音楽情報科学　39－20
（2001. 2. 23）

発話機構モデルによる声道形状逆推定法を用いた音韻と発話様式の分析

西塘憲一[1),2)]　党建武[1)]　本多清志[1)]

[1)]ATR 人間情報通信研究所
〒619-0288 京都府相楽郡精華町光台 2-2-2

[2)]長岡技術科学大学
電気系電気電子システム工学科

E-mail: xkenichi@isd.atr.co.jp, jdang@isd.atr.co.jp, honda@isd.atr.co.jp

あらまし　本研究では，発話機構モデルによる声道形状逆推定法を用いた音韻と発話様式の分析を試みた。人間の発話機構のメカニズムを忠実的に再現する調音モデルの使用により音声波形から声道形状の推定過程に人間の生理学的拘束条件を導入した。調音器官の動きと音響パラメータとの対応関係をX線マイクロビーム日本語調音データベースに基づいて分析したところ，舌背の前後方向における位置は母音の第一・第二フォルマントの差とほぼ一意的な関係を持つことがわかった。この関係を新たな拘束条件として適用することにより，声道形状の推定精度が明らかに上昇することを確認した。提案した発話機構モデルを用いる逆推定法により子音を含む音節の音声波形から声道形状系列を推定した。X 線マイクロビームデータを用いて評価したところ，推定結果は妥当であることを確認した。発話の口調，速度及び強さの組合せによる音声資料を用い，逆推定方法により発話様式の推定と分析を行った。推定により得られた下顎と舌背の運動軌跡はX線マイクロビームデータと同様の傾向を示した。

キーワード　逆推定，生理学的調音モデル，発話様式，音声生成，音声合成

Application of the inverse estimation method with a physiological articulatory model to analyze speech style

Ken'ichi Nishigaki[1),2)], Jianwu Dang[1)], Kiyoshi Honda[1)]

[1)]ATR Human Information Processing Research Labs.
2-2-2 Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0288, Japan.

[2)]Nagaoka University of Technology
Department of Electrical Engineering

Abstract　This study attempted to analyze speech styles using an inverse estimation method with a physiological articulatory model, where physiological constraints of human were introduced into the inverse estimation processing via the plausible physiological model. Relations of articulatory movements and formant patterns were investigated based on an X-ray microbeam Japanese database. It is found that the anterior-posterior position of the tongue dorsum has a nearly one-to-one relationship with the frequency difference between F1 and F2. It is confirmed that the inverse estimation of vocal tract shapes became more accurate by implementing the relation in the physiological model as a constraint. Vocal tract shapes for producing syllables with a consonant were estimated using the proposal method, and were evaluated using articulatory data from the microbeam system. Speech materials with different rhythms, speech rates and stresses were used in the analysis of speech styles. The articulatory movements obtained from the inverse estimation showed the same tendency as those observed in X-ray microbeam data.

key words　Inverse estimation, Physiological articulatory model, Speech style, Speech production, Speech synthesis

1.はじめに

音声生成過程では，発話意図に基づいた運動指令に従い，発話器官を駆動して声道形状の時間変化を生成し，音源による音波がその声道形状に調整され口唇または鼻孔から放射する。音声生成メカニズムの解明に関する研究では，上記の過程とは逆に，音声波形から声道形状または調音状態を推定し，さらにさかのぼって運動指令を推定する手法が用いられる。このような推定過程は一般的には逆推定と称され，一つの音韻識別法として音声認識の高度化への適用が期待されている。

音声波形から声道形状を推定する作業において最大の問題は，音声波形から声道形状を一意的に決定できないことである。換言すれば，音声波形から声道形状を求める推定過程は，多意的で(一から多への)非線形のマッピングである[1]。すなわち，理論的には一つの音声波が生成できる声道形状は無限に存在する。このようなマッピングの冗長性は，音声生成過程に固有な特性によるものである。しかし，逆推定の複雑さは音響パラメータと調音パラメータの選択にも依存する。逆推定に付随する多意性を抑えるため，従来の研究ではさまざまな拘束条件が導入されてきた。それらの拘束条件を大別すると，形状学的(空間的)な拘束条件と動的(時間的)な拘束条件に分けられる。前者は，静的な拘束条件であり，推定した声道形状の合理性を保証する。後者は発話器官の運動の連続性と均衡にかかわっている。

これまでに逆推定に用いられた手法には，音響的特徴量に等価な音響管を推定するものと，より強い拘束条件を持つ調音モデルを用いて調音パラメータを推定するものがある。前者の場合調音パラメータは統計的なもので，生理学的な拘束条件は明示的には入れられておらず，特に動的な発話を模擬する際に，人間の生理学的な拘束条件を正確に反映するという保証はない。このような点からみると，推定信頼度と精度を高めるには，人間の発話機構を忠実に再現する調音モデルを用いることが望ましい。本研究では，後者の方法にならい，三次元の生理学的調音モデルを用い音声波形から声道形状への推定を試みた[7]。さらに，音声の入力，逆推定，声道形状の表示を統合し，人間のいわば「ものまね」を模擬したシステムを作り，感情表現などの違いを含む発話の様式の分析を試みた。

2.生理学的調音モデルのもつ拘束条件

本研究における三次元生理学的調音モデルは，一名の成人男性話者の３次元 MRI データに基づいて構築し，筋の収縮によりモデルを駆動するものであり，形状学的および生理学的な面で人間の発話機構を比較的に忠実に再現している[6～8]。ここでは，このモデルの使用によりもたらされる利点及び逆推定の多意性を抑える方法などを検討する。

2.1 生理学的モデルと従来の拘束条件

従来の研究では，逆推定の多意性を抑えるために，空間的な拘束条件と時間的な拘束条件の片方または双方が用いられている。生理学的調音モデルの使用により，静的および動的な拘束条件以外に生理学的な拘束条件も導入される。

我々が構築したモデルは，生理学的に人間の発話機構と同様に筋の収縮力により発話器官を駆動する。それを実現するため，高解像度のMRI画像を用い解剖学の資料を参照しながら，筋の配置を抽出してモデル化したものである。モデルの制御には，調音目標を指定することによって筋電信号と同様の筋の収縮パターンを推定して発話器官を駆動する方式を用い，この過程で人間の生体メカニズムが再現されるように設計されている。このモデルにおいて実現される生理学的なメカニズムを，「あいあい」という音素系列の発話の例を用いて説明する。例えば，人間が「あいあい」をゆっくりと発話すると，下顎は大きな開閉動作を行い，舌も前上方と後下方に大きく動く。しかし早口で発話する場合，下顎はほとんど動かず，舌はより中性的な狭い範囲の動作に変化する。また，人間の発話器官の運動速度には限界があるので，調音目標が指定されても持続時間が短い場合その目標に到達する前に次の目標に向けて動き出す。このような人間の発話機構固有の生理的特徴は生理学的調音モデルにより再現することができるが，従来のモデルでは実現しにくい。

従来の推定方法では静的および動的な拘束条件はそれぞれ別々の数学的手法により導入されるのに対して，我々の生理学的なモデルではこれら２つの拘束条件が統合されている。前述のように，３次元の生理学的調音モデルは舌，口唇，下顎の可動器官と，声道壁および鼻腔の不変構造からなる。このような構成のモデルでは，下顎と舌上の数個の点（例えば

文献[7]で用いられた3つの制御点）により，声道形状をほぼ正確に記述することができる。したがって調音目標に向かう制御点の移動を正確に実行すれば，形状学的な拘束条件を実現することができる。また，連続発話を生成する場合，生理学的モデルの構造的特徴により，制御点が調音目標に向かい次から次へと移動することにより声道形状は連続的に変化する。これにより，声道形状変化の動的な拘束条件が実現される。

2.2 生理学的調音モデルの拘束条件

生理学的調音モデルの制御に用いた筋ベクトル空間は，モデルの幾何学的座標により指定される調音空間との間に一対一の対応関係をもつ。従来の研究より，舌の調音位置は母音の第一フォルマント(F1)と第二フォルマント(F2)からなる音響空間と類似のパターンを持つことが知られている[3～5]。もし舌背の制御点により舌の位置を指定するならば，母音のフォルマントパターンと舌の調音位置との関係を直接調音モデルの制御に応用することができる。

Table 1 Speech material used in this investigation

V-V Sequence	/ae/, /ai/, /au/, /ea/, /ei/, /ia/, /ie/, /iu/, /ou/, /ua/, /ui/, /uo/
V-C-V Sequence	/aka/, /ata/, /asha/, /apa/, /ara/, /aza/, /aba/, /ada/, /ama/, /aha/, /awa/, /aya/, /acha/, /ana/, /aga/, /asa/

音響と調音との関係を定量化するための資料として，ウィスコンシン大学のX線マイクロビーム装置を用いて構築したX線マイクロビーム発話データベース（詳細は[2]を参照）を使用した。このデータベースより,男性話者6名の母音連鎖(V-V)と母音-子音-母音(V-C-V)連鎖からなる発話データを選択した。分析に用いられた音声資料をTable 1に示す。

音響パラメータと調音器官の空間位置との関係を求めるため，F1とF2を音響パラメータとし，下顎と舌上面に置かれたペレットの水平位置(X 軸)と垂直位置(Y 軸)を調音パラメータとした。フォルマントとペレットの位置は母音の安定区間から切り出したセグメントを用いて求めた。このときの正中断面から見た各ペレットの位置を図 1に示す。

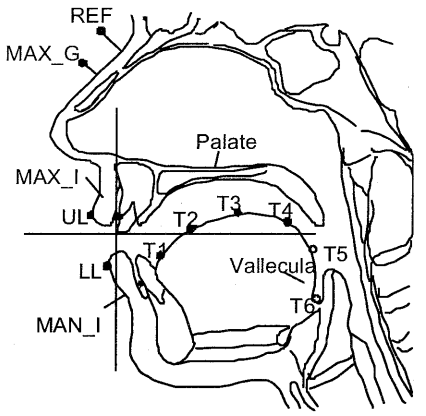

図 1 Pellet placements used in the x-ray microbeam experiment.

舌背(舌尖より後部約5.7cmの点)の前後位置はF1とF2の差との相関を分析した。舌背の前後位置とF1-F2の対数との相関を図 2に示す。左側は目標話者のデータで，右側は他の話者のデータである。舌背の前後位置と F1-F2 との相関は，目標話者では-0.956, 他の話者では-0.942である。図 2に示すように舌背の位置とフォルマントの差とがほぼ線形の関係をもち，標準残差は約0.02である。この結果は，音響パラメータと調音パラメータとは相互に線形マッピングが可能であることを示唆している。

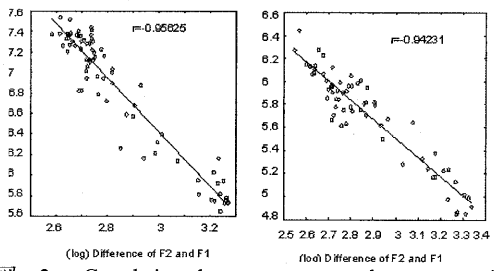

図 2 Correlation between tongue dorsum anterior-posterior position and the (log) difference of F1 and F2. The left panel is the data obtained from the target speaker, and the right panel is for another speaker.

日本人男性話者6名について，舌背の前後位置とF1-F2との相関係数を求めた。その結果を図 3に示す。すべての話者に対して,相関係数は-0.91から-0.96の範囲にあり，相関係数の平均値は-0.942，平均標準残差は0.03である。フォルマント周波数を用いて舌背の位置を予測した場合，6 名話者に対する予測の平均誤差は 0.18cm であり，予測の最大誤差は0.244cmである。この分析により，選択された音響

パラメータと調音パラメータとの関係を線形とした場合の標準残差は小さく，音響パラメータから調音パラメータへの予測誤差も小さいことが明らかとなった。この結果は，フォルマントから舌位置へのマッピングが音声波形から声道形状への逆推定において拘束条件として使用できることを示唆している。

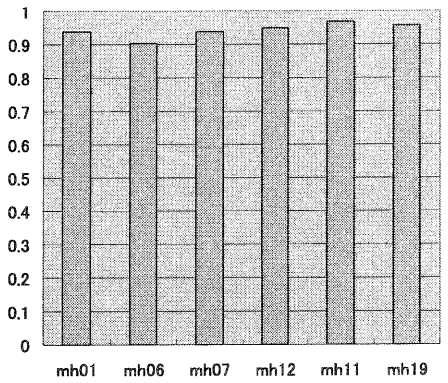

図 3 Correlations between anterior-posterior positions of the tongue dorsum and the difference of F1 and F2 for six Japanese male speakers.

3.音声波形から声道形状への逆推定

上述の分析は，我々の提案した生理学的調音モデルを導入することにより逆推定の精度を改善できる可能性を示している。本節では，生理学的調音モデルを用いる音声波形から声道形状への推定方法について論じる。

3.1 パラメータの選択

音声波形から声道形状への推定は，多意的な非線形のマッピングであるが，音響パラメータと調音パラメータの選択により推定の複雑さを減少させることが可能である。我々は調音モデルの構造的な特徴と，制御及び拘束条件導入の容易さなどを考慮に入れた上，次式に示す 10 個の調音パラメータを選択した。

$$X=(J_x,J_y,T_x,T_y,D_x,D_y,L_a,L_l,G_h,V_a)^T \qquad (1)$$

ここで，はじめの3つの対(J_x,J_y)，(T_x,T_y)，(D_x,D_y)は，それぞれ下顎，舌尖および舌背の XY 座標であり，調音モデルの3つの制御点に対応している。これらの値は発話器官モデルの調音動作の生成に用いられる。L_a と L_l はそれぞれ口唇を短管とした場合の断面積と長さであり，G_h は喉頭の高さ，V_a は軟口蓋の開口面積である。この4つのパラメータは，調音動作の計算には使用されないが，声道形状の形成と音響特性の計算に用いられる。これらの値は直接設定されるかまたは逆推定過程で修正されることにより声道形状の調整に関与する。

これまでに，我々はフォルマントパターンを用い母音と母音連鎖の声道形状の推定を試みている[7,8]。フォルマントパターンは子音を含む音節に対応する調音目標の推定には適用できないので，MFCC(メルケプトラム係数)を音響パラメータとして用いる。MFCCは人間の聴覚特性に近いメル尺度と一致するので,逆推定に適切なパラメータであると思われる。本研究では，音声波のエネルギー（振幅）および12次の MFCC 係数を音響パラメータとして用いる。前節で記述したように，F1 と F2 の差は舌背位置と高い相関があるので，母音発話時の舌位置の拘束条件として使用される。そのため，母音のフォルマントパターンも求めている。

3.2 音響から調音へのマッピング方法

生理学的調音モデルを用いて音声波形から声道形状を推定するには，分析合成法(Analysis-by-Synthesis, AbS)に基づいて，モデルによる合成音声の音響パラメータを入力された音声の音響パラメータに近づけるように，発話機構モデルの調音目標をステップ毎に更新する。二組の音響パラメータの差があらかじめ設定した誤差範囲より小さくなった時の声道形状を話者の声道形状とする。したがって，音声波形から声道形状を推定する過程は本質的に音響パラメータから調音パラメータへのマッピングである。

音声生成の立場から見ると，音声の音響パラメータは調音パラメータの関数であると考えてもよい。調音パラメータを X で表記すると，音響パラメータは $f(X)$ で表すことができる。ここで，入力音声の音響パラメータを f_r で表記する。音声波から調音状態を推定することは，音声波から得られた音響パラメータ f_r に対して，それに近似する $f(X)$ を求めることである。これは次の重みつき2乗誤差の和を評価関数とする最小化問題である。

$$J(X)=\|f_r-f(X)\|_Q^2+\|X-X_0\|_R^2+\|X-X_P\|_W^2 \qquad (2)$$

ただし，f_r と $f(X)$ は 13 次元ベクトルである。$\|X\|_R^2$ は 2 次形式 X^TRX を表す。各項の重み行列 $Q(13\times13)$, $R(10\times10)$, $W(10\times10)$ は簡単化のため対角

行列とした。X_0は調音目標の参照値で，中性母音/e/の調音目標とした。X_pは前フレームでの推定値で，連続音声発話時の声道形状を推定する際の拘束条件の一つとして用いられる。

評価関数$J(X)$の最小値を与える$\hat{X}$を求めるには，まず，$\hat{X}$の第k回の近似解を$\hat{X}_k$として，関数$f(X)$をその周りで線形化する。

$$f(X) \cong f(\hat{X}_k) + \left.\frac{\partial f(X)}{\partial X}\right|_{X=\hat{X}_k} \bullet (X - \hat{X}_k) \qquad (3)$$

次に，式(3)を式(2)に代入し，Xで偏微分して零とすれば，第k+1回の近似解$\hat{X}_{k+1}$は次式で与えられる。

$$\hat{X}_{k+1} = \hat{X}_k + \lambda\left\{A^T Q A + R + W\right\}^{-1}\left\{A^T Q(f_r - f(\hat{X}_k)) + R(X_0 - \hat{X}_k) + W(X_p - \hat{X}_k)\right\} \qquad (4)$$

ただし，$A = \partial f(\hat{X}_k)/\partial X$である。係数$\lambda$は，$f(\hat{X}_{k+1}) \le f(\hat{X}_k)$を満足するように決められる。

式(4)から求めた調音目標の推定値$\hat{X}$を新しい調音目標として調音モデルを駆動する。重み行列Rは，

$$R_k = (R_0 - R)\gamma^k + R \qquad (5)$$

として徐々にRの値に漸近させる。これより，調音目標の初期推定値の精度が悪い場合でも，推定値の発散を防ぐことができる。

また，関数$f(X)$の偏導関数Aを解析的に求めることは調音モデルの構造上煩雑となるため，$\hat{X}_k$に微小変動を与えて$f(X)$の変化を計算する。偏導関数は差分計算によって求めた。重み行列の値は，$\mathbf{Q} = \mathbf{I}$，$\mathbf{R}_0 = 0.05\mathbf{I}$，$\mathbf{R}_0 = 0.01\mathbf{I}$（$\mathbf{I}$は単位行列である）として，$\gamma = 0.6$にした。

4.逆推定のシステム

生理学的調音モデルを用いる逆推定システムの構成を図 4に示す。本システムは Windows98 上で動作するアプリケーションで，大きく分けると次の 4 つの部分からなる。

音声入力部: PCに接続されたサウンド入力デバイスより音声を録音する。サウンド入力デバイスはマイクロフォン等の汎用のものが利用でき，またサンプリングレートや量子化ビット数も選択できるように設計した。編集の目的で音声区間を選択して切り出し，ファイルに保存することが可能である。X線マイクロビームシステムにより収録した音声データを処理する機能も設計した。雑音が含まれている音声については，環境の雑音を自動で判別し，音声信号の前後にあるノイズを切り捨て，発話音声のみを取り出すことができる。

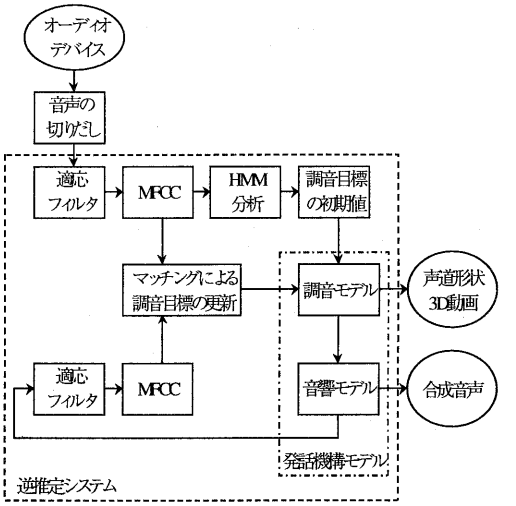

図 4 A flow chart of the estimation of vocal tract shapes from speech sounds.

逆推定システム部: 逆推定では，入力音声を16kHzでサンプリングし，プリエンファシスを施して放射特性を補正する。次に適応フィルタによりスペクトルを平坦化する。その後，音響パラメータとして MFCC 係数を求める。さらに音響パラメータに基づいてあらかじめ設定した日本語 5 母音と 17 子音に対応する基準調音目標より初期調音目標を選択する。

調音目標に従って調音モデルにより得られた声道形状から声道面積関数をもとめ，音響モデルにより音声を合成する[6]。この合成音声に入力音声と同じ処理を施して音響パラメータを求める。合成音声と入力音声の音響パラメータとを接近させるようにモデルの調音目標を更新する。

上記の計算を繰り返すことにより推定した声道形状は話者の声道形状に漸近する。合成音声と入力音声の音響パラメータとの差があらかじめ設定した許容誤差範囲より小さい場合，計算が停止する。このときの声道形状を話者の声道形状とみなす。

声道形状表示部: 逆推定部より得られた声道形状系列を用い声道の3次元動画を作成して表示する。異なる方向から声道形状を観測することにより，発話過程の理解と学習に有益であると思われる。この

目的で，利用者が動画の表示に対して自由回転,拡大縮小,透過処理,ワイヤーフレームなどの処理を施すことができる機能も装備した。Direct3D の使用によりスムーズなリアルタイム 3D アニメーションを実現した。アニメーションファイルの入出力機能を搭載しており，過去に推定されたデータから再現することもできる。

合成音声出力部: 音響モデルにより得られた合成音声を出力する。合成音は 3D アニメーションと同期して再生することもできる。合成した音声データは wav 形式で出力でき，他のアプリケーションでも利用することを可能にした。

5.逆推定方法の評価

先行研究では，我々は母音を用いてこの逆推定法の有効性を評価した[8]。これには，モデルによる合成音声を用いた評価及びX-線マイクロビーム装置による観測データを用いた評価を行った。その結果，発話機構モデルを用いた逆推定法は母音に対して有効であることがわかった。本研究では，モデルのフォルマントと舌背位置との拘束条件に基づき，母音連鎖と子音を含む音節を用いて逆推定の精度を評価する。

5.1. 拘束条件の導入と逆推定精度

この評価に用いる音声資料はTable 1 に示した母音連鎖のみである。調音データはX線マイクロビーム装置による観測データであり，音声信号が同時に収録されているので，音声信号より推定された声道形状を評価することができる。この評価は発話機構モデルの目標話者のデータのみを行った。

調音データと音声データはともに観測データと模擬した連鎖母音の定常部分から抽出したものである。声道形状の比較に用いられたペレットは下顎のペレットと舌上面のペレット T1 から T4 である（ペレットの位置は図 1 を参照）。推定した声道形状と観測した声道形状との差を求めるため，上記の 5 個のペレットの位置を，推定した声道形状の正中矢状断面に投影した。下顎のペレットの位置と下顎の制御点との距離，および舌の各ペレットから舌表面までの最短距離を求めた。それらの距離の平均値を声道形状の推定誤差として図 5に示す。左側のグラフは逆推定を行う際に舌位置の拘束条件を使用せずに得られた結果である。すべての音声データに対して，平均誤差は 0.2cm となっている。右側のグラフは，舌位置の拘束条件を用いて推定した結果である。拘束条件の導入により，推定精度は全面的に改善され，推定誤差は 0.1cm 程度に減少した。

4 つの低次フォルマントについて推定誤差を図 6 に示す。左側は舌位置の拘束条件を用いない場合の結果であり，右側はその拘束条件を導入した場合の結果である。前者の推定誤差は約 0.9%から 4.2%の範囲であり，平均誤差は 0.25%である。声道形状の推定結果と同様に，舌位置の拘束条件の導入により推定精度が全面的に改善されている。

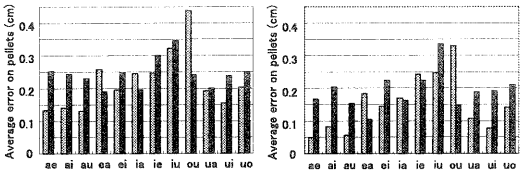

図 5 Averaged values of the distances from the tongue pellets to the model tongue surface and from the jaw pellet to the jaw control point for the vowels in the vowel sequences from the target speaker. (a) without the articulatory constraint and (b) with the constraint

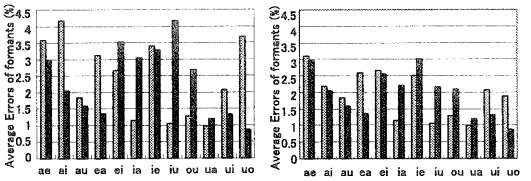

図 6 Simulation errors averaged over the four lower formants for the vowels in the vowel sequences from the target speaker. (a) without the articulatory constraint and (b) with the constraint.

5.2. 子音を含む音節に対する逆推定の精度

本節では子音を含む音節を用いて逆推定方法を評価する。母音連鎖の場合と同様に，下顎のペレットの位置と下顎の制御点との距離，および舌の各ペレットから舌表面までの最短距離を求めて，それらの平均値を推定誤差とする。ただし，連鎖母音における誤差が波形の定常区間の一点のみから求められたのに対して，この誤差は音声区間のすべてフレームに対して求める。そのため，X線マイクロビームのデータはモデルの声道形状のフレームレート（50FPS）に合わせてダウンサンプリングされたものを用いた。声道形状の推定誤差を図 7示す。左側は初期調音制御点を"真"（既知）の調音目標の近辺の値になった

場合の結果で，右側は与えた初期調音目標を"真"の調音目標より若干外れた場合の結果である。前者の平均誤差は 0.31cm，後者の平均誤差は 0.37cm であった。すべての音声データに対して，平均誤差は 0.34cm となっている。その誤差はペレット T2,T3 に対して比較的大きい。モデルの声道形状系列は推定した 6 つの調音目標セットにより計算したものであり，推定誤差はフレームごとの声道形状の差異とモデルのダイナミクスによるフレームのずれの影響との累積である。この誤差は全体的に母音連鎖の誤差より大きく見えるが，モデルの動的特性の影響を考慮すると，この推定は妥当な結果と思われる。特に初期調音目標は推定しようとする目標に若干離れても推定過程は的確な目標へ収束する。この結果より，本推定方法は子音を含む音節の逆推定にも有効であることがわかる。

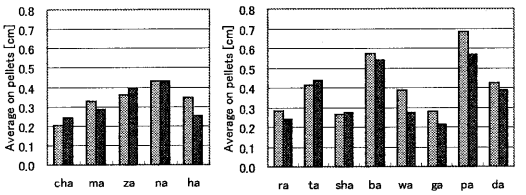

図 7 Averaged values of the distances from the tongue pellets to model tongue surfaced and from the jaw pellet to jaw control point for syllables. (The left bars show consonants and the right ones for vowels).

6.逆推定法により発話様式の分析

本節では，提案した逆推定法を用いて発話様式の相違に伴う発話動作の相違を検出する。日本人男性話者 1 名から収録した発話様式（パラ言語的要素）の異なる文「そうですか」の文末の音節/ka/を切り出して，逆推定システムに適用した。使用した音声試料の発話様式とその特徴をTable 2に示す。それらの音声資料の口調は，発話速度，及び強さの組合せである。本節では，発話器官の中の下顎と舌背が観測点として用いられる。

/ka/の発話時に下顎の上下位置の時間変化を図 8 に示す。ここで，下顎の観測点は，X 線マイクロビームデータでは下顎のペレットの位置（ペレット位置は図 1 を参照）とし，逆推定では下顎の制御点とした。左側は X 線マイクロビームにより観測した下顎の軌跡で，右側は音声波形から逆推定した下顎の上下動である。失望口調と感心口調で発話する場合，母音区間に下顎が大きく開くのに対して，疑問口調と相槌口調の場合，文末に向かうにつれ下顎が上方に動く。逆推定により得られた下顎の動きは同様の傾向を示している。

Table 2 Speech styles for producing /ka/

sample No.	impression	speed	stress
081_ka	がっかりした	ゆっくり	弱く
082_ka	感心して	速い	強い
083_ka	疑って	ゆっくり	普通
084_ka	軽い相槌	速い	弱い

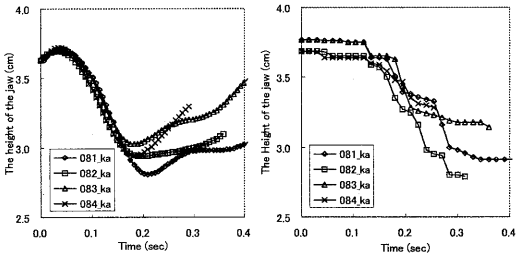

図 8 Trajectories of the jaw movement in producing /ka/ with different speech styles (left: observations; right: simulations).

/ka/の生成過程では，舌背は破裂子音/k/の調音時の高い位置から/a/の調音の低い位置まで幅広く移動する。ここで，舌背の動きに対して顎の場合と同様の比較を実施する。舌背の位置について，モデルでは推定された舌背の最高点の位置の座標とし，X線マイクロビーム測定時のペレット T3 の座標を舌背位置と見なす。舌背の上下方向の時間変化を図 9に示す。左側は舌背のペレットの軌跡で，右側は音声波形により推定された舌背の位置変化である。音節の開始時に，子音/k/の閉鎖点を形成するため，発話様式の違いに関係なくすべての発話では舌背を上方へ移動して同じ調音点に到着する。文末に推移するにつれて発話器官は母音/a/の声道形状を形成するように動作する。この推移過程で発話様式による舌背の運動に差が見えてきた。舌背の高い位置にとどまる時間が失望口調，疑問口調，相槌口調にくらべ，感心口調では短くなっている。発話器官が調音目標の強調方向により早く移動することにより，はっきりした音が発話できるからである。また，文末での舌背の高さが，発話様式により異なる。これは発話様式により語尾の母音のニュアンスの違いによるものである。同じ変化は推定した舌の動きにも見られた。

さらに逆推定した声道形状により得られた合成音を用い発話様式の差異による影響を聞き比べてみた。発話器官の動きの差異による発話様式への影響を明確にするため，音源の基本周波数F0と音源の強さは調整しなかった。ピッチの変化や声の大きさに違いがないのにもかかわらず，強調の度合いを判別することができた。それは発話様式は音源だけでなく，舌，顎，口唇などすべての発話器官が関与しているためといえる。

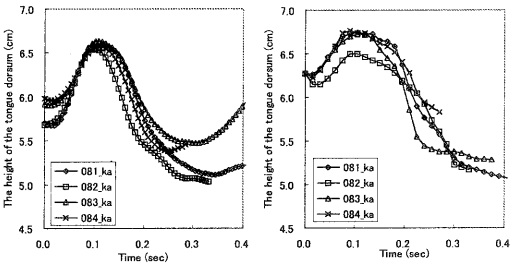

図 9 Trajectories of the tongue dorsum movement in producing /ka/ with different speech styles (left: observations; right: simulations).

7.まとめ

本研究では，3次元生理学的調音モデルを用いて音声波から声道形状を推定する手法とその応用を検討した。この生理学的調音モデルは人間の発話機構のメカニズムを忠実的に再現するため，逆推定の過程には人間の生理学的拘束条件が実現される。また，従来の手法により取り扱われた時間的と空間的な拘束条件はこのモデルの基本特性として統合されている。

さらに推定精度を高めるため，X線マイクロビーム日本語調音データベースを用いて調音器官の運動と音響パラメータと対応関係を調べた。F1とF2との差は舌背の前後位置とかなり高い相関をもっていることが明らかとなり，この結果を舌位置の拘束条件として調音モデルに導入した。舌位置の拘束条件の有無による推定精度への影響を比較し，この拘束条件の導入により声道形状の推定精度が明らかに上昇することを確認した。

本発話機構モデルを用いる逆推定法を子音を含む音節への適用により評価した。調音目標を用いモデルの計算により得られた声道形状系列をX線マイクロビームのペレット運動軌跡と全音声区間にわたって比較し，妥当な推定結果が得られた。

本研究では，感情の違う4種類の音声資料を用い，逆推定方法により発話様式の推定と分析を行った。推定により得られた下顎と舌背の運動軌跡はX線マイクロビームデータと同じ傾向を示した。

本逆推定法は音声波形から発話器官を視覚化することにより，耳の不自由の方の発話訓練と外国語の習得への応用を期待することができる。しかし，この推定方法には子音の初期調音目標の設定と推定誤差の削減など数多くの課題が残っている。

謝辞

この研究の一部はCREST（科学技術振興事業団戦略的基礎研究推進事業）の支援により行われた。

参考文献

[1] Atal, S., Chang, J., Mathews, J., & Tukey, W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," Journal of the Acoustical Society of America, 63, 1535-1555.

[2] Hashi, M., Westbury, J. & Honda, K. (1998) "Vowel posture normalization," Journal of the Acoustical Society of America, 104, 2426-2437.

[3] Moore, C. (1992). "The Correspondence of the vocal tract resonance with volumes obtained from magnetic resonance image," Journal Speech and Hearing Research, 35, 1009-1023.

[4] Stevens, K. & House, A. (1955). "Development of quantitative description of vowel articulation," J. Acoust. 24, 175-184.

[5] Stevens, K. (1999) *Acoustic Phonetics* (MIT Press, Cambridge, MA).

[6] 党, 本多, (1999, 9), "生理学的調音モデルを用いる音声合成法", 音講論, 243-244

[7] 党, 本多, (1998, 3), "生理学的調音モデルに基づく3次元声道形状の生成", 音講論, 265-266

[8] Dang, J. and Honda, K. (2000, 5). "Estimation of vocal tract shape from speech sounds via a physiological articulatory model," Proc. International Workshop of Speech Production, 233-236, (Munich, Germany).