音声生成過程の可視化と色表示について*

白井克彦**, 小林哲則**

Visualization of Speech Production Process and Color Representation of Phonetic Information

Katsuhiko SHIRAI and Tetsunori KOBAYASHI

1. は し が き

音声という聴覚に関する情報を、視覚情報に変換し可視化することは、音声の基礎研究を支援するうえからも、また、発話障害あるいは聴覚障害者に対する訓練・聴覚代行手段を提供するうえからも重要である。本稿では、音声の可視化に関する話題の中から、筆者らが関係してきた音声生成過程における調音運動の可視化と音韻情報の色表示についてとりあげ、それらの概要を紹介する。

可視化手段の中には、多かれ少なかれ、その対象に対するモデルが適切な形で盛り込まれている。音声についても例外ではなく、その可視化にあたっては、生成過程のモデルが大きな役割を演じることになる。本稿では、まず、音声生成過程のモデルに基づいて、調音運動の様子を推定し可視化することについて述べる。

人間の音声言語による通信能力は、際だって高いものであるが、それゆえに大変複雑でもある。音声の生成過程は、脳において発話したい単語や文章が生じた後、それに対応する音韻系列に従って調音器官を適切に動かし、そこに肺から出した呼気流を通すことによって音声に変換するプロセスである。これらの過程において、出力である音声の観測についてはマイクで受ければよいので容

易であるが、それ以外の筋電、神経指令などの観測は大変困難である。調音運動についても、口唇の動きはある程度外から観測でき、超音波 CT, X線、微小磁石を装着する方法などによって、舌面や顎の運動を見ることが試みられているものの、簡単とはいえない。

本稿では、複雑な調音運動を大量の発話データに対して分析可能にすることを主目的として研究された調音運動の推定方法を紹介する。本方法は、音声生成モデルを設けて、その出力である音声波から調音器官の運動を求める、いわゆる逆問題を解くものである。これによって観測困難な音声生成時の調音運動を可視化することができる。もち論、この手法によって意図された音韻系列の推定もできるので、本手法は音声認識における特徴抽出法としても有効なものと考えられる。

つぎに聴覚障害者に対する補助手段として、音声の色表示の可能性について若干述べる。音韻情報の可視化は、古くから研究されており、重要な問題であるが、まだ適切な提示方法は得られていないのが現状である。われわれの研究もまだ緒に付いたばかりであるが、現状での成果を簡単に報告することにする。

2. 調音運動の可視化

われわれが音声を発声する際には、調音器官と呼ばれる顎、舌、唇等の器官を動かすことによ

*原稿受付 1991年8月16日

**早稲田大学理工学部

てさまざまな口腔の形状を作りだし、声帯で発生された声帯波または口腔の狭めによって発生した乱流を口腔内で共鳴させることによって、さまざまな音色の声を作り出すことになる。たとえば、母音 /i/ を発声するためには、顎を閉じ、舌を前に出して唇を平らにするし、/o/ を発声するためには、顎を開け、舌を後ろに引っ込め、唇を丸める、といった具合である。

このような調音器官の運動の様子を可視化することは、高性能の音声認識装置や音声合成装置を構築するうえでの基礎データ解析のために重要なばかりでなく、発話障害者に対して有用な教示データを与える観点からも重要な意味を持つ。

口腔の形状を調べるためには、超音波やX線等を使った計測装置を利用することがまずあげられるが、マイクを使って収集することができる音声波に現れる情報のみを用いて口腔形状が推定できるならば、より自然な発声環境における調音状態を推定できる（特殊な計測装置を使うとき、一般に発声は影響を受ける）と共に、より気軽なデータ収集環境が与えられるという意味で価値がある。本稿では、この音声波から調音状態を推定する問題に関するいくつかのアプローチを紹介する。

2.1 調音モデル

音声波に含まれる情報から調音状態を推定する問題は、一般には調音状態を表す調音モデルを用意し、そのモデルのパラメータを音声波の分析の結果得られる音響的特徴を用いて推定する問題と

して定式化される。調音モデルとしては、単に口腔形状の等価音響管モデルを効率的に表現することを主眼としたものから、調音器官の解剖学的な構造を具体的に考慮したものまで、さまざまなものが提案されている^{1)~6)}が、ここでは、後者の代表的モデルである、白井・菅田の調音パラメータ^{5),6)}を取り上げて説明する。

図1に白井・菅田の調音モデルの構成を示す。モデルは、舌の形状を表す X_{T1} , X_{T2} , 顎の開閉の程度を表す X_J , 唇の突き出しおよび丸め具合を表す X_L , 声門の開きの程度を表す X_G , および鼻腔との結合の度合を表す X_N の六つからなっている。これらのパラメータは口腔の正中断面のX線写真のデータの実測を行い、これらの座標を効率よく表現するよう設定されたものである。

2.2 調音パラメータと音響パラメータ

調音パラメータ $\mathbf{x}$ が与えられれば、口腔内の形状が決定でき、これを基に等価音響管モデル（図2参照）の各断面積を決定できる。音響管の各断面積が決定されれば、その音響管全体の伝達関数は、隣り合う音響管の断面積比によって決まる各音響管境界での反射係数や音響管内での減衰係数により決定でき、この伝達関数を基にその口腔形状から発声される音声の任意の音響パラメータ $\mathbf{y}$ を算出することができる。これら種々のパラメータ間の関係を図3に示す。

このように、調音パラメータ $\mathbf{x}$ から音響パラメータ $\mathbf{y}$ への変換は、非線形な処理を含むものの、一本道に従った比較的容易な過程に従っている。しかしながら、その逆の過程、すなわち音響パラメータから調音パラメータへの変換過程は、非線形関数の逆関数を求める問題となるため、非常に

図1 調音モデル

図2 声道の等価音響管モデル

図3 調音パラメータと音響パラメータとの関係

複雑な過程となる。ここでは、この逆問題を解く代表的な三つの方法を紹介する。

2.3 モデル・マッチング法による調音パラメータの推定⁶⁾

逆問題を解くための最も基本的な方法は、合成による分析 (A-b-S: Analysis by Synthesis) を基礎として構成されたモデル・マッチング法と呼ばれる方法である。

先にも述べたように、調音パラメータ x が与えられれば、音響パラメータ y は x の関数 $y = h(x)$ によって求める (合成する) ことができる。よって与えられた音声波に対する x を求めるには、 x から y を合成してみたとき、 y が音声波から求められる実際の音響パラメータ y_s に近くなるような x を捜せば良いことになる。実際には、 y と y_s との誤差に加えいくつかの罰金項を加えた。

$$J = (y_s - h(x))^t P (y_s - h(x)) + x^t Q x + (x - x_0)^t R (x - x_0) \quad (1)$$

を評価関数として、これを最小化する x を勾配法によって漸近的に求めることになる。このとき、第 k 回近似解を x^k とすれば、第 $k+1$ 回近似解 x^{k+1} は

$$x^{k+1} = x^k - \lambda \cdot \left\{ \frac{\partial h(x^k)^t}{\partial x} P \frac{\partial h(x^k)}{\partial x} + Q + R \right\}^{-1} \left\{ \frac{\partial h(x^k)^t}{\partial x} P (y_s - h(x^k)) - Q x^k - R (x^k - x_0) \right\} \quad (2)$$

として求めることができる。ただし、ここで P, Q, R は重み行列を表し、 x_0 は前フレームでの推定値

を表すものとする。

2.4 テーブル・ルックアップによる調音パラメータの推定⁷⁾

前節に述べたモデル・マッチング法は、繰り返し計算によって解を求めるため計算コストが高い。そこで高速性が要求される場合に用いられるのがテーブル・ルックアップ法である。

テーブル・ルックアップ法では、まず調音パラメータ空間をベクトル量子化し調音パラメータのコードベクトルを求める。次にこの調音パラメータのコードベクトルに対応する音響パラメータを計算し、音響パラメータ—調音パラメータの変換表を作る。音声波から調音パラメータを求める際には音声波を分析して得た音響パラメータと最もよく整合する音響パラメータを持つ項目が表から検索され、対応する調音パラメータが推定結果として求められる。表の検索は単純なフルサーチで行うこともできるが、一般に音声信号は時間的相関が高いことから、前のフレームにおける調音パラメータの推定値を基に探索領域を限定することが有効である。このような領域限定を行った場合、モデル・マッチング法に比べ計算コストを1/50以下にすることができると示されている。

2.5 ニューラルネットワークによる調音パラメータの推定⁸⁾

テーブル・ルックアップ法は非常に高速ではあるが、ベクトル量子化によってパラメータの推定精度が制限されてしまう。そこで中庸をいくものとして提案されているのがニューラルネットワークを用いる方法である。多層のフィードフォワード型ニューラルネットワークは任意の非線形変換を学習することが原理的には可能とされている⁹⁾ ことから、音響パラメータから調音パラメータへの変換関数をこれに学習させるのである。

4層のニューラルネットワークを用いて、そのノード数を入力層から順に12, 24, 24, 4とした場合 (出力としては X_{T1} から X_L までの4パラメータのみを対象としている)、計算量を1/15としたうえで良好なパラメータ推定ができることが確認されている。また、モデル・マッチング法は勾配法をベースとしているため時として極小値に収束して極

端な推定誤りを生じることがあるが、ニューラルネットは大局的なマッピングを実現するためこのような危険性が少ないという利点も有することが確認されている。

ニューラルネットを用いて求めた調音運動の様子をモデル・マッチング法によるものと比較して図4に示す。

以上、調音パラメータを音声波から推定する三つの方法について述べた。これらの方法によって

図4 /aimai/ における推定された調音パラメータの時間変化とスペクトル (……: ニューラルネット, —: モデルマッチング)

図5 /o/, /i/ における推定された口腔内の形状

推定された /o/, /i/ に対する口腔の推定形状を図5に示す。

3. 音韻情報の色表示

連続して発声された音声を認識してこれを文字表記に変える技術は、音声タイプライタと呼ばれ古くから音声認識技術の一つの目標として掲げられてきた。通常の音声対話で生ずる音声は、発音が曖昧であったり通常の文法に従わなかったりするだけでなく、その会話の状況や対話者同志の特別な共有知識などに依存したりもするので、簡単には音韻認識できない。つまり、語彙や文法などの知識を補っても現状では音声を正確な文字表記に直すことは難しい。しかし、完全ではないものの、現在HMM(Hidden Markov Model)法^{10),11)}やNN(Neural Network)法¹²⁾などによってかなり高い音韻認識率が達成されつつあり、音韻に近い符号に変換することは可能となっている。そこで、この曖昧性を持つ疑似音韻系列を人間の感覚に訴える形で提示することができれば、これを聴覚障害者とのコミュニケーションの媒介手段として利用できることが期待できる。このような観点から、検討されているのが、音声の疑似音韻系列への変換とその色表示の問題である。

図6に本システムで用いている階層的クラスタ化による音韻認識部の概要を示す¹³⁾。本方式の特徴は音声パワーの変化によって音韻性を大分類し、その後メルケプストラムによって音韻認識を進める。高品質な特定話者の音声(ATR音声データ

ベース)においては母音97.9%, 子音85.5%の認識率が得られている。

音韻認識部によって得られた音韻系列には、認識誤り、挿入、脱落が含まれているが、これをそのまま色表示する。

色表示における問題は、各音韻に対する色の割当てとその表示方法である。これまでに、日本語の5母音に対する色の割当てについ

の場合、人間の音声発話に含まれる多様な情報をわかりやすい形で映像表現することが求められる。音声は音圧という物理パラメータの時間変化によってある言語情報を伝える。これを高度な音声認識技術を駆使して文字列で置き直せば、その言語情報を伝えることができ、これは確かに可視化の一形態である。しかし、ここでは感情など音声に含まれる重要な情報がかなり失われたものになる。より高度な聴覚代行手段を提供するためには、生の音声をできるだけ忠実に表現した可視化が望まれることになろう。

参考文献

- 1) K. N. Stevens and A. S. House: Development of a quantitative description of vowel articulation, J. Acoust. Soc. Am., 27, 484-493 (1955).
- 2) C.H. Coker and O. Fujimura: Model of specification of the vocal-tract area function, J. Acoust. Soc. Am., 40, 1271 (1966).
- 3) B. Lidblom and J. Sunberg: Acoustical consequences of lip, tongue, jaw and larynx movement, J. Acoust. Soc. Am., 50, 1167-1179 (1971).
- 4) P. Marmelstein: Articulatory model for the study of speech production, J. Acoust. Soc. Am., 53, 1070-1082 (1973).
- 5) 白井, 菅田: 調音機構のモデル化と非線形重回帰分析による調音パラメータの推定, 信学論, 59-A, 8, 668-675 (1976).
- 6) 白井, 菅田: 音声波からの調音パラメータの推定, 信学論, 61-A, 5, 409-416 (1978).
- 7) K. Shirai, T. Kobayashi and J. Yazawa: Estimation of articulatory parameter using table-lookup method, Proc. ICASSP 86 (1986).
- 8) T. Kobayashi, M. Yagyu and K. Shirai: Application of neural networks to articulatory motion estimation, Proc. ICASSP 91 (1991).
- 9) K. Funahashi: On the capabilities of neural networks, ATR Workshop on neural networks and parallel distributed processing (1988).
- 10) K. Lee H. Hone and R. Reddy: An overview of the SPHINX speech recognition system, IEEE Trans. ASSP, 36, 1, 599-609 (1990).
- 11) T. Hanazawa, K. Kita, S. Nakamura, T. Kawabata and K. Shikano: ATR HMM-LR continuous speech recognition system, Proc ICASSP 90 (1990).
- 12) A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang: Phoneme recognition using time-delay neural networks, IEEE Trans. ASSP 37, 4, 328-339 (1989).
- 13) K. Shirai, E. Kitagawa and T. Endo: Optimal construction of context sensitive quantizer for phoneme recognition in continuous speech, Proc. Eurospeech 91 (1991).