…平成15年度助成研究より…

声道物理モデルの機械系による構築と聴覚フィードバックに基づく発話獲得に関する研究

香川大学　工学部　知能機械システム工学科
助教授　博士(工学)　澤　田　秀　之

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響である。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、これらが互いに適当な位置や形状を形成することで言葉が生成される。

発声機構は大きく分けて、声帯振動による音源の生成と、声道による共鳴特性の付加という二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。音声の生成においては各組織の複雑な協調動作が必要であり、このような能力は幼児が生まれてから数年のあいだに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されるものである。また音声獲得後であっても、人間は練習によって声まねをすることができる。これは、発声のために必要な器官が生まれながらに備わっている一方で、発話技術は学習によって後天的に獲得されるものであるためと考えられる。

人間どうしのコミュニケーションにおいて、言葉は最も重要な情報伝達手段の一つであり、音声生成メカニズムや音声認識手法などが古くから研究されてきた。特に最近のヒューマンインタフェース技術には、人間らしい音声による情報提示や、音声による入力デバイスが不可欠な要素となっている。計算機を用いた音声の生成は、生音のサンプリングや物理モデルに基づいたアルゴリズムによる手法が研究の主流となっている。その一方で、人間の発声機構を物理的に構成することにより、より人間に近い手法で人間らしい音声を生成できると考えられる。

人間の音声を機械的に作り出そうとする試みは意外と古く、1791年のWolfgang von Kempelenによるものが最初と言われる。これは、吹子でリードを振動させ、革製の共鳴筒を手で変形させて音響を作り出すものであった。その後も幾つかの研究がみられ、1937年のRieszによる音声合成器は、電子的に発話動作を再現した。現在、人間との共存、対話を可能とする人間型ロボットの研究開発がさかんに進められている。しかしこれらの多くは形態、動作の模倣に関するものであり、特に音声の生成においては、ロボット自らに創発的に音声生成手法を獲得させようとする研究は見られない。

我々は、発話動作をおこなう器官を声道物理モデルに基づ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機械系自らが発話手法を獲得し、音声生成をおこなう研究を進めている。これまでに、機械式声道物理モデルを構築し、母音及び一部の子音音声の学習と生成が可能となった。

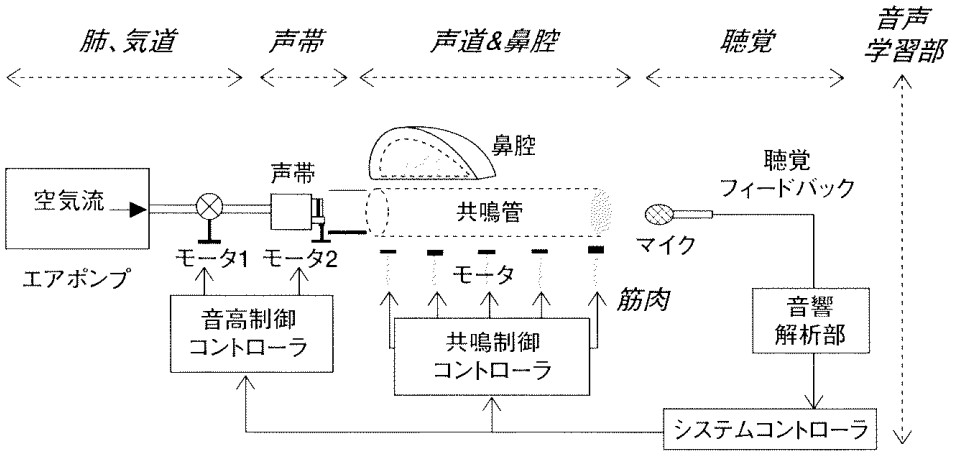


図１　声道物理モデルの構成図

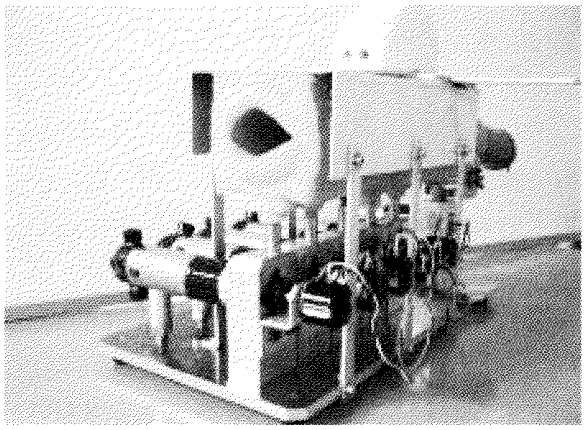


図２　声道部

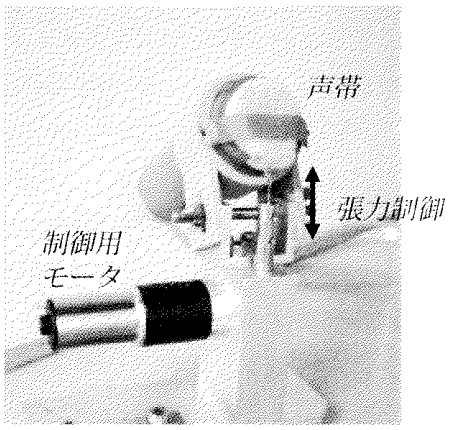


図３　声帯部

本研究で構築した声道物理モデルおよびその制御機構を図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、声道共鳴管、鼻腔共鳴部、マイクロフォン、音響アナライザから成り、それぞれ人間の肺、気道、声帯、声道、鼻腔、聴覚部に対応する。エアポンプから送られる空気流が人工声帯を振動させることで原音（音源波）を生成し、声道モデルの共鳴管形状をモータにより変形させることによって任意の共鳴特性を持った音声を生成する。機械モデルから生成される音声は、マイクロフォンから音響アナライザに入力される。音声から抽出される音響特徴をモータ制御量と関連付けることにより、発話動作と音声を適応的に獲得する聴覚フィードバック学習を実現した。声道部を図２に、声帯部を図３に示すが、共にシリコーンゴムによって人間の皮膚程度の柔度で成形されており、その変形操作にDCモータを用いている。

人間が歌の練習においてピッチを習得していく過程では、自分の出している声の高さを耳で聞き、目標となる理想の高さと比較することによって誤差をなくすように発声手法を学習していく。本システムは、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、獲得をおこなう。作成した人工声帯は、シリコーンゴムで成形された２枚の羽根による振動で音源を生成しているため、得られ

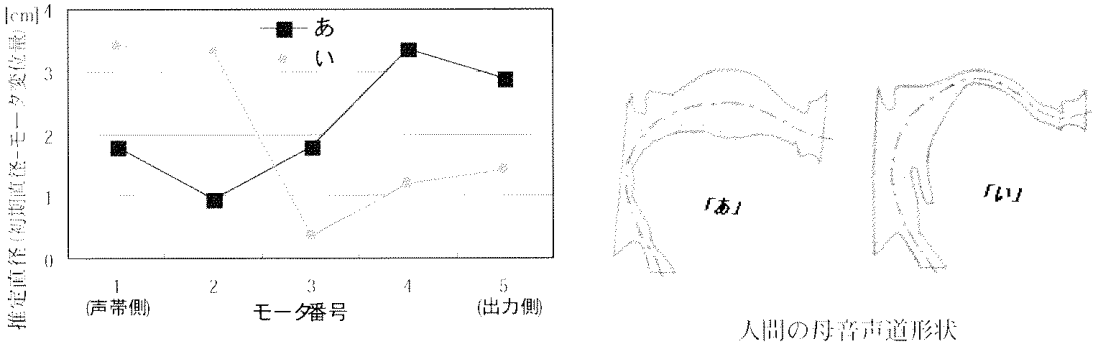
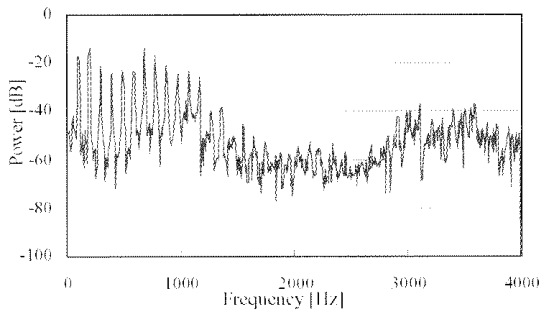
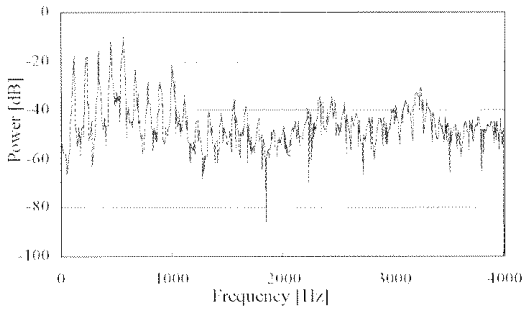


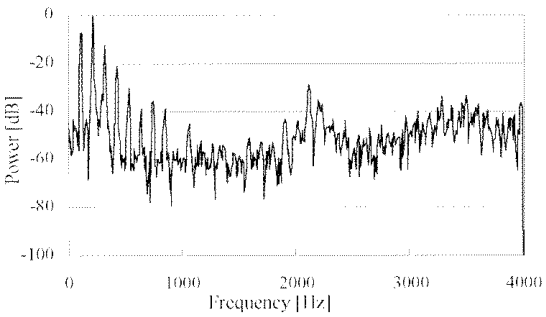
図４　声道共鳴管の推定直径と母音声道形状



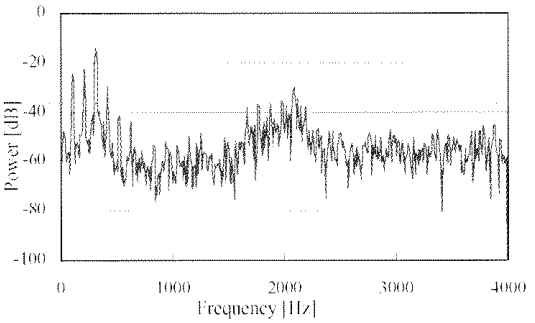
（ａ－１）人間の「あ」のスペクトル



（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

図５　声道物理モデルによる生成音声

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。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（空気流量制御および声帯張力制御）の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制御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。

母音・子音の発話動作の学習には、ニューラルネットワーク（NN）を適用した。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

形状および、発話時の声道制御手法の獲得が可能となった。まず学習時において、NNはシステムが生成した音声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、特定の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に生成したい音声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、望む音声を生成することが可能となった。

上述の聴覚フィードバック学習により、人間が目標となる音声を与えるだけで、自己学習による発話動作の獲得が可能となった。システムの評価として、日本語5母音の実音声を目的の音響とし、本システムによって発話動作獲得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。それぞれの音声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ニューラルネットワーク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、音声を生成させた。図4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5に、生成音のスペクトルを人間の実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母音に共通の特徴が見られ、本システムが学習によって声道形状を良好に獲得出来たことがわかる。

今後更に明瞭な発声をおこなうために、機械系各部の形状、自由度および、その制御に必要なモータ点数について検討をおこなう必要がある。また聴覚フィードバックによる音声獲得過程の解析、目標とした人間の音声特徴と学習の結果獲得された音声の特徴の比較などをおこない、人間の学習能力の解明への手がかりとしていきたい。更に本機械モデルについて、音声を通したヒューマンインタフェースデバイス構築のための要素技術、可能性の面からも検討をおこなっていく。機械システムが人間のように音声から発話動作を適応的に学習し、再現することにより、人間どうしのように発話動作をともなった音声コミュニケーションが実現できるばかりでなく、聴覚障碍や発話障碍を持った患者が、ロボットの発話動作を見ながら対話的に発声訓練できるシステムの提案も可能となると考えている。

最後に、本研究遂行にあたり助成を賜りました財団法人サウンド技術振興財団に、心より感謝申し上げます。