声道物理モデルによる音響生成

東本 敏男　　澤田 秀之

香川大学 工学部 知能機械システム工学科
〒761-0396 香川県高松市林町 2217-20
sawada@eng.kagawa-u.ac.jp

概要：声道モデルを機械系によって構成することにより、人間の発声機構を再現する試みをおこなっている。これによって人間のような発話動作に基づいた音声の生成が可能になると考えている。本モデルは主に、肺、声帯、声道部、聴覚部から成り、ピッチを変化させながら音響を生成することを可能とした。ピッチの生成には複雑な声帯振動の解析が必要であるが、ここでは喉頭摘出者が用いる人工声帯を使用している。調音のための共鳴管をシリコーンゴムで形成し、これを外側から棒で押し引きして声道断面積を変化させることにより、共鳴特性を変化させて母音音声を生成することが可能である。本論文ではまず機械系による声道物理モデルの構成について紹介し、次に聴覚フィードバックを用いたピッチの制御およびニューラルネットワークによる母音音声の学習アルゴリズムについて述べる。

Mechanical Model of Human Vocal System and Its Voice Production with Auditory Feedback Control

Toshio HIGASHIMOTO and Hideyuki SAWADA

Faculty of Engineering, Kagawa University
2217-20, Hayashi-cho, Takamatsu-city, Kagawa, 761-0396

Abstract: While various ways of vocal sound production have been actively studied so far, we are constructing a phonetic machine having a vocal chord and a vocal tract based on a mechatronics technology. The mechanical construction of a human vocal system is considered to generate natural voice, and we have so far developed a pitch and vowel generation part to speak and sing in a human-like way. In the voice generation, the analysis and mechanical realization of the behaviors of the vocal chords and the vocal tract are required. Furthermore, the fluid mechanical system is less stable to make the control difficult. Several motors are employed to manipulate the mechanical vocal system. With its self-learning ability, the machine is able to start speaking under the auditory feedback control. This paper introduces the mechanical human vocal system together with a pitch and vowel learning algorithm by neural networks.

1. はじめに

人間の音声は、音声生成器官の複雑な働きによって作られる音である。発声器官は主に、肺、気道、声帯、声道、舌、口蓋とそれらを動かす筋肉などから成り、互いに適当な位置や形状を形成することにより言葉が生成される。声は大抵の動物にあるが、言葉は人間にしかないコミュニケーション手段であり、音声生成のメカニズムや音声の認識手法などが古くから研究されてきた[1]。特に最近のマンマシンインタフェース技術には、人間らしい音声による情報の提示や、音声による入力デバイスは不可欠な要素となっている。

計算機による音声合成では、サンプリング法、規則合成、物理モデルといったソフトウェアによる手法が主流となっているが、それらの多くは波形レベルでの音声生成であるために、歌声のように発話動作から得られるような自然性が重視され

る音声には適した方法とはいえない。人間と同等の発声器官を機械系によって構成することにより、人間の様に学習によって人間らしい発話動作と音声を獲得することができると考えられる。本論文では、発話動作をおこなう器官をエアーポンプ、人工声帯、声道モデルなどを用いて機械的に構成し、計算機による聴覚フィードバック制御によって音声を生成する試みについて述べる。

我々はこれまでに、物理モデルによって生成される音声のピッチや大きさを聴覚フィードバックにより制御することで、ハミング音声の生成に成功している[2]。この他にも機械的に音声生成を行う試みはいくつか報告されているが[1],[3]-[5]、人間の発話動作が充分に解明されていないため、その制御は試行錯誤的なものが多い。本研究では、その音色をニューラルネットワーク(以下NN)を用いて機械モデルが自ら発話動作を学習し、音声を生成することを目的としている。

本モデルは、入力された音響のスペクトルを模擬した音響を生成することを目的としている。そのために、音声から発声動作を逆推定する処理にNNを用いた。これにより、個別の声道の形状と生成される音響との対応関係を学習させて、望む音響を生成することが可能となった。

2. 音声生成システムの概要

2.1 音声生成システムの構成

人間の発声は大きく分けて、声帯振動による音源の生成と共鳴によるホルマントの付加という２つの働きによって構成されている。肺からの呼気流は気道を通って声帯の振動を引き起こし、音源を生成する。更にこの声帯音源波に対して声道が音響フィルタの役割を果たすことによって、音素が構成される。このフィルタの伝達特性は声道内壁及び舌の形状などによって決まるが、主として顎や舌の非定常な動作によって引き起こされる変化から子音が生成され、母音は定常的な声道形状を形成することによって生成される。更に器官組織の湿り気や粘性なども生成される音質に大きな影響を与えるが、これは風邪を引いて声が嗄れるといったときに経験することからもわかる。一方、声の高さを決める基本周波数（ピッチ）及び声の大きさは、声帯音源波が持っている情報であるが、これらは肺からの呼気流量や声帯の形状と弾性などによって調節されている。

以上のような人間の発声機構を模擬して構築した、声道物理モデルから成る音声生成システムの構成を図1に示す。このシステムは、エアポンプ、エア調整弁、人工声帯、調音用共鳴管、マイクロフォン、音響アナライザからなり、それぞれ人間の肺、気管、声帯、声道、聴覚部に対応する。エアポンプから送られた空気流が人工声帯を振動させることで原音(音源波)が生成され、声道モデルの共鳴管形状をモータにより変形させて共鳴特性を加えることによって調音される。更にピッチおよび音量制御のために、2つのモータを用いている。

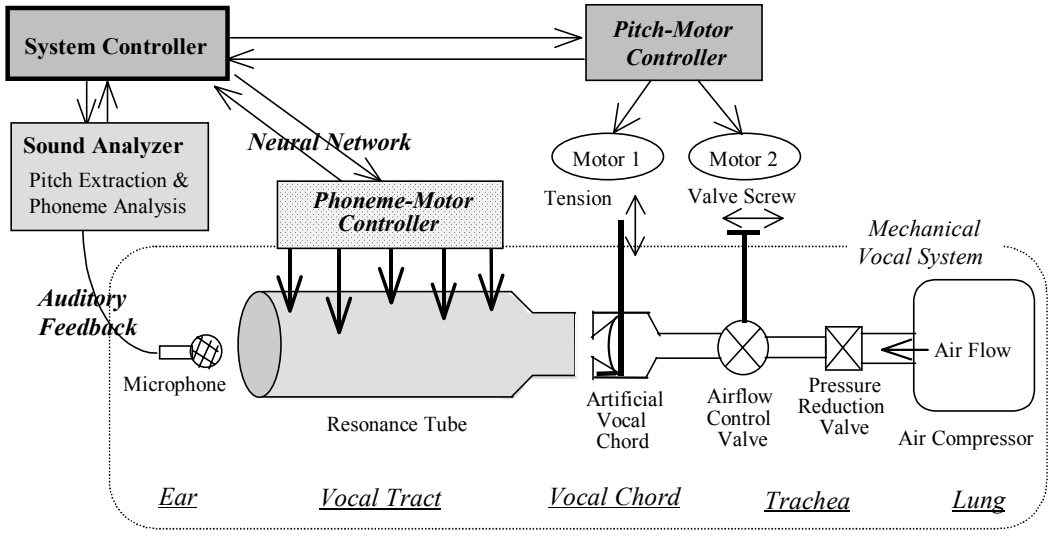


図１　声道物理モデルの構成

モータ１は人工声帯ゴムの張力操作用であり、モータ２は流量調節弁の開閉量調節用としている。また聴覚部において、音響の周波数特徴を用いて生成した音響の解析を行う。

音声の生成においては発声器官どうしの複雑な働きが必要であり、このような能力は幼児が言語を習得していく過程において発声と聞き取りの試行錯誤を繰り返すことによって獲得していくものである。そこで本システムでは、機械系から生成される音声をマイクロフォンに入力し音響アナライザで解析を行う聴覚フィードバック機構を実現している。これにより、人間が目標となる音声を与えるだけで、自己学習による音声の獲得が可能である。また獲得した発声手法によって連続的に発声させることにより、ソフトウエア合成では難しい音素間のつながりの自然や非定常な子音音声生成に対しても、本手法が有利であると考えている。

以下に声道物理モデルの各部について説明する。

2.2 人工声帯

音声のピッチと大きさは声帯音源波の特性によってが決まるが、音源波は2枚の筋肉とそれを覆う粘膜から成る声帯の複雑な動きにより生成されている。このような声帯の運動を計算機によりシミュレーションする試みもなされている[6]が、ここでは声帯摘出者が疑似声帯として用いる口内笛を用いて声帯音源の生成を行った。図2に示すように構成がシンプルな上に、ゴムの張力調整によって比較的簡単に振動特性が変えられるという利点もある。図3に人工声帯による音源波形の一例を示す。声帯音源波は、孤立三角波の繰り返しとして近似されることが多く、本人工声帯による波形は比較的実物に近いものとなっている。また空気の流量を10～14[*l*/*min*]と変化させたときの、張力とピッチの関係の一例を図4に示す。張力および空気の流量を変化させることによって、基本周波数が約110Hzから350Hzまで変化しており、1オクターブ以上のダイナミックレンジを持つことがわかった。

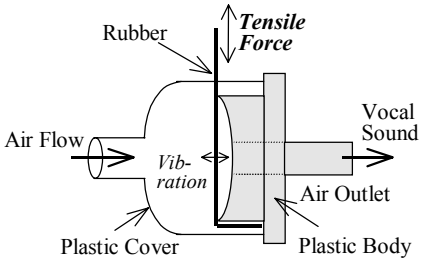


図2 人工声帯

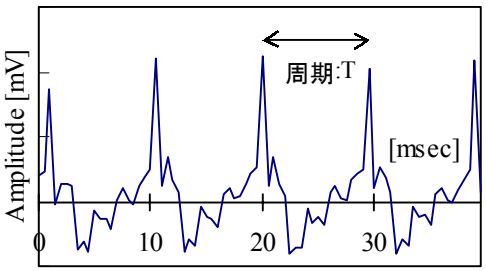


図3 人工声帯の波形

2.3 声道モデル

声道モデル部の調音用共鳴管は、図5に示すようにシリコーンゴム(スキンモールドSC)を人間の皮膚程度の柔らかさで成形して作成した。成人男性の声道の平均長を考慮し、内法寸法 180mm(声道方向の長さ) x 36mm(直径)の円筒型とした。この筒

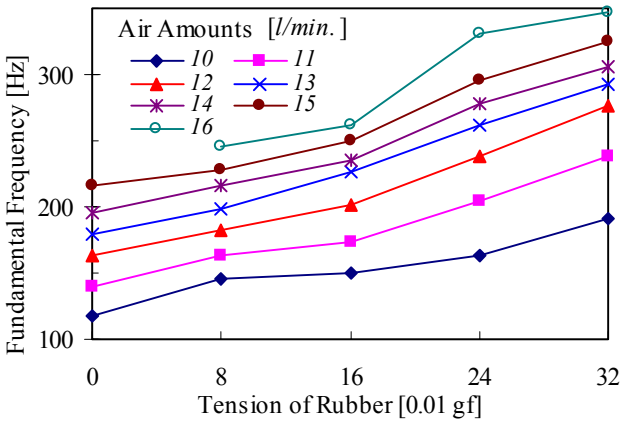


図4 ゴムの張力と基本周波数の関係

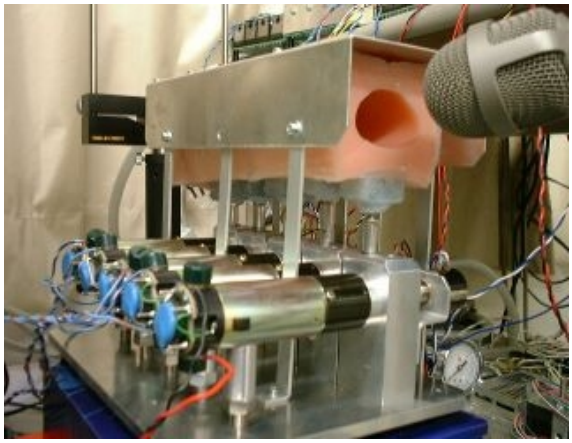


図5 調音用共鳴管と声道モデル

の下側に声道変形のためのステンレス棒を取り付けている。声帯側から声道の方向に20mm, 60mm, 100mm, 140mm, 180mmの5ヶ所で、それぞれモータ駆動によって上下動をおこなう。ステンレス棒を押し引きすることによって、断面積が約0cm^{2}から約10.2cm^{2}のあいだで共鳴管に変形を加えることが可能である。

2.4 聴覚部

生成された音響に対しFFTによる周波数解析を行うことによって、その共鳴特性および基本周波数を求める。それぞれの測定条件を以下に示す。なお音声信号は8bitで入力し、ハミング窓をかけてFFTをおこなっている。

共鳴特性：
サンプリング周波数: 8.0[kHz]
FFTのデータ点数: 1,024

基本周波数抽出：
サンプリング周波数: 1.0[kHz]
FFTのデータ点数: 1,024

3. 声道モデルの制御

音声信号から声道断面積関数を推定する手法としては、PARCOR分析法がよく知られている[7]。しかし、PARCOR分析法は1次元モデルに基づいたものであり、声道断面の2次元的形状の影響は考慮していない。また、未学習の音声を生成するための共鳴管変形のモータ量推定に不適切な値を出力してしまうという問題も見られた[2]。

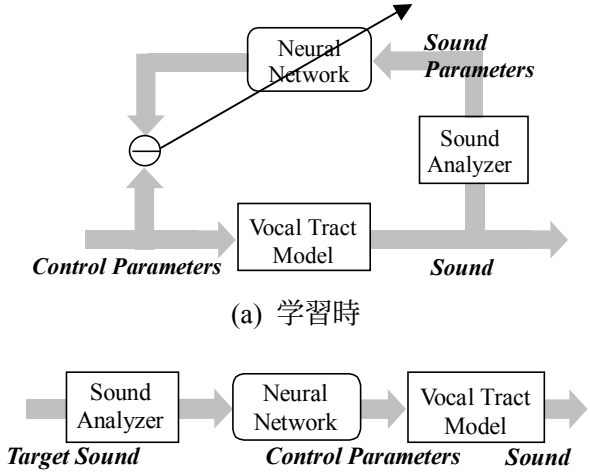


図6 ニューラルネットワークの動作

そこで、NNを用い、ある音響とそのときの声道の形状との対応関係を学習させることで、本声道モデルに特化した声道断面積関数を導き出すことを試みた。

3.1 ファジー推論ニューラルネットワークの導入

学習時と音響生成時におけるNNの構成を図6に示す。学習時において、NNはシステムが生成した音響の音響パラメータを入力とし、そのときのモータ制御パラメータを教師信号とすることで、ある音響を生成するために必要な声道の形状（モータ制御信号）を学習する。学習後は、NNを声道モデルに対し直列につなぐ。NNに生成したい音響の音響パラメータを入力するとモータ制御量が出力され、声道の形状を変化させることによって、望む音響を生成することが可能となる。

本システムでは、音響パラメータとモータ制御パラメータの対応関係の学習と学習後の音響生成に、Kohonenの自己組織化マップ(Self-organizing Map: SOM) [8]を用いて学習データからルールを自動生成するファジー推論ニューラルネットワーク(Fuzzy Inference Neural Network: FINN) [9]を用いた。図7にFINNの構造を示す。FINNは入出力層とルール層からなり、入力層とルール層間、ルール層と出力層間は全結合となっている。入力層とルール層間にはメンバシップ関数が配置されており、メンバシップ関数から各入力に対する適合度μが得られる。またルール層と出力層間には後件部定数w'が配置されている。つまり、入力層とルール層がif-then型のファジー推論規則の前件部を表し、ルール層から出力層が後件部を表すことになる。

FINNの学習は、SOMによるルール抽出過程とLeast Mean Square (LMS)学習過程からなる。ここでSOMの入力に音響パラメータとモータの制御パラメータを用いることで、ある音響を生成する

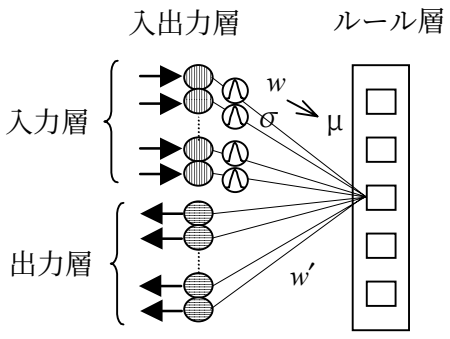


図7 FINNの構造

ために必要な声道形状（モータ制御信号）を学習することができる。更に学習後のFINNに生成したい音響の音響パラメータを入力すると、モータ制御量が出力されて声道形状を変化させ望む共鳴特性を与えることが可能となる。またSOMとNNの可塑性により、未学習パターンに対しても獲得した入出力関係を反映した出力が得られることが期待できる。

なお、ピッチの変化によってスペクトル包絡も変化するため、FINN学習時の音響の基本周波数は、成人男性の平均的な基本周波数である120Hzとしている。

3.2 FINNによるモータ制御量の学習

3.2.1 SOMについて

SOMは入力ベクトル$\boldsymbol{I}=(I_1, I_2, ..., I_m)^t$から、競合層上の各ユニット$\boldsymbol{v}_i=(v_{i1}, v_{i2}, ..., v_{im})^t$への写像を自己組織的に生成することが出来る。ここで用いたSOMは入力層の各ユニットと競合層上の各ユニットが全て結合しているものである。

SOMの学習は、競合層上の重みベクトルを変更することによって行う。まず、入力ベクトル$\boldsymbol{I}$と競合層上にある全てのユニット$\boldsymbol{v}_i$のユークリッド距離を求め、距離が最小となる最整合ユニットを$\boldsymbol{c}$とする。

$$\boldsymbol{c}=\arg\min_i\{\|\boldsymbol{I}-\boldsymbol{v}_i\|\} \qquad (1)$$

次に重みベクトルが、マップ生成時の時間tにおける入力ベクトルを$\boldsymbol{I}(t)$ 重みベクトルを$\boldsymbol{v}(t)$ とした時、最整合ユニット$\boldsymbol{c}$を用いて以下の式により更新される。

$$\boldsymbol{v}_i(t+1)=\boldsymbol{v}_i(t)+\boldsymbol{h}_{ci}(t)[\boldsymbol{I}(t)-\boldsymbol{v}_i(t)]$$

$$\boldsymbol{h}_{ci}(t)=\begin{cases}\alpha(t)\cdot\exp\left(-\dfrac{\|\boldsymbol{r}_c-\boldsymbol{r}_i\|^2}{\sigma^2(t)}\right) & (\boldsymbol{i}\in\boldsymbol{N}_c)\\ 0 & (\boldsymbol{i}\notin\boldsymbol{N}_c)\end{cases} \qquad (2)$$

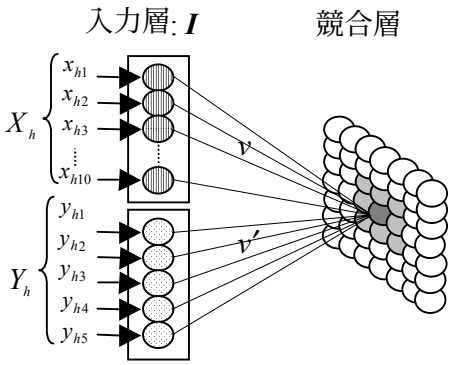


図8 本システムにおけるSOMの構成

ここで、$\boldsymbol{r}_i$はユニットiの位置ベクトル、N_cはユニットcの近傍集合、$\alpha(t)$は学習率係数、$\sigma(t)$はユニットc近傍の広がり幅を表す係数である。$\sigma(t)$と$\alpha(t)$は時間と共に減少する関数である。

以上の手順を全ての入力ベクトルに対して繰り返し行うことで、競合層上のユニットの重みベクトルは、類似したもの同士が近くに集まり、相違しているもの同士は遠くに配置される。

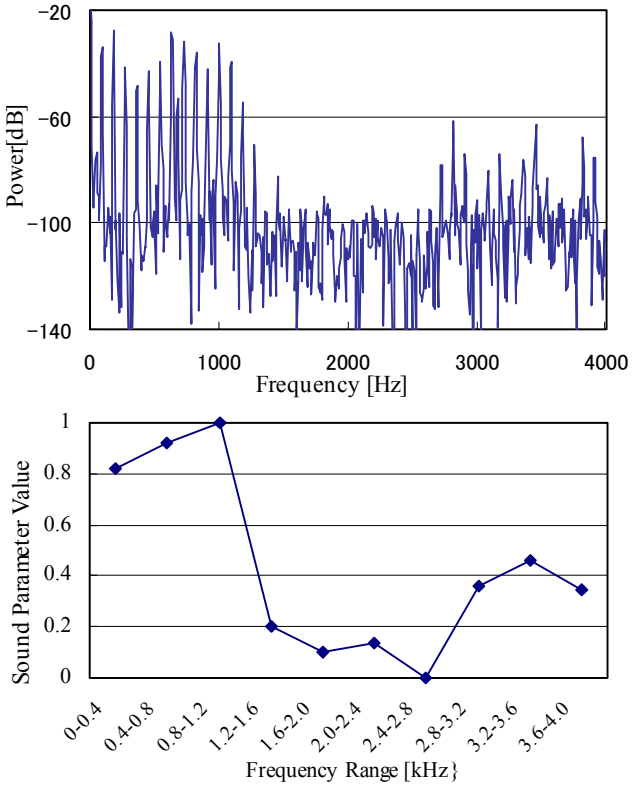
3.2.2 SOMによるルール抽出過程

本システムでは、図8に示すようにSOMの入力ベクトル$\boldsymbol{I}_h$に、10次元の音響パラメータ$\boldsymbol{X}_h$と5次元のモータ制御パラメータ$\boldsymbol{Y}_h$を統合したものを用いる。

$$\boldsymbol{I}_h=[\{\boldsymbol{X}_{hj}\},\{\boldsymbol{Y}_{hk}\}]^t \qquad (3)$$

$$\boldsymbol{j}=1,\cdots,10 \quad \boldsymbol{k}=1,\cdots,5$$

ここでいう音響パラメータとは、音響スペクトルの周波数領域(0～4kHz)を10等分し、400Hz毎のパワーの平均をそれぞれの領域で求め、これを最大値と最小値を元に規格化したものである。声道物理モデルから生成された音響のスペクトルとその音響パラメータの例を図9に示す。またモータ制御パラメータも、5つのモータの制御電圧値を規格化したものである。



(b) 音響パラメータ

図9 原音のスペクトルと音響パラメータの例

競合層上のユニットには、

$$\boldsymbol{v}_i = \left[\{\boldsymbol{v}_{ij}\}, \{\boldsymbol{v}'_{ik}\}\right]^t \quad (4)$$

$$j = 1,\cdots,10 \qquad k = 1,\cdots,5$$

の重みベクトルが配置されている。$\boldsymbol{v}_i$ と $\boldsymbol{v}'_i$ は、それぞれ X_h と Y_h に結合している重みベクトルであり、この値がメンバシップ関数の中心値と後件部定数となる。

以上の条件でSOMによる学習を行うことによって、マップ上では類似したルールを持つユニット同士が近くに配置される。更に、ルールの少ないシステムを構築するために学習後、任意の二つのユニットの重みベクトルを、ユークリッド距離の差がある閾値以下の場合に統合する。二つの類似したユニットの重みベクトルを$\boldsymbol{v}_1$, $\boldsymbol{v}_2$とし、統合された重みベクトルを$\boldsymbol{u}$とした時、統合は式(5)により行っている。

$$\boldsymbol{u} = \frac{\boldsymbol{v}_1 + \boldsymbol{v}_2}{2} \quad (5)$$

この操作をすべてのユニットの組合わせにおいて、ユークリッド距離の差が閾値より大きくなるまで繰り返す。

以上のアルゴリズムにより、学習終了後の各ユニットの重みベクトルから音響パラメータとモータ制御パラメータの関係を表すルールが生成されることになる。

3-2-3 LMS学習過程

SOMで仮決定したメンバシップ関数の中心値と後件部定数を図7のFINNに配置する。メンバシップ関数の中心値と後件部定数の微調整を行う為に、SOMによる学習時に用いた音響パラメータとモータの制御パラメータを教師信号としてLMSアルゴリズム[9]による学習をおこなった。また同時にメンバシップ関数の分散の調整を行う。これにより、システムの出力誤差が最小になるようメンバシップ関数の中心値と分散、後件部定数が最適化される。この際、過学習によりSOMから得られたルールを失わない為に、メンバシップ関数の中心値と後件部定数の変更定数を小さく設定した。またメンバシップ関数の分散の初期値は、小さすぎると学習の収束は早いが汎用性に乏しいシステムになってしまうので大きめの値に設定した。

4. ピッチ制御

4.1 聴覚フィードバックによるピッチ制御

人間が歌の練習においてピッチを習得して行く過程では、自分の出しているピッチを耳で聞き、目標となる理想のピッチと比較することによって誤差をなくすように学習して行く。本システムでは、聴覚フィードバックによってこの過程を模倣することにより適応的にピッチの学習を行う。ここで構成した機械システムは、人工声帯に取り付けているゴムの振動によって音源を生成しているため、得られるピッチは必ずしも再現性の良いものではない。また、ある基本周波数の音声を保持する場合にも、空気流量の変動や張力のわずかな変化に対してピッチが変動してしまう。このような外乱に対して安定した音声を生成するためにも、フィードバック制御は不可欠であると考えられる。

まずピッチ学習モードにおいて、出力音響の基本周波数とそれを与える2つのモータ制御値の対応関係を記述するマップを獲得する。発声実行モードでは、マップを参照しながらモータコントローラに位置データを送出して音響が生成されるが、出力は常に音響アナライザによって監視されており、ピッチのずれは適応的に修正される[2]。

4.2 学習モードにおけるピッチの学習

ピッチの学習では、目標のピッチと出力音声のピッチとの誤差を無くすようモータ操作量の演算を行いながら、ピッチの学習を進めていく。学習結果は、ピッチ-モータ対応マップとしてシステムコントローラ内に保存される。この聴覚フィードバックプロセスによって得られたピッチの例を図10に示す。なお図中のC～Gは音名のド～ソであり、それぞれ130.8, 146.8, 164.8, 174.6, 196.0Hzである。

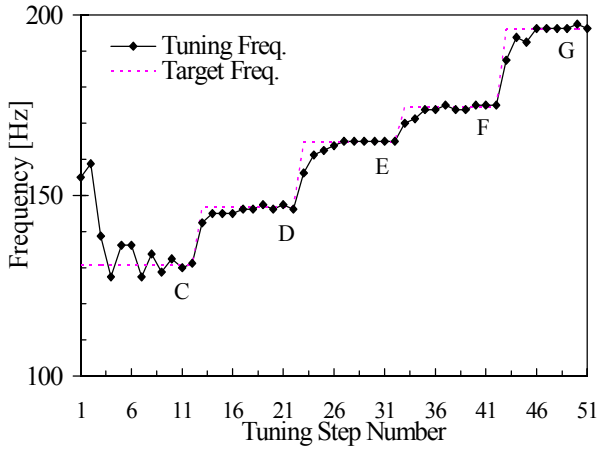


図10　ピッチ学習の結果

5. 音響生成

システム評価として、人の5母音の実音声を目的の音響とし、本システムによって音声の学習をおこなった。まずランダムに200パターンのモータ制御パラメータを生成し、これを声道物理モデルに与えて音響を出力させた。それぞれの音響から音響パラメータを求め、モータ制御パラメータと音響パラメータの組をFINNによって学習させた。

学習終了後、人間が「あ」～「お」の5母音を声道物理モデルに与え、音響を生成させた。図11に声道物理モデルによる母音「あ」、「い」発声時の声道の直径（共鳴管の初期直径(36mm) - モータ変位量）を、母音声道形状と共に示した。また図12に母音「あ」、「い」、「う」、「え」に対する生成音のスペクトルを人間の音声のスペクトルと共に示した。各スペクトルにおいてそれぞれの生成母音の特徴が見られ、モデルが学習によって声道形状を獲得していることがわかる。

精度の向上のためには、制御点数の増加、聴覚フィードバックによる適応的な制御などが更に必要と考えている。また、より自然な音声を生成させるためにも、子音発声動作獲得手法の検討、声道モデルの構成やその材質、声帯についての改良が必要である。

6. まとめと今後の課題

人の発声器官を物理的に構築し、その音声生成にニューラルネットワークを用いたシステムについて報告した。音源に人工声帯を用い、聴覚フィードバックを使って調音用共鳴管および声帯部を適応的に制御することによって、母音音声の獲得と音高の制御が可能であることを示した。また人間が音声を入力することにより、その音響スペクトルを模倣してリアルタイムで発話するシステムについて述べた。

本システムにより発話時の声道形状の静的および動的な変形を物理的に見ることが可能であり、例えば聴覚障碍者や発話障碍者の発話訓練に利用できると考えている。

今後は、音響の大きさと音高の制御を含んだ学習手法、子音発声のための発話動作の獲得、外乱などに対して安定した音声を生成させるための聴覚フィードバック機構の検討を進めて行く必要がある。また、声質を更に人間に近づけるために、声帯の構造や振動機構を詳細に解析して再現すること、声道共鳴官の材質を改良することなどを検討中である。

人間の発声器官を機械系により実現し、発話動作獲得のための学習機構を解析することにより、音声を通した新しいヒューマンインタフェース構築への足掛かりとなると考えている。

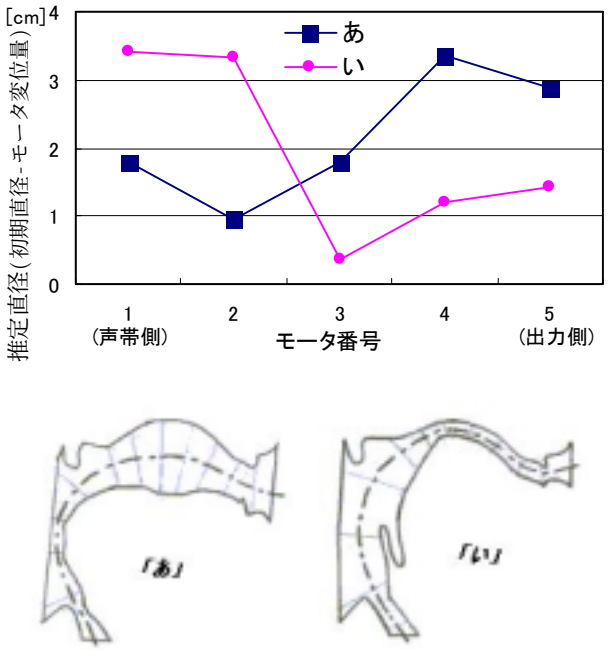


図11 共鳴管の推定直径と母音声道形状

参考文献

[1] James L. Flanagan: "Speech Analysis, Synthesis and Perception", Springer-Verlag, (1972)

[2] Hideyuki Sawada and Shuji Hashimoto: "Mechanical construction of a human vocal system for singing voice production", Advanced Robotics, International Journal of Robotics Society of Japan, Vol.13, No.7, pp.667-661 (2000)

[3] 梅田規子、寺西立年: 「声の韻質と声質－音響的声道模型による音声の合成」, 日本音響学会誌 第 22 巻, 第 4 号, pp.195-203 (1966)

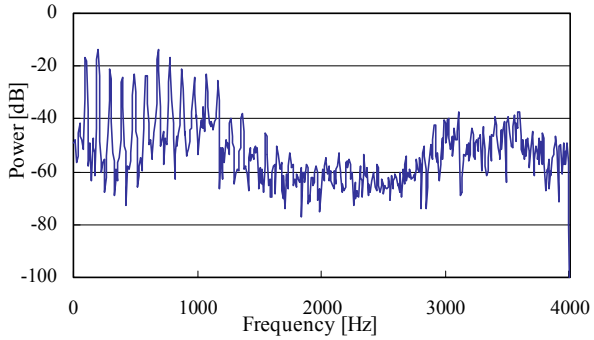
[4] 大須賀公一、荒木祐之、澤田謙次、小野敏郎: 「機械式音声合成装置の実現に向けて-第1報:構音器官の三次元形状の再現-」, 日本ロボット学会誌, Vol.16, No.2, pp.189-194 (1998)

[5] Kazufumi Nishikawa, Kouichirou Asama, Kouki Hayashi, Hideaki Takanobu and Atsuo Takanishi: "Development of a Talking Robot", CD-ROM Proceedings of IEEE/RSJ Int'l Conference on Intelligent Robots and Systems (2000)

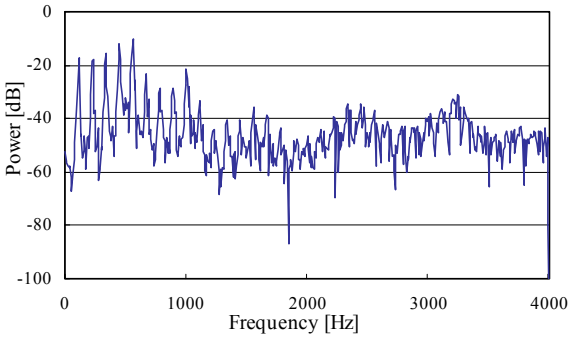
[6] Kenzo Ishizaka and James L. Flanagan: "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords", Bell Syst. Tech. J., 50, 1223-1268 (1972)

[7] 例えば、古井貞熙著: 「音声情報処理」森北出版 (1998)

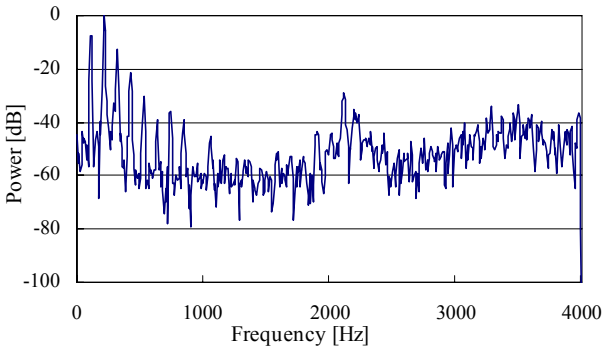
[8] Teuvo Kohonen: "Self-Organizing Maps", Springer-Verlag Berlin, 1995
[9] T. Nishina and M. Hagiwara: "Fuzzy interface neural network", Neurocomputing, Vol.14, pp.223-239, 1997



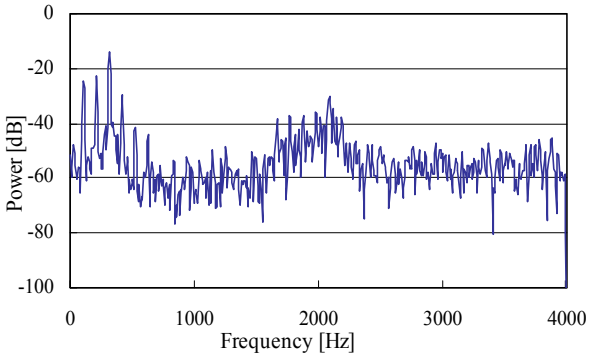
(a-1) 人間の「あ」のスペクトル



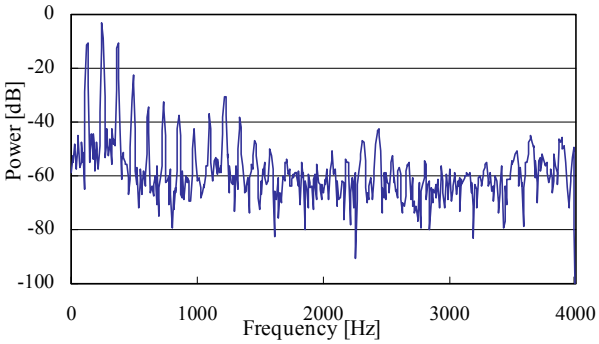
(a-2) 声道モデルによる「あ」のスペクトル



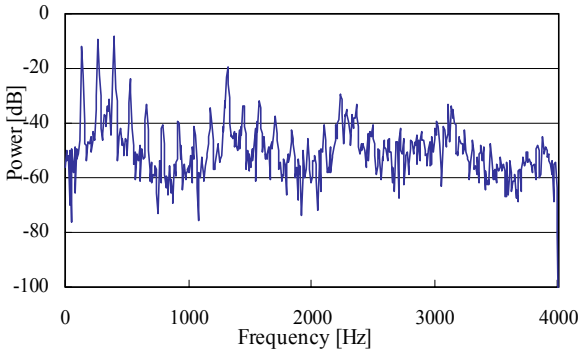
(b-1) 人間の「い」のスペクトル



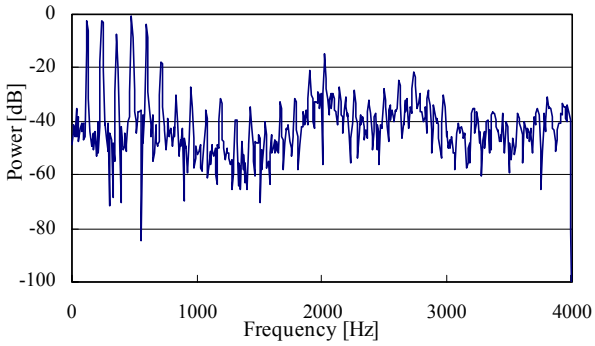
(b-2) 声道モデルによる「い」のスペクトル



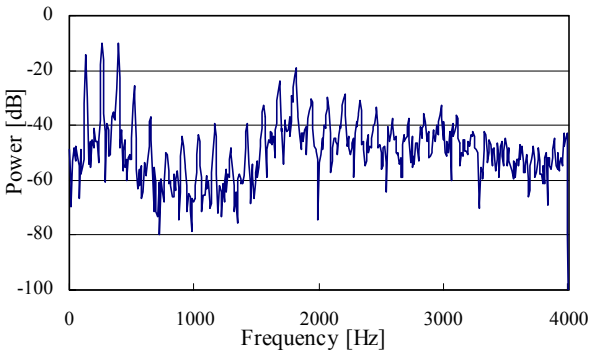
(c-1) 人間の「う」のスペクトル



(c-2) 声道モデルによる「う」のスペクトル



(d-1) 人間の「え」のスペクトル



(d-2) 声道モデルによる「え」のスペクトル

図12 声道物理モデルによる生成音声