMSE は、その 2 パターンについて平均した値とする。

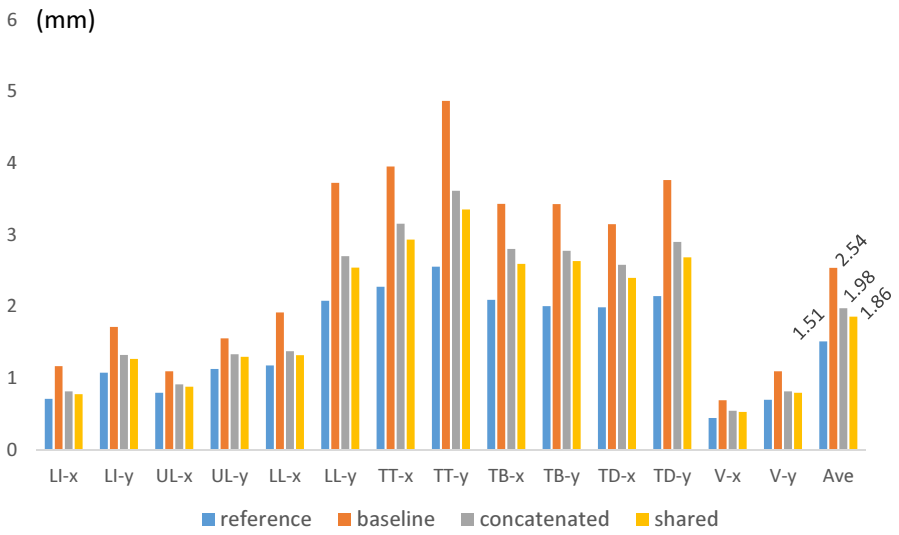

図 3.17: 測定点ごとの変換誤差

3.4.3　結果

変換誤差として、評価データの全ての発話について RMSE を計算し、それを平均したものを図 3.17 に示す。図の横軸は、調音運動データの各測定点を表している (LL=下唇、UL・LL=上口唇・下口唇、TT・TB・TD=舌尖・舌背・舌の後方、V=軟口蓋、-x,-y は水平方向・垂直方向を表す)。縦軸は、変換誤差の値 (単位は mm) である。また、Ave は全ての測定点の変換誤差を平均したものである。変換誤差の各値はモデル話者が男性の場合と女性の場合の結果を平均した値である。図 3.17 の reference と baseline を比較すると、baseline の変換誤差が増大していることがわかる。この変換精度の悪化は、入力話者がモデル話者と異なることで、入力音声と変換モデルの間に音響的なミスマッチが生じることが原因だと考えられる。reference の値に注目すると、舌上の測定点 (TT,TB,TD) の変換誤差が他の点と比較して大きい（変換精度が悪い）ことがわかる。これらの点は、baseline での変換精度の悪化の度合も大きい。舌は調音器官の中でも運動が複雑なため、逆推定が難しい傾向があり、さらに話者依存性の影響も受けやすいと言える。二つの提案手法 (concatenated, shared) はどちらも、baseline に比べ、変換精度が大きく改善されている。このことから、話者変換を利用した話者正規化によって変換モデルの話者依存性が緩和されたことがわかる。

二つの提案手法を比較した場合、全測定点における変換誤差の平均 (Ave.) において、分布共有モデルが連結モデルよりも、 0.12mm 上回っている。また、これは、入力音声から話者変換音声を経由せずに調音運動へ直接変換することで、多段の変換による誤差の蓄積が避けられた結果だと考えられる。

図 3.18 と図 3.19 に、各音素ごとの変換誤差を示す。横軸は各音素を示しており、縦軸は変換誤差 (mm) を表している（音素の表記については、付録 C を参照されたい）。ここでの

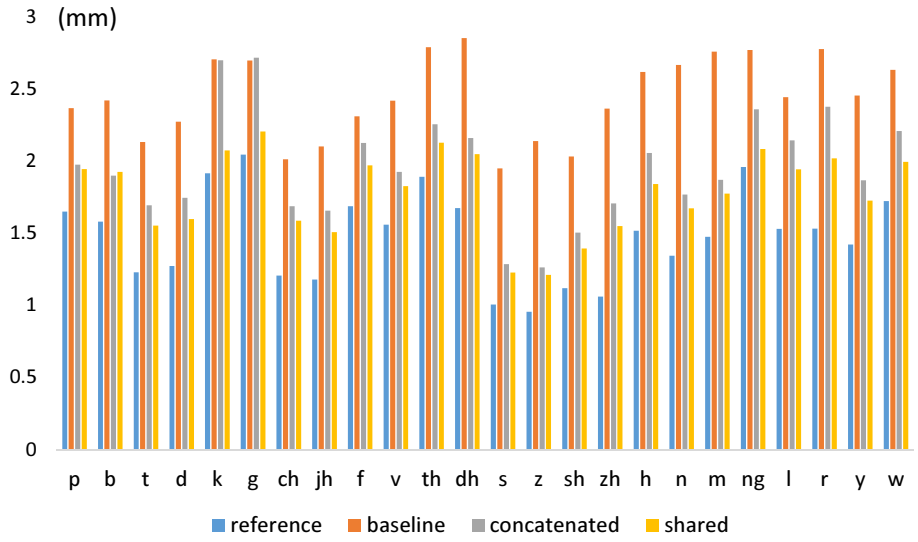

図 3.18: 音素ごとの変換誤差 (子音)

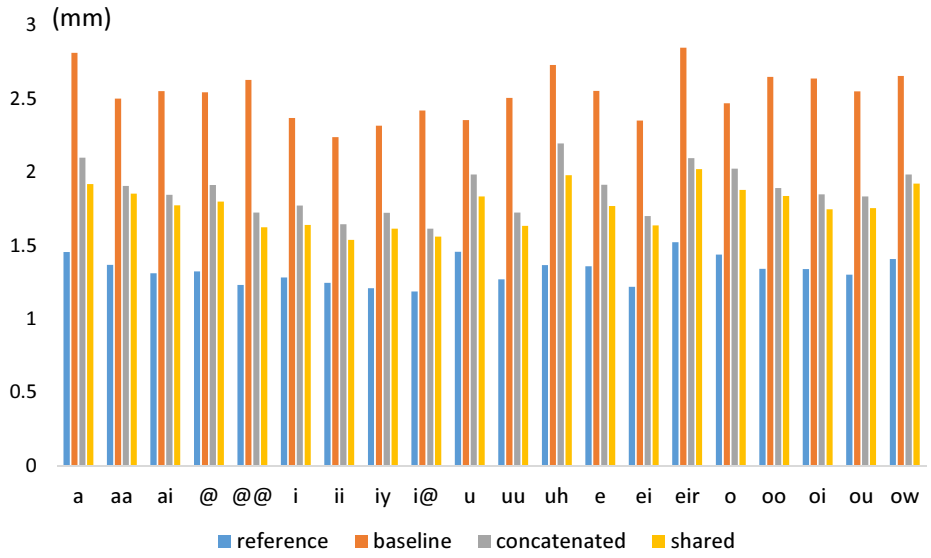

図 3.19: 音素ごとの変換誤差 (母音)

変換誤差は、各音素について、全ての調音点の水平方向・垂直方向の変換誤差を平均したものである。母音に対する変換精度は連結モデル・分布共有モデルを用いることで、baseline に比べ変換精度がいずれの音素についても向上することがわかる。一方、子音に注目すると、軟口蓋閉鎖音[3]/k//g/では、連結モデルを用いても変換精度が baseline から改善されて

[3]舌の中部を軟口蓋に接触させ閉鎖を作ることで発音される音素

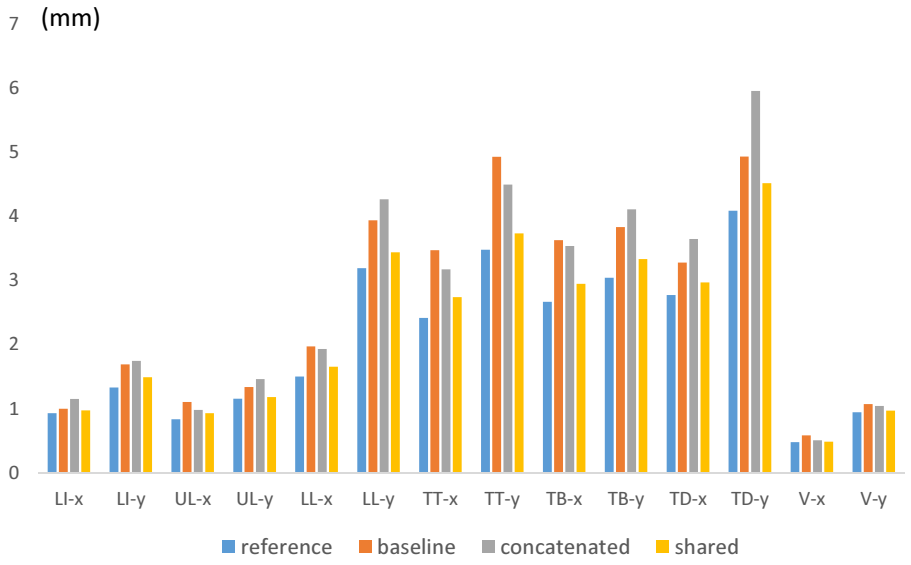

図 3.20: 音素/k/における測定点ごとの変換誤差

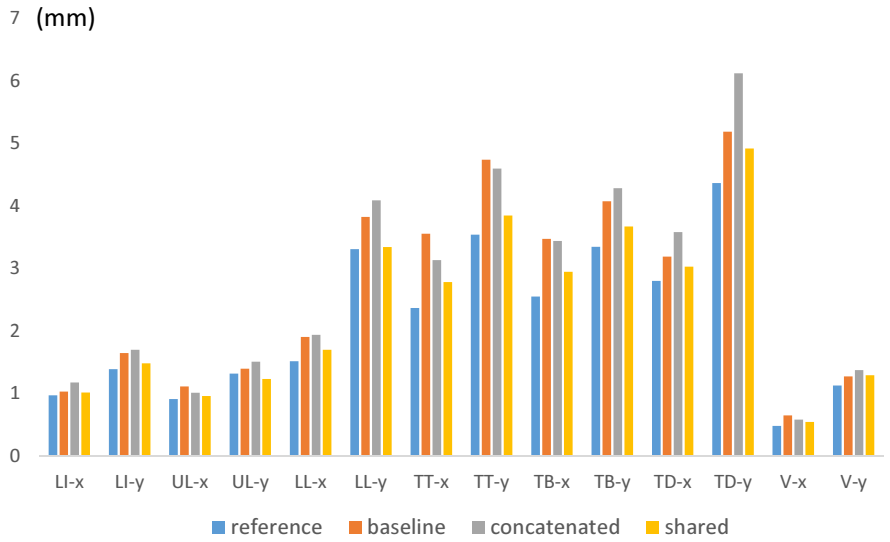

図 3.21: 音素/g/における測定点ごとの変換誤差

いない。しかし、分布共有モデルを用いることで、変換精度が向上し、reference とほぼ変わらない精度で推定可能であることが分かる。

図 3.20 と図 3.21 に、音素/k//g/における調音点ごとの変換誤差を示す。この図は図 3.17 と同じ形式だが、変換誤差を求める際に音素/k/または/g/のフレームのみを用いた。図舌上のセンサー、特に後方の 2 つのセンサー（TB,TD) に注目すると、連結モデルは baseline

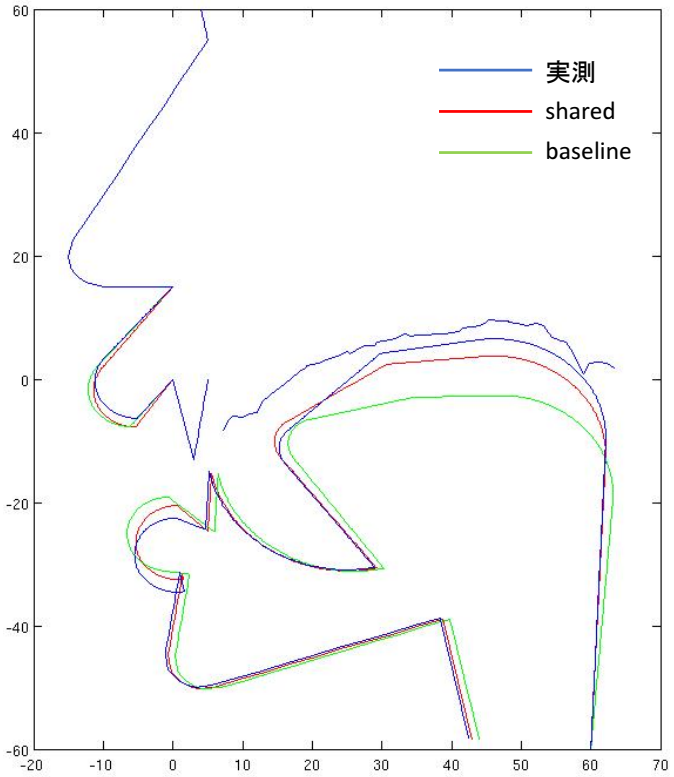

図 3.22: 音素/i/の推定結果

からの改善がほとんど見られないが、分布共有モデルは reference と同等の精度が得られていることがわかる。/k//g/は、どちらも軟口蓋閉鎖音であるため、調音点に近い TB・TD において、分布共有モデルを用いることで良好な変換精度が得られるということは、発音トレーニングなどへの応用を検討する上で非常に重要である。

推定結果を用いて、3.3 節の調音モデルを駆動した例を図 3.22 に示す。図 3.22 は、図 3.11 と同じく単語 “thieves” の音素/i/を発音した時の調音運動の様子である。図中の調音運動は、男性話者の実測データ（正解データ）と、baseline 条件と shared 条件（モデル話者・男性話者、入力話者・女性話者）における推定結果で調音モデルを駆動した結果である。baseline 条件の調音運動に注目すると、実測データと比べて舌背が持ち上がっておらず、/i/の発音に必要な狭めが形成されていないことがわかる。一方で、shared 条件では、舌背が持ち上がり狭めが形成されている。提案法の推定結果の妥当性が、調音モデルを用いることで視覚的に示された。

3.5　まとめ

本章では、音声–調音マッピングの問題点の一つである、

- 特定話者のデータから構築された変換モデルは話者依存モデルとなり、多様な話者の音声を対象とした推定が難しい。

という課題に取り組んだ。

話者依存モデルの問題の原因は、入力話者の音声とモデル話者の音声の間に生じる音響的なミスマッチである。そこで、音声–調音マッピングの前処理として、話者変換を用いて、任意の入力話者の音声をモデル話者の音声に変換（正規化）することで、音響的なミスマッチを軽減することを試みた。

話者正規化音声–調音マッピングモデルとして、連結モデルと分布共有モデルという二つのモデルを提案した。連結モデルは、話者変換モデルと音声–調音マッピングモデルを縦続接続したモデルで、入力音声に各モデルの変換が多段に作用する変換器となっている。一方で、分布共有モデルは、話者変換と音声–調音マッピングの 2 つ変換モデルの GMM において、モデル話者の音声の特徴量空間が一致しているという仮定のもと、その特徴量空間を周辺化することで、2 つ変換モデルの GMM を統合する。2 つのモデルを統合することで、一度の変換で入力音声からモデル話者の調音運動へ変換することが可能になった。また、分布共有モデルの構築法として、欠損値を含む EM アルゴリズムを提案した。

MOCHA-TIMIT を用いて、入力話者とモデル話者が異なる話者という条件で、音声から調音運動を推定する実験をおこなった。その結果、従来の音声–調音マッピングモデルに比べ、提案手法（連結モデル・分布共有モデル）は高い変換精度を示した。二つの提案手法を比べた場合、分布共有モデルの方が、高い変換精度を示した。これは、連結モデルにおける多段の変換よって生じていた変換誤差の蓄積が、音声から調音運動へ直接変換することができる分布共有モデルによって解消されたと考えられる。

提案手法は、入力音声に関して話者正規化をおこなうが、これは任意の入力話者の音声の音響空間をモデル話者のものと一致させる操作である。したがって、音声から推定される調音運動は、モデル話者の調音運動空間に基づくものである（模擬調音運動）。提案手法は、入力話者自身の調音運動を求めることはできないが、モデル話者に注目した場合、その逆推定の対象をモデル話者自身の音声から任意の話者の音声へ拡張することに成功した。

任意の入力話者の音声から、模擬調音運動ではなく、その話者自身の調音運動を推定するためには、音声–調音マッピングの出力空間として入力話者の調音運動空間を構築する必要がある。入力話者の調音データが存在しない条件で、入力話者の調音運動空間を構築するというのは非常に難しい問題であるが、解決が望まれる今後の課題の一つであると言える。

また、提案手法では，入力話者とモデル話者の音声パラレルデータの収録において、両者が同一の調音運動をしていることを前提としている。互いに同一言語を母語としている場合であっても、地方訛りなどの違いによって必ずしも同一の調音運動をしていない可能性もある。両者の調音運動に差が存在するまま音声パラレルデータとして使った場合、入力話者特有の喋りの特徴（例えば、訛り）は話者変換によってモデル話者の喋りの特徴に変換されてしまい、音声–調音変換の結果として得られる調音運動は、入力話者の特徴を反映していないと考えられる。発音トレーニングのような応用では、推定するべき調音運動は入力話者自身の喋りの特徴を十分に反映した調音運動である。したがって、提案法の枠

組みにおいて、発話者の喋りの特徴を反映した調音運動を推定するためは、入力話者の喋りの特徴を保持したまま話者変換をおこなう手法が必要となる。これについても、今後検討するべき課題と言える。

第4章

音声の構造的表象を用いた未観測音素の調音運動の推定

4.1　はじめに

GMM に基づく音声–調音マッピングは、モデル話者の音響空間と調音運動空間の対応をモデル化し、それに基づき入力音声から調音運動を推定する技術である。これは、入力音声から音響空間における座標を求め、モデル化した対応関係に基づき、その座標を調音運動空間上に写像するという操作と言える。前章で提案した話者正規音声–調音マッピングにおいても、本質的には従来のマッピングと変わらない。その変換過程は、入力音声から入力話者の音響空間上における座標を求め、それを話者変換によってモデル話者の音響空間に写像し、さらに音声–調音マッピングによって最終的な出力空間である調音運動空間に写像する操作である。これらの音声–調音マッピングは、入力として音声が与えられ、その音響空間における座標が求まるということを前提にした技術である。つまり、推定対象として音声（音響情報）が存在しているということを前提にしている。この前提は、音響理論に基づく方法や、HMM や ANN に基づく音声–調音マッピングにおいても同様である。

一方で、「音声が存在しない発話」というものを考えることができる。それは、発話者の母語の音素体系に存在しない音素（発音）の発話である。そのような音素は、発話者の自然な発話からは観測することが難しい（以下、未観測音素と呼ぶ）。例えば、日本語しか発話できない話者にとって、日本語の音素に含まれない英語音素（例えば、/æ/や/v/）は、未観測音素となる。未観測音素の調音運動は、発話者の音声が存在しないため、音響情報が存在することを前提とした従来の逆推定問題としての取り扱いは難しい。しかし、未観測音素の調音運動（未観測調音運動）の推定は、語学学習や発話トレーニングにおいて重要な課題となっている。2.4.2 節の iii) で紹介した調音運動情報を用いた発音トレーニングでは、学習者に対して自らの調音運動と目標となる調音運動を視覚的にフィードバックし、発音の誤りに対する直感的な理解を助けることを目指している。このとき、目標となる調音運動は、学習者にとって未観測調音運動になると考えられるが，その推定法は確立されていない。そこで、本章では、1) 当該話者以外の話者の音声より，2) 音響情報以外の情報を用いて、未観測調音運動を推定する方法を検討する。

本章では、未観測調音運動の推定を通して、2.5 節で述べた逆推定問題の課題の一つである、

- 音声の音響情報のみに注目した推定であるため、適用できる問題が限られている。

という課題に取り組む。ここで、音声の音響情報以外の情報として注目するのが音声の言語的情報である。言語的特徴は、音声の音響情報が内包する情報の一つの側面である。調音運動の役割は、音声に言語的特徴を付加することであることを考慮すれば、言語的情報は調音運動と非常に関わりが深い情報だと言える。

本章で検討する未観測調音運動の推定法では、音声の言語的特徴として、未観測音素に関する音声の構造的表象 [43] を利用する [41][42]。音声の構造的表象は、音声から話者性などの非言語的な情報をそぎ落とすことを狙った表現方法であり，これは音声に含まれる言語的情報に対する表現方法の一つである。音声の構造的表象では、音響事象間の相対的な配置（距離関係）を抽出することができ、それは理論的には変換不変量となるが，これを

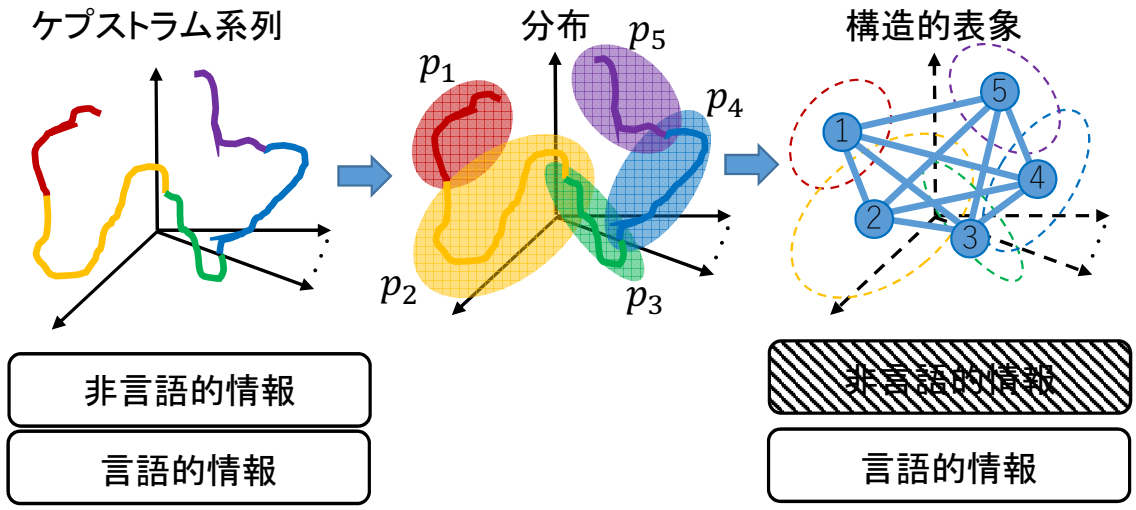

図 4.1: 音声の構造的表象

話者差に起因する音響バイアスの除去を目的として適用する。未観測音素を発話可能な話者から、音声の構造的表象によって未観測音素と他の音素に関する距離関係を抽出し、その距離関係を満たす音響事象を与える調音運動を、元話者の調音空間内で探索することで未観測調音運動を推定する。

4.2　音声の構造的表象を用いた未観測調音運動の推定

4.2.1　音声の構造的表象

我々が耳にする音声は、その生成過程・伝達過程・聴取過程において不可避的に混入する非言語的な歪みの影響を受けたものである。音声の構造的表象は、音声から非言語的な歪みを取り除き、言語的な特徴のみを抽出することを目的としている [43]。

音声に混入する非言語的な歪みは、乗算性歪みと線形変換性歪みに大きく分けられる。乗算性歪みは、マイクロフォンの音響特性に代表される歪みで、ケプストラム空間では加算演算 $c' = c + b$ として表現される。一方で、線形変換性歪みは、声道長や聴覚特性の違いによる歪みに対応し、ケプストラム空間では線形変換 $c' = Ac$ によって表現される [44]。したがって、非言語的な歪みは、ケプストラム空間においてアフィン変換 $c' = Ac + b$ によって表現される。音声の構造的表象では、音声を音響事象ごとに分布化し、各分布の絶対的な座標を捨て、分布間の相対的な関係によって構造的に表すことにより、アフィン変換に対して不変、即ち，マイクの違い，声道長の違いに対して不変な（つまり，非言語的な情報が除去された）音声の記述を実現する (図 4.1)。今、二つの音響事象のケプストラム空間における確率密度関数をそれぞれ $p_1(x), p_2(x)$ とする。この二つの分布間の Bhattacharyya

距離（以下 BD）は、

$$BD(p_1, p_2) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \tag{4.1}$$

となる。二つの分布に対して共通のアフィン変換が作用した場合、BD はその変換について不変となる [45]。つまり、幾つかの音響事象からなる音響事象を、各事象間の BD によって記述すれば、それは非言語的な歪み（共通かつ静的のアフィン変換）に対して不変、つまり言語的な情報のみを保持した記述と言える。

音声の構造的表象を特徴量として用いる場合には、まず音声に対して音響事象の単位を定め（例えば音素）、その事象の確率密度関数をガウス分布によってモデル化する。ガウス分布に対する BD は以下の式で与えられる。

$$BD(p_1, p_2) = \frac{1}{8}\boldsymbol{\mu}_{1,2}^{\top}\boldsymbol{V}_{1,2}^{-1}\boldsymbol{\mu}_{1,2} + \frac{1}{2}\frac{|\boldsymbol{V}_{1,2}|}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}} \tag{4.2}$$

ここで、分布 p_1,p_2 は、それぞれ $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$,$\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ に従うガウス分布であり、$\boldsymbol{\mu}_{1,2} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$,$\boldsymbol{V}_{1,2} = \frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}$ である。N 個の音響事象からなる事象群に対して、各事象間の BD を計算することによって、$N \times N$ の距離行列を構築する。この距離行列は、音声の言語的情報のみを保持する特徴量として用いることが可能であり、これまでに、話者の違い・録音環境の違いに頑健な外国語の発音分析や音声認識への応用が試みられている [46][47]。

音声の構造的表象は、音声を音響事象間の距離行列で表現するため、もはや各音響事象は音響空間における絶対的な座標を持たない。つまり、構造的表象によって表された音響事象は、もはや音響的な実体を持たない。このことから音声の構造的表象は、極めて抽象度の高い音声の表現と言える。

4.2.2　音声の構造的表象を用いた音声合成

音声の構造的表象から音声を合成する手法が検討されている [48]。前節で述べたように、音声の構造的表象によって表現された音響事象は、音響空間内における絶対的な座標を捨て去っており、音響的な実体を持たない。したがって、音声の構造的表象から音声（音響情報）を得るには、音響空間における絶対的な座標を与える必要がある。音声の構造的表象を音響空間上に定位する操作は、構造的表象を抽出の過程で音声から捨て去った話者の身体性を改めて付加することに対応する。齋藤ら（2007）は、身体性を与える代わりに、構造的表象内のいくつかの音響事象に音響空間における絶対的な座標（=分布, 音響実体）を初期条件として与えた。そして、座標が与えられた事象と与えられていない事象間の距離を拘束条件として、残りの音響事象を定位させ音声を合成した。例えば、図 4.1 の構造的表象において、音響事象（ノード）1,2,3,4, の座標を音響空間上で与えれば、それのノードはアンカー（固定点）となり、そこからの距離を拘束条件としてノード 5 の座標を推定することができる（幾何学的には、4 つのアンカーを中心として構成される超球の交点を探索する問題となる）。

今、ノード 1,2,3,4 の音素は発音できるが、ノード 5 の音素を発音できない話者のノード 5 の音素（ターゲット）の音声を合成することを考える。このとき、ノード 1,2,3,4,5 の音

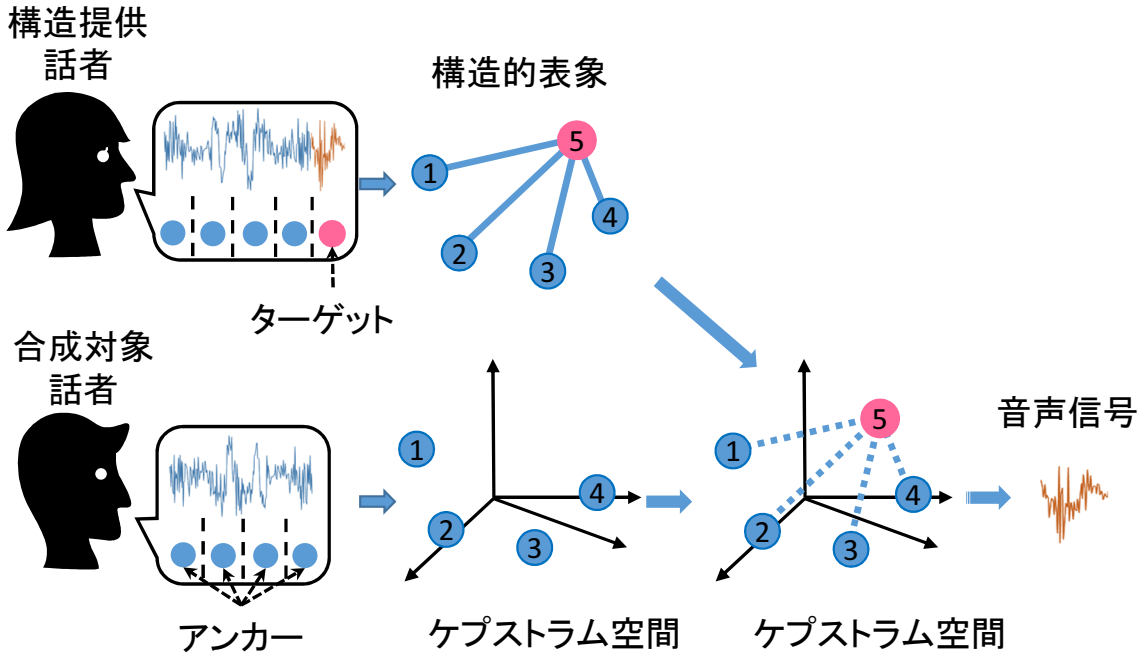

図 4.2: 構造的表象を用いた音声合成

素を全て発音できる話者（構造提供話者）がいれば、その音声からノード 1,2,3,4,5 の音声の構造的表象を抽出することができる。このとき、音声の構造的表象は話者非依存の特徴であるため、抽出した構造を合成対象話者に適用することを考える。合成対象話者の音響空間内のアンカー音素 1,2,3,4 の座標、および、ターゲット–アンカー間の距離を制約として、合成対象話者の音響空間内のターゲット音素の座標（分布）を推定することができる（図 4.2）。この推定は、以下の式で表せる。

$$\hat{p}_5 = \arg\min_{p_5} \sum_{n=1}^{4} \left\{ BD(p_n, p_5) - d_{n,5}^{(r)} \right\}^2 \tag{4.3}$$

ここで、p_i は合成対象話者の音響空間における音素 i の分布を表しており、$d_{i,j}^{(r)}$ は構造提供話者 (reference speaker) から抽出した構造における音素 i と音素 j の BD である。

4.2.3　未観測調音運動の推定

齋藤ら（2007）の手法に基づき未観測調音運動の推定法を提案する。提案法では、合成法と同様に、アンカー音素を定義し推定対象となる話者の音響空間上の座標を用意する。さらに、未観測音素（ターゲット音素）とアンカー音素の音声の構造的表象を他話者（構造提供話者）から抽出する。推定対象話者の未観測音素の音響空間上の座標は、アンカー音素と構造的表象から推定できるが、本手法の推定対象は調音運動である。したがって、音響空間におけるアンカー音素の座標と構造的表象を拘束条件として、それを満たす音響事象を与える調音運動を探索すること考える。このとき、拘束条件は音響空間に存在する

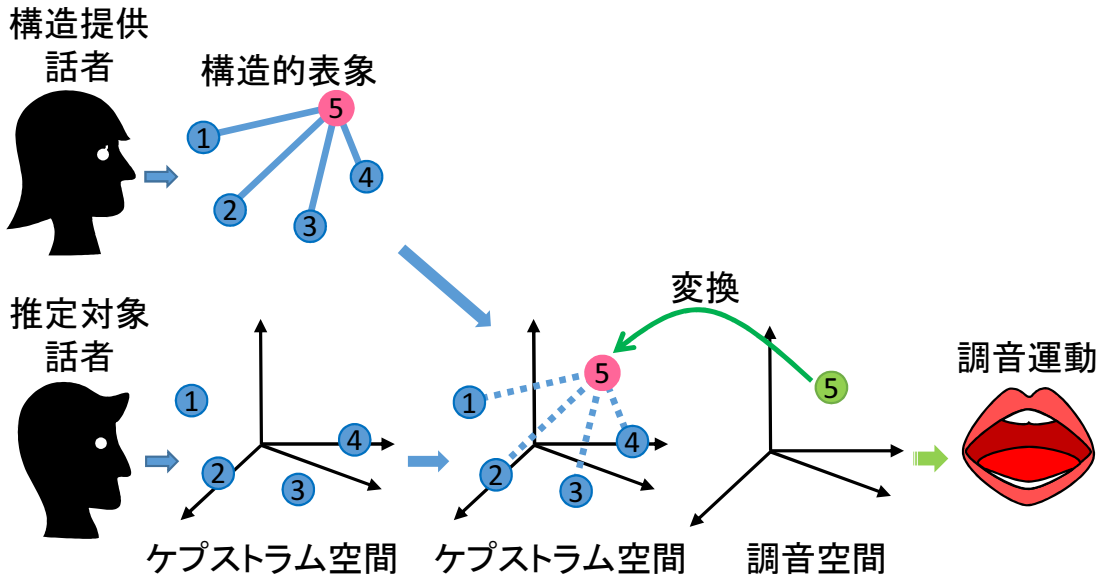

図 4.3: 構造的表象を用いた調音運動推定

が、探索するべき空間は調音空間となる。そこで、調音運動を音響空間に写像しながら、音響空間における拘束条件を満たす調音運動を推定する（図 4.3）。調音運動を $\boldsymbol{x}$、調音運動を音響空間に写像した結果を $\mathcal{F}(\boldsymbol{x})$ とすると、図 4.3 における未観測調音 $\boldsymbol{x}_5$ は、

$$\vec{x}_5 = \arg\min_{\vec{x}} \sum_{n=1}^{4} \left\{ BD(p_n, \mathcal{N}(\mathcal{F}(\vec{x}), \boldsymbol{\Sigma})) - d_{n,5}^{(r)} \right\}^2 \tag{4.4}$$

$$= \arg\min_{\vec{x}} J(\vec{x}) \tag{4.5}$$

ここで、$\mathcal{N}(\mathcal{F}(\vec{x}), \boldsymbol{\Sigma})$ は、音響空間における平均ベクトル $\mathcal{F}(\vec{x})$, 共分散行列 $\boldsymbol{\Sigma}$ のガウス分布であり、平均ベクトルは調音運動から写像した音響特徴量となっている。また、式 4.5 は推定式におけるコスト関数を $J(\cdot)$ で表したのもである。

提案法では、調音空間から音響空間への写像 $\mathcal{F}(\cdot)$ として、GMM による調音－音声マッピング [11] を用いる。GMM による調音－音声マッピング [11] は、2.4.5 節で述べた GMM による調音－音声マッピングの入力と出力を入れ替えたものである。今、音声特徴量を $\boldsymbol{y}$、調音運動特徴量を $\boldsymbol{x}$、二つの特徴量の結合ベクトルを $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ とすると、結合ベクトルの確率分布は GMM によって以下のようにモデル化できる。

$$p(\boldsymbol{z}; \lambda^{(\boldsymbol{z})}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu_m}^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{4.6}$$

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(x,x)} & \boldsymbol{\Sigma}_m^{(x,y)} \\ \boldsymbol{\Sigma}_m^{(y,x)} & \boldsymbol{\Sigma}_m^{(y,y)} \end{bmatrix} \tag{4.7}$$

ここで、m は混合成分のインデックス、α_m は混合成分の重みである。二乗誤差最小化基準による調音運動から音声へのマッピング関数は以下の式で表される。

$$\hat{\boldsymbol{y}} = \sum_{m=1}^{M} p(m|\boldsymbol{x}; \lambda^{(\boldsymbol{z})}) \left\{ \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(y,x)} {\boldsymbol{\Sigma}_m^{(x,x)}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)}) \right\} \tag{4.8}$$

ここで、$p(m|\boldsymbol{x}; \lambda^{(\boldsymbol{z})})$ を定数 $\gamma_{m,x}$ で近似することで、

$$\hat{\boldsymbol{y}} = \sum_{m=1}^{M} \gamma_{m,x} \left\{ \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(y,x)} {\boldsymbol{\Sigma}_m^{(x,x)}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)}) \right\} \tag{4.9}$$

となり、$\boldsymbol{x}$ の線形式という簡単な形になる。これを式 (4.4) における写像 $\mathcal{F}(\cdot)$ として用いる。

式 (4.4) の推定では、推定対象話者–構造提供話者間において、未観測音素に関する構造的な歪み（距離の差）を最小化するように、推定対象話者の調音運動 $\boldsymbol{x}$ が音響空間への写像を通して探索される。この過程は、推定対象話者（例えば外国語学習者）が構造提供話者（例えば教師）の発話から言語的な特徴を取り出し、その特徴を自分の調音器官を通して再現しようとしていると捉えることができる。ここで、音声の構造的表象は言語的な制約、調音–音声マッピングによる写像は話者の身体的、または物理的制約として機能する。

4.2.4　音声の構造的表象の修正

提案法は、未観測音素とアンカー音素に関する音声の構造的表象が、推定対象話者と構造提供話者の間で厳密に一致していることを前提としている。しかし、二話者が同じ言語を共有しているとしても、個人固有の言語的な特徴（訛りに代表されるその話者特有の癖[1]）によって構造的表象に差が生じる可能性がある [50]。その差は推定結果に悪影響を与えると考えられるため、推定の前処理として話者間の構造的表象の差を取り除く手法を導入する。

前節と同様に図 4.3 の場合について考える。構造的提供話者からは、音声の構造的表象として、未観測音素と 4 個のアンカー音素に関する 5×5 次元の距離行列 $\boldsymbol{D}^{(r)} = \{d_{i,j}^{(r)}\}_{1 \leqq i,j \leqq 5}$ を得ることができる。ここで、$d_{i,j}^{(r)}$ は音素 i と音素 j の BD である。一方で、推定対象話者から得ることのできる構造的表象は、アンカー音素に関する 4×4 次元の距離行列 $\boldsymbol{D}^{(t)} = \{d_{i,j}^{(t)}\}_{1 \leqq i,j \leqq 4}$ である。そこで、構造的提供話者の距離行列においてアンカー音素のみで構成された部分行列 $\boldsymbol{D}_{(4)}^{(r)} = (d_{i,j}^{(r)})_{1 \leqq i,j \leqq 4}$ に注目し、それを推定対象話者の距離行列 $\boldsymbol{D}^{(t)}$ に一致させることを考える（構造的正規化）。つまり、推定話者の構造を基準として構造提供話者の構造を修正することになる。

今、$\boldsymbol{D}_{(4)}^{(r)}$ に対して、対角行列 $\boldsymbol{S}_{(4)} = \mathrm{diag}\{s_1, s_2, \cdots, s_4\}$ を両側から掛けた行列 $\boldsymbol{S}_{(4)} \boldsymbol{D}_{(4)}^{(r)} \boldsymbol{S}_{(4)}$ を考える。この演算により、$d_{i,j}^{(r)}$ は、$s_i s_j d_{i,j}^{(r)}$ となり、これは i 番目のアンカー音素と j 番目のアンカー音素の BD を $s_i s_j$ 倍したものを表す。以下の式に従って、$\boldsymbol{S}_{(4)} \boldsymbol{D}_{(4)}^{(r)} \boldsymbol{S}_{(4)}$ と $\boldsymbol{D}^{(t)}$

[1]日本語では地方訛りはアクセントやイントネーションなど韻律に出やすいが，外国語では，地方訛りは例えば母音の声色の変化，即ち，母音群の配置の違いとして検討されることが多い [51]

の差を最小にする変換行列 $\boldsymbol{S}_{(4)}$ を求める。

$$\hat{\boldsymbol{S}}_{(4)} = \arg\min_{\boldsymbol{S}_{(4)}} \sum_{i=1}^{3} \sum_{j=i+1}^{4} (s_i s_j d_{i,j}^{(r)} - d_{i,j}^{(t)})^2 \tag{4.10}$$

上式によって、アンカー音素に関する二話者間の構造の差を修正する変換行列 $\hat{\boldsymbol{S}}_{(4)}$ を求めることができる。しかし、式 (4.4) の推定法では、構造提供話者のターゲット–アンカー間の距離が必要となる。そこで、$\hat{\boldsymbol{S}}_{(4)}$ に基づいて、構造提供話者のターゲット–アンカー音素間の距離を修正する。$\hat{\boldsymbol{S}}_{(4)}$ を拡張した対角行列 $\hat{\boldsymbol{S}} = \mathrm{diag}\{\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_4, 1\}$ を考える。この対角行列は、4 次元目までの対角要素が $\hat{\boldsymbol{S}}_{(4)}$ と同値で、5 次元目の対角要素の値が 1 である 5×5 次元の変換行列である。その行列を修正行列として $\boldsymbol{D}^{(r)}$ に作用させ、$\hat{\boldsymbol{S}}\boldsymbol{D}^{(r)}\hat{\boldsymbol{S}}$ を得る。距離行列 $\hat{\boldsymbol{S}}\boldsymbol{D}^{(r)}\hat{\boldsymbol{S}}$ において、ターゲット–アンカー間の BD は $\hat{s}_i d_{i,5}^{(r)} (1 \leqq i \leqq 4)$ となる。これを構造提供話者の修正後のターゲット–アンカー音素間の BD として用いる。

この修正は、構造提供話者が推定対象話者の訛りも含めて真似た場合の発話の構造的表象を求めていると考えることができる。例えば、推定対象話者が構造提供話者と比べてアンカー音素の発音が曖昧な話者だとする。この場合、アンカー音素同士の音響的特徴が似たものになり、その結果、アンカー音素間の BD は短く縮退したものになる。その発話を真似て構造提供話者がアンカー音素とターゲット音素を含んだ発話をおこなった場合、アンカー同士はもちろんターゲット–アンカー間の BD も短く縮退したような発話になるのが妥当だと考えられる。これは、アンカー音素のみに注目して得られた $\hat{\boldsymbol{S}}_{(4)}$ の各要素 $\{\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_4\}$ が 1 よりも小さい値となり、その結果 $\hat{s}_i d_{i,5} < d_{i,5}$ となるため、修正後のターゲット–アンカー間の BD が本来の BD よりも縮退したものになることによって説明できる。

4.2.5　話者正規化音声–調音マッピングと関係

前章で述べた、話者正規化音声–調音マッピングは、任意の入力話者から特定のモデル話者の調音運動を推定するもであった。ここで、今回の課題である未観測調音運動の推定を、話者正規化音声–調音マッピングを用いておこなうことを考える。今、構造提供話者を入力話者、推定対象話者をモデル話者とすれば、入力話者に未観測音素を発話してもらい、それを話者正規化音声–調音マッピングを用いてモデル話者の調音運動に変換することで、未観測調音運動を推定できる。しかし、この推定結果と、本章での推定法で求まる推定結果は、本質に異なるものである。話者正規化音声–調音マッピングを用いて未観測調音運動を推定した場合、その未観測調音運動は、本来の言語的特徴を保持している保証がない。なぜならば、話者正規化で用いる話者変換は、音声の構造的表象を歪ませてしまうからである。式 (4.2) で表される BD はアフィン変換には不変[2]だが、話者変換はアフィン変換ではない。したがって、話者変換による話者間の音響特徴量の写像は音声の構造的表象を歪ませることになり、未観測調音運動の言語的特徴は本来のものとは異なるものとなる。

一方で、本章の提案手法は、未観測音素の言語的特徴に基づいた手法である。構造提供話者と推定対象話者の間に、訛りの差が生じる場合、前節の修正法を用いて構造を修正、

[2]本来のバタチャリヤ距離は、あらゆる連続な全単射である変換に対して不変だが、分布としてガウス分布を仮定した場合、アフィン変換に対して不変な距離となる。

表 4.1: 音素表
MOCHA-TIMIT 内での音素表記とその音素を含む単語例（当該音素を太字で示す）

音素表記	@	a	e	i	iy	o	u	uh
単語例	**a**bout	ex**a**m	y**e**ll	s**i**mple	mon**ey**	**o**ften	g**oo**d	s**u**n

すなわち、歪ませることになるが、その操作もやはり言語的特徴に基づいたものとなっている。つまり、本章の推定法は、話者正規化音声–調音マッピングと異なり、全ての処理において未観測音素が持つ言語的特徴の妥当性が保証されたものとなっている。

4.3　実験

4.3.1　実験条件

音素を音響事象の単位として、単独の音素を対象とした調音運動の推定実験を行った。実験用データは、前章の実験でも用いた MOCHA-TIMIT である。

実験では、コーパス内のどちらかの話者を推定対象話者とし、同一話者もしくは、もう一方の話者を構造提供話者とする。また、コーパス内の発話に現れる単母音 8 種（表 4.1）に注目して、その中の 1 音素を疑似的な未観測音素（ターゲット音素）、その他の 7 音素をアンカー音素として実験を行う。まず、構造提供話者の全音声データから、音素ラベルを用いてターゲット音素とアンカー音素を発声しているフレームを切り出し、それぞれの音素を分散共分散行列が対角行列であるガウス分布によってモデル化した（図 4.4）。つまり、各音素のモデルは、連続発話において様々な前後音素の影響を受けて発声された当該音素を平均化したモデル（分布）と言える。その後、その分布を用いてターゲット–アンカー間の BD を抽出した。推定対象話者についても同様に、音声データからアンカー音素の分布をモデル化する。さらに、推定対象話者の音声–調音パラレルデータを用いて調音–音声マッピングモデルを構築した。このとき、推定対象話者は未観測音素を発音できないという状態を再現するため、音声–調音パラレルデータから、音素ラベルを用いてターゲット音素を発声しているフレームを取り除いた。調音–音声マッピングモデルの GMM の混合数は 128 とした。また、推定目標として推定対象話者の調音データにおけるターゲット音素のガウス分布を求めた。この分布の平均ベクトルが推定すべき調音運動（正解データ）となる（図 4.4）。

音響特徴量と調音運動特徴量は前章での実験と同じく、音響特徴量は 24 次元のメルケプストラム、調音運動特徴量は 14 次元の調音運動データとした。また、調音運動特徴量に関しては、主成分分析を適用し直交化したものを特徴量として用いた[3]。また各特徴量のフレーム幅は 10ms とした。これらの特徴量を用いて、各事象のガウス分布、調音–音声マッピングを構築した。

[3]直交化されていない状態で、後述する最急降下法による探索を行うと、探索結果が十分に収束しないことが事前の検討で明らかになった。

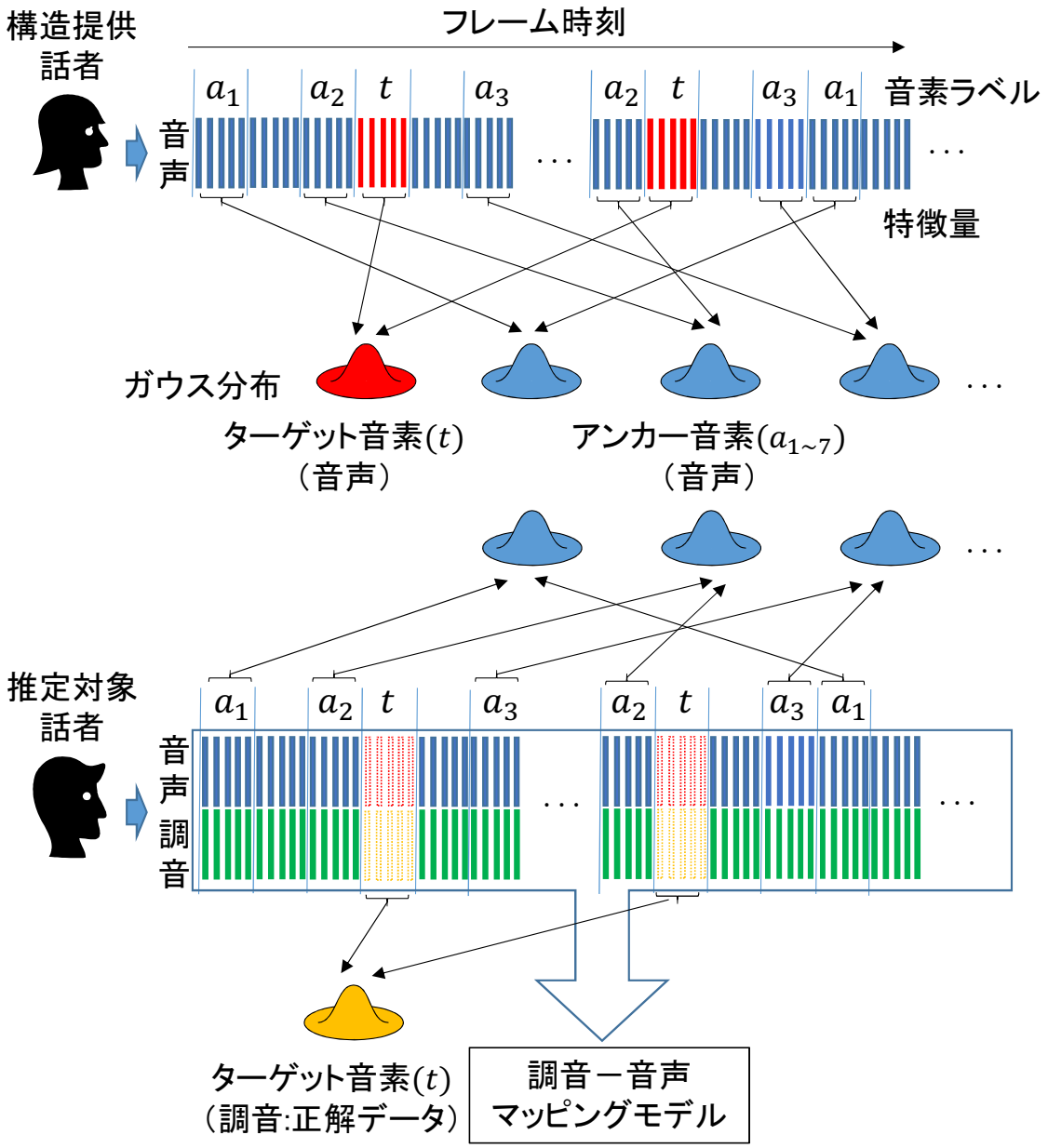

図 4.4: 実験設定

調音運動の推定は、式 (4.4) に示すように調音運動 $\boldsymbol{x}$ に関する最適化問題となっている。そこで、最急降下法を用いて最適な $\boldsymbol{x}$ を探索する。最急降下法では、$\boldsymbol{x}$ の初期値が必要となる。そこで、構造提供話者から抽出されたターゲット–アンカー間の BD（7 個のアンカー音素とターゲット音素間における 7 個の BD）において、最も小さい BD となるアンカー音素を調べ、その音素の推定対象話者の調音運動（ガウス分布の平均ベクトル）を探索の初

期値とした。この初期位置は、音響空間においてターゲット音素に最も発音が似ているアンカー音素の調音運動を意味している。また、式 (4.4) におけるターゲット音素の分散行列 $\boldsymbol{\Sigma}$ は、初期位置の音素の音響空間上の分散とした。探索においては、分散は固定し、平均ベクトル $\boldsymbol{x}$ のみを更新した。式 (4.9) において定数 $\gamma_{m,x}$ で近似されている混合成分の事後確率 $p(m|\boldsymbol{x};\lambda^{(\boldsymbol{z})})$ は、$\boldsymbol{x}$ の更新ごとに更新後の $\boldsymbol{x}$ の値を用いて計算し直した。

式 (4.4) をそのまま用いて探索を行う場合、アンカー音素の数に比べ探索空間の次元が高いため推定結果が安定しない、探索空間に調音的制約が考慮されていないため探索結果の調音運動が物理的に不適切なもの（舌が口蓋を抜ける等）になる、などの現象が観測された。そこで、探索空間を制限するために以下の拘束項（第二項と第三項）をコスト関数に追加した。

$$\begin{aligned} J'(\boldsymbol{x}) = J(\boldsymbol{x}) &+ \alpha_1(\boldsymbol{x}-\boldsymbol{x}_0)^\top \boldsymbol{\Sigma}_{(a)}^{-1}(\boldsymbol{x}-\boldsymbol{x}_0) \\ &+ \alpha_2(\boldsymbol{x}-\boldsymbol{x}_c)^\top \boldsymbol{\Sigma}_{(a)}^{-1}(\boldsymbol{x}-\boldsymbol{x}_c) \end{aligned} \tag{4.11}$$

ここで、$\boldsymbol{x}_0$ は探索に用いる調音運動特徴量の初期位置、$\boldsymbol{x}_c$ はアンカー音素群の調音運動特徴量の平均ベクトル、$\boldsymbol{\Sigma}_{(a)}$ は調音–音声マッピングの構築に用いた調音運動データから求めた分散共分散行列である。$J'(\boldsymbol{x})$ の第二項は探索範囲をターゲット事象に最も発音が近い音素を与える調音運動の周辺に限定することを意味し、第三項は探索範囲を平均的な母音の調音運動の周辺に限定することを意味する。α_1, α_2 は、それぞれの拘束の強さを制御するパラメータである。

4.3.2　結果

各ターゲット事象における推定誤差を図 4.5 に示す。推定誤差は、推定結果と正解データの調音運動特徴量を磁気センサシステムの座標データに変換し、二乗平均平方根誤差（Root Mean Squared Error, RMSE）を算出したものである。図 4.5 は、推定対象話者と構造提供話者の組み合わせごとの結果である。各結果には探索の初期位置と正解データの RMSE（初期位置）、式 (4.4) をそのまま用いて推定した場合の推定誤差（推定結果（拘束なし））、式 (4.11) の拘束条件を適用した場合[4]の推定誤差（推定結果（拘束あり））、さらに推定対象話者と構造提供話者が異なる場合は、推定の前処理として 4.2.4 節の修正法を用いた場合の推定誤差（推定結果（拘束あり・修正あり））が示されている。また、図中の Ave. は全ターゲット音素の RMSE を平均した値である。

Ave. の値に注目すると、以下のことがわかる。

初期位置と推定結果（拘束なし）の比較

構造提供話者：男性と推定対象話者：女性の組み合わせ以外の結果で、推定結果（拘束なし）のほうが推定誤差が大きい（推定精度が悪い）ことがわかる。初期位置は、ターゲット音素に最も発音が近いアンカー音素の調音運動である。したがって、この傾向は提案手法に基づいた推定を行うよりもターゲット音素に発音が最も近い別音素の調音運動を推定結果としたほうが正解に近いということを表している。

[4]今回の実験では、$\alpha_1 = \alpha_2 = 0.2N$ とした。ここで N はアンカー音素の数である

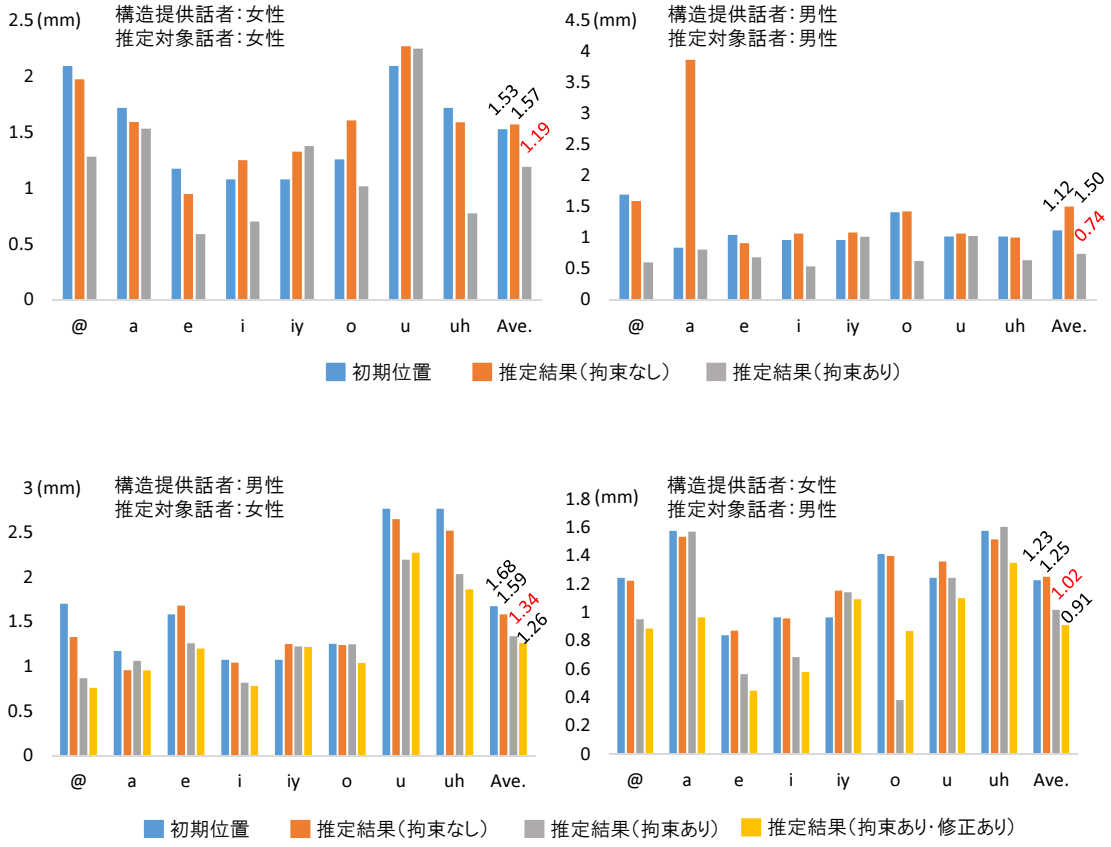

図 4.5: ターゲット音素ごとの推定誤差

推定結果（拘束なし）と推定結果（拘束あり）の比較　すべての話者の組み合わせにおいて推定結果（拘束あり）のほうが推定精度が高く、初期位置と比較しても、それを上回る精度となっている。これは、一切の制限がない調音空間を探索するのは難しかったが、適切に空間を制限したことによって正確な推定が可能になったことを表している。この結果から、調音空間を適切に制限すれば、アンカー音素と未観測音素の間の言語的な距離から未観測調音運動を推定できることが示された。

構造提供話者と推定対象話者が異なる場合について　構造提供話者が推定対象話者と異なる場合、同一話者を用いた場合よりも推定精度が悪化していることがわかる（図 4.5 の赤字、推定対象話者女性 1.19mm → 1.34mm, 推定対象話者男性 0.74mm → 1.02mm）。この原因は、話者間における構造的表象の差だと考えられる。しかし、推定結果（拘束あり・修正あり）に注目すると、推定精度が改善していることがわかる。つまり、4.2.4 節で提案した修正法が話者間の構造的表象の差を緩和したと考えられる。

各ターゲット音素における結果に注目すると、すべての条件において母音/iy/の推定結果が初期位置よりも悪化していることがわかる。この結果の原因として、母音/iy/が前舌

狭母音であり、他の母音（アンカー音素）から音響空間的にも調音空間的にも離れた位置に存在することが探索（推定）を難しくしていることがあげられる。これは、今回採択した初期値設定の限界であると考えられ、今後の課題である。

4.4　まとめ

本章では、未観測音素の調音運動の推定という新しい問題を通して、音声–調音マッピングの課題の一つである、

- 音声の音響情報のみに注目した推定であるため、適用できる問題が限られている。

という課題に取り組んだ。推定対象話者の音声が存在しない未観測音素の調音運動を推定するために、構造提供話者の発話から音声の構造的表象によって表される言語的特徴量を用いて推定する方法を提案した。構造提供話者の未観測音素を含む幾つか音素の音声から構造的表象を抽出し、それを拘束条件として、推定対象話者の調音運動を探索する。

MOCHA-TIMIT を用いた英語単母音を対象した実験をおこなった結果、ほぼ全ての単母音で提案法の有効性が確認できた。その一方で、探索の初期位置が推定結果に影響を与えることが明らかになった。

従来の音声–調音マッピングは、推定対象となる音響事象（音素）における音響情報のみを頼りに推定をおこなうため、推定対象の音響情報が存在しない音響事象（未観測音素）の推定をおこなうことが難しかった。一方、提案法では、複数の事象から抽出される情報を用いることで、音響情報が存在しない音響事象の調音運動の推定を可能にした。つまり、従来法では考慮されていなかった推定対象以外の音響事象を考慮することで、音声–調音マッピングの適応範囲を拡大することができたと言える。

今回の実験では、単母音のみを対象にした検討を行ったが、今後の課題として全ての音素を対象にした検討をおこなう必要がある。全ての音素に対して適用できれば、任意の話者の様々な音素を含む連続発話から発話系列を表す構造的表象を抽出し、それを特定の話者の音響空間上に調音空間を通して再配置することで、前章の話者正規化音声–調音マッピングシステムとは異なる、話者の言語的特徴に注目した話者非依存な音声–調音マッピングシステムへの応用も期待できる。

第5章

結論

5.1　まとめ

本論文では、音声から調音運動を復元する逆推定法、特に音声–調音マッピングの適用範囲の拡大のために、任意話者の音響情報、および言語情報に基づく調音運動の推定法を検討した。音声–調音マッピングでは、音声–調音パラレルデータから両者の特徴量空間の対応関係をモデル化し、そのモデルに基づいて入力された音声を調音運動に変換する。このとき、音声–調音パラレルデータが特定話者（モデル話者）のデータであった場合、変換モデルは話者依存モデルとなる。この話者依存モデルは、モデル話者の音声にしか適用できない。この限定された適用範囲は、音声–調音マッピングを実用的な応用から遠ざけている一つの要因である。

そこで、本論文では、適用範囲を拡張するために、まず話者正規化音声–調音マッピングを提案し、任意の話者の音響情報から調音運動を推定する手法を提案した。この話者正規化音声–調音マッピングでは、任意話者の音声の音響特徴量を話者変換によってモデル話者の音声特徴量に変換（正規化）する。この正規化は、入力音声と話者依存モデルの間での音響的ミスマッチを軽減し、任意の話者の音声に対する推定精度の向上させる狙いがある。話者変換と音声–調音マッピングを縦列に接続した連結モデルと二つの変換モデルを一つの変換モデルに統合した分布共有モデルを提案し、MOCHA-TIMIT を用いた実験をおこなった。その結果、分布共有モデルを用いることで、モデル話者以外の話者の音声に対して良好な推定精度で調音運動が推定できることが明らかになった。

また、従来の音声–調音マッピングは音声の音響特徴量のみに基づいて調音運動を推定していため、音響情報の存在しない音声の調音運動を推定することが難しいという問題があった。そこで、音声に含まれる言語情報に注目して、音響情報の存在しない音声である未観測音素の推定に取り組んだ。未観測音素の調音運動の推定は、音声の構造的表象によって表される言語的特徴量に基づいておこなわれる。提案法では、構造提供話者から未観測音素に関する構造的表象を抽出し、それを制約として推定対象話者の調音空間を探索することで、推定対象話者の未観測調音運動を推定する。MOCHA-TIMIT を用いた実験の結果、探索空間である推定対象話者の調音空間を適切に制限することで、未観測調音運動の推定が良好におこなえることが明らかになった。

5.2　今後の展望

本論文では、音声–調音マッピングの適用範囲を、1) 特定話者の音響情報から任意の話者の音響情報へ、2) 音響情報が存在する音声から音響情報が存在しない音声へ、拡張するための検討をおこなった。これらの検討は、逆推定問題における推定対象側（入力側）の音声に関する適用範囲を拡張するものであった。一方で、推定目標側（出力側）の調音運動に関しては、音声–調音パラレルデータが存在する話者に限定されたままである。特定話者の音声–調音パラレルデータのみから、任意の話者の調音運動空間を推定するためには、音響空間における二話者間の対応から、調音運動空間における対応を推定する必要がある。調音運動データが存在しない任意の話者の調音運動空間を推定するのは、非常に難しい課

題であるが、もし実現できれば、音声–調音マッピングの適用範囲をさらに拡大することができため、今後の研究が期待される。

謝辞

本研究ならびに本論文の執筆にあたり、多大なる御指導、ご鞭撻を賜りました指導教員である峯松信明教授、齋藤大輔講師に深く感謝いたします。また、廣瀬啓吉名誉教授には短い期間でしたが多くのことを学ばして頂きました。修士課程での指導教員であった九州大学の鏑木時彦教授、若宮幸平助教には、博士課程でも大変にお世話になりました。札幌保健医療大学の末光厚夫准教授には、磁気センサシステムの測定や研究の議論など、様々な場面で助けて頂き、感謝の念が絶えません。NTT コミュニケーション科学基礎研究所の廣谷定男先生には、修士課程でのインターンでお世話になった後も、様々な場面で貴重な助言を頂きました。深く感謝いたします。

峯松・齋藤研究室の皆さまの御蔭で、大変に充実した研究生活を送ることができました。博士課程の先輩として私を力強く導いて頂いた柏木陽佑氏、研究室において後輩の私に暖かく接して頂いた笠原俊氏、実質同期として最も長い時間共に研究に励んで頂き、磁気センサシステムの測定の助手としても力を貸して頂いた橋本哲弥氏に深く感謝いたします。また、後輩の島田智大と外山翔平氏には、磁気センサシステムの被験者として実験に参加して頂き、非常に貴重なデータを測定することが出来ました。心から感謝いたします。峯松研究室の学生の皆様は、常に研究に励み素晴らしい成果を生み出し続けていました。その姿勢が、私の研究生活の原動力となっていました。皆様、本当に感謝いたします。

最後に私の長きに渡る学生生活を辛抱強く支えて頂いた家族に感謝いたします。

2016 年 12 月 1 日
内田秀継

参考文献

[1] H. Zhen, K. Richmond, and J. Yamagishi, "Articulatory Control of HMM-Based Parametric Speech Synthesis Using Feature-Space-Switched Multiple Regression," *IEEE Trans. Speech Audio Process.*, **17(6)** 1171-1185 (2013).

[2] V. Mitra, H. Nam, C. Y. Wilson, E. Saltzman, and L. Goldstein"Articulatory Information for Noise Robust Speech Recognition," *IEEE Trans. Speech Audio Process.*, **19(7)**, 913-1924 (2011).

[3] P. Lumban Tobing, K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Articulatory Contrallable Speech Modification Based on Gaussian Mixture Models with Direct Waveform Modification Using Spectrum Differential," In *Proc. Interspeech2015*, 3350–3354 (2015).

[4] P. Badin,A. Youssef, G. Bailly, F. Elisei, and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," In *L2SW, Workshop on "Second Language Studies: Acquisition, Learning, Education and Technology,"* (2010).

[5] S. Fujita, J. Dang, N. Suzuki, K. Honda, "A Computational Tongue Model and its Clinical Application," *Oral Science International*, **4(2)**, 97-109 (2007).

[6] T. Kaburagi, K. Wakamiya, and M. Honda, "Three-dimensional electromagnetic articulography," *Journal of the Acoustical Society of America*, **118**, 428–443 (2005).

[7] 竹本浩典, 北村達也, "ＭＲＩに基づく音声生成の研究手法の概要," 電子情報通信学会誌, Vol. 94, No.7, pp.585-590 (2011).

[8] B. H. Story, "Technique for tuning vocal tract area functions based on acoustic sensitivity functions," *Journal of the Acoustical Society of America*, **119**, 715–718 (2006).

[9] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulatotion and recohnizing gestures from acoustics with a neural network trained on x-ray microbeam data," *Journal of the Acoustical Society of America*, **92(2)**, 688–700 (1992).

[10] S. Hiroya, and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech and Audio Processing*, **12**, 175–185 (2004).

[11] T. Toda, W. A. Black, and K. Tokuda, “Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model”, *Speech Commun.*, **50**, 215–227 (2007).

[12] B. Uria, S. Renals, and K. Richmond, “A deep neural network for acoustic-articulatory speech inversion,” In *Proc. NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning* (2011).

[13] C. P. Browan, and Louis Goldstein, “Articulatory Phonology: An Overview,” *Haskins Laboratories Status Report on Speech Research*, SR-111 / 112, pp. 23–42, (1992).

[14] T. Kaburagi, “Morphological and Acoustic Analysis of the Vocal Tract Using a Multi-Speaker Volumetric MRI Dataset,” In *Proc. INTERSPEECH2015* (2015).

[15] Liberman, A.M., Cooper, F.S., Shankweiler, D.P. and Studdert-Kennedy, M., ”Perception of the speech code”, *Psych. Rev.,* **74(6)**, 853–870 (1967).

[16] Liberman, A.M. and Mattingly, I.G., ”The motor theory of speech perception revised”, *Cognition*,**21**, 1–36 (1985).

[17] 鏑木時彦 編著 (2010), 「音声生成の計算モデルと可視化」, コロナ社.

[18] 北村達也, 竹本浩典, 本多清志, “母音発話 MRI データに基づく声道模型の音響特性,” 信学技報, EM2007-89, pp. 19–24 (2007).

[19] S. Masaki, M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto, Y. Nakamura, and N. Ninomiya, “MRI-based speech production study using a synchronized sampling method,” *Journal of the Acoustical Society of Japan*, **E20(5)**, 375–379 (1999).

[20] D. Whalen, K. Iskarous, M. Tiede, D. Ostry, H. Lehnert-Lehouillier, E. Vatikiotis-Bateson, D. Hailey, “The Haskins optically corrected ultrasound system (HOCUS)”, *Journal of Speech, Language, and Hearing Research,* **48(3)**, 543-553 (2005).

[21] T. Hueber, G. Chollet, B. Denby, M. Stone, “Acquisition of Ultrasound, Video and Acoustic Speech Data for a Silent-Speech Interface Application,” *In Proc. 8th International Seminar on Speech Production,* (2008)

[22] P. W. Schönle, K. Gräbe, P. Wening, J. Höhne, J. Schrader, and B. Conrad, “Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract,” *Brain and Language,* **31**, 26–35 (1987).

[23] T. Kaburagi and M. Honda, “ Dynamic articulatory model based on multi-dimensional invariant-feature task representation,” *Journal of the Acoustical Society of America*, **110**, 441–452 (2001).

[24] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta and M. T. T. Jackson, “ELecromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements,” *Journal of the Acoustical Society of America*, **92(6)**, 3078–3096 (1992).

[25] “NDI Measurement science”,http:// www.ndigital.com/msci/products/wave-speech-research/, (参照 2016-11-29).

[26] I. Toshima, S. Hiroya, T. Mochida, and H. Gomi, “Motor command invariance during speech production investigated by physiological perioral dynamics model,” In Proc., International Seminar on Speech Producion, (2011)

[27] S. Aryal and R. G-Osuna, “Reduction of non-native accents through statistical parametric articulatory synthesis,” *Journal of the Acoustical Society of America*, **137(1)**, 433–446 (2015).

[28] A. Suemitsu, J. Dang, T. Ito, M. Tiede, ”A study on effect of real-time articulatory feedback presentation in American English pronunciation learning”, *In Proc., Acoustic society of Japan Autumn Meeting 2013* (2013)

[29] T. Kaburagi, “A method for estimating vocal-tract shape from a target speech spectrum,” *Acoustical Science and Technology,* **36(5)**, 428-437 (2015).

[30] P. Zhu, L. Xie, Y. Chen, “Articulatory Movement Prediction Using Deep Bidirectional Long Short-Term Memory Based Recurrent Neural Networks and Word/Phone Embeddings,” *In Proc., Interspeech2015*, (2015).

[31] K. Richmond, “Preliminary Inversion Mapping Results with a New EMA Corpus,” *In Proc., Interspeech2009*, (2009).

[32] K. Richmond, “Estimating articulatory parameters from the acoustic speech signal,” Ph.D. dissertation, the Centrefor Speech Technolgy Research, Edinburgh University (2002).

[33] A. Wrenh, “The MOCHA-TIMIT articulatory database,” http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html (参照 2016-11-13).

[34] “TIMIT Acoustic-Phonetic Continuous Speech Corpus,” https://catalog.ldc.upenn.edu/ldc93s1 (参照 2016-11-29).

[35] “Welcome to mngu0,” http://www.mngu0.org/ (参照 2016-11-29).

[36] A. Kain, and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” *In Proc., International Conference on Acoustic, Speech and Signal Processing*, (1998).

[37] T. Toda, A. W. Black and K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory,” *IEEE Trans. Audio, Speech, and Language Processing*, **15(8)**, 2222–2235 (2007).

[38] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Many-to-many eigenvoice conversion with reference voice” *In Proc., Interspeech2009*, (2009).

[39] P. Birkholz, C.Jackèl, B. J. Kröger, “Construcion and control of a three-dimensional voccal tract model,” In *Proc. ICASSP 2006*, (2006).

[40] 徳田恵一, 小林隆夫, 深田俊明, 斎藤博徳, 今井聖, “メルケプストラムをパラメータとする音声のスペクトル推定” 電子情報通信学会論文誌 A, **J 74-A**, 1240–1248, (1991).

[41] 内田秀継, 齋藤大輔, 峯松信明, “音声の構造的表象を用いた未観測調音運動の推定法に関する検討,” 信学技報, vol. 115, no. 392, SP2015-86, pp. 7–12 (2016).

[42] H. Uchida, D. Saito, N. Minematsu, “Prediction of the articulatory movements of unseen phonemes of a speaker using the speech structure of another speaker,” In *Proc. INTERSPEECH2016* (2016).

[43] N. Minematsu, S. Asakawa, and M. Suzuki, Y. Qiao, “Speech structure and its application to robust speech processing,” *Journal of New Generation Computing*, **28**, 299–319, (2010).

[44] M. Pitz and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” *IEEE Trans. Speech and Audio Processing,* **13**, 930–944 (2005).

[45] Y. Qiao and N. Minematsu, “A study on invariance of fdivergence and its application to speech recognition,” *IEEE Trans. on Signal Processing*, **58(7)**, 3884–3890 (2010).

[46] S. Kasahara, S. Kitahara, N. Minematsu, H. -P. Shen, T. Makino, D. Saito, and K. Hirose, “Improved ans robust prediction of pronunciation distance for individual-basis clustering of world Englishes pronunciation,” In *Proc. ICASSP2014* (2014).

[47] M. Suzuki, G. Kurata, M. Nishimura, and N. Minematsu, “Discriminative re-ranking for automatic speech recognition by leveraging invariant structures,” *Speech Commun.*, **72**, 208–217 (2015).

[48] 齋藤大輔, 朝川智, 峯松信明, 廣瀬啓吉, “構造的表象からの音声生成に関する基礎的検討”, 信学技報, SP2007-80, pp.55-60 (2007).

[49] 見原隆介 , 齋藤大輔, 峯松信明, 廣瀬啓吉 “音声の構造的表象に基づく異言語間・異話者間の音声変換手法”, 信学技報, SP2009-71, pp.55–60 (2009).

[50] 朝川 智, 峯松信明, 広瀬啓吉,“音声の構造的表象に基づく英語学習者発音の音響的分析,” 信学論 (D), **J90-D**, 1249-1262 (2007).

[51] R. Hagiwara, “Dialect variation and formant frequency: The American English vowels revisited,” *Journal of the Acoustical Society of America*, **102(1)**, 655–658 (1997).

[52] 匂坂芳典，浦谷則好, “ATR 音声・言語データベース,” 音響学会誌, 48 巻, 12 号, pp. 878-882 (1992).

[53] “音声資源コンソーシアム,” http://research.nii.ac.jp /src /phoneticallybalanced.html (参照 2016-11-13).

[54] S. Maeda, “An articulatory model of the tongue based on a statistical analysis,” *Journal of the Acoustical Society of America*, **65(S22)**, (1979).

発表文献

雑誌論文

[1] Hidetsugu Uchida, KoheiWakamiya, Tokihiko Kaburagi, “Improvement of measurement accuracy for the three-dimensional electromagnetic articulograph by optimizing the alignment of the transmitter coils,” Acoustical Science and Technology, 2016.

国際会議論文

[2] HidetsuguUchida, Daisuke Saito, Nobuaki Minematsu, “Prediction of the articulatory movements of unseen phonemes of a speaker using the speech structure of another speaker,” Interspeech 2016(accepted).

[3] Y. Yang, HidetsuguUchida, Daisuke Saito, Nobuaki Minematsu, “Voice Conversion Based on Matrix Variate Gaussian Mixture Model Using Multiple Frame Features,” Interspeech 2016(accepted).

[4] HidetsuguUchida, Daisuke Saito, Nobuaki Minematsu, Keikichi Hirose, “Statistical Acoustic-to-Articulatory Mapping Unified with Speaker Normalization Based on Voice Conversion,” Interspeech 2015.

[5] Kohei Wakamiya, Hidetsugu Uchida,Tokihiko Kaburagi, “The Effect of Additional Transmission Channel in Three-Dimensional Electromagnetic Articulography,” KYJCA 2015.

[6] Hidetsugu Uchida, KoheiWakamiya, Tokihiko Kaburagi, “A Study on the Improvement of Measurement Accuracy of the Threedimensional Electromagnetic Articulography,” Interspeech 2014.

国内研究会・全国大会

[7] 内田秀継, 斉藤大輔, 峯松信明, “音声の構造的表象を用いた未観測調音運動の推定に関する検討,” 電気情報通信学会音声研究会, 2016-1.

[8] 内田秀継, 斉藤大輔, 峯松信明, “ 音声の構造的表象を用いた未観測調音運動の推定に関する実験的検討, ”日本音響学会秋季研究発表会, 2015-9.

[9] 内田秀継, 斉藤大輔, 峯松信明, 広瀬啓吉, “ 統計的音声－調音マッピングにおける声質変換を利用した話者正規化法の検討, ” 電気情報通信学会音声研究会, 2014-11.

[10] 内田秀継, 斉藤大輔, 峯松信明, 広瀬啓吉, “ 話者変換音声を対象とした音声－調音マッピングに関する実験的検討, ”日本音響学会秋季研究発表会, 2014-9.

[11] 若宮幸平, 内田秀継, 鏑木時彦, “ 3 次元磁気センサシステムにおける送信チャネル数に関する検討 , ”日本音響学会秋季研究発表会, 2014-9.

[12] 内田 秀継, 若宮 幸平, 鏑木 時彦, “ 三次元磁気センサシステムにおける送信コイルの配置の検討の最適化についての検討, ” 電子情報通信学会音声研究会, 2013-11.

[13] 内田 秀継, 若宮 幸平, 鏑木 時彦, “ 三次元磁気センサシステムにおける送信コイル配置の検討と精度評価, ” 日本音響学会, 秋季研究発表会, 2013-9.

[14] 内田 秀継, 若宮 幸平, 鏑木 時彦, “ 三次元磁気センサシステムにおける送信コイル配置の検討, ” 日本音響学会, 秋季研究発表会, 2012-9

学位論文

[15] 内田秀継 “三次元磁気センサシステムによる発話運動解析に関する研究,” 修士論文, 九州大学大学院芸術工学府 (2014)

[16] 内田秀継 “三次元磁気センサシステムを用いた位置推定における非決定性の検討,” 学士論文, 九州大学大学院芸術工学府 (2012)

付録 A

日中二カ国語話者の音声–調音パラレルデータの収録

研究の一環として、日中二カ国語話者の日中二カ国語話者の音声–調音パラレルデータの測定をおこなった。その測定手順と分析結果をまとめる。

被験者

測定の被験者は、中国（北京市）出身の 50 代の男性である。14 歳から日本語の勉強を始め、現在は日本（兵庫県）で音声学（日本語教育・中国語教育）の研究を行っている（詳細な情報は表 A.1 参照)。日本語の発音は、極めて流暢であり母語話者相当と言える。

表 A.1: 被験者データ

性別	男性
年齢	58 歳
出身地	北京市（40 歳まで在住）
出身地以外で長期滞在 （1 年以上）した 中国の地域とその期間	天津市（1.5 年間）, 河北省（8 年間）
日本語を勉強し始めたときの年齢	14 歳
日本で長期滞在した県とその期間	東京（1.5 年間）, 大阪市（2 年間）, 兵庫県（18 年間）

収録環境

調音観測システムとして、北陸先端科学技術大学院大学所有の EMA(AG-500, Carstens 社) を使用した。AG-500 は、キューブ型のフレームに 6 個の送信コイルが配置された三次元磁気センサシステム (3D-EMA) である。被験者は椅子に腰かけた状態で頭部をキューブの内部に入れた状態で発話を行う。被験者の前方には、音声収録用のマイクと読み上げ文表示用のディスプレイが設置されている。

EMA のセンサは、上下の口唇 (UL,LL)、下顎切歯（顎, LI)、舌尖 (TT) および舌背の 2 点 (TB,TD) の計 6 点に取りつけた (図 A)。これらのセンサは、およそ正中矢状面に沿って取り付けられた。さらに、頭部の運動の補正用として、鼻の付け根、左右の耳の付け根の計 3 点、測定データの座標系の原点を定義するために、上顎切歯に 1 点 (UI)、合計で 10 個（6+3+1）のセンサを被験者に取り付けた。

読み上げ文

日本語の読み上げ文は、ATR 音素バランス 503 文 [52] のサブセットの A セットおよび B セットからなる計 100 文、中国語の読み上げ文は、合文法無意味文のみで構成された名工大中国語セット [53] の 100 文を採用した。

測定手順

被験者がセンサの装着感に十分に慣れた状態で、収録を始めた。中国語と日本語を 5 文ずつ交互に発話していき、読み誤った文は最後にまとめて再収録した。読み上げ文の収録後、舌尖で口蓋をなぞってもらい、口蓋形状をトレースした。最後に、測定データの座標系の軸を定義するために噛み合わせ面の測定を行った。

データ処理

EMA の測定データは 3 次元運動データとして収録される。この 3 次元データは、頭部の運動情報、上顎切歯のセンサ情報及び、噛み合わせ面の情報を用いて、被験者の正中矢上面が x-y 平面、噛み合わせ面が x-z 平面、上顎切歯のセンサが原点となる座標系に変換される。変換した結果、本来、噛み合わせ面に直交するはずの y 軸が傾いていることが明らかになった。これは、噛み合わせ面の測定が正常に行われなかったこと示唆している。そこで、下顎切歯のセンサの x-y 平面での運動に主成分分析 (Principal Component Analysis; PCA) を行い、その第一主成分が x-z 平面と直交するように x-y 平面を回転した。これは、顎の開閉方向を y 軸として新たに定義したことになる。

また、同時収録された音声データを参考に、手作業で発話の前後の無音区間を取り除いた。最後に、EMA の計測誤りを含む発話を取り除いた結果、中国語 98 文、日本語 99 文の音声–調音パラレルデータが作成された。

分析

表 A.2: Maeda 調音モデルの基底の抽出順

-x,-y は各パラメータの x 座標,y 座標を表す

抽出した順番 (N)	1 ∼ 2	3 ∼ 4	5 ∼ 6	7 ∼ 8	9 ∼ 12
パラメータ	UI-x,UI-y	TD-x,TD-y	TB-x,TB-y	TT-x,TT-y	UL-x,UL-y,LL-x,LL-y
対応する調音器官	顎	舌の後部	舌の中部	舌尖	両唇

収録した調音運動データを、統計的手法を用いて分析し、同一話者によって発話された日本語及び、中国語の調音運動の特性を明らかにする。分析法として、PCA と Maeda 調音モデル [54] を用いる。Maeda 調音モデルは、PCA と同じく、観測値を幾つかの基底とそれに対する重みによって表す線形モデルである。PCA との相違点は、基底を特定の調音器官（または部位）の運動に注目して抽出する点である。また、Maeda 調音モデルの基底の抽出は、一つずつ順に行う。時刻 t における観測値を $\boldsymbol{x}_t \in \mathcal{R}^{D_x} (t = 1 \sim T)$ としたとき、N 番目の基底 $\boldsymbol{b}^{(N)}$ は、以下の式によって抽出される。

$$\boldsymbol{b}^{(N)} = \arg\min_{\boldsymbol{b}} \left(\sum_{t=1}^{T} \bar{\boldsymbol{x}}_t^{(N-1)} - \boldsymbol{b} w_t^{(N)} \right)^2 \tag{A.1}$$

ここで、

$$\bar{\boldsymbol{x}}_t^{(N-1)} = \boldsymbol{x}_t - \boldsymbol{B}^{(N-1)}\boldsymbol{W}_t^{(N-1)} - \bar{\boldsymbol{x}} \tag{A.2}$$

である。$w_t^{(N)}$ は、N 番目の基底の重みであり、$\boldsymbol{W}_t^{(N-1)} = (w_t^{(1)}, \cdots, w_t^{(N-1)})^\top$ である。また、$\boldsymbol{B}^{(N-1)} = (\boldsymbol{b}^{(1)}, \cdots, \boldsymbol{b}^{(N-1)})$ である。$\bar{\boldsymbol{x}}$ は、観測値の平均ベクトルである。重み $w_t^{(N)}$ は、$\boldsymbol{x}$ の特定のパラメータに関して PCA を行うことで得られる、第一主成分に対する重みとする[1]。つまり、この手順で抽出される N 番目の基底は、N 番目に着目したパラメータの運動の特性を恣意的に抽出したものとなっている。今回の分析では、調音運動データの矢状面における運動のみを分析対象する。つまり、6 つのセンサの x-y 平面での運動を考えるため、$D_x = 12$ となる。Maeda 調音モデルの基底の抽出に用いたパラメータを表 A.2 に示す。

分析の結果を図 A.2 に示す。PCA 及び、Maeda 調音モデルを用いて得られた基底に重み（±7.5）を掛けて得られたパラメータ（センサ位置）と、全データを平均して得られたパラメータがぞれぞれドットで示されている。ここで、Maeda 調音モデルの基底に関しては、舌のセンサに着目して抽出した基底のみを示した。PCA の基底に注目すると、第一主成分では、二つの言語間で大きな違いは見られないが、第二主成分では舌の運動に違いが見られる。この結果から各言語に特有の運動パターンが存在することが示唆される。一方で、Maeda 調音モデルの基底に注目すると、いずれの基底においても大きな違いは見られない。Maeda 調音モデルの基底が舌の各部位における生理学的特性、例えば、筋肉の構造などを反映した運動を抽出していると考えることができる。これらの分析結果から、今回の二カ国語話者が舌の各部位の運動を組み合わせて、言語特有の舌の運動を生成している様子が確認できた。

[1]複数のパラメータに着目しても構わない。また、同じパラメータに複数回着目することもある。

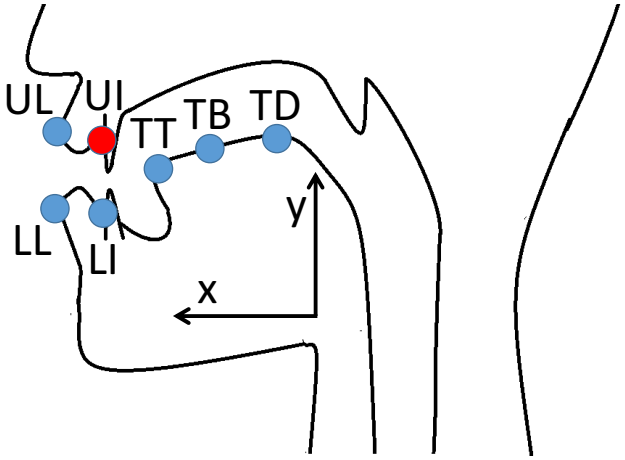

図 A.1: センサー（測定点）の位置

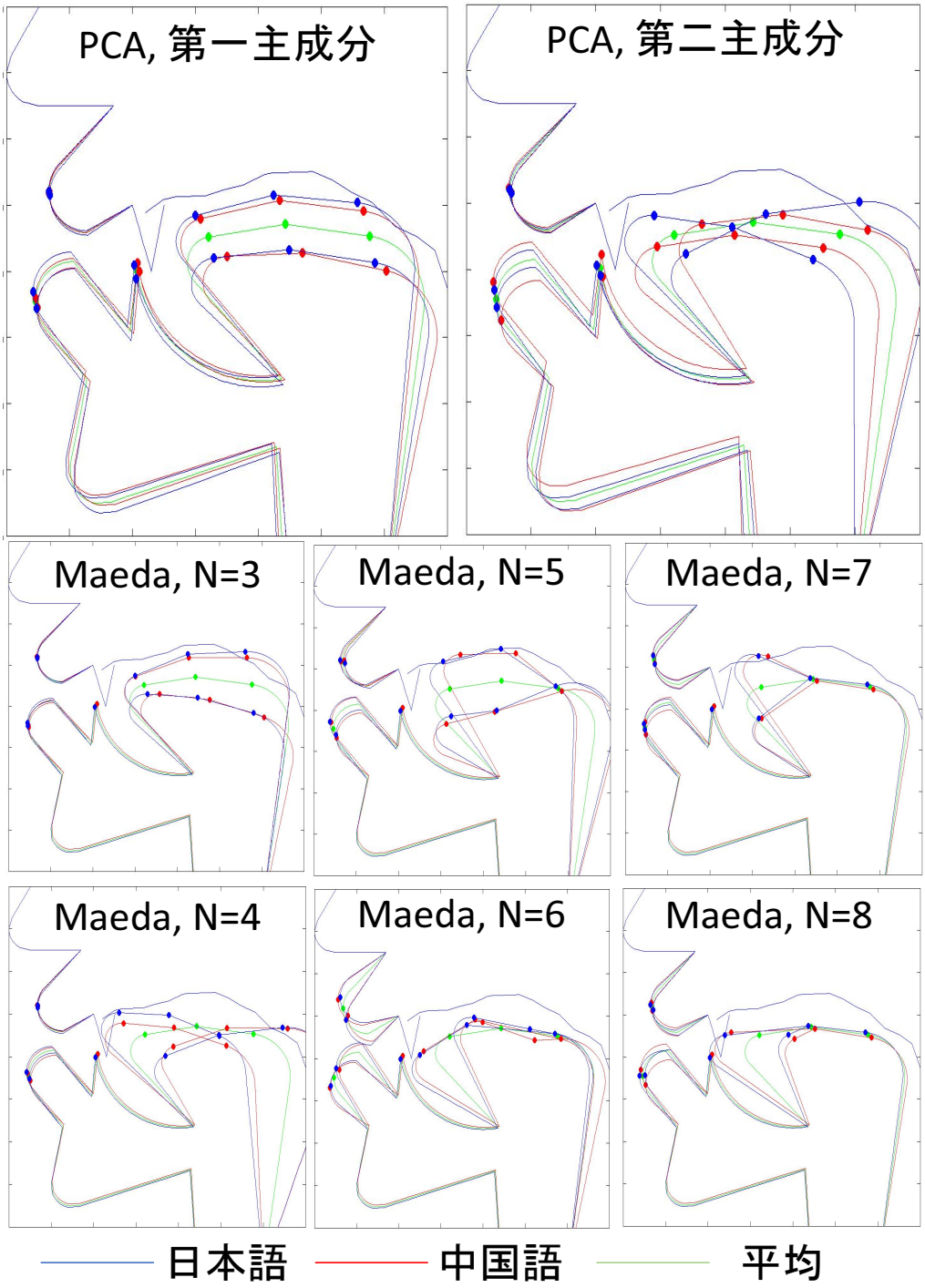

図 A.2: PCA 及び Maeda 調音モデルによる分析結果

付録B

調音モデルの調整パラメータ

図 3.10 の調音モデルにおける主要な調整パラメータを以下に示す。

表 B.1: 調音モデルの調整パラメータ (1)

パラメータ名	半径 (mm)
TTc	4
TDc	16
LLc	6
ULc	6

表 B.2: 調音モデルの調整パラメータ (2)

パラメータ名	x 座標 (mm)	y 座標 (mm)
P1	4	60
P2	0	15
P4	5	0
P11	60	-60

口蓋形状 (PLs-PLe) は、全測定データにおける舌上のセンサー（TT,TD,TB）の座標データから、x 軸に沿って 1mm 毎に最大の y 座標を持つ点を抽出し、その点を繋いだ曲線とした。

付録C

音素表

本研究で用いた音素表記とその音素を含む単語例（当該音素を太字で示す）を以下の表に示す。なお、単語表記は、MOCHA-TIMIT に同梱されている音素ラベルデータに則したものである。

表 C.1: 音素表 (子音)

音素表記	単語例
p	**p**ick
b	**b**efore
t	pe**t** rol
d	**d** ad
k	**k**eep
g	**g**rade
ch	**ch**ocolate
jh	**j**ewel
f	**f**ail
v	sa**v**e
th	**th**irty
dh	**th**is
s	s**t**ole
z	oo**z**e
sh	di**sh**
zh	plea**s**ure
h	**h**ome
n	**n**aer
m	**m**ost
ng	si**ng**er
l	**l**ive
r	**r**oll
y	**y**oung
w	**w**orry

表 C.2: 音素表（母音）

音素表記	単語例
a	ex**a**m
aa	pl**a**nt
ai	br**i**ght
@	**a**bout
@@	f**ur**rier
i	s**i**mple
ii	**ea**stern
iy	mon**ey**
i@	n**ear**
u	g**oo**d
uu	n**ew**
uh	s**u**n
e	y**e**ll
ei	pl**a**ce
eir	w**here**
o	**o**ften
oo	p**oo**r
oi	v**oy**age
ou	**o**cean
ow	l**ow**