

発話ロボットの発話動作獲得と歌声の生成

中村 光宏[†] 林 恭守[†] 澤田 秀之[†]

[†] 香川大学 工学部 〒761-0396 香川県高松市林町 2217-20

E-mail: [†] sawada@eng.kagawa-u.ac.jp

あらまし 発話機構を全て機械系によって構成することにより、人間のような発話動作に基づいた音声の生成が可能となる。主に肺、声帯、声道部、鼻腔、聴覚部から構成される本発話ロボットは、声のピッチを変化させながら発話を生成することが可能である。ピッチ制御には複雑な声帯振動の解析が必要であるが、ここではシリコンゴムを皮膚程度の柔度で成形した二枚の振動体の張力と、空気の流量を、モータにより操作することによって実現した。一方、調音のための共鳴管をシリコンゴムで成形し、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や子音声を生成することが可能である。本稿では、まず発話ロボットの構成について紹介し、次にニューラルネットワークを用いた聴覚フィードバック学習に基づく音声、音階の獲得と、楽譜作成インタフェースを用いた歌声生成について述べる。

キーワード 発話ロボット、聴覚フィードバック制御、発話動作、音階学習、歌声生成

Autonomous speech acquisition of a talking robot and its singing performance

Mitsuhiro Nakamura[†], Yasumori Hayashi[†] and Hideyuki Sawada[†]

[†] Faculty of Engineering, Kagawa University

2217-20, Hayashi-cho, Takamatsu-city, Kagawa, 761-0369, Japan

E-mail: [†] sawada@eng.kagawa-u.ac.jp

Abstract A talking and singing robot which has been constructed by referring to the human vocal system will be introduced in this paper. Although various ways of vocal sound production have been actively studied so far, mechanical construction of human vocal system is considered to advantageously realize natural vocalization with its fluid dynamics. The robot has several motors to manipulate the vocal tract and the vocal cords. The robot autonomously learns the relations between motor control parameters and the generated vocal sounds using an auditory feedback learning with neural networks, and sings a song by mimicking a human vocalization. This paper presents the construction of the talking robot and its singing performance, together with the adaptive control for the pitch and phoneme learning.

Keywords Talking robot, Auditory feedback, Human vocal system, Pitch acquisition, Singing performance

1. はじめに

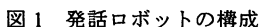
人間の発声器官は主に、肺、気道、声帯、声道、舌、口蓋とそれらを動かす筋肉などから成り、互いに適当な位置や形状を形成することにより言葉が生成される。人間は言葉を有効に利用して、コミュニケーションをおこなっている。これは、音声特別な道具なしに情報を伝え合うことができる、最も容易で効率的な手段であるからと考えられる。また音声には、言語的情報だけでなく、話し手が誰であるかという個人性情報や、喜怒哀楽のような話し手の感情を表す情緒表現などの様々な情報が含まれてお

り、これらを即時的情報伝達手段として有効に利用できるという利点がある。

このように人間は、音声を巧みに利用しながら、他者とのコミュニケーションを円滑におこなっており、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]。18 世紀末に von Kempelen による機械式音声合成器が作られ、音声合成に関する研究が始まったと言われている。初期の音声合成器はいずれも機械式のものであり、声帯の振動数（基本周波数）の操作や、声道形状の連続変形方法が課題であった。しかし 20 世紀に入り、ベル

現在の主流となっている音声合成は、ソフトウェアを用いた手法で、大きく分けて録音再生・編集方式、パラメータ編集方式、規則合成方式がある。しかしこれらの手法では、個人性の付加や自然な抑揚のある音声を生成することは困難である。これは、人間の音声は、個人の身体を用いた物理的構音動作により生成されるものであるためである。ソフトウェア合成により、波形レベルで音声に個人性や自然な抑揚を付加する手法は未だ確立されていない。近年、MRIを用いて計算機上に声道の3次元モデルを作成し有限要素法により音声を生成するもの、機械系で人間の発声機構を再現して物理的に音声を生成するものなど、人間の発話動作に基づき音声を生成しようとする研究が行われている^{[2]-[4]}。これらの研究は、音声を生成するだけでなく、人間の発声原理を解明することにも繋がると考えられる。

本ロボットは主に、エアーコンプレッサ、人工声帯、共鳴管、鼻腔、マイクロフォンで構成されている。これらはそれぞれ、人間の肺、声帯、声道、鼻腔、聴覚に対応している。マイクロフォンは、構音部により生成された音声を基に、発話ロボットが聴覚フィードバックにより自律的に発話手法を獲得するために使用するものである。



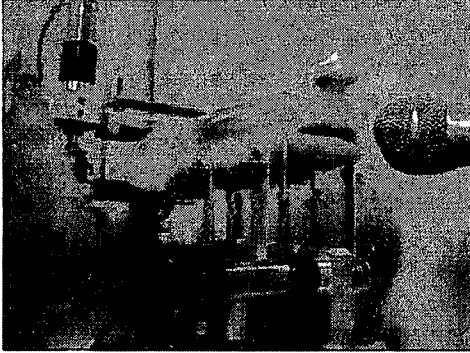


図2 発話ロボットの声道および鼻腔部

本ロボットは、声道モデルから生成される音声をマイクロフォンより入力し、音響解析部で解析を行う聴覚フィードバック機構を実現しており、音階の学習、発話動作の獲得をそれぞれ自律的に行うことが可能である。これにより、人間が目標となる音声を与えるだけで、ロボットが自律的学習によって発話手法を獲得することが可能となる。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、ソフトウェア合成では難しい音素間のつながりの自然性や、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。さらに、人間が楽譜情報を与えることにより、ロボットが自ら獲得した発話手法を用いて歌唱の生成を行うことが可能となった。発話手法の獲得と楽譜入力インタフェースについて、次章に詳述する。

3. 歌唱の生成

我々は、生活の中で自然に音楽や歌に慣れ親しんでいる。経済的文化的に十分に成熟した社会を迎え、人と人との対話コミュニケーションや、趣味としての音楽や芸術の必要性が、癒しや精神的な安定の面から今後ますます必要となっていくと考えられる。

近年、人間型ロボットやペットロボットなどが各所で提案され、いくつかは市販されているが、これらは身体動作の再現に主眼が置かれ、音声に関してはソフトウェア合成をもとにスピーカーから出力されている。発話ロボットに関する報告もいくつか見られるが^[4]、人間の発声機構や発話動作が十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。

そこで本研究では、発声器官を全てメカニカルに構築した発話ロボットが、自律的に発話手法を獲得し、歌唱をおこなう手法の検討を進めている。

人間の音声は発声器官の複雑な動きによってつくられるものである。人間には生まれながらにして

発話器官が備わっているが、発話手法は、乳幼児期の言語獲得期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される。本発話ロボットは、この過程を学習によって自律的に再現し、歌声を生成することが可能である。

歌唱の生成には主に、

- ① 歌詞を明瞭に発音できること
- ② 声の高さを楽譜にあわせて変化させること
- ③ リズムに合わせて音声が発話できること

の3つが必要となる。本研究では、声道形状の変形によって①を、声帯と空気の流量の制御によって②を、ロボット制御システムのクロック管理によって③を実現することとした。

3.1 音素生成のための声道形状獲得

人間の発話においては、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、一方の子音は顎や舌、軟口蓋などの非定常な動作から引き起こされる変化によってつくられる。つまり、ある「音」を発声させるための声道形状を決めることが出来れば、ロボットがその発声器官を使用して口を動かしながら発話することが可能となる。また、似たような音響特徴を持つ音は、似たような声道形状によってつくられることは、我々の発話中の口内形状からもわかる。

そこで本研究では、音の音響的な特徴と声道形状との関連付けに、ニューラルネットワーク(NN)を適用することとした^{[5],[6]}。ロボットから生成された音声の特徴パラメータと、その時のモータ制御パラメータとの対応関係を適応的に学習することにより、ある音響を生成するための声道の形状を推定することが可能となる。学習終了後には、NNを声道物理モデルに対して直列につなぐことにより、生成したい音声の音響パラメータを入力すると、NNが声道の形状を推定してモータ制御パラメータが想起されることで、望む発話動作を生成することができる。ここでは、自己組織化ニューラルネットワーク(SONN)による学習を提案する。

3.1.1 SONN による発話学習

SONNの構成を図3に示す。音響パラメータの関連付けに自己組織化マップ(SOM)を、音響パラメータからモータ制御パラメータの想起に3層パーセプトロンを適用することにより、音響パラメータとモータ制御パラメータを柔軟に関連付けることが可能となった。またSOM上では、類似した音響パラメータは近傍同士に、相違した音響パラメータは遠くに配置されるため、未知の音響に対しても、近傍のパターンを基にモータ制御パラメータが推定されることになる。

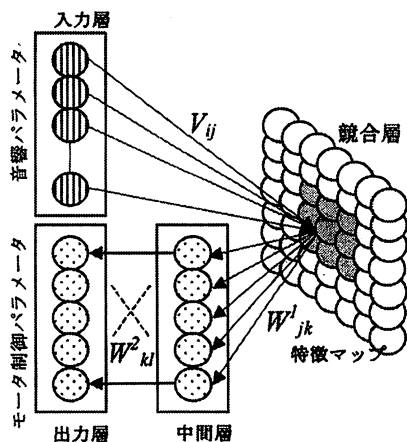


図3 自己組織化ニューラルネットワークの構成

SONNは、入力層、競合層、中間層と出力層をもち、重み係数を持つ結線により全結合している。入力には16次のケプストラム係数から成る音響パラメータを、教師信号には6つのモータ制御パラメータ(共鳴管:5 鼻腔:1)を用いる。3層パーセプトロンの学習には、バックプロパゲーションを用いた。SONNの学習後、ロボットに音声を入力すると、その音響パラメータが競合層上に写像される。競合層上の距離が閾値以内のセルを選択し、それらの距離に応じてセルの出力を決定する。競合層のパターンにより、出力層からの推定声道形状が求められ、ロボットの発話器官から音声が生産される。

3.1.2 発話文章入力のためのインタフェース

前述の学習で得た声道形状を用いて、日本語の文章を容易に作り発声するためのインタフェースの実装をおこなった。図4に示すように、50音パネルを用いて任意の音素をマウスで選択することにより、

容易に文章や歌詞の入力を行うことが可能である。さらに発話にゆらぎを持たせたり、歌唱表現を与えるために、以下のような細かなパラメータ設定が可能とした。

- 音素ごとに発声時間、空気の流量などのパラメータを設定できる。選択音素が子音の場合には、子音部と母音部のパラメータを別々に設定することにより、揺らぎや抑揚を与えることが出来る。
- 各音素間のモータ変化速度を細かく調整できるため、ゆっくりと音が変わっていく、滑舌を良くするなどの多彩な発声が可能となった。
- 各音素間の区切りの有無を選択することができる。空気流量設定の機能により、連続発声した場合のアクセントの強弱が多彩につけられるようになる。
- 作成したパラメータを保存、呼び出しすることが出来るため、編集が容易となる。

3.2 学習によるピッチの獲得

人工声帯の張力制御モータ及び空気流量制御モータにより、発話音声のピッチのコントロールが可能である^[7]。今回はシステム簡略化のため、空気流量コントローラのみを使うこととした。空気流量とピッチ変化の関係の例を図5に示すが、流量制御によりピッチが85Hz~165Hz(F1~E2)までほぼリニアに変化していることがわかる。

本ロボットは、空気流による二枚の人工声帯の振動によって音源を生成するため、得られるピッチは必ずしも再現性の良いものではない。ある基本周波数の音声を保持する場合にも、エアーコンプレッサの圧力変動による空気流量の変動や声帯張力のわずかな変化に対してピッチが変動してしまう。そのため、あらかじめ獲得したモータとピッチの関係を保持して利用することができない。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

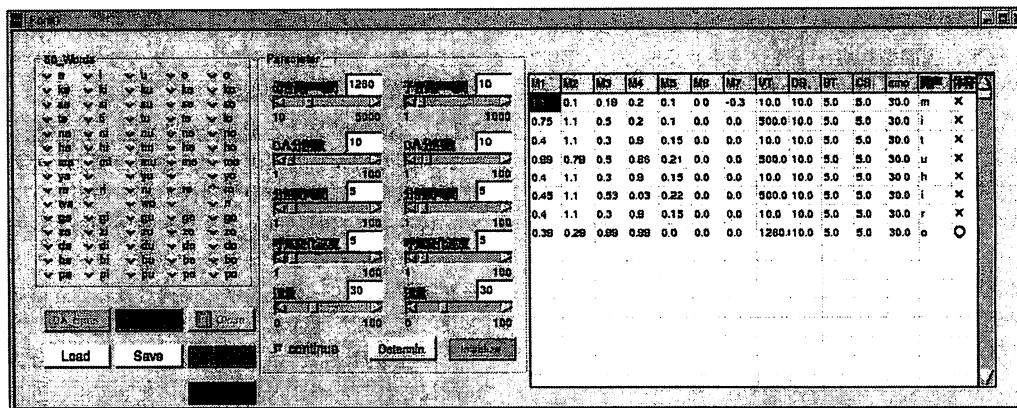


図4 日本語文章入力のためのインタフェースパネル

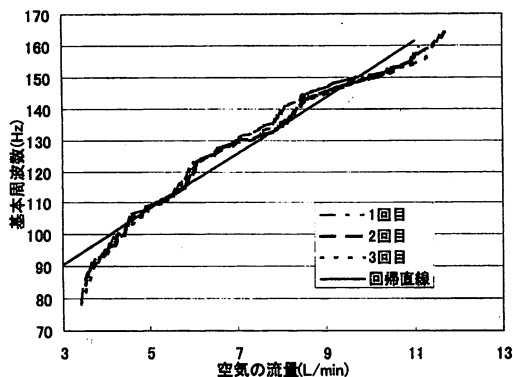


図5 モータ制御量とピッチの関係(空気圧 1 kg/cm²)

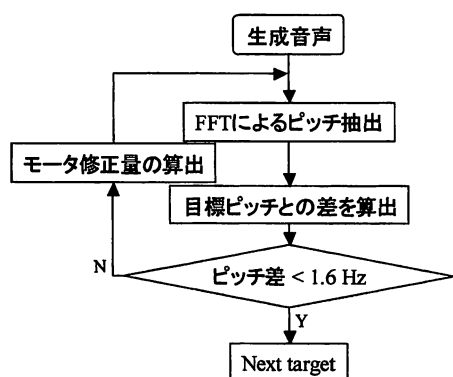


図6 音階の適応学習実行のためのフローチャート

バック制御により適応的にモータ制御量を修正していくことが不可欠となる。

そこで歌唱の前に、ロボットが発声練習をして、モータ制御量と声の高さを記述した「モータ＝ピッチ」マップを獲得し、これを参照しながら歌声を生成するシステムを構築した。歌声生成実行過程では、楽譜情報に基づいてモータコントローラに制御データを送出し、歌声が生成される。この時のピッチは常に監視され、ピッチのずれが適応的に修正される。

聴覚フィードバックによるピッチ獲得におけるフローチャートを、図6に示す。以下の手順によってモータ＝ピッチマップが獲得されることになる。

- ①システムコントローラが空気流量制御モータに任意の値を送出し、声帯振動が観測させてから学習が開始。同時に、目標となるピッチ値をコントローラに設定。
- ②生成された音声の基本周波数をFFTにより算出。
- ③生成音声のピッチと目的のピッチを比較し、ピッチ差を算出。
- ④ピッチ差に基づき、図5の関係から空気流量を推定。ピッチ誤差を減少させるモータ制御量を

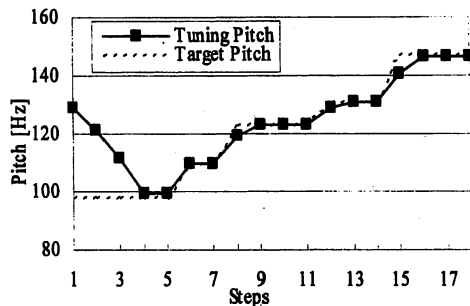


図7 音階学習の結果

ロボットに送出し、②に戻る。

⑤あらかじめ決められたピッチ差以下となった時のモータの制御量を、モータ＝ピッチマップに格納。

⑥次の目標ピッチをシステムに提示し、②に戻る。

以上を繰り返し、歌唱に必要な全ての音階を獲得することにより、モータ＝ピッチマップが得られることになる。

音階学習結果の例を図7に示す。この学習結果を用いて歌声生成を行うことにより、コンプレッサの空気圧変動や外乱の影響を抑えて安定した歌唱が行えることになる。

3.3 楽譜作成インタフェースと歌声生成

聴覚フィードバック学習によって得られる声道部および声帯の制御により、発話ロボットは発話と歌唱の生成が可能となる。

前述の文章作成のためのインタフェースに、ロボットが音階を自律的に獲得するためのプログラムを組み込み、更に、音階・音長などの楽譜情報をユーザが容易に作成できるインタフェースを実装した。インタフェースパネルを図8に示す。このパネルには、左から歌詞、音階、音長選択部があり、右側にはユーザが選択入力した楽譜情報が逐次表示されていく。歌詞選択部分は、前述の声道形状学習アルゴリズムで得た声道形状が関連付けされており、音階選択部分はピッチ学習アルゴリズムで得られた音階が反映される。これによりユーザは、任意の音素の選択と音階、音長の入力が容易にできる。またテーブルに楽譜が記載されていくので、直感的に楽譜として理解できる。

システムコントローラは、メトロノームに対応するタイマを内部に持ち、まず楽譜情報ユニットの音長情報を参照して、演奏テンポのプランニングを行う。楽譜情報を基に歌声出力信号が生成され、学習によって得られたマップを基にモータ制御信号が送出される事によって、ロボットが歌唱をおこなう。

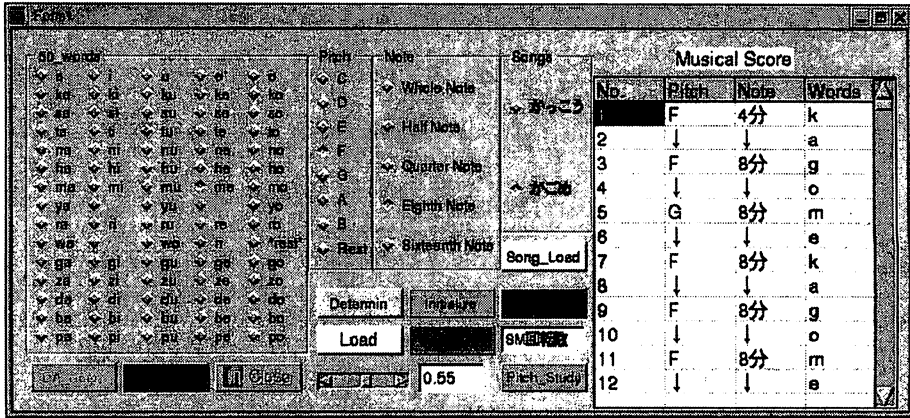


図8 歌声生成時のインタフェースパネル

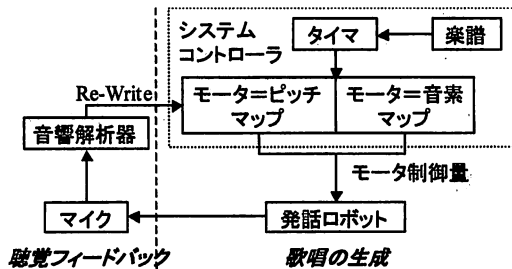


図9 歌声の生成とピッチ制御

3.4 歌唱実行とピッチ制御

歌声生成の流れを図9に示す。楽譜作成インタフェースを用いてユーザが入力した楽譜情報は、音素生成信号と音階制御信号を生成し、ロボットに送信される。これにより、声道部と声帯部が独立にモータによって駆動され、ピッチを変えながら歌声を生成することが可能となった。

演奏中には、空気圧の変動や声帯の張力の変化などが原因となって、出力音声のピッチが変動してしまう。そこで、聴覚フィードバックによるピッチの適応制御をおこなう。音響解析器によって出力音声のピッチを算出し、楽譜音名とのズレを常に監視しており、補正が必要と判断した場合にチューニング信号を出力し、モータ=ピッチマップの補正をおこなう。これにより、外乱や空気流の変動に対してロバストに歌声を生成することが可能となった。

4. まとめと今後の課題

本稿では、発話ロボットの構成を紹介し、聴覚フィードバック学習に基づく音声、音階獲得と、楽譜作成インタフェースを用いた歌声生成について述べた。

今後は、外乱に対して更に安定した音声を生成させるため、聴覚フィードバック機構の改良を進めて行く必要がある。また、声帯の構造や振動機構の解析、声道共鳴官の材質の改良、舌機構の付加といった発話器官の拡張により、発話の明瞭度を上げ、声質を更に人間に近づけることが可能となると考えている。また本ロボットは、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事が可能であり、聴覚障害者や発話障害者の発話訓練にも利用できると考えている。

文 献

- [1] James L. Flanagan: "Speech Analysis, Synthesis and Perception", Springer-Verlag, 1972.
- [2] 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.
- [3] 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.
- [4] Kotaro Fukui, Kazufumi Nishikawa, Shunsuke Ikee, Eiji Shintaku, Kentaro Takada, Hideaki Takanobu, Masaaki Honda, Atsuo Takanishi: "Development of a Talking Robot with Vocal Cords and Lips Having Human-like Biological Structure", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.2526-2531, 2005.
- [5] 東本敏男、澤田秀之:「声道物理モデルによる音響生成」, 情報処理学会 インタラクショナル論文集, pp.125-132, 2002.
- [6] Hideyuki Sawada and Mitsuhiro Nakamura: "A Talking Robot and Its Singing Skill Acquisition", International Conference on Knowledge-Based Intelligent Information and Engineering Systems, pp.898-907, 2005.
- [7] Toshio Higashimoto and Hideyuki Sawada: "A Mechanical Voice System: Construction of Vocal Cords and its Pitch Control", International Conference on Intelligent Technologies, pp. 762-768, 2003.