博 士 論 文

人間の音声生成メカニズムに基づく音声合成方式に関する研究

指導教官　党 建武 教授

北陸先端科学技術大学院大学
情報科学研究科知識情報処理学専攻

平井 啓之

2008 年 1 月 23 日

要旨

近年，コーパスベースの波形編集型音声合成技術の進歩により，高品質な音声の合成が可能となってきた。しかし，この方式では収録されていない声質の音声の合成は不可能である。よって，前後の音素の環境や声の高さが異なると音声の周波数特性が変化する為，合成する環境に合った全ての音素をデータベースに用意しておく必要がある。また，感情や個人性などの多様な音声の生成を行う場合は異なる発話様式毎に異なるデータベースを用意する必要がある。その結果，データベースのサイズの増大と言う欠点が生じことになる。今後，ますます複雑になる様々な機器と人間とのインターフェースとして音声合成を用いる為には，機器への組み込みが可能なコンパクトでかつ多様な音声を合成することのできる音声合成方式の開発が不可欠であろう。このような方式の１つに人間の音声生成メカニズムに基づく音声合成方式がある。この方式は 1) パラメータの補間特性が良く大幅なパラメータ圧縮を行うことが可能， 2) 人間の音声生成メカニズムを分析し解明した結果を直接モデルに反映することが容易である為に従来データベースに蓄積せざるを得なかった様々な特性を持つ音声を計算により生成することが可能，と言った特徴がある。そのため，これまでに幾つものモデルが提案されてきた。しかし，それらのモデルが実用化に至っているとは言い難い。その要因は，1) 音声生成機構には未だ明らかにされていない生理機構が存在すること，2) 音声生成機構はソース (音源)・フィルタ (声道形状) の結合としてモデル化できるがそれぞれを独立した要素としてモデル化が進められてきたこと，3) モデルの構築に用いる声道形状データの量と質が不十分であったこと，などが考えられる。そして，これらの要因は発話時の声道の観測が困難であった為にこれまで解決することができなかったのではないかと考えている。この問題に対し，近年 MRI の技術の進歩により様々な条件で発話時の声道の撮像が可能となってきた。よって我々は最新の MRI 技術を用いてこれまで抱えてきた問題を再考することで，この方式の実用化を目指すことにした。

本論文では，人間の音声生成メカニズムに基づく音声合成方式に関する我々の研究を論じる。ここで我々は，最新観測手段によりソースおよびフィルタの問題をそれぞれ再考し，両者の相互関係を取入れ２次元生理学的モデルを構築するこ

とにより人間のメカニズムに基づいた音声合成方式を試みる。さらに，本方式の実用化へのボトルネックとなる音質の問題を解決するため，３次元動画計測手法を用い声道断面積を正確に計測し声道形状モデルを高精度化することで高品質化を目指す。これらを行うため，本論文では，まずソースにおける F0 調節の生理機構について分析し，咽頭筋の活動以外の新たな F0 調節機構について提案する。つぎに，F0 調節と声道形状制御を１つのモデルに統合した発話器官の２次元モデルを構築し，提案した F0 調節機構により引き起こされる F0 調節と声道形状制御との相互作用の存在を明らかにする。最後に，この F0 調節機構を実装することが容易で，かつ高品質な音声の合成ができ，テキスト音声合成の実用化への応用に適した声道断面積モデルを用いた音声合成方式を提案する。

F0 下降の生理機構の解明では，複数の被験者が降下音階で持続発声を行った時の喉頭の正中矢状断面画像を MRI を用いて撮像し，喉頭周辺の器官の位置変化を調べることで F0 調節機構の分析を行った。その結果，輪状軟骨の後板が頸椎の自然湾曲に沿って下降する結果，声帯を短くする回転が生ずる現象を確認した。この生理機構によりこれまで問題とされてきた F0 下降に伴う喉頭下降の現象や外喉頭筋の活動の理由を説明することができる。また，この F0 調節機構を考慮すると，F0 調節はその力が舌骨を介し舌形状を変化させ，反対に調音動作は咽頭の位置を変化させ F0 が変化するといった，ソースとフィルタ間の相互作用が生じる。よって，本研究ではつぎに，先に示した F0 調節機構を考慮し，舌と喉頭を含む全ての器官の動作を同時に計算することで，この相互作用を実現する発話器官の２次元計算モデルを提案する。本モデルでは，各筋の活動量を入力として舌・喉頭・顎などの発話器官に加わる全ての力が釣り合うときの各器官の位置が計算され，その位置より得られた声道形状および声帯長を用いて音声の生成が行われる。合成実験により，先に示した F0 調節機能が F0 調節と舌形状制御の相互作用を引き起こすこと，本モデルでその相互作用が実現できることが明らかとなった。また本モデルでは，全てのパラメータを特定の１名の話者の生理学的データと MRI 画像を用いて同話者に最適化することで，個人性を有する合成音声の再現を目指した。しかし，個人性の再現には至らなかった。この結果より，この合成方式の実用化には音質がボトルネックになると考えられる。よって最後に，フィルタについて，これまで行ってきた２次元断面上での声道形状の模倣を３次元に拡張することで，

本方式を用いても高品質な音声の合成が可能であることを明らかにする。具体的には近年開発された３次元 MRI 動画データと実音声を用いて声道形状を推定することで高精度化を行った。声道形状のモデルには声道を任意の幅の円錐台の連続体で近似する声道断面積モデルを用いた。ただし，将来的に先に解明した F0 調節機構などの音声生成の生理機構を組み込むことができるように，声道断面積モデルは物理的な声道の器官の位置との対応が容易にとれることを考慮し構築した。幾つかの単語の合成実験の結果，声道の３次元動画データを用いることで，人間の音声生成メカニズムに基づく音声合成方式を用いても，高品質な音声の合成が可能であることが確認された。今後はこの声道断面積モデルに，先に示した F0 調節の生理機構を実装し，また，任意の音声を合成するために必要な音素を含む幾つかの単語の分析実験を追加することで，コンパクトで自然な音声を合成できるテキスト音声合成システムを開発する予定である。

abstract

Recently with the advancements in the corpus-base speech synthesis technology, it becomes to be possible to obtain synthetic speech sounds with a high quality. Generally different context environment and F0 changes possibly induce some changes in the frequency characteristics. For this reason, it is necessary to collect huge speech data to cover all phoneme environment changes. This is not feasible in fact. Moreover, to synthesize the emotional speech and/or personalized voice, it requires database to have variety of speech with each style. To build a friendly user interface between human and familiar equipments it requires us to develop a compact speech synthesis method that can synthesize the desired voice quality and speech styles. For this purpose, a human mechanism based speech synthesis method is one of the solutions. This method has some definite advantages. One of them is that the database can be compressed greatly because the parameters used in the method vary slowly with time as human articulation. The other advantage is that this model is able to produce various synthetic speech sounds by manipulating the parameters instead of collecting the speech data. So far, a number of human mechanism based models have been proposed, unfortunately, few models can be used practically. One of the problems is that the physiological mechanism of speech production has not fully understood yet because of the difficulty in measuring speech organs during speech. The other problem is that the sound source and the vocal tract shape have been modeled independently, the interaction between them was not considered sufficiently. In addition, there are not enough the vocal tract data with high quality for constructing the model. Due to the development of the MRI technology in recent years, the measurement of the vocal tract shape under various conditions has become to be possible. Therefore, we attempt to develop a practical speech synthesis method based on human speech production mechanism by means of the advanced MRI technology.

In this paper, we proposed a novel speech synthesis method based on human speech production mechanism, where the human speech production is modeled as a combination of a sound source and a filter, the resonance property of the vocal

tract. Based on the new observations using MRI technology, we refine the sound source model and the filter (vocal tract) part respectively, and develop a speech synthesis method by taking the interaction between the source and filter using a 2D physiological speech organs model. Furthermore, we challenge the bottleneck, the sound quality, towards practical use of such a method. To break the bottleneck and develop a practical system, we proposed a vocal-tract area function model by applying an accurate 3D measurement method on dynamic vocal tract shapes. The study was carried out in the following procedures. At first we investigated F0 control mechanism by analyzing the laryngeal complex and proposed a control method. Then, we proposed a physiological model of the speech organs and used it to confirm our observation of F0 control mechanism. Finally, we proposed a speech synthesis method which is suitable for practical application of text-to-speech synthesis system by using vocal-tract area function model.

To investigate F0 control mechanism, the MR images were measured during phonations with different F0 levels. It is found that a rotation of the cricoid cartilage was always associated with laryngeal descent during lowering F0. The function of the rotation is to shorten the vocal folds, and the mechanism was realized by vertical sliding motion of the posterior plate of the cricoid cartilage along the physiological curvature of the cervical vertebrae. This mechanism showed that the laryngeal descent and strap muscle activity are responsible for F0 lowering. Based on this mechanism, changes in F0 may cause a change of the tongue shape, and vice versa. To investigate this relation, a physiological model of speech production was designed to represent the interaction between F0 change and tongue shape change. The position of the speech organs was computed by driving their static equilibrium using muscle forces. Speech was synthesized based on the calculated vocal-tract shape and the length of vocal-fold. Simulation results of this model demonstrated that the proposed model can represent the observed F0 control mechanism and realize the interaction between F0 control and articulatory activity.

However, the proposed model cannot reproduce an acceptable individual speech.

Poor sound quality is the bottleneck for this method to be a practical one. To solve the problem we extend the optimized parameters of vocal-tract from 2D to 3D. Accurate 3D vocal tract shapes were estimated from 3D MRI Movie data with reference to the recorded speech data. The vocal-tract shape was represented by a vocal-tract area function model. The results of a comparison between synthesized and recorded speech sounds showed that the proposed method can provide high quality speech sounds by using 3D MRI data. In the future, a compact text-to-speech system is planed to be developed on the proposed approach by taking the F0 adjustment mechanism into account.

目 次

第 1章

まえがき

音声はコミュニケーションにおいて最も重要な道具の１つである。人と機械とのコミュニケーションの役割が高まってきている現代において，音声合成は機械からの情報発信方法として重要な技術の１つと言える。近年，合成音声の品質の向上はめざましく，街中でも様々な所で利用が広がってきている。しかし，それらは主に大きなシステムに組み込まれた音声合成サーバからの出力であり，携帯電話やカーナビなど組み込み機器の合成音声は音質の面で未だ十分とは言えない。また，音声合成サーバからの出力音声についても，流暢にはなったが，感情のない単調な合成音声であるため，コミュニケーションの本来の意味である互いの意識共有の道具という点から考えると，音声合成の技術はまだまだ不十分であると感じられる。様々な機械との自然なコミュニケーションを取るには，目的に応じて感情や声質など多様な音声を表現できるコンパクトな音声合成の開発が必要であろう。

現在のテキスト音声合成では，大量に録音された音声データベースの中から最適な音素片を選択し，最小限の変形を行い接続するというコーパスベースの波形編集合成と呼ばれる手法 [1][2] が主流である。このような手法では，データベースの規模を大きくすることで，録音された音声波形を原音に近い形で出力することが可能となる。そのため近年の記憶メディアの技術進歩に伴いデータベースの大規模化が進むことで，高品質な音声の合成が可能となってきた。しかし，収録されていない声質の音声の合成は不可能であるため，前後の音素の環境の違いや声の高さ (F0) が異なることにより音素の周波数特性が変化する場合，利用される環

境に合った音素を全てデータベースに用意しておく必要がある。その為，データベースのサイズに制約がある携帯電話などの組み込み機器に用いる場合は音質の劣化が生じる。また，感情や個人性を含む多様な音声を生成するにも，目的に応じた声の種類毎にデータベースを用意する必要があり，この収録に費やす時間と労力についても問題となってきている。

それに対し，人間の音声生成メカニズムに基づく音声合成という手法がある[3][4][5][6][7]。この手法では，人間の発話動作を模倣することで音声の生成が行われる。このモデルでは，声帯や声道の形状や特性を表す物理的意味を持つ変数がパラメータとして用いられる。これらのパラメータは時間変化がゆっくりであることや，補間特性が良いことから大幅なパラメータ圧縮を行うことができると言われている。さらに，人間の音声生成メカニズムを解明した結果を直接モデルに反映することが容易であることから，人間と同じ特性を持った合成音声を新たに生成することも可能であると思われる。よって，波形編集方式でデータベース増加の要因となった前後の音素環境の違いやF0の違いなども，それらの生じる物理的・生理的要因を解明し合成方式に反映することで，蓄積された標準的なパラメータから利用される環境に応じた多様な音声に変換し合成することが可能となると思われる。また，感情やは個人性なども生理的な要因により生じる現象と考えられるため，同様に生成するアルゴリズムを実現できる可能性が高いと思われる。

しかし，このような優れた特徴を有するにもかかわらず，これまで実用化された例はほとんどない。その原因は，人間の音声生成機構は非常に複雑であり，未だ明らかにされていない機構が存在すること，そのため音源と声道との制御は互いに独立した動作として単純化・モデル化が進められてきたこと，モデルの構築に用いた声道形状データの量と質の不足のため合成音声の品質が不十分であったことなどが考えられる。そしてこれらの問題は，実際に音声を発話している時の音源および声道の発声・発話機構の観測がこれまでは非常に困難な問題であったため解決することができなかったと思われる。それに対し，近年MRIの技術の進歩により，様々な条件で発声・発話動作の撮像が可能となってきている。よって、われわれは最新のMRI技術を基に，これまで抱えてきた問題を再考したうえで，発声・発話の各モデルを高精度化することで，人間のメカニズムに基づく音声合成方式の実用化を目指し研究を行う。

1.1 本論文のフィロソフィ

本論文では、人間の音声生成メカニズムに基づく音声合成方式に関する我々の研究を論じる。ここで、我々は、人間の音声生成機構を，ソース (音源)・フィルタ (声道形状）の結合としてモデル化し、最新観測手段によりそれぞれのモデルについて問題点を再考し、両者の相互関係を取入れ２次元生理学的モデルを構築することにより人間のメカニズムに基づいた音声合成方式を試みる。さらにこのような合成方式の実用化へのボトルネック（合成音質）を解決するため、最新の３次元の計測手法を用い声道断面積を正確に計測し声道形状のモデルを高精度化することにより、提案合成方式の高音質化を目指す。この目標を達成するために，まずソースにおける F0 調節の生理機構について分析し，咽頭筋の活動以外の新たな F0 調節機構について提案する。つぎに，F0 調節と声道形状制御を１つのモデルに統合した発話器官の２次元断面モデルを構築し提案した F0 調節機構の存在を明らかにする。最後に，この F0 調節機構を実装することが容易で，かつ高品質な音声の合成ができ，テキスト音声合成の実用化への応用に適した声道断面積モデルを用いた音声合成方式を提案する。以下、図 1.1 を用いて本研究の概略構成を示す。本研究の最終的な目的は，図中の最下部に示す人間の音声生成メカニズムに基づく音声合成方式を用いたテキスト音声合成の実用化である。本論文では，図中のグレーで囲まれた部分について論じる。

人間の音声生成メカニズムに基づく音声合成方式では、未だ明らかになっていない音声生成機構を解明しモデル化を進めることで合成音声の自然性が向上することが期待される。本研究では、はじめに、ソースに関して F0 下降の生理機構の解明 (図 1.1 の左上) を行う。F0 を上昇させるメカニズムについては，以前より，直接作用する輪状甲状筋の存在が知られている [8][9]。一方，F0 下降については直接作用する筋が存在せず，筋の弛緩により説明されてきた。しかし，F0 下降時には，舌骨に付着する筋の活動や喉頭が下降する現象が観測されており [10][11]，輪状甲状筋の弛緩だけでは説明できない問題がある。内喉頭筋 (輪状甲状筋) 以外の筋の活動は舌形状に影響を与える可能性が高いことから，このような F0 調節機構が明らかになれば，コーパスベースの波形編集方式におけるデータベースの増大の要因の１つにもなっている F0 変化と音声の周波数特性の変化との関係が明らか

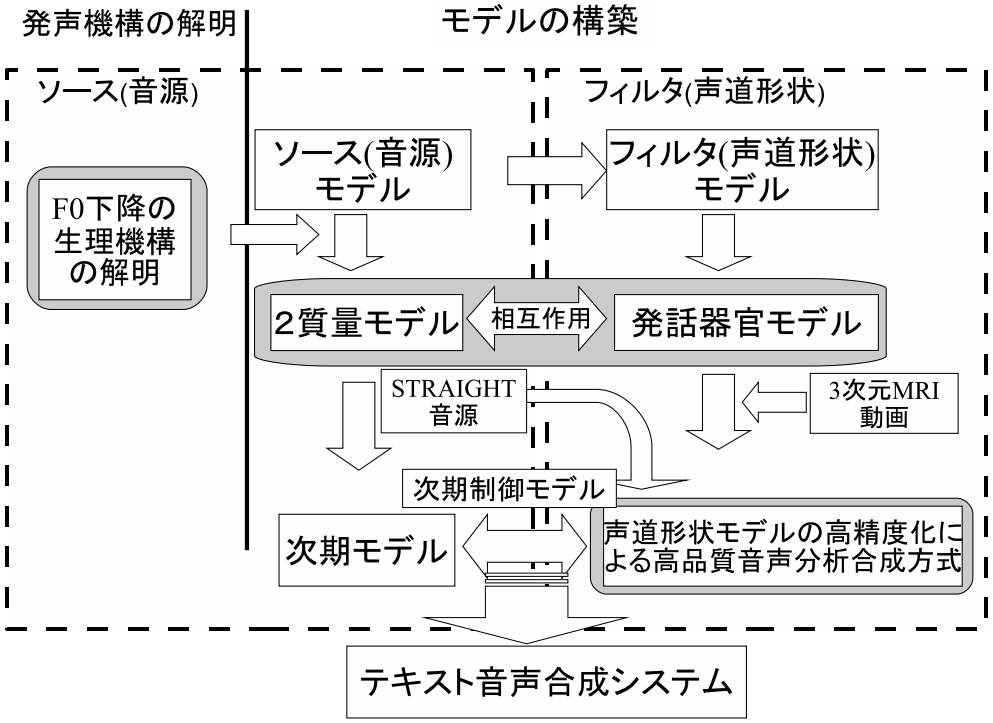


図 1.1: 本研究の構成

になると考えられる。本研究では，F0 を下降させた時の声道形状を磁気共鳴画像 (MRI) の高速撮像法を用いて撮像し分析することで，輪状甲状筋以外の F0 調節機構について明らかにする。

次に，先に明らかにした F0 下降の生理機構を考慮した F0 調節機構を有する発話器官の 2 次元計算モデル (図 1.1 の中ほどのグレー) を提案する。この研究の目的は 2 つある。第 1 の目的は，先に明らかにした F0 調節機構を考慮することによりソース (喉頭による F0 調節) とフィルタ (声道形状制御による音声の周波数特性の変化) との間に相互作用が生じること，および，この提案モデルでその相互作用を表現できることを明らかにすることである。一般に音声生成メカニズムに基づく音声合成では，ソースとフィルタは独立したモデルとしてモデル化されている。しかし，先に明らかにした F0 下降の生理機構を考慮すると，ソースを制御する喉頭とフィルタを制御する舌との間には舌骨を介した力の授受が生じることがわかる。提案モデルでは，筋活動を入力とし舌と喉頭を含む全ての発話器官間の力の授受を考慮し声道形状を計算することにより相互作用を実現する。第 2 の目的は，提案方式による音声合成の音質の評価である。この研究では，モデルのパラメー

タを特定の話者に最適化することにより特定話者の個人性を有する合成音声の生成を目指す。そして，合成音声と同話者の実音声とを比較することで提案方式の音質評価を行う。

最後に，フィルタのモデルについてパラメータを 3 次元 MRI 動画を用いて推定することで高精度化し，高品質な音声の合成を可能にする音声の分析合成方式 (図 1.1 の右下のグレー) を提案する。先に提案した発話器官の 2 次元モデルは，定性的には F0 調節と声道形状の制御との相互作用などを再現できる結果が得られたが，定量的には合成音声と実音声のフォルマント周波数の比較結果から大きな誤差を含むことがわかった。そのことから，本方式の実用化へのボトルネックの１つは音質であると考えられる。先のモデルでは MRI の矢状断面の 2 次元声道形状の再現を目標にパラメータの最適化を行なった。しかし，音声の合成には 3 次元の声道形状データが必要である。また，声道形状の２次元から３次元への変換は未だ確立された方法が存在するわけではない。よってこの誤差は，主に 2 次元から 3 次元へ変換する際に生じた可能性が疑われる。先のモデルを含めて声道形状のモデルには一般的に 2 次元のモデルが用いらることが多い。その理由の１つは，これまでは発話時の 3 次元声道形状データの計測が非常に困難であったことが挙げられる。これに対し，近年 3 次元 MRI 動画撮像法が開発された。よって，本研究ではこの 3 次元 MRI 動画を用いることでフィルタ部の高精度化を図る。発話器官の動作を再現する力学的モデルの 3 次元への拡張は既に党らによって進められており [12]，舌の有限要素法の改良などにより高精度に発話器官の 3 次元の動作を再現することが可能になってきている。しかし，音質については，個人性を再現できるまでには至っていない。よって，本研究では，計算コストも考慮し，テキスト音声合成システムとしての実用化へのアプローチとして声道断面積関数を高精度で近似するモデルを提案する。提案方式では，3 次元 MRI 動画 [42] と実音声を入力として，実音声のスペクトルを再現できる実際の声道に近い声道断面積関数の推定を行う。本方式で得られた声道断面積関数を入力とし高品質分析合成システム (STRAIGHT)[25] の合成手法を応用することで，発話メカニズムに基づく声道のモデルを使用しているにも関わらず，高品質な音声の合成を行うことができることを明らかにする。ただし，このモデルでは F0 を直接入力することにより韻律の制御が行われる。つまり，これまで提案してきた F0 調節と声道形状制御との

相互作用を実現する機構は含まれていない。これは，この研究がフィルタ部について注目し，音質を改善することを主目的とした為である。評価実験では３次元MRI 動画の測定と同じ単語を同じ様に発話した時の実音声から抽出した F0 を用いて音声の合成を行うため，相互作用を含んだ自然な音声を合成することができる。しかし，自由な文を生成するには，党らのような３次元の発話器官モデルに拡張するか，発話器官モデルのシミュレーション結果に基づいて得られた知見を基に声道断面積モデルと F0 を同時に制御する機構を別に設ける必要があると考えている。先に２次元のモデルで明らかにしたように喉頭を含む発話器官モデルでは原理的に自然な形で相互作用を表現することが可能である。しかし，実用化を目的とする場合，現状では計算コストなどを考慮し，声道断面積モデルを用いる後者の方が有望であると考えている。よって，本モデルでは将来的にそのようなシステムへ拡張することを念頭に置き，発話器官の位置と声道断面積モデルのパラメータとの変換が容易になるように，声道中の基準となる位置とパラメータとの関連ができるだけ失われないことを考慮してモデルを作成した。

本論文では、ここまでについて論じる。今後は，最後に述べたような声道形状制御と F0 調節の相互作用を考慮し双方を同時に制御する新たな制御モデル（図中の次期制御モデル）を構築し，また，さらに任意の音声を合成するために必要な音素を含む単語の分析結果を追加して行くことで，自然な音声を合成できるコンパクトなテキスト音声合成システムの開発が可能になると考えている。

1.2　本論文の構成

本論文の構成を以下に示す。

第 2 章では，種々の音声合成方式について述べ，今後の音声合成方式として人間の音声生成メカニズムに基づく音声合成方式が有望であることを示す。また，一般的な音声生成メカニズムに基づく音声の合成方法についても記載する。

第 3 章では，人間の音声生成メカニズムに基づく音声合成方式を構築する上で最も重要な手段となる声道形状の計測方法について記載する。また，第 6 章にて用いる最新の 3 次元 MRI 動画撮像法の概要についても紹介する。

第 4 章では，MRI を用いて行った F0 調節の生理機構の分析について述べる。こ

の中で，従来明らかにされていなかった新たな F0 下降のメカニズムを提案する。

第 5 章では，発話器官の生理学的な 2 次元計算モデルについて述べる。従来の音声合成モデルではソースとフィルタは独立した要素としてモデル化されてきた。しかし、第 4 章の結果を考慮すると，自然な音声を合成するには相互作用を考慮する必要があると考えられる。よってソースとフィルタとの相互作用を実現する発話器官モデルを提案する。このモデルを用いた音声の合成実験により，第 4 章の F0 下降機構によって F0 調節と声道形状制御との相互作用が生じること，および，このモデルでその相互作用が表現できることを明らかにする。

第 6 章では，本方式の実用化へのボトルネックである音質の問題を解消するため、声道形状モデルのパラメータを 3 次元 MRI 動画データを用いて推定することで高精度化し，高品質な音声の合成を可能にする音声分析合成方式を提案する。本方式を用いた音声合成実験より人間の音声生成メカニズムに基づく音声合成方式を用いても実用可能な品質の音声の合成が可能であることを明らかにする。

第 7 章では，まとめとして，本研究の要約と今後の課題を示す。

第 2章

種々の音声合成方式

本章では，テキスト音声合成 (Text-To-Speech system) に用いられる種々の音声合成方式について，その概要および特徴ついて述べる。そして，これからの音声合成方式として声道モデルが有望な合成方式の１つであることを示す。また，一般的な声道モデルを用いた音声の合成方法についても記述する。

2.1　テキスト音声合成

はじめに，現在実用化されている多くのテキスト音声合成において用いられているコーパスベースのテキスト音声合成の処理について説明する。図 2.1 に構成を示す。

入力されたテキストは言語処理部にて，単語，文節の区切り解析，係り受け解析，アクセント付けなどを行い，どのように読み上げるかを指定する読み記号列に変換される。音声合成規則部では，どの音素片をどのような韻律で合成するかを決定する。韻律生成部では指定された読み記号列の情報に従い読み上げる音声の物理量である F0，パワー，音素の継続時間長の推定が行われる。音声コーパスには，各音素片毎に F0，パワー，時間長などの韻律情報と音声波形もしくは音声を符号化したパラメータ列が記憶されている。素片選択部では，音声コーパスに含まれる音素片の中から，読み記号列で指定される音韻系列に適合し，韻律生成部にて推定された F0, パワー，音素の継続時間長に最も近く，また，接続したときの歪みが最も少なくなる素片の組み合わせが選択される。音声合成部では，選

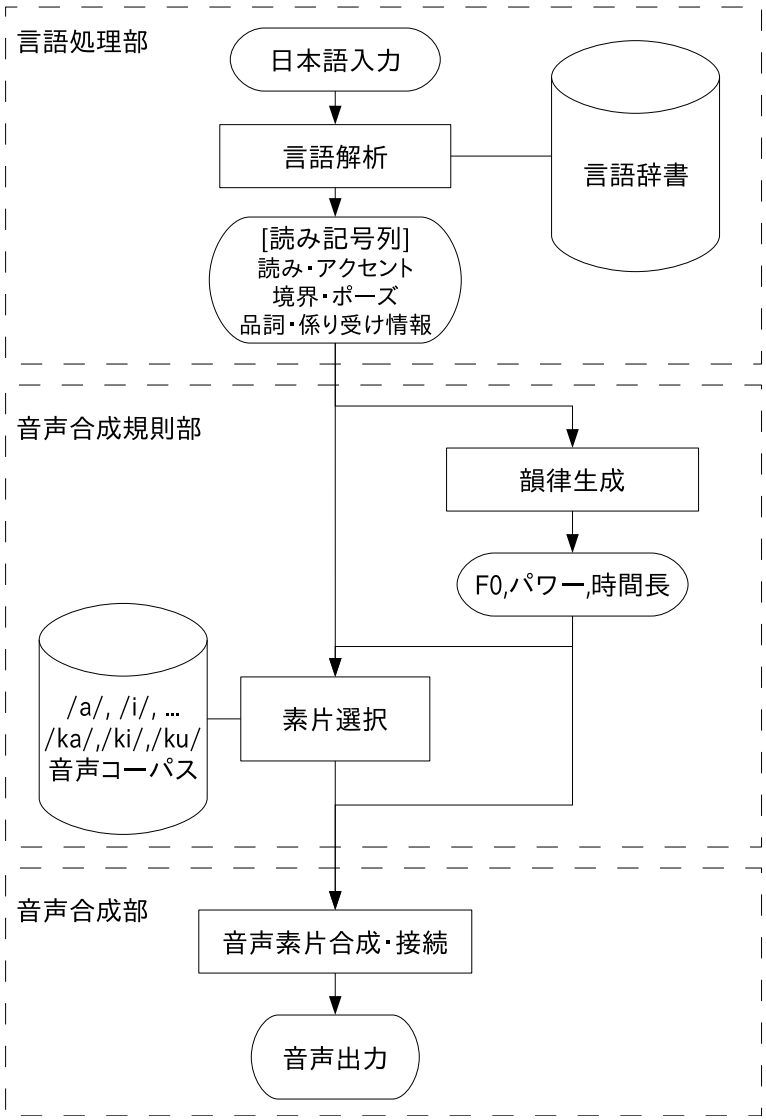


図 2.1: テキスト音声合成システムの構成

択された音素片の波形もしくは符号化されたパラメータ列を用いて，推定されたF0，パワー，音素の継続時間長と等しくなるように音声波形を合成し接続することで音声が生成される。音声合成部に用いられる音声合成方式は，音声コーパスに蓄積されるパラメータの形式により幾つかの方式が提案されている。次節にて代表的な幾つかの方式について簡単に説明を行う。

2.2 音声合成方式

2.2.1 波形編集合成方式

比較的容易に高品質の合成音声を得ることができることから，近年，商用を目的としたテキスト音声合成において一般的に広く使用されている方式である。実音声より切り出した音素片を波形データとして蓄積しておき，時間領域で韻律の変形を行い，接続することで音声の合成が行われる。音素片を大量に持つことで合成時に韻律を全く変更しない方式 [2] と，PSOLA(Pitch Synchronous Overlap Add)[20] などを用いて韻律生成部で推定された韻律に変更する方式 [1] がある。ただし、どちらの方式を用いてもスペクトルの変形を行うことはできない。自然な音声の発話では，各音素のスペクトルはその前後の調音動作の影響を受け変化する。これは調音結合と呼ばれる現象である。また，F0 やアクセントの有無，文中の位置などの発話環境の違いによってもスペクトルは異なる。よって，高品質な音声を合成するには，各音素毎に様々な組み合わせの調音結合の影響を含み，かつ多くの種類の発話環境を含むコーパスを構築する必要がある。そのため，近年，高品質化を目指してコーパスの大規模化が進んでおり，XIMERA[21] では数十時間以上のコーパスが使用されている。

この方式の特徴は，実音声波形を合成に使用することから合成音声の品質が非常に高いことが挙げられる。コーパスのサイズを大きくし，韻律の変更を行わないことで実音声と同等の品質の合成音声を得ることも可能である。欠点は，コーパスのサイズが大きいことである。そのため，携帯電話やカーナビなどのように容量に制限のある機器への組み込みでは品質が大きく劣化する。また，感情音声や個人性などの異なるスピーチスタイルを追加するためには，異なるスピーチス

タイルで発話したコーパスを全て再収録する必要があるという問題もある。

2.2.2 スペクトルパラメータによる分析合成方式

VOCODER 方式とも呼ばれるこの方式は，線形に分離されたスペクトル包絡の形状と音源情報をパラメータとして持ち，音源情報から生成された音源波形にスペクトル包絡をフィルタリングすることで音声を合成する方式である。スペクトルパラメータには主に，PARCOR[22]，LMA[23]，LSP [24] などが使用される。これらの方式は，波形編集方式で広く用いられている PSOLA と比較し，F0，パワー，時間長の変形だけでなく，スペクトルの変形も可能であるという特徴がある。そのため，スペクトルの形状の異なる素片を滑らかに接続すること，時間方向の間引き補間を行うことでデータを圧縮することなどが可能な方式である。各合成単位毎に１つの音声素片データを持ち、それらを変形しつなぎ合わせることで任意の音声を合成するという古い世代のテキスト音声合成方式では広く用いられていた方式である。しかし，スペクトルの変形により生成される音声が物理的に存在する音声となる保証はなく，大幅な変形に対しては自然性が大きく劣化する。この方式の利点は，コーパスのサイズを小さくできるという点である。そのため，現在でも組込み用に用いられることが多い。しかし，波形編集方式と比較し音質が悪いという欠点がある。

近年，河原らによって開発された STRAIGHT[25] もこの方式の１つである。精度の高い基本周波数の推定，および，基本周波数に適応する相補的な時間窓と周波数領域での平滑化により，音源の周期性の影響をスペクトルから除去することを可能にした方式である。その結果，VOCODER 方式であるにもかかわらず，高品質な音声の合成を可能にしている。しかし，他の手法では１フレームあたりのスペクトルの形状は数十次のデータで表現されるが，この方式では，周波数スペクトルの値をパラメータとするため数百次のデータとなり，データ量が大きいという欠点がある。

2.2.3　人間の音声生成メカニズムに基づく音声合成方式

本論文で研究の対象とする方式である。人間の音声生成のメカニズムを真似て音声の合成を行う方式である。声道の物理的な計算機モデルを作成し音波の伝達特性を計算することで音声の合成が行われる。声道断面積を直接モデル化する方式 [26] から発話器官の構造的なモデル [5],[27],[28],[29] まで古くから多くの研究が行われている。

この方式では，パラメータに声帯や声道の形状や特性を表す物理的意味を持つ変数が用いられるため，以下のような特徴を持つことが知られている。

1. パラメータの変形に対してロバストである。
   前節で示したスペクトルパラメータを用いてもパラメータの変形は可能である。しかし，２組のパラメータ (たとえば/a/と/i/) の中間の音声を補間により求めた場合，スペクトルパラメータの場合は人間の発声できる音声となる保証はない。それに対しこの方式では幾何学的な声道形状の中間値を求めることで物理的に意味のある中間の音 (たとえば中性母音) を生成することが可能である。
2. パラメータの時間変化がゆっくりである。
   音声のスペクトルの変化に対し発話器官の位置変化は遅いことが知られている。
3. 人間の音声生成メカニズムを分析することで得られた発話器官の物理的・生理的特性を音声生成モデルに組み込むことで，それらの特性を合成音声に容易に反映することができる。

1.,2. の特徴により，スペクトルパラメータを用いた分析合成方式よりもパラメータの圧縮率を高くすることが可能であることがわかる。また，波形編集合成方式では，前後の音素や F0 などの発話環境が異なると音声のスペクトルが異なるため，自然性の高い音声を合成するには全て発話環境で発話された音声をデータベースに蓄積する必要があった。しかし，3. の特徴により，発話環境の違いによりスペクトルの変化が生じる要因を解明しモデルに実装することで，蓄積されているデータから計算により作り出すことが可能であることがわかる。このことにより，自然性を維持しながらデータベースに含まれる音素の数を大幅に削減することが可能となり，LSI などへの組み込み用の音声合成の品質を向上させることができると考

えられる。また，兄弟が音声が良く似ていることからもわかるように音声に含まれる個人性は発話器官の物理的形状に起因するところが大きいと考えられる。そして，感情音声も人間の生理から自然に生成されるものであり，発話器官の変化から生じていることが予想される。よって，これらについても人間の音声生成メカニズムの解析を進めることで個人性や感情などの制御が可能となり，多様な音声を合成できる音声合成を開発できる可能性があると考えている。

しかし，このように，多くの特徴を有しているにもかかわらずこの方式による音声合成が実用化された例はほとんどない。その要因は，実際の人間の音声生成機構は非常に複雑であり未だ明らかにされていない機構が存在すること，そのために過度な単純化・モデル化が進められてきたこと，またモデルの構築に用いる声道形状データの量と質を十分に確保することが困難であったことなどが考えられる。これらは，音声から声道形状への変換が 1 対多になるため，発話器官の動作を観測することが容易ではなかったことにより引き起こされたのではないかと考えている。また，音声の合成を VOCODER 方式で行うためスペクトルパラメータによる分析合成と同様に音質が劣化したことも要因の 1 つであろう。近年，観測の問題に対して，発話動作の計測に適した MRI を用いた新しい撮像法の開発が進み，様々な条件で発声・発話動作の撮像が可能となってきている。また，音質の問題に対しては，VOCODER 方式にも関わらず高品質な音声が可能な STRAIGHT が提案されている。よって，これら最新の技術を用い，これまで抱えてきた問題を再考したうえで音声生成モデルを高精度化することにより，人間のメカニズムに基づく音声合成方式を実用化することが可能になると考えている。

2.2.4 HMM 音声合成 [30][31]

HMM 音声合成はテキスト音声合成方式の 1 手法である。音声データのスペクトルおよび韻律情報を HMM を用いてモデル化し，音声合成時には与えられた HMM の状態列に対し出力確率が最大となるスペクトルパラメータおよび韻律のパラメータを求めることにより音声を合成する手法である。大量の学習データから得られた HMM に対し決定木に基づくコンテキストクラスタリングを行い，合成時にはその中から最適な状態を選択することで自然性の高い音声の合成を行うことがで

きる。また，合成に必要なパラメータを全て HMM でモデル化しているため音声認識の話者適応の技術を応用することにより，話者変換，感情音声合成，発話スタイル制御など，多様な音声の合成も可能なシステムである。しかし，合成のパラメータとして，メルケプストラムなどスペクトルパラメータによる分析合成方式が用いられるため音質が波形編集方式より劣るという欠点がある。人間の音声生成メカニズムに基づく音声合成方式をパラメータとして用いる HMM 音声合成については，将来，取り組みたいテーマの１つである。

2.3 音声生成メカニズムに基づく音声合成方式による音声生成方法

2.3.1 音声の生成機構

はじめに，人間の音声の生成機構について示す。音声の生成は，人間の発話器官の協調動作によって行われる。図 2.2 は人間の発話器官の断面図である。肺から押し出された空気は，声門もしくは気管の極端な狭め等で音源となる空気振動に変換される。声門では，有声帯振動波形 (声帯音源) と呼ばれる母音などの有声音の音源が生成される。有声音を発声する際は，無発声時に大きく左右に開いている声帯を接近させる。その結果，空気が高速で声帯の間を通過することになり，声帯を閉じようとする空気力学的な力が発生する。声帯が閉じられると空気の流れが無くなり閉じようとする力も消滅し，声帯が再度開かれる。この動作が繰り替えされることで声帯が振動し，周期的な空気の断続が生じて有声音が生成される。声門以外で生成される音源には 2 通りの音源がある。1 つは，声道に生じる狭めを一時的に完全に閉鎖し，高まった気圧を一度に開放することで生成される急激な空気流による無声音源である。破裂音の音源となる。もう 1 つは，極端な狭めを空気が通り抜ける際に生成される乱流による無声音源である。摩擦音が生成される。生成された音源波が，喉頭，咽頭，口腔，鼻腔，口唇により形成される管状の空間 (声道) を伝搬し、口唇および鼻腔から体外へ放射されることにより音声の生成が行われる。声道の形状は、顎，舌，口唇などによる調音運動により変化する。音波は管を伝搬する際に共振・反共振が生じるが、声道の形状に応じて共振・

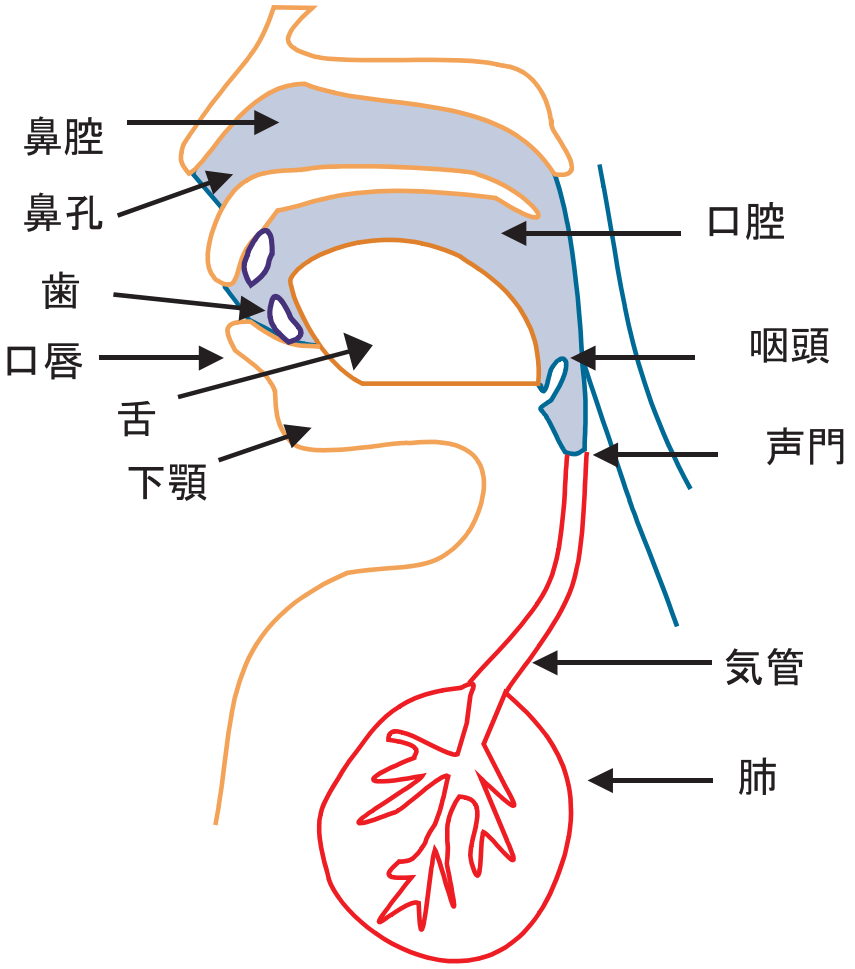


図 2.2: 人間の発話器官

反共振の生じる周波数が異なる。そのため調音運動により種々の音声を発声することが可能となる。また，軟口蓋は上下させることで，鼻腔との境界を閉じたり開いたりすることができる。開いた場合，音波は鼻腔と口腔の両方へ伝搬することになり，鼻音が生成される。

2.3.2 音声生成モデル

計算機により音声生成機構をモデル化し音声の生成を行う代表的な方法について説明する。音声生成機構はソース (音源) とフィルタ (声道形状による伝達特性) の結合としてモデル化することができる。ソース部では，声の高さ，大きさ，発声の種類 (通常の発声やささやき声など) に応じた音源波形が生成される。フィルタ部では，声道の形状に応じた音源から放射端までの声道の周波数伝達関数が計算される。ソース部で生成された音源波形を，フィルタ部の周波数伝達関数に通すことで音声波形を生成することができる。また，フィルタ部の伝達関数を計算するために，声道を音響管による音響フィルタと放射モデルの 2 つの要素で表現することで，音声生成モデルは，音源・音響フィルタ・放射モデルの 3 つの要素により構成される線形システムとしてを表すことができる。ここで，音響フィルタを伝送特性モデルで表すと，代表的な発話様式は図 2.3 のようにモデル化できる。図中の，Kt は声門から口唇までの伝送特性，Kb は狭めの音源から声門側の伝送特性，Kf は狭めの音源から口唇までの伝送特性，Kl は声門から鼻腔の接続位置までの伝送特性，Ko は鼻腔の接続位置から口唇までの伝送特性，Kn は鼻腔の接続位置から鼻孔までの伝送特性を示す。Ug は声帯音源，Pf は狭めに挿入される無声音源を示す。Zl は口唇から放射される時の放射インピーダンス，Zn は鼻孔から放射される時の放射インピーダンスを示す。図 2.3 に示すモデルを用いることで各発話様式に於ける音源から放射端までのフィルタの伝達関数を計算することができる。

2.3.3 音源

一般に声帯音源には，Rosenberg 波 [32]，LF モデル [33] などの音源波形の近似モデルや，石坂・フラナガンの 2 質量モデル [34] に代表される声帯振動の物理モデルが用いられる。

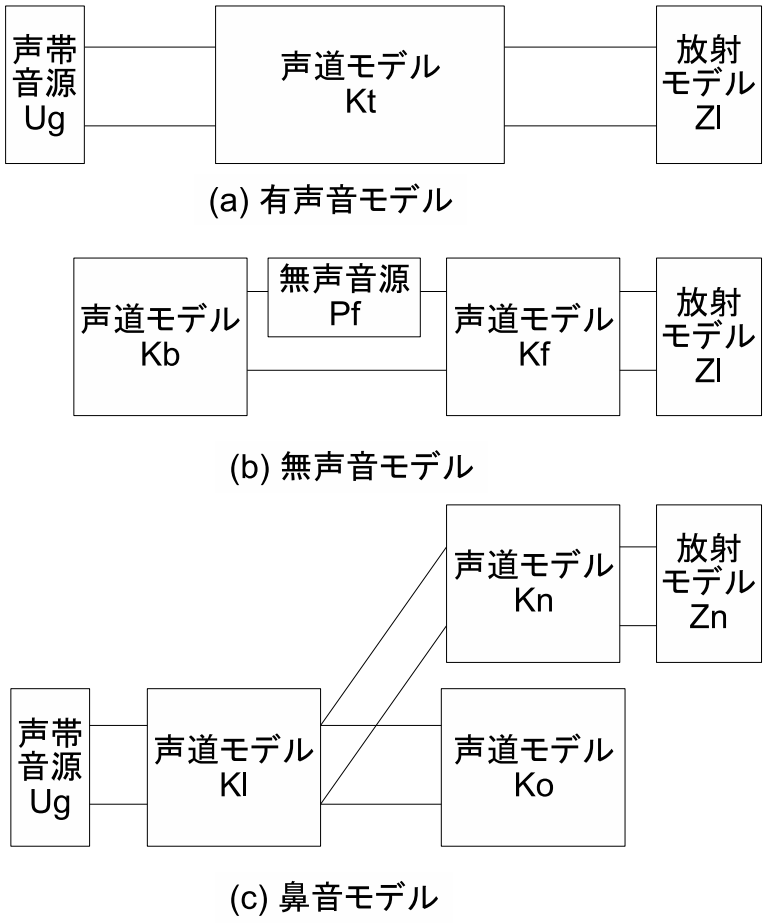


図 2.3: 音声生成のモデル

たとえば，Resenberg 波は，次式で示されるモデルである。

$$\begin{aligned} f(t) &= a\left\{3\left(\frac{t}{\tau_1}\right)^2 - 2\left(\frac{t}{\tau_1}\right)^3\right\},\ 0 \le t \le \tau_1 \\ &= a\left\{1 - \left(\frac{t-\tau_1}{\tau_2}\right)^2\right\},\ \tau_1 < t \le \tau_1 + \tau_2 \end{aligned} \tag{2.1}$$

τ_1 が 1 周期の 40%, τ_2 が 16%の時に一般に好まれる音質になることが知られている。LF モデルは Rosenberg 波よりパラメータが多く，自由度の高いモデルをである。このような自由度の高いモデルは，多様な声質の音声を合成する試みに用いられることが多い。

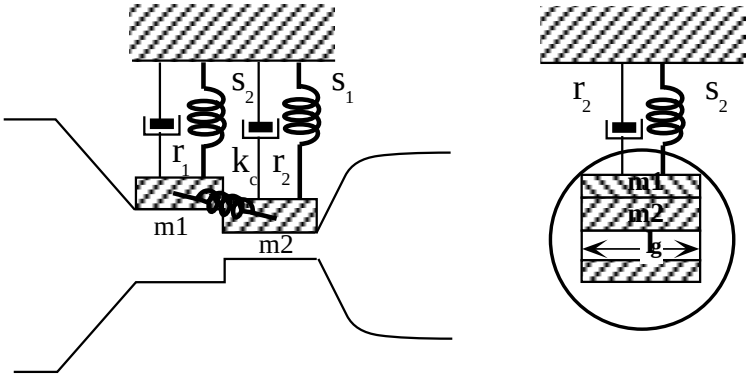


図 2.4: 2 質量モデル

2 質量モデルは，図 2.4 に示すように声帯を上下 2 つの部分に分け，それぞれをバネとダンパーで接続した機械振動系のモデルである。声門下圧と声門間隙に生じるベルヌイ圧によって振動が起こる自励振動モデルである。2 質量モデルは，上下の部分の開閉の位相差の再現や，声道と音源波形との相互作用など，実際に近い声帯振動を再現することができ，自然な音声を生成できる優れたモデルである。しかし，完全な自励振動モデルであるため，制御が困難という問題がある。

無声音源については，狭めの位置に音圧源として挿入される単純な白色雑音もしくはスペクトルに傾斜をもつ有色雑音が用いることが多い。

2.3.4 音響管による音響フィルタ

図 2.3 中の声道モデルと記載されている部分は，声道形状から計算される伝送特性を持つフィルタとして実現される。声道形状から伝送特性への変換は，声道の

半径が大きくないことから，周波数の低い領域 (4〜5kHz 以下) では声道を 1 次元音響管で近似し計算する。一般には，短い円筒管の連続体で近似し，計算されることが多い。本論文では，第 5 章において円筒管を用いた伝送特性モデルの計算と，第 6 章においてそれを拡張した円錐台を用いた伝送特性モデルの計算の 2 通りを用いている [35][36]。ここでは，円錐台を用いた場合の計算式を示す。

声道のある地点 x_1 から x_2 までを 1 つの円錐台で近似した時の地点 x_1, x_2 それぞれの音圧 p と粒子速度 u の関係を示す伝達マトリクスは以下のようになる。

$$\begin{bmatrix} p_2 \\ u_2 \end{bmatrix} = L^{-1}(x_2) M L(x_1) \begin{bmatrix} p_1 \\ u_1 \end{bmatrix} \tag{2.2}$$

ここで

$$L(x) = \begin{bmatrix} x & 0 \\ -\frac{1}{Z_v} & x \end{bmatrix}, \tag{2.3}$$

$$M = \begin{bmatrix} \cosh\Gamma(x_1 - x_2) & \zeta \sinh\Gamma(x_1 - x_2) \\ \frac{1}{\zeta}\sinh\Gamma(x_1 - x_2) & \cosh\Gamma(x_1 - x_2) \end{bmatrix} \tag{2.4}$$

また，伝達定数 Γ および特性インピーダンス ζ は，粘性により生じる単位長さあたりの直列インピーダンス Z_v，熱交換による単位長さ当たりの並列アドミッタンス Y_t，壁振動による単位長さ当たりの並列アドミッタンス Y_w を用いて，

$$\Gamma = \sqrt{Z_v(Y_t + Y_w)}, \tag{2.5}$$

$$\zeta = \sqrt{\frac{Y_t + Y_w}{Z_v}} \tag{2.6}$$

となる。また，Z_v, Y_t, Y_w は以下のようにモデル化することができる。

$$Z_v = i\omega\rho\left(1 + \frac{2}{r_w}(1-i) - \frac{3i}{r_v^2}\right) \tag{2.7}$$

$$Y_t = \frac{i\omega\rho}{\rho c^2}\left[1 + (\gamma - 1)\left(\frac{\sqrt{2}}{r_t}(1-i) + \frac{i}{r_t^2}\right)\right] \tag{2.8}$$

$$Y_w = \frac{i\omega}{\rho c^2}\frac{\omega_0^2}{b + i\omega a - \omega^2} \tag{2.9}$$

ここで，ω は角周波数 [rad/s]，r_v はパイプ半径 r と粘性的境界層厚との比 $r_v = r\sqrt{\omega\rho/\mu}$，$r_t$ はパイプ半径 r と温度的境界層厚との比 $r_t = r\sqrt{\omega\rho C_p/\lambda}$ を示す。

各定数については，以下の値を用いた [35]。

音速，　$c = 331.45\,(1 + 0.0018T)$　m/s.

空気密度，　$\rho = 1.2929(1 - 0.0037T)$　kg/m^3.

粘性定数，　$\mu = 1.708 \times 10^{-5}\,(1 + 0.0029T)$　$kg/(s \cdot m)$.

熱伝導率，　$d = 5.77 \times 10^{-3}\,(1 + 0.0033T)$　$cal/(m \cdot s \cdot ℃)$.

定圧比熱，　$C_p = 240cal/kg \cdot ℃$.

特性比熱比，　$\gamma = \dfrac{C_p}{C_v} = 1.402$.

声道の両端を閉じた時の最低角共鳴周波数，　$w_0 = 406/\pi\ rad/s$.

声道壁の減衰係数と質量との比，　$a = 130\pi\ rad/s$.

機械的共鳴角周波数の自乗，　$b = (30\pi)^2\ (rad/s)^2$

音源から放射モデルまでの声道形状を円錐台の連続体に分割し，全ての円錐台の伝達マトリクスの積を求めることで，音源から放射モデルまでの伝達マトリクスを計算することができる。

2.3.5　放射モデル

出力端での音響放射は、無限大バッフルに囲まれた円形ピストンからの放射としてモデル化することができる。放射インピーダンスは，出力端の開口面積を A_{out} とすると，以下の近似式で表現できる [37]。

$$Zr \equiv \frac{P_{out}}{U_{out}} = \frac{\zeta}{A_{out}} \frac{i\omega R_r L_r}{R_r + i\omega L_r}, \tag{2.10}$$

ここで，

$$R_r = \frac{128}{9\pi^2}, L_r = \frac{8\sqrt{A_{out}}}{3\pi^{3/2}c} \tag{2.11}$$

である。

第 3章

MRIによる声道形状の計測

人間の音声生成メカニズムに基づく音声合成方式の研究を行う上で大きな問題の1つは，如何にして発話時の声道を計測するかということである。古くから様々な方法により声道形状の計測は行われてきた。しかし，簡便に十分な精度で声道形状を計測することは困難であった。それに対し，近年，核磁気共鳴画像 (MRI) を用いた様々な試みがなされている。本研究においても声道形状の撮像法として主に MRI を用いている。本章では，本研究に用いた MRI による声道形状の撮像法について説明する。

3.1 種々の声道形状計測方法

現在，研究用途として使用されている方法としては，X 線シネ画像 [38]，X 線マイクロビーム [39][40]，磁気センサシステム EMMA(Electro-magnetic Midsagittal Articulograph)[41]，MRI などがあげられる。X 線マイクロビームとは，測定ターゲットとなる位置に鉛玉を着け，その鉛玉の位置を X 線を用いて追尾することで口腔内の位置を計測する装置である。EMMA は測定点に微小の受信コイルを装着し，複数の送信コイルにより交流磁場を発生させ，受信コイルに発生した信号から受信コイルの位置を計算するものである。これらの測定方法は，1) 高速な測定，高いサンプリングでの計測，2) 静かな環境で音声と同時に測定可能，などの特徴を有するが，計測用の鉛玉，受信コイルを装着した点の座標しか測定できないという問題がある。また，測定用素子の装着が困難な喉頭などは計測できない。

それに対し，X 線シネ画像や MRI は声道を画像として計測する方法である。その中で MRI は，放射線の問題もなく，声道の 3 次元の形状をボリュームデータとして測定することができるため，最も有用な手段の１つと言える。本論文では主に MRI を用いて声道形状の撮像を行った。

3.2 核磁気共鳴画像法 (MRI)

MRI とは，体内に含まれる水素の核磁気共鳴現象を利用して内部の情報を画像化する手法である。強い静磁場に被測定物を挿入すると被測定物内の原子核スピンの向きが揃う。その状態に特定の周波数のラジオ波パルスを照射することで，原子核スピンが傾く。その後パルス照射を停止すると電波を出しながら定常状態に戻る。そのときの戻り方の違いから組織の違いを画像化する手法である。

特徴としては,

1. 放射線被曝がない。
2. 測定に用いるパラメータを変更することで，対象とする組織の種類や，SNR，測定時間などを調整することが可能である。実験の目的に合わせて，これらを調整することで，筋肉や，軟骨組織などの調音器官を測定することができる。
3. 任意の位置の任意の方向の断面の計測，および 3 次元画像の撮像も可能である。

などがあげられる。

しかし,

1. 歯列と空気はほぼ同じ程度の輝度となるため分けることができない。
2. 動きに弱く，基本的に 1 セットのデータを撮像するために時間がかかるため動画の撮像が困難。
3. 測定時に大きなノイズが発生する。

などの欠点がある。

第 4 章の計測実験では，測定に用いるパラメータを調整することで，対象とする部位 (軟骨) の識別が可能な範囲で画像の鮮明度を犠牲にし４秒/スライスで撮像できる高速撮像法を用いた。撮像に用いたパラメータは第 4 章に記載する。

第 6 章では，上記問題 1,2 に対して近年開発された計測手法 [42] を用いて発話動作時の声道形状の抽出を行った。以下その手法について簡単に紹介する。

3.3 歯列の補填

声道形状の計測では，空気の流れる領域を抽出する必要があるが，MRI では空気と歯列がほとんど同じ輝度として計測されるため境界の特定が困難である。そのため，他の方法にて撮像した歯列をはめ込む必要がある。本研究の開始当初 (第 4 章の声道形状計測実験) は，閉口状態で舌を歯列に押し当てて撮像した MRI 画像から得られた歯列のデータを，わずかに撮像される歯根の形状を参考に，目視にて位置を調整し，計測された MRI データに重ね合わせていた。その後，歯形を用いる方法 [43][44]，歯冠プレート [45] を用いる方法などが提案された。第 6 章の研究では，竹本らの歯列補填法 [46] を用いた。この方法では，ジェル状のジュースを口腔に含んだ状態で MRI 撮像を行い歯列をあらかじめ撮像しておき，歯根や顎などの部分の情報を用いて，発話時に撮像した MRI データの歯列の位置に合致するように回転および移動量を計算し，スーパインポーズすることで歯列の情報が追加される。

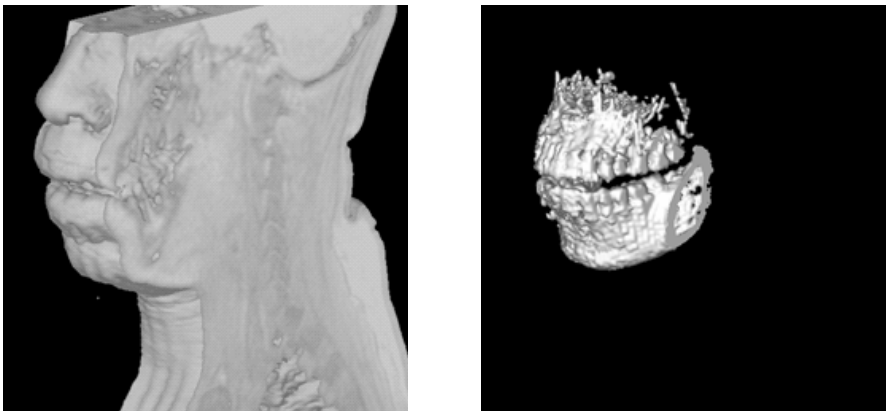
以下に手順を示す。

1. ジェル状ジュースを含んだ状態の MRI の 3 次元画像から上顎骨および下顎骨を含めて上下の歯列を別々に抽出する。
2. 上下の歯列の 3 次元画像それぞれについて，上顎骨，下顎骨に付着する組織の特徴的な部位 (輝度が回りと比べて特に明るいもしくは暗い点) を基準点として 3 つ以上選択する。
3. 歯列を補充したい 3 次元画像について，2 で設定した基準点に対応すると思われる点をそれぞれ選択する。
4. 対応する全ての基準点間の距離の総和が最も近くなるように，上下の歯列をそれぞれ回転量，移動量を計算する。
5. 4) で求められた回転量，移動量を初期値として，上下の歯列のデータとスーパインポーズされる 3 次元画像の，上顎骨および下顎骨また付着する組織の

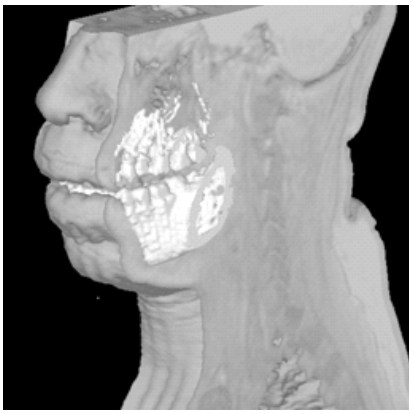
形状が一致するようにシミュレーテッドアニーリングを用いて上下それぞれの歯列のデータの回転量，移動量の最適化を行う。

6. 計算された位置に歯列のデータをスーパインポーズする。

補填した例を図 3.1 に示す。



(a) 補填前の画像　(b) 補填する上・下顎



(c) 補填後の画像

図 3.1: 歯列補充例 (竹本らのスライドより抜粋)

3.4　３次元 MRI 動画

MRI は撮像対象の内部構造や立体的な形態解析などを行うことのできる優れた方式であるが，動態の観察には不向きな撮像法と言える。本研究開始当初は，動画撮像が可能な方法がなかったため，観測ターゲットとなる動作を少しづつ変化

させて撮像した複数の MRI 静止画を連続観察することで発話動作の解析を行っていた。第 4 章のデータはこの手法により解析したものである。撮像時間は画像の明瞭度とのトレードオフとなる。喉頭の撮像では正中矢状断面のみの撮像で 1 画像あたり約 4 秒の持続発声が必要であった。この手法は特殊な装置も必要とせず容易に撮像できる方法であるが，明瞭度を犠牲にしても 1 画像の撮像に数秒程度かかるため，音素間の渡りの部分などのように持続発声で再現できない運動は撮像できなかった。

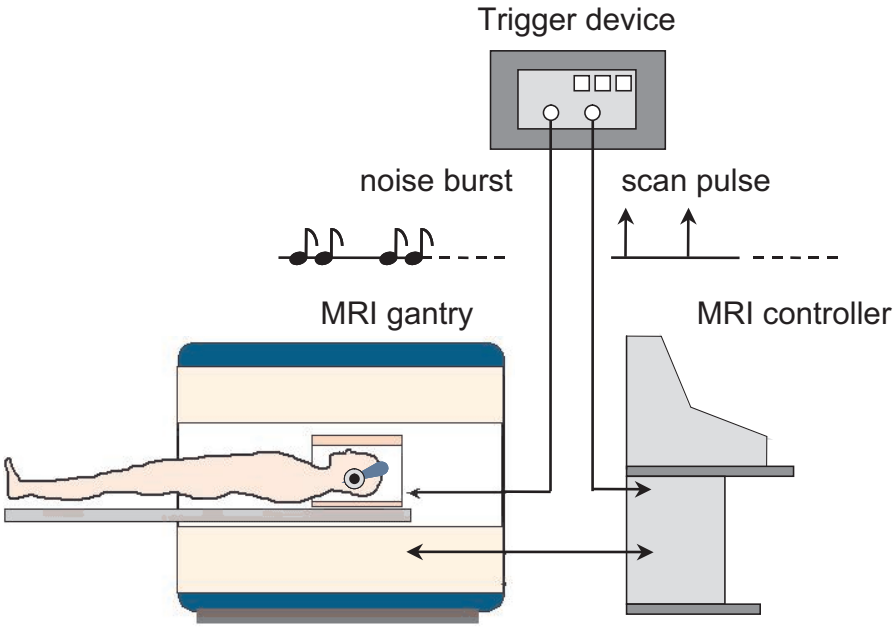


図 3.2: 3 次元 MRI 動画撮像法 [42]

その後，正木らによって 2 次元 MRI 動画 [47]，竹本らによって 3 次元 MRI 動画 [48][42] が開発されることにより，発話動作時の動画の撮像が可能となった。これらは，心臓の動作を撮像するためのシネ撮像を応用した方法で，同じ発話タスクをトリガーに同期し複数回繰り返し発話することで動画を撮像するものである。通常の静止画像の撮像では，256 × 256 画素の画像を計測するために少なくとも 128 回のパルスの照射が行われる。また，3 次元画像を得るためには複数 (N) プレーンを撮像するため N 倍のデータ計測が必要となる。よって 3 次元 MRI 動画撮像法では，図 3.2 に示すように，パルス照射の開始基準となるトリガを外部より MRI 装置に入力し，そのトリガに同期したトーンバーストを同時に被験者に提示する。

被験者がトーンバーストに同期して発話動作を繰り返すことで，複数回の発話動作から各時間に対応する 3 次元画像構築に必要なデータを計測することができる。

現在は 256 回の反復発話を 1 セッションとして，セッション間に休憩をはさみ 3 セッションもしくは 4 セッションの撮像で，毎秒 30 フレームの時間分解能で声道形状を撮像することができる。

3.5 声道断面積関数の抽出

3 次元 MRI データから，声道断面積関数を抽出する方法について述べる。声道断面積の抽出で問題となるのは，平面波の伝搬する道筋を示す中心線の決定方法である。中心線の引き方により，声道断面積も声道長も変化する。本研究では，竹本らの方法 [42] を用いた。この手法では声道形状は正中矢状断面を中心し左右対称と仮定し，正中矢状断面の画像を用いて中心線の決定を行う。以下に手順を示す。

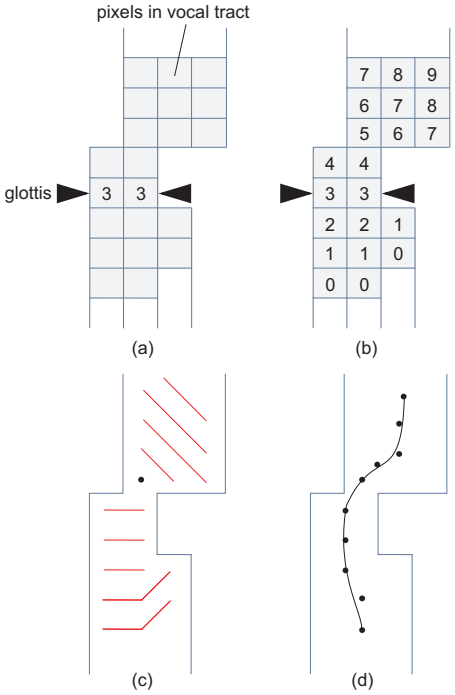
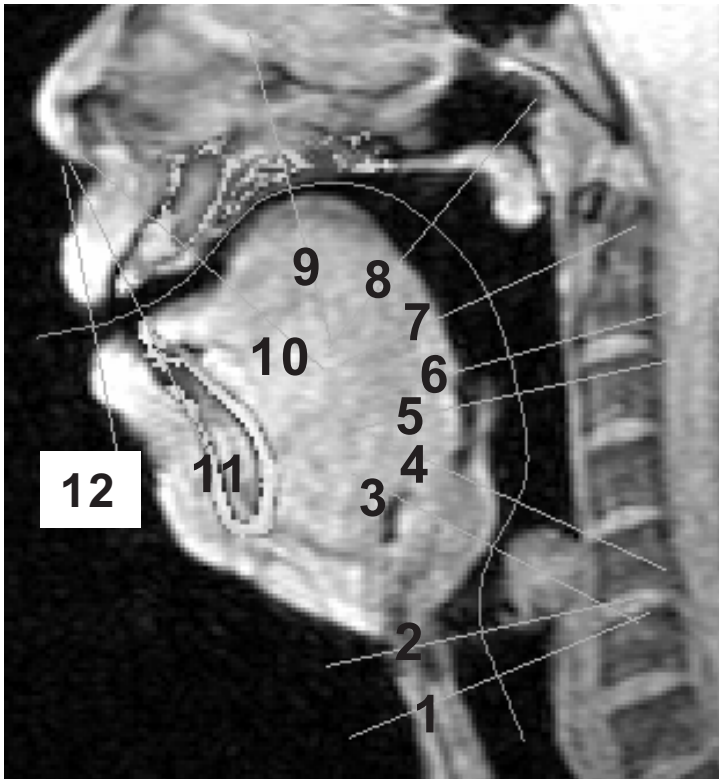


図 3.3: 声道中心線抽出 [42]

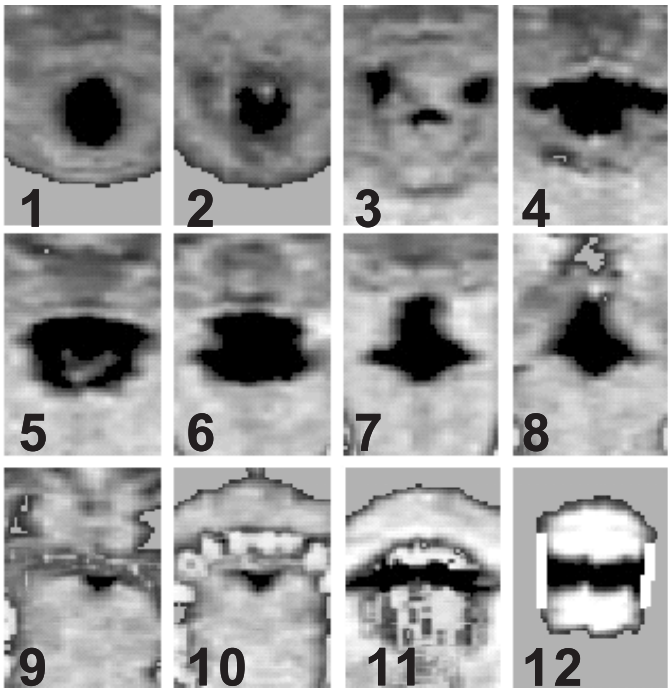
1. 正中矢状断面画像より声帯を目視で抽出し，声帯と重なる線を設定する。

2. 図 3.3 (a) に示すように声帯に沿った線上の全ての画素に任意の同じ数字を割り当てる。図 3.3 では 3 を用いた。
3. 図 3.3 (b) に示すように，現在数字の埋められた画素の上下左右の画素に数字を割り当てる。声帯から口唇側は 1 加算して数字を割り当てる。肺側へは 1 減算した数値を割り当てる。
4. 同じ数値を示す等高線を求める。図 3.3 (c) に示す。
5. 図 3.3 (d) に示すように，等高線の中点を求め，スプライン補間により声道の中心線を求める。

声道断面積関数は，声道の中心線に沿って等間隔の位置で，中心線に垂直な平面の声道断面を 3 次元 MRI データより求め，その面積を計算することで得ることができる。図 3.4 に母音/i/を発声したときの中心線および声道断面の抽出結果の例を示す。また，図 3.5 に 5 母音「あいうえお」を発話した時の 3 次元 MRI 動画から抽出した声道断面積の測定例を示す。



(a)正中矢状断面



(b)声道断面

図 3.4: 声道中心線および声道断面の抽出例

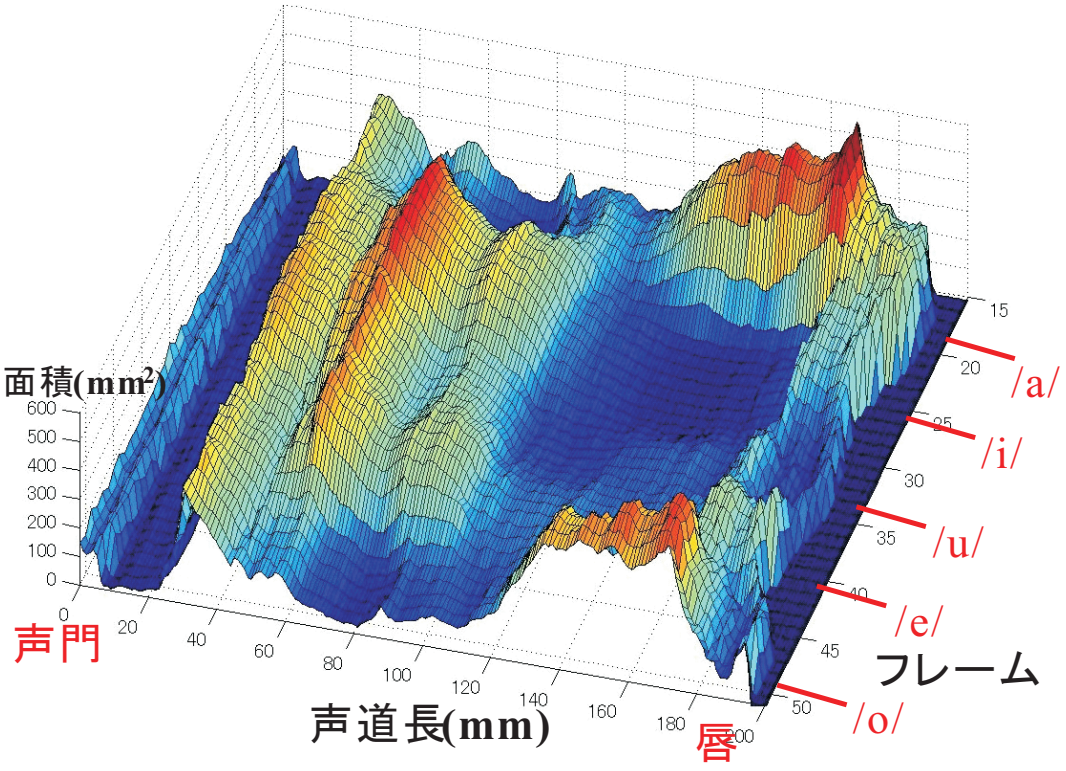


図 3.5: 声道断面積の抽出結果

第 4章

ソース(音源)におけるF0調節機構の分析

音声生成メカニズムに基づく音声合成方式により自然性の高い音声の合成を行うには，音声生成の生理機構を解明することが重要な課題の１つである。特にF0調節に伴う音声の周波数特性の変化をもたらす生理機構の解明は，テキスト音声合成システムのデータベースに含まれる冗長な音素の削減につながるため，自然な音声を合成できるコンパクトな音声合成システムの開発に大きな役割を果たすと考えられる。よって，本章ではF0調節機構について分析を行う。

F0を上昇させる仕組み(F0上昇機構)は，直接関与する喉頭筋の存在が広く知られているが，F0を下降させる仕組み(F0下降機構)については十分に理解が得られているとは言いにくい。これまでF0下降は，F0を上昇させる喉頭筋の弛緩により説明されてきたが，F0を下降させる時には喉頭の位置が下降することや，F0の下降時に舌骨下筋の活動が生じることなどが知られており，喉頭筋の弛緩以外にもF0を調節する機構が存在すると思われる。

本章では，F0を変化させて持続発声した時の喉頭の正中矢状断面をMRIを用いて撮像し，喉頭軟膏や周囲構造の相対変化を調べた。その結果，問題となるF0下降を説明できる生理機構を含め，幾つかのF0調節に関する喉頭機能を確認したので報告する。

4.1 喉頭筋による F0 調節の生理機構

音声の基本周波数 (F0) を調節する生理機構には，多くの喉頭筋が複雑に関与して声帯を引き伸ばすことにより声帯粘膜の張力を変化させるような喉頭の筋活動に由来する要素と，発声時の呼気努力の変化に伴う声門下圧変動に依存するような空気力学的な要素があるといわれる。これら二つの要素は，発話時には相互に依存する性質を持っており，喉頭筋の収縮によって設定された声帯張力にふさわしい声門下圧が与えられるような生理的調節が行われているように思われる。従来，F0 調節に関与する喉頭機能の中で最もよく知られている機構は，輪状甲状筋 (CT : cricothyroid muscle) の収縮により輪状軟骨 (cricoidcartilage) と甲状軟骨 (thyroid cartilage) とが関節を中心に回転して声帯を前後に引き伸ばす仕組みである [8][9]。現在のところ，この仕組みの作用機序があまりにも自明であるために，F0 調節機構のほとんどが輪状甲状筋の機能で説明できるような印象を与えているようである [49]。しかし，筋収縮の程度と声帯長あるいは F0 との関係についてあまり精密な研究が行われていないばかりでなく，輪状甲状筋の収縮によって変化する F0 の範囲は大きく見積もっても 1.5 オクターブ程度ではないかとする推測もある [50]。また，輪状甲状筋以外の内喉頭筋の収縮が F0 に影響を及ぼす可能性も十分に高い [51]。これに対して，F0 を変化させるときに喉頭の上下方向の位置変化が起こる傾向が知られ，喉頭軟骨で形成される枠組みに対する外喉頭筋の補助的作用が多くの研究者によって調べられている。F0 の上昇時には舌骨上筋 (舌骨に上方から付着する筋肉群) が舌骨を前方に引くことによって声帯の伸張を促進する [52]。一方，F0 の下降時には舌骨下筋 (舌骨に下方から付着する筋肉群) の活動が起こることが知られており [10][11]，喉頭を引き下げる作用が何等かの仕組みにより声帯の短縮をもたらすと考えられている。しかし，喉頭の下降そのものが声帯の短縮を引き起こす直接的な仕組みは見当たらず，舌骨下筋の収縮がどのような生理機構の連鎖によって F0 を下降させるかは，F0 下降の現象を考える際に常に問題となっている。

4.2 実験

4.2.1 実験方法

磁気共鳴コンピュータ断層装置を用いて頭頸部の正中矢状断面を撮像した。使用した装置はSMT-100GUX(島津) であり，静磁場強度は１テスラ [T] である。撮影方法は，スライス厚 10mm，TE=12ms，TR=30ms のフィールドエコー法であり，撮像時間は１画面あたり４ｓである。この撮像法の特徴は撮像時間が比較的短く，またプロトン密度画像に近い撮像法なので軟骨組織が比較的高輝度で撮像される。従って，本研究のような発声時の喉頭軟骨の形態計測に適している。撮像時には被験者は仰臥位をとり，頭頸部の動きを防ぐため頭部と腰部の２箇所を軽くベルトで固定した。また，F0 抽出に用いるため，被験者の頭上に取り付けられた伝声管より取り出した音声を隣室のスピーカから出力し，その音声をマイクを通して DAT に記録した。被験者は 28 歳〜43 歳までの４人の健康な成人男性である。

4.2.2 実験手順

あらかじめ被験者の楽に発声できる音域を調べ，その範囲内で高い周波数から下限の周波数まで音階に合わせて母音/a/を持続発声させ撮像を行った。周波数範囲は約１〜1.5 オクターブである。1 回の撮像ごとに，まず発声すべき高さの楽器音を被験者に聞かせ，同じ高さの音声を練習のため 1 度発声し，その後，再度発声を開始した直後に撮像を開始した。MRI の撮像中は装置から非常に大きな周期音が発生するので，音声の F0 は撮像前後の区間から抽出した。発声方法は，3 人の被験者 (A, B, C) に対しては自由な発声とした。1 人の被験者 (D) に対してはすべての周波数に対して喉頭の高さをできるだけ変化させない発声（以下 D-1 とする）と，反対に積極的に喉頭の位置を変化させる発声 (以下 D-2 とする) の 2 回の実験を行った。

4.2.3 画像の分析

上記実験により得られた画像を，計算機に取り込み各器官の抽出を行った。分析の対象とした器官は，下顎骨 (mandible)，舌骨 (hyoid bone)，甲状軟骨 (thyroid cartilage)，輪状軟骨 (cricoid cartilage)，披裂軟骨 (arytenoid cartilage)，頸椎 (cervical spine) である。記録された画像は自動的な画像抽出には適していないため，肉眼による輪郭の検出を行った。頸椎以外に対しては各被験者ごとに記録された画像のなかで最も鮮明に得られた画像を選び，各器官の正中面での基準の輪郭を抽出した。他の画像に対してはその抽出した輪郭が変化しないものとして，最もよく基準の輪郭のあてはまる位置及び傾きを探索した。頸椎は第 2 頸椎の上端から第 7 頸椎の下端までの前壁をトレースした。各器官の位置を数値化するため，基準の輪郭に 2 か所の基準点を設定しておきその点の (x,y) 座標値により表すことにした。ただし，頸椎については輪状軟骨の二つの基準点を結ぶ線分に垂直な 2 本の直線を引き，それぞれの線と頸椎のトレース線との交点を求めて頸椎の基準点とした。図 4.1 に例を示す。図中の点線は，抽出した輪郭を基に解剖学的知識を用いて描いた舌骨，甲状軟骨，輪状軟骨の側面図で，×印は各器官の基準点である。

4.3 実験結果

実験により得られた MRI 画像を用いて F0 変化における各器官の相対変位を分析した。また，F0 を調節する基本的な機構は輪状甲状関節を中心とする喉頭軟骨の回転であるため，各器官の角度の変化も求めた。

4.3.1 抽出結果

図 4.2 に MR 画像及び各器官の抽出結果を示す。これは被験者 A の例であり，96〜260Hz の範囲で発声したときの輪状軟骨の輪郭及び頸椎前壁のトレースを最も高い音声を発声したときの MRI 画像の上に重ねて表示してある。また，図 4.3 に被験者 A, B, C の舌骨，甲状軟骨及び輪状軟骨の軌跡を示す。

この場合の軌跡とは各器官の二つの基準点の中点の軌跡を示している。これらの図から喉頭全体が F0 の変化と共に上下しており，また輪状軟骨は頸椎の自然湾

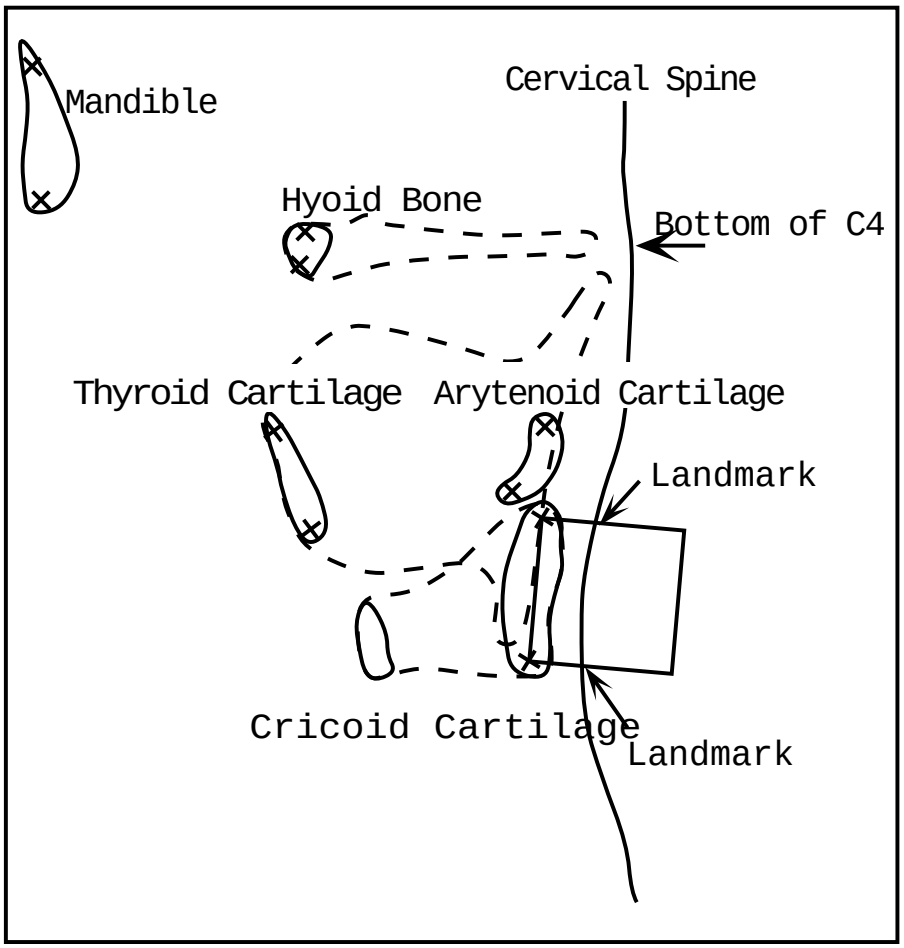


図 4.1: The contours of laryngeal components and related structures traced from an MRI picture. Each trace of the rigid structures has two landmarks to measure positional changes. The landmarks on the cervical spine are difined by cricoid position.

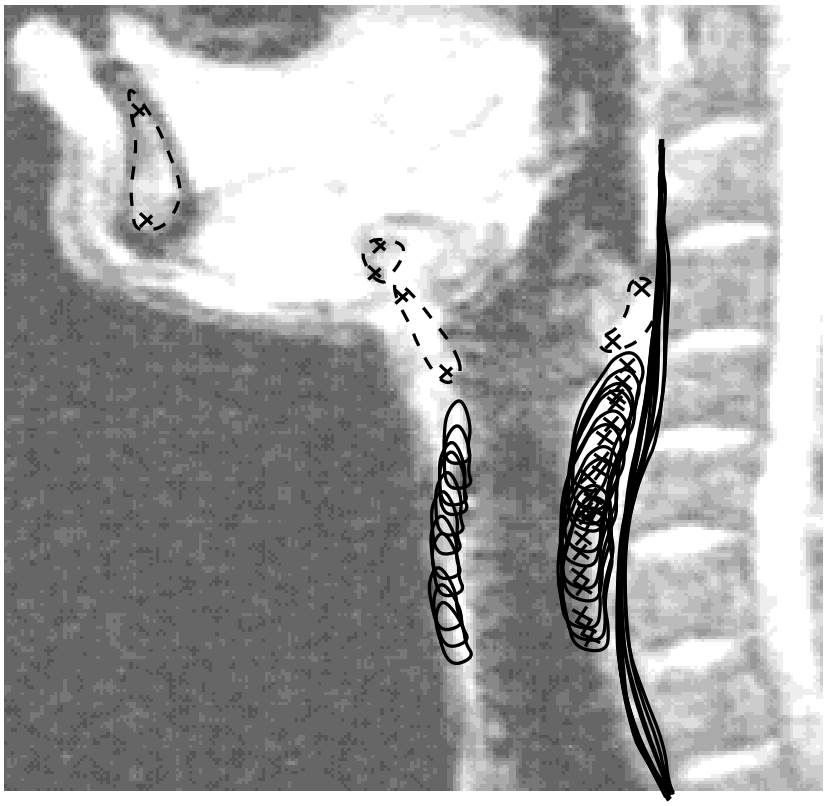
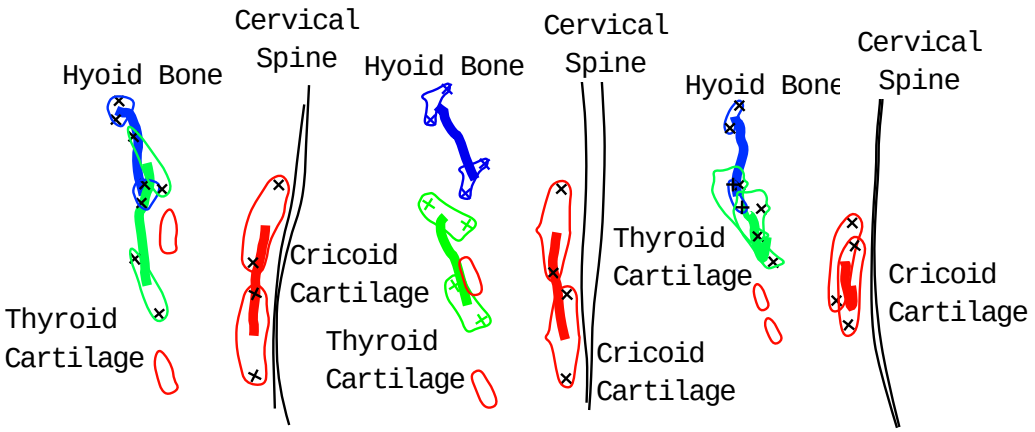


図 4.2: Changes of the position of the cricoid cartilage and the shape of the cervical spine in one of the subjects. The contours for all F0 levels are superimposed on the MRI picture for the highest F0. The other rigid structures, as shown in dotted lines, are also for the highest F0.



(a) SUBJECT A (b) SUBJECT B (c) SUBJECT C

図 4.3: Trajectories of positional changes of the rigid structures. Each trajectory (shown in bold line) indicates positional change of two landmarks on the contours.

曲に沿って上下していることが観測される。図 4.4 は被験者 A の各周波数での各器官の抽出結果を輪状軟骨が重なるよう移動と回転により重ねて表示したものである。二つの円は甲状軟骨の二つの基準点がすべて乗るように描いた同心円である。この図より甲状軟骨は輪状軟骨上のある点を中心に回転運動を行っていることが分かる。この回転の中心は輪状甲状関節であると想像されるが，解剖学的知識と比べると調節の位置がわずかに下方へずれている。この関節位置と回転中心とのずれは輪状軟骨と甲状軟骨とが回転運動だけでなく水平移動の成分を持っていることを示している。つまり，甲状軟骨の基準点は回転運動と水平移動のために楕円軌道上を移動し，その軌道と重なる円の中心は楕円の中心より下方へずれた位置に観測されたのである。

4.3.2 喉頭の上下運動

図 4.5 の上段に喉頭の高さの変化を，また下段に計測した喉頭の高さの説明を示す。図の下段に示すように，輪状軟骨の二つの基準点の中間位置を喉頭の高さと決め，最も高い音声を発声したときの画像における第 4 頸椎の下端を基準 (0) として，喉頭の相対的な位置を表している。図 4.5(a) の被験者 A，B，C の結果より単純な音階を自然に発声した場合，F0 の高さと喉頭の高さには正の相関があるこ

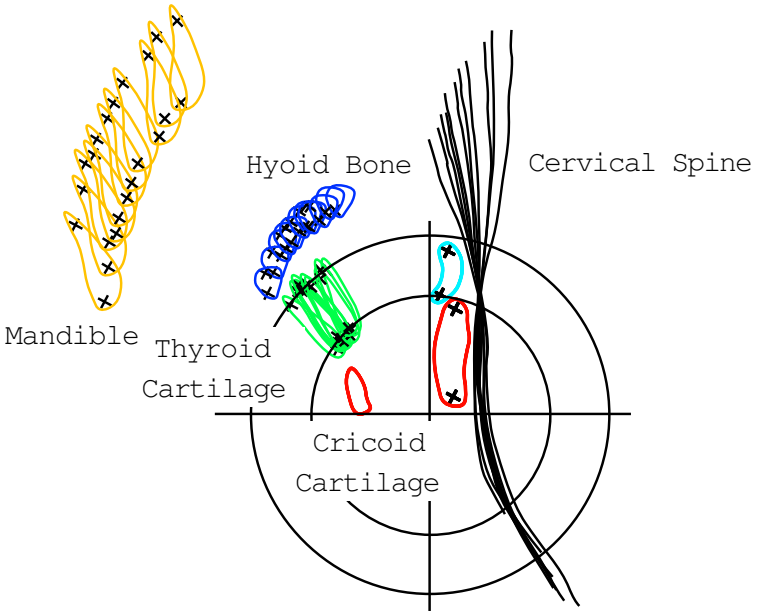


図 4.4: Movements of the rigid structures relative to the cricoid cartilage. The contours are rotated and translated into the coodinate system defined by the cricoid cartilage. Two circles illustrate approximate circular trajectory of the landmarks on the thyroid cartilage. The center of the circles indicates the functional axis of the cricothyroid joint.

とが分かる。喉頭の高さの変化は 1 オクターブ当たり，被験者 A では 19.4mm，被験者 B では 29.9m，被験者 C では 10.8mm であった。被験者 C では，低い F0 で逆に喉頭が上昇している。これは，撮像中の仰臥位の姿勢では，あらかじめ自然な立位で測定した周波数範囲の発声ができずに，低い F0 の範囲で無理な発声を行ったためであると考えられる。立位の状態に比べ撮像中の仰臥位の状態の方が低い F0 を発声しにくいという傾向は他の被験者でも同様に見られた。また，図 4.5(b) は被験者 D に喉頭の高さを変えない発声 (1) と変える発声 (2) を指示したときの結果であり，D-1 では高さがほとんど変化せず，D-2 で高さが十分変化していることから，指示どおりの発声が行われていることが確認できる。

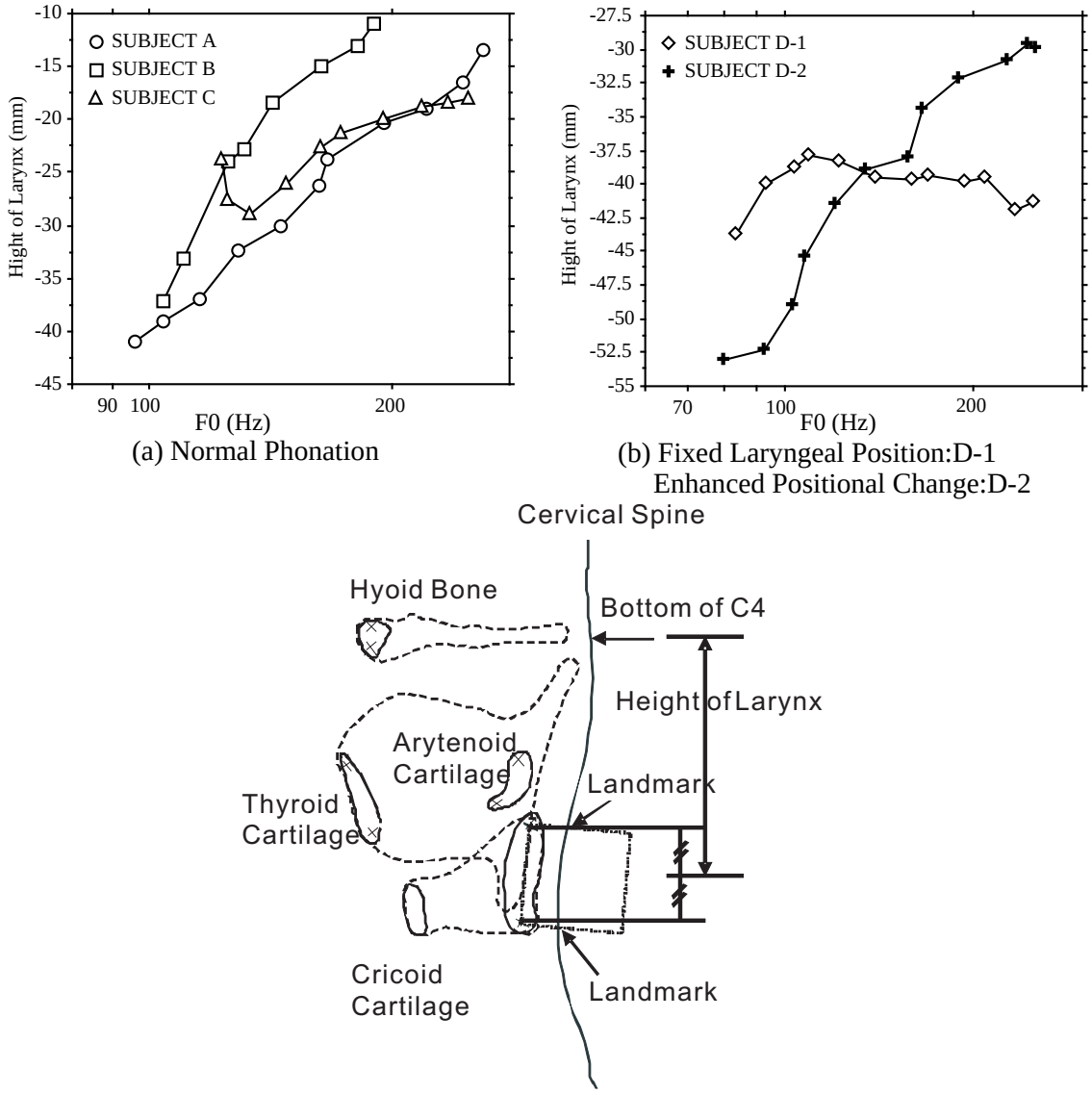


図 4.5: Vertical movement of the cricoid cartilage in all F0 levels in four male subjects. The subject A, B and C produced normal phonation, and the subject D demonstrated two different phonation types (D-1; fixed larynx position, and D-2; enhanced larynx movement). The height of the cricoid cartilage is relative to the lower edge of the fourth cervical vertebra (C4). The values are plotted as a function of the logarithm of F0, which is consistent in the following plots.

4.3.3　輪状軟骨と甲状軟骨の相対角度変化

声帯の長さの調節は主に輪状甲状関節を中心とする輪状軟骨と甲状軟骨との位置（角度）関係によって行われていると考えられる。よって，F0 変化に対する輪状軟骨と甲状軟骨の相対的な角度変化を求めた。輪状軟骨の角度は輪状軟骨後板上の二つの基準点を結んだ線分の傾き，甲状軟骨の角度は甲状軟骨の二つの基準点を結ぶ線分の傾きとした。図 4.6 は，それらの角度の差を最も低い F0 での値を基準 (0) として相対変化と，その計測した角度の説明を示したものである。

この結果から，喉頭の上下動を行わない場合を含めすべての被験者で F0 の変化と輪状軟骨と甲状軟骨の相対角度とは強い相関を持っていることが分かる。また，楽に発声できる範囲で 1 オクターブ当たり平均で 18.5 °，最大で 33.7 °の回転が観測された。図 4.6 においても被験者 C では，低い F0 の領域で他の被験者と異なり，F0 が変化しているにもかかわらず角度の変化が見られない。この結果は，輪状甲状関節を中心とする回転を使用しない F0 の調節の可能性を意味している。

4.3.4　頸椎の湾曲と輪状軟骨の回転

図 4.2 と図 4.3 から分かるように輪状軟骨は頸椎に沿って移動していると考えられる。輪状軟骨と頸椎との関係を調べるため，輪状軟骨の後板の傾斜角と共に，輪状軟骨の高さにおける頸椎前壁の傾斜角を求めた。図 4.7 に輪状軟骨の後板及び頸椎の傾斜角の変化，および計測した角度の説明を示す。図に示すように，輪状軟骨の後板の傾斜角は輪状軟骨後板上の二つの基準点を結んだ線分の傾きで表した。頸椎の傾斜角は図 4.1 中の Landmark と示された輪状軟骨の高さにおける頸椎前壁上の二つの基準点を結んだ線分の傾きで表した。最も F0 の高いときの傾斜角を基準 (0) として，それ以外の周波数に対しては傾斜角の相対変化として表してある。図 4.7 の結果から被験者 B を除くすべての被験者で，輪状軟骨の後板と頸椎の傾斜角の変化は同様の傾向を持っており，F0 の増加と共に角度が減少しているのが分かる。無理なく発声できる音域で発声した場合，頸椎の湾曲の程度にはほとんど変化がないので，図 4.7 の示す頸椎の傾斜角の変化は，各々の F0 を発声するときに輪状軟骨の高さが変化することによって，輪状軟骨が接する位置での頸椎の傾斜角が変化することを示している。よって図 4.7 の結果は，F0 の下降に伴って

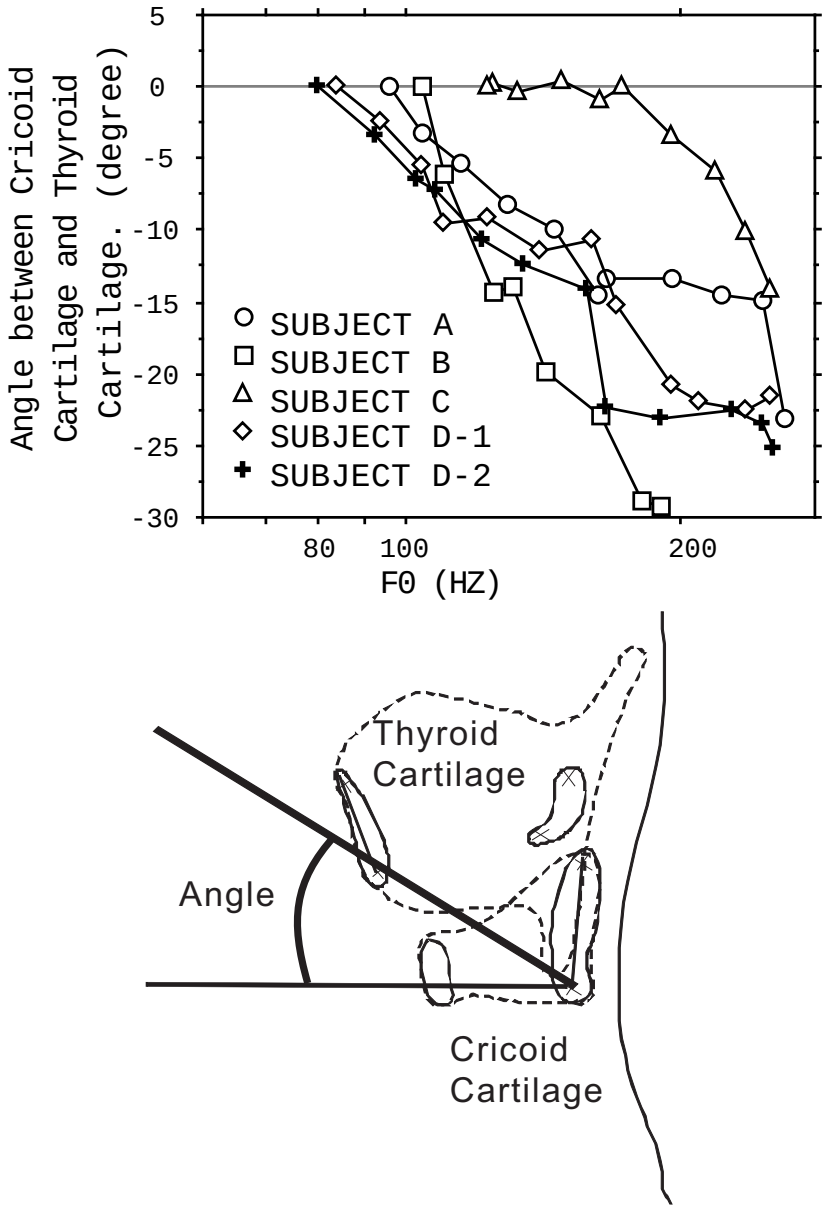


図 4.6: Relative change of the angle between the cricoid and the thyroid cartilages. The measures are standardized at the value for the lowest F0.

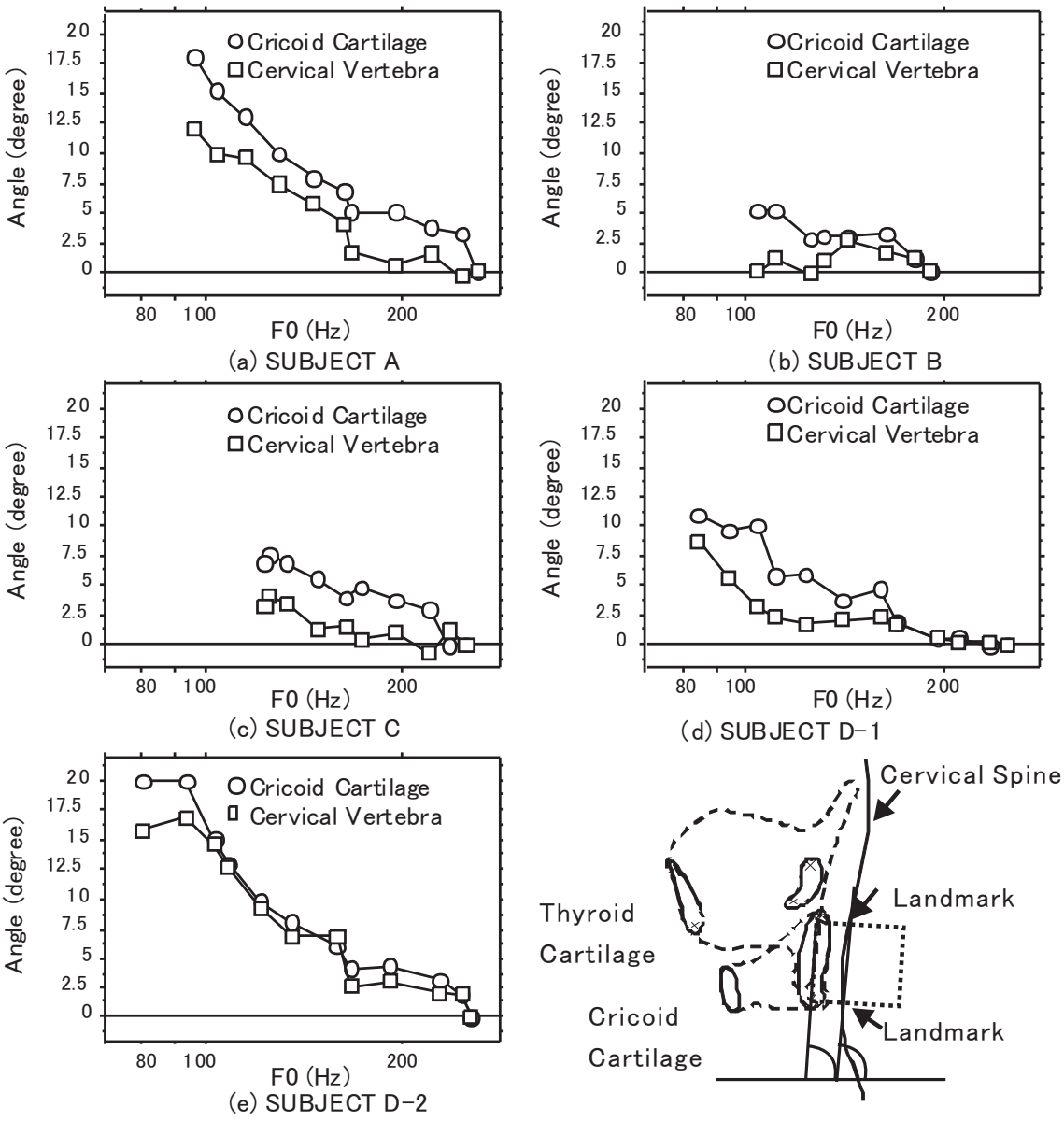


図 4.7: Relative tilt of the posterior plate of the cricoid and the contour of the cervical spine at the position of the cricoid cartilage. The measures are standardized at the value for the highest F0.

喉頭が下降し，輪状軟骨が頸椎の湾曲に沿って回転することを示している。これに対して，図 4.7(b) の被験者 B では，F0 が変化しているにもかかわらず頸椎の傾斜角はほとんど変化せず輪状軟骨の回転もわずかである。この現象は，F0 調節には頸椎の自然湾曲に沿う輪状軟骨の回転以外にも次節で述べるような他の要素があることを示唆している。また，頸椎の傾斜角の変化より輪状軟骨の角度変化の方がわずかに上回る傾向は被験者 B を含めすべての被験者で見られた。

4.3.5 甲状軟骨と舌骨の位置変化

声帯の長さの調節に直接関与するもう一方の器官である甲状軟骨の位置（角度）の変化と，甲状軟骨の回転に影響を与える舌骨との距離の変化を求めた。図 4.8 は甲状軟骨の角度変化，および計測した角度の説明を表したものである。この場合も甲状軟骨の二つの基準点を結ぶ直線の角度を計測し，F0 の最も低いときの角度を基準として角度の相対変化を求めた。被験者 B と D-1 では F0 の変化と共に甲状軟骨は回転しているが，それ以外の被験者では高い F0 の範囲でわずかに回転しているのみである。被験者 B は頸椎の湾曲が少なく，D-1 では喉頭の上下動が起こらない。従って，頸椎の湾曲により輪状軟骨が十分回転できる場合には甲状軟骨はほとんど回転せず，輪状軟骨の回転だけでは声帯長の変化が不十分な場合に甲状軟骨が回転して補償していると考えることができる。図 4.9 は，F0 変化に伴う舌骨と甲状軟骨との距離の変化，および計測した距離の説明を示したものである。図に示すように，計測点は舌骨体部下端付近と甲状軟骨切痕付近の基準点であり，この 2 点間の距離を求めた。図に示した距離は F0 の最も高い時の距離を基準として相対変化を表したものである。舌骨と甲状軟骨との距離の変化が大きかった被験者は A と D-2 である。これらの被験者では，輪状軟骨が頸椎の自然湾曲に沿って十分回転しており，甲状軟骨は F0 の低い範囲でほとんど回転していない。従って，自然湾曲による輪状軟骨の回転によって F0 を調整できる領域では甲状舌骨筋 (TH) の弛緩により舌骨と甲状軟骨間の距離が拡大し，輪状甲状関節の回転が促進されているのではないかと考えられる。

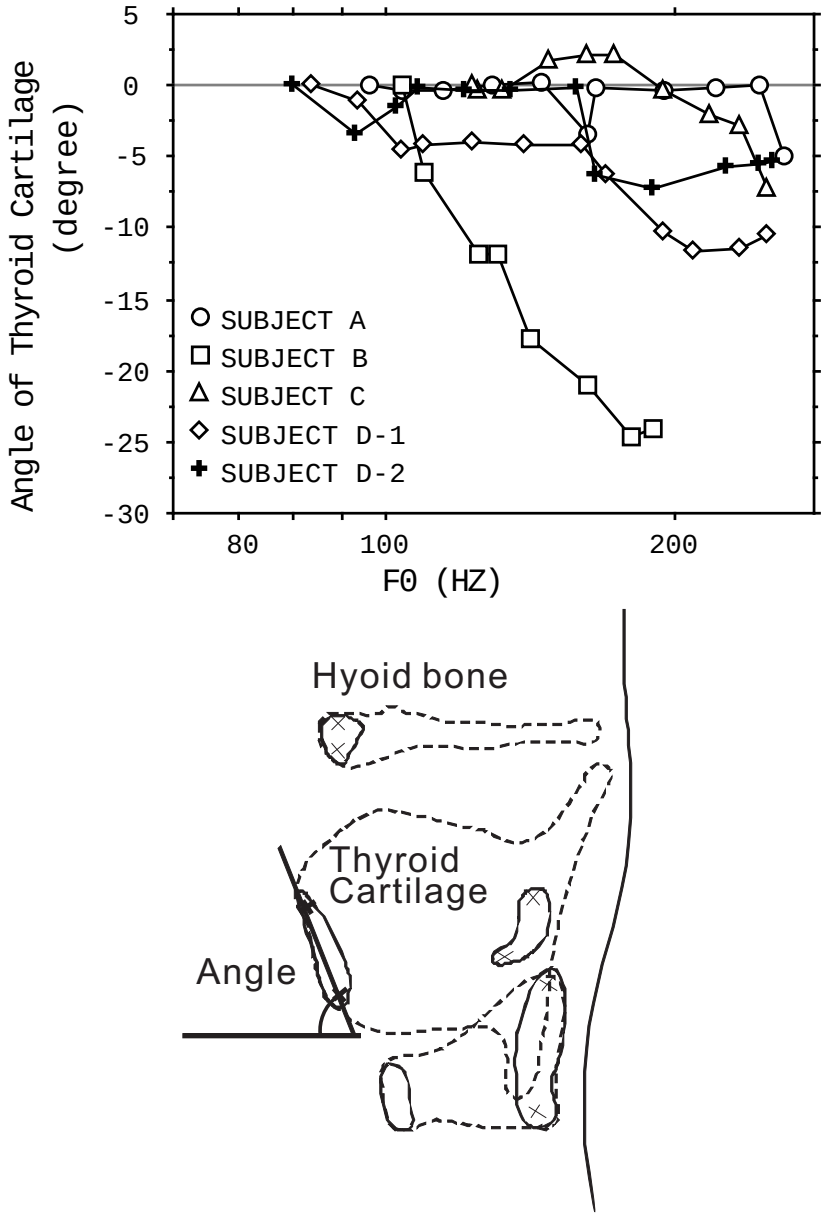


図 4.8: Relative tilt of the thyroid cartilage, standardized at the value for the lowest F0.

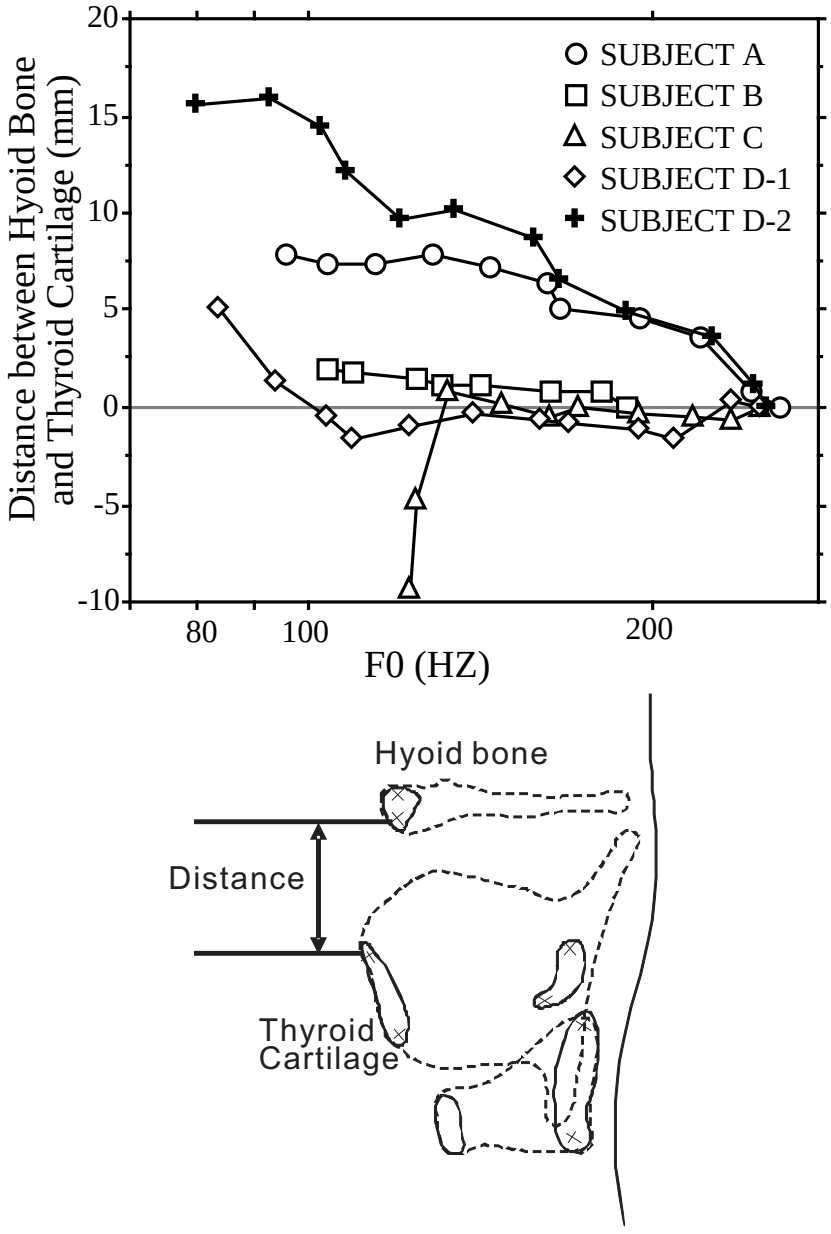
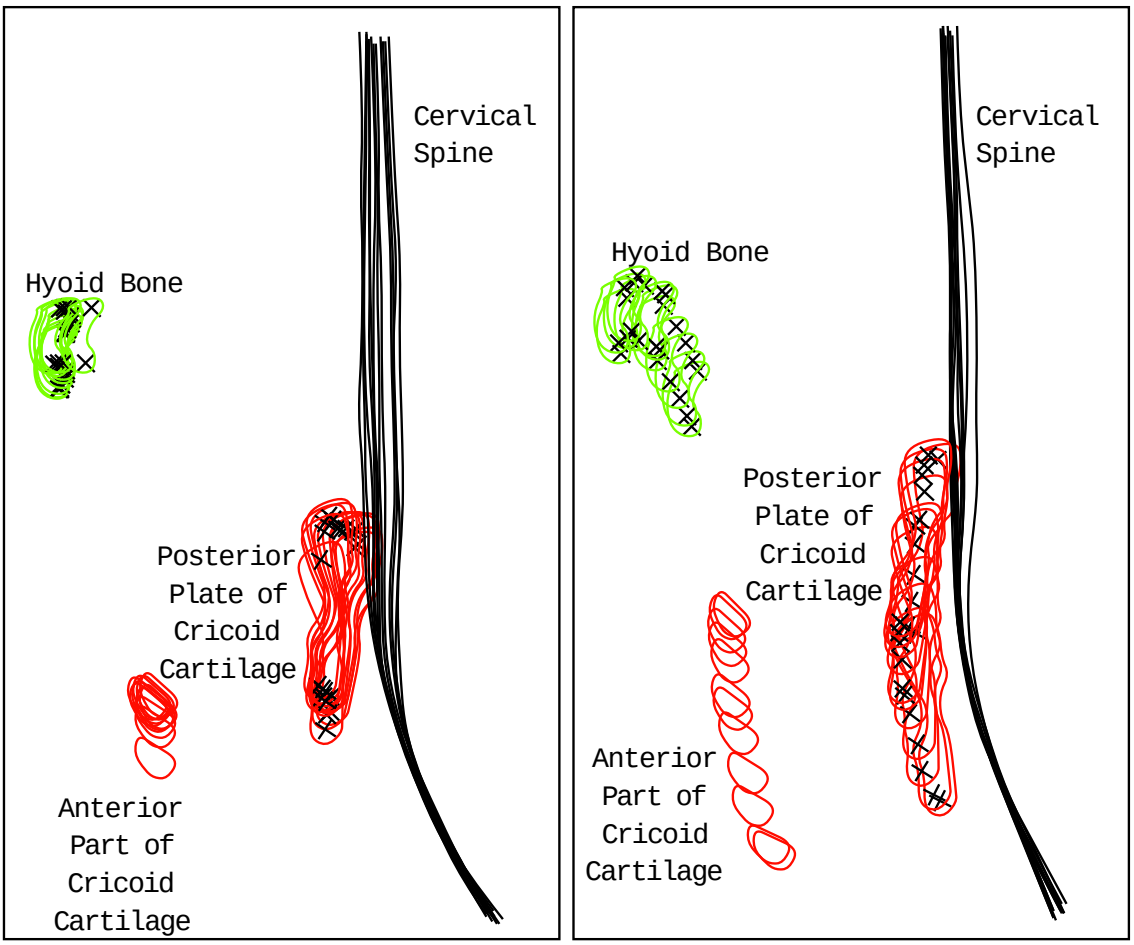


図 4.9: The distance between the hyoid bone and the thyroid cartilage, standardized at the value for the highest F0.

4.3.6　喉頭の上下運動を伴わない場合

喉頭の高さを固定した F0 変化 (D-1) と喉頭の下降を伴う F0 変化（D-2）を比較したときの，喉頭の位置変化の差異について調べた。図 4.10 は，D-1 と D-2 において，頸椎の前壁のトレース，輪状軟骨，及び舌骨の輪郭の抽出結果をすべての F0 について重ね合わせたものである。喉頭の高さを固定した場合は，F0 下降に伴



(a) Fixed Laryngeal Position.　(b) Enhanced Positional Change.

図 4.10: Midsagittal images of the cricoid cartilage, the hyoid bone and the anterior contour of the cervical spine. The subject D produced the same range of F0 with fixed laryngeal position (D-1; 250Hz〜84Hz) and enhanced positional change (D-2; 253Hz〜80Hz).

い第 2〜5 頸椎が大きく前方に移動しているのが分かる。輪状軟骨の後板の上半部

は第 5 頸椎が接しており，輪状軟骨の上半部が前方に押し出されることによって輪状軟骨が声帯長を減少させる方向に回転すると考えられる。

喉頭の上下運動を伴う場合と伴わない場合の，披裂軟骨と甲状軟骨との距離と F0 との関係，および計測した距離の説明を図 4.11 に示す。図に示すように，披裂

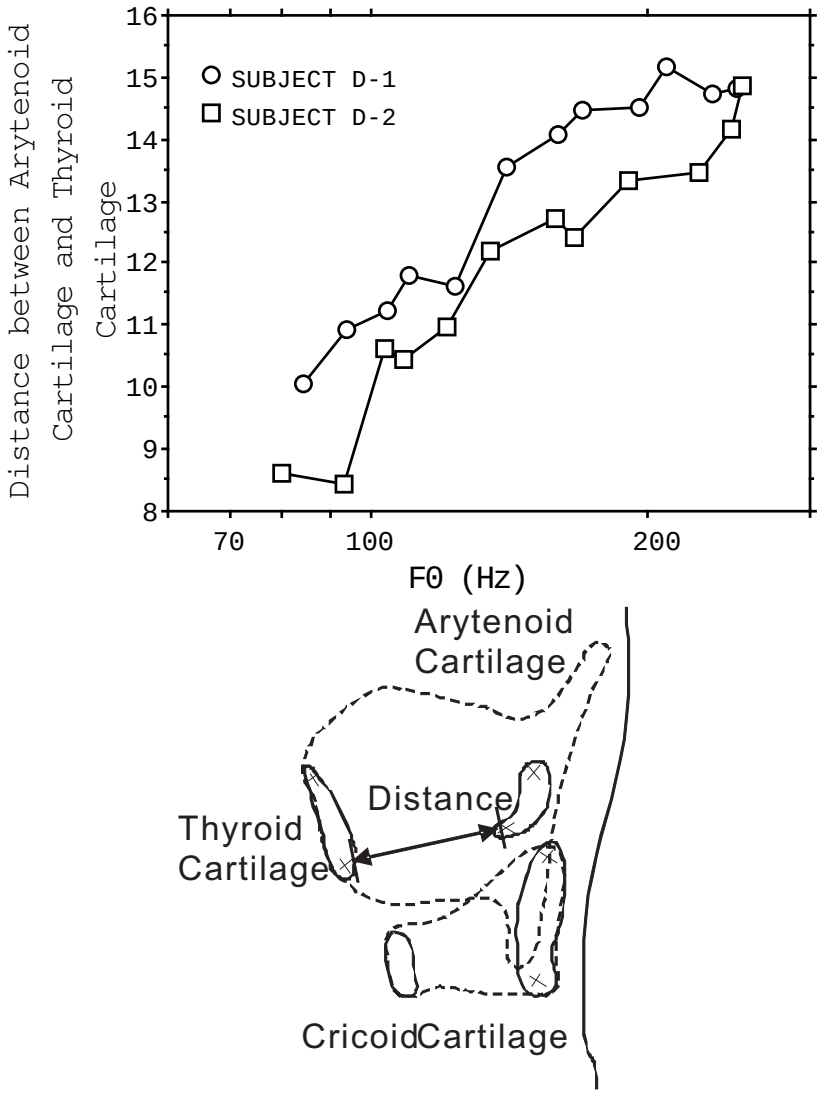


図 4.11: The distance between the arytenoid and the thyroid cartilages in the subject D. The distance indicate the change which is in proportion to length of the vocal fold.

軟骨と甲状軟骨との距離とは披裂軟骨の下端付近の基準点と甲状軟骨の下端付近の基準点との距離であり，声帯長の真値と比例関係にあると見なすことができる。

図より喉頭の上下運動の有無にかかわらず F0 の対数に比例して声帯長が増加することが分かる。しかし，喉頭の高さを固定した場合には，同一の F0 でも声帯長は全体的に長くなっている。この結果は，F0 の調節は主に声帯長の長さの調節によって行われているが，長さ以外にも F0 の調節に関与する要素が存在することを示している。

4.4 考察

4.4.1 頸椎の自然湾曲による輪状軟骨の回転

今回の実験で行われた下降音階で持続発声したときの輪状軟骨の回転による F0 調節の生理機構を図 4.12 に示す。F0 を下降させる場合は，喉頭を下降させること

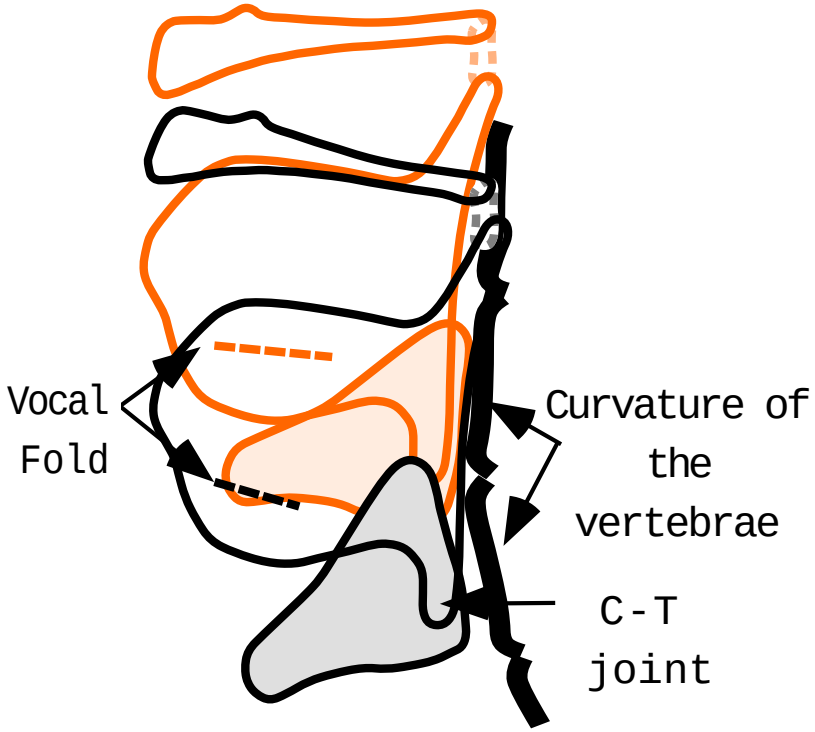


図 4.12: Proposed mechanism of F0 control using larynx lowering along the curvature of cervical spine. When the curvature of the vertebrae in available for F0 control, vertical movement of the larynx along the cervical induces the rotation of the cricoid cartilage.

により，輪状軟骨を頸椎の自然湾曲に沿って回転させ声帯長を短くする。頸椎の自然湾曲が十分大きい場合は，甲状舌骨筋を弛緩させ甲状舌骨と甲状軟骨との距離を増加させることによって，舌骨の影響を減少させ喉頭の下降を助長する場合もある。また，頸椎の湾曲が十分でない場合は，頸椎の湾曲を変形させ輪状軟骨を回転させる場合もある。

この頸椎の自然湾曲は人間の２足歩行に伴う代償性脊柱湾曲の一部をなし，構音器官が成人の形態に近づく５歳頃までに形成されること [53] を考慮すると，頸椎の自然湾曲が音声生成に関与する現象は興味深い。被験者 B には頸椎の湾曲があまり見られないが，これは今回の実験では仰臥位で撮像を行ったため直立姿勢の状態に比べ頸椎の湾曲が少なくなったためと考えられる。自然な状態で発話する場合は，喉頭の高さで頸椎の自然湾曲が存在するため，一般的にはこの機構が利用されていると考えられる。また，仰臥位の姿勢では立位の姿勢と比べて，音域が狭くなる現象も，この回転機構と仰臥位での湾曲の減少を考慮すると簡単に説明することができる。

頸椎と輪状軟骨後板の傾斜角の比較において，すべの被験者で頸椎の傾きの変化の程度より輪状軟骨の角度の変化の程度が大きいという結果が得られている。これは，輪状軟骨の回転を頸椎の自然湾曲だけでは十分に説明できない場合があることを示唆している。頸椎の湾曲以外の輪状軟骨を F0 下降の方向に回転させる要因としては，輪状甲状筋 (CT) の弛緩が考えられる。しかし，周波数の低い領域でも同様に見られることから，輪状咽頭筋 (CP) の収縮によって生じた膨らみによる輪状軟骨の回転 [54] など，その他の要素も存在すると考えられる。

4.4.2　舌骨による甲状軟骨の回転

F0 の下降に伴って甲状軟骨も声帯長の減少の方向に回転する現象が認められた。ほとんどの場合，頸椎の湾曲による輪状軟骨の回転に比べ甲状軟骨の回転する程度は小さく，また回転の見られない周波数領域もあるため，甲状軟骨の回転は輪状軟骨の回転を補う要素と考えられる。甲状軟骨が大きく回転する被験者では舌骨と甲状軟骨との距離が増大せず，そのような被験者は舌骨が下降しながら後方へ移動していることが分かる。それらのことから，図 4.13 に図を示すように，舌

骨の後方移動に伴って甲状軟骨の上部が後方に回転することによって生じた結果ではないかと考えられる。ただし，Sonninen[55] が指摘しているように，甲状舌骨

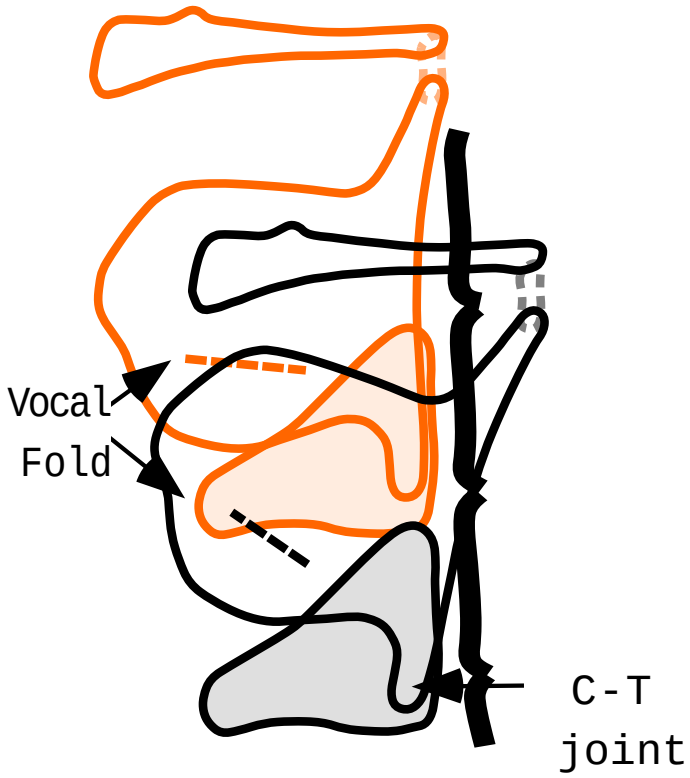


図 4.13: Proposed mechanism of F0 control using horizontal movement of the hyoid bone. When the curvature of the vertebrae is not available for F0 control, horizontal movement of the hyoid bone can cause the rotation of the thyroid cartilage.

筋と胸骨甲状筋の同時収縮によってもこのような甲状軟骨の回転をもたらすという可能性も残っている。

4.4.3 喉頭の高さが変化しないときの F0 下降機構

喉頭の高さを固定した状態であっても F0 を下降させることはある程度の範囲で可能である。喉頭の高さを変える発声を行うときは喉頭を上下させるだけで頸椎の自然湾曲に従って輪状軟骨が回転するが，喉頭の高さを変えない発声では，頸椎の変形によって輪状軟骨を回転させる現象が見られた。従来，声道の後壁は形

の変化しないものとして取り扱われてきたが，声の高さを変えるために咽頭後壁が変形する現象は興味深い観測結果である。しかし，この頸椎の変形による F0 の調節は喉頭の上下動による調節法より余分な発話努力を要するものであり，これも頸椎の自然湾曲を利用できない場合の補助的な要素の一つであると考えられる。また，同じ被験者で同じ F0 を発声しているにもかかわらず，喉頭の上下動を伴わず輪状軟骨の回転が少ない場合の方が全体に声帯長が長くなることから，声帯長以外にも F0 を変化させる要因があることが推測された。その要因としては従来から報告されているように，声帯筋や甲状披裂筋 (声帯筋) の収縮，披裂軟骨の内転力の増加による声帯の実効振動長の減少など [56] が考えられるが，MRI の正中矢状断面のみではどの要因が影響しているのかを判断することは困難である。

4.5 本章のまとめ

本章では，磁気共鳴画像法 (MRI) を用いて F0 を下降させたときの喉頭周囲構造の位置変化を分析することにより F0 調節の生理機構について分析した。その結果，咽頭筋の活動以外に大きく分けて以下に示す 2 つの F0 の調節機構が存在することが確認された。

1. 輪状軟骨後板の接する位置の頸椎の角度変化による輪状軟骨の回転。
   頸椎の自然湾曲に沿って喉頭が下降すると輪状軟骨が F0 を下降させる方向に回転する。
2. 舌骨の前後動による甲状軟骨の回転。

これらの F0 調節機構は，F0 の変化と舌骨の位置の変化が関連していることを示しており，F0 調節による声道形状変化の生理機構の解明につながると考えられる。また，これら以外にも，声帯長の伸縮以外の機構による F0 調節への影響なども推測される結果であった。

本章では，複数の話者について各発話器官の位置と F0 との関係を実験から数値として得ることができた。しかし，話者ごとに頸椎の形状が異なることや，上記 1.,2. 以外の機構の存在の可能性や示されたことから F0 制御の統計的な予測モデルの構築は行わなかった。物理的なモデルについては次章にて示すことにする。

第 5章

ソースにおけるF0調節とフィルタにおける声道形状制御との力学的相互作用を考慮した発話器官の２次元モデル

本章では，人間の音声生成メカニズムに基づく発話器官の 2 次元モデルの構築，および，構築したモデルを用いた音声の合成実験について述べる。本章の目的は以下の２つである。第１の目的は，前章で提案した F0 調節機構により F0 調節と声道形状の制御との間に相互作用が生じること，および，本モデルにてその相互作用を再現できることを明らかにすることである。前章で提案した F0 調節機構では，F0 調節時に咽頭筋の活動だけでなく喉頭の上下動や外喉頭筋の活動が生じるため，これらの活動が舌骨を介して舌形状に影響を与えると考えられる。また反対に，外舌筋による調音動作も舌骨を介して喉頭に影響を与える。その結果，F0 調節と声道形状制御との間に相互作用が生じると思われる。よって，本章では，舌骨を介した喉頭と舌との力の授受を考慮した発話器官の２次元正中矢状断面のモデルを提案する。提案モデルを用いて F0 調節と声道形状制御との相互作用を再現する音声合成実験を行い，相互作用の存在を明らかにする。第２の目的は，合成音声の音質の評価である。本モデルの構築では，全てのパラメータを特定の話者の生理学的データと MRI 画像を用いて同話者に最適化することで，個人性を有する合成音声の再現を目指した。本モデルにより生成された音声と同話者の実音声から計測されたフォルマント周波数を比較することで音質の評価を行った。

5.1 F0 調節と声道形状制御との相互作用

現在の音声生成理論では，一般に喉頭における F0 の調節と調音器官による声道形状の変化とは互いに独立した要素としてモデル化されている [57]。しかし，両者の間には相互作用があり，必ずしも完全な独立性があてはまらない。例えば，母音の調音動作によって声の高さが変化する現象が見られる。これは母音の固有ピッチ [58] と呼ばれ，舌などの調音器官の変化が喉頭に力の作用を及ぼす機構が直接的ないし間接的原因であると考えられている。反対に，F0 を変化させるための動作が同じ仕組を介して母音の調音状態にわずかの変化をもたらす。この相互作用は，現在までの喉頭筋による F0 の調節のメカニズムに関する研究 [8],[59] および前章の結果から，以下に示すような原因で発生すると考えられる。

F0 の変化は輪状軟骨と甲状軟骨との相対的な回転により声帯の長さが変化することによって決まる。輪状軟骨と甲状軟骨との角度が狭まると，声帯が伸長され F0 が上昇する。反対に，この角度が広がると，F0 の下降が起こる。この回転運動は主に輪状甲状筋 (CT) によってもたらされると考えてられているが，それ以外に図 5.1 に示すような喉頭の枠組に作用するメカニズムが存在する。図 5.1(a) は，

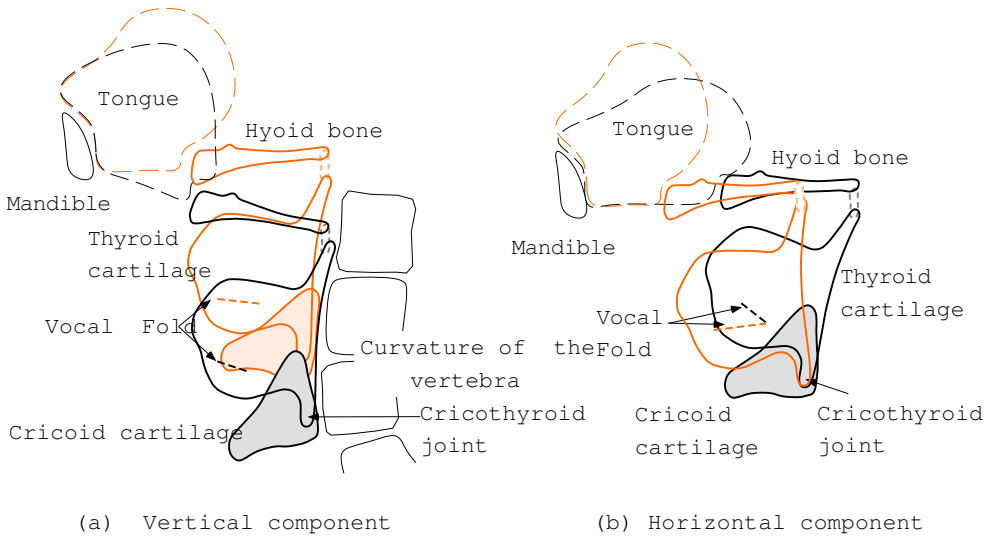


図 5.1: Mechanisms of tongue-larynx interaction.

輪状軟骨の回転による F0 の調節機構を示している。輪状軟骨の角度変化は頸椎の前壁の傾斜角度と密接な関係を持っており，頸椎の湾曲に沿って喉頭が上下する

ことにより輪状軟骨の回転が生ずる。一方，図 5.1(b) は，甲状軟骨の回転による F0 の調節機構を示している。甲状軟骨は，靱帯などにより舌骨と結び付いているため，舌骨の前後方向の移動により甲状軟骨の回転が生ずる。これらのメカニズムは，舌骨の位置変化が喉頭機能に影響を及ぼす機構を意味し，母音調音のための舌の変形によって舌骨が移動し F0 の変化がもたらされる現象を説明できる。その逆に，舌骨と舌が直接結合されているために，前述した F0 調節に使われる筋の活動により舌骨が移動し，その結果として舌が変形する現象も存在する。従って，F0 調節と声道形状制御との間で双方向の相互作用が存在することが理解できる。

5.2 モデルの作成

F0 調節と声道形状制御との相互作用は，個々の発話器官が筋や靱帯などの組織により複雑に接続されていることに起因すると考えられる。このような相互作用を表現することのできる発話器官の計算モデルとして，図 5.2 に示すような発話器官の相互接続を考慮した発話器官モデルを作成した。このモデルは，各器官の発話時の形状を生成する「調音モデル部」と，生成された声道形状から音声を生成する「音声合成部」に分けられる。「調音モデル部」では，筋の活動量を入力として定常母音における発話器官の形状が正中断面上で生成される。「音声合成部」では，声道断面から計算された伝達関数と 2 質量モデル [34] で生成された音源波形を用いて音声の合成が行われる。

5.2.1 調音モデル部

調音モデルは，(1) 舌の変形を計算する有限要素法モデルと，(2) その他の硬性器官の位置変化を計算するマススプリングモデルとにより構成される。表 5.1 に，本モデルに導入されている要素を示す。また，主要な筋の走行を図 5.3 に示す。

舌の変形は，舌筋の収縮をパラメータとする有限要素法により 2 次元あるいは 3 次元のモデル化が可能である。本モデルは，草川らの 2 次元モデル [16] を参考にして，舌を舌骨と下顎骨を結ぶ連続弾性体として有限要素法を用いてモデル化した。本モデルで用いた方法は，2 次元弾性問題の解析に用いられる有限要素法

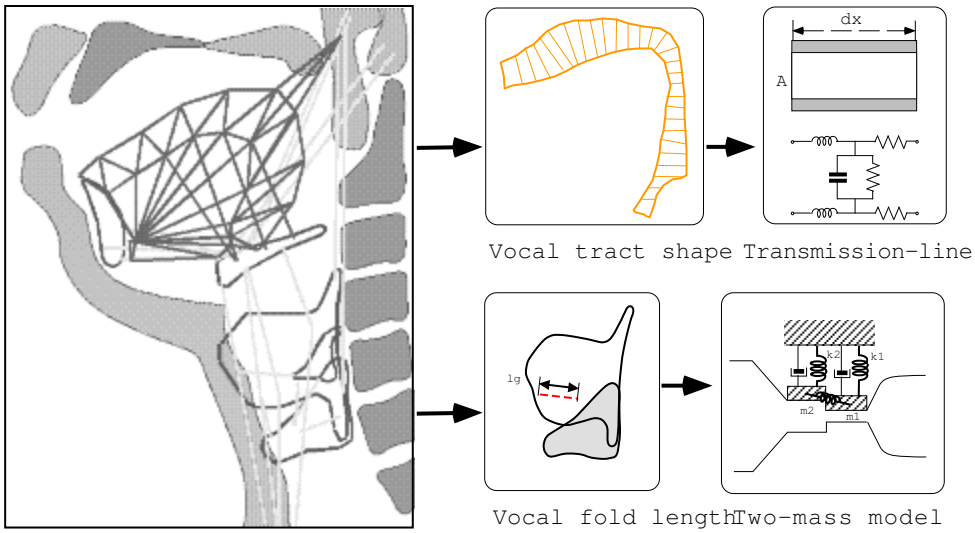


図 5.2: A physioloial model of speech organs and the method for vowel synthesis. (a) The articulatory part of the model deforms according to the mass-spring effects of all the muscles acting on each components, deriving vocal tract shape and vocal fold length. (b) Vowel synthesis involves estimation of the area function and the two-mass model parameters.

表 5.1: Speech organs and muscles in the model.

Organs (fixed)	Palate, Cervical spine, Sternum
Organs (mobile)	Hyoid bone, Thyroid cartilage, Cricoid cartilage, Arytenoid cartilage, Mandible
Extrinsic tongue muscles	Genioglossun (GGa, GGm, GGp), Styloglossus (SG), Hyoglossus (HG)
Intrinsic laryngeal muscles	Cricothyroid (CTa, CTp), Vocalis
Suprahyoid muscles	Digastric, Stylohyoid, Geniohyoid (GH)
Infrahyoid muscles	Sternohyoid (SH), Sternothyroid (ST), Thyrohyoid (TH)
Others	Cricothyroid joint, Cricoarytenoid joint, Ligaments, Mandibular muscles, Stylopharyngeus

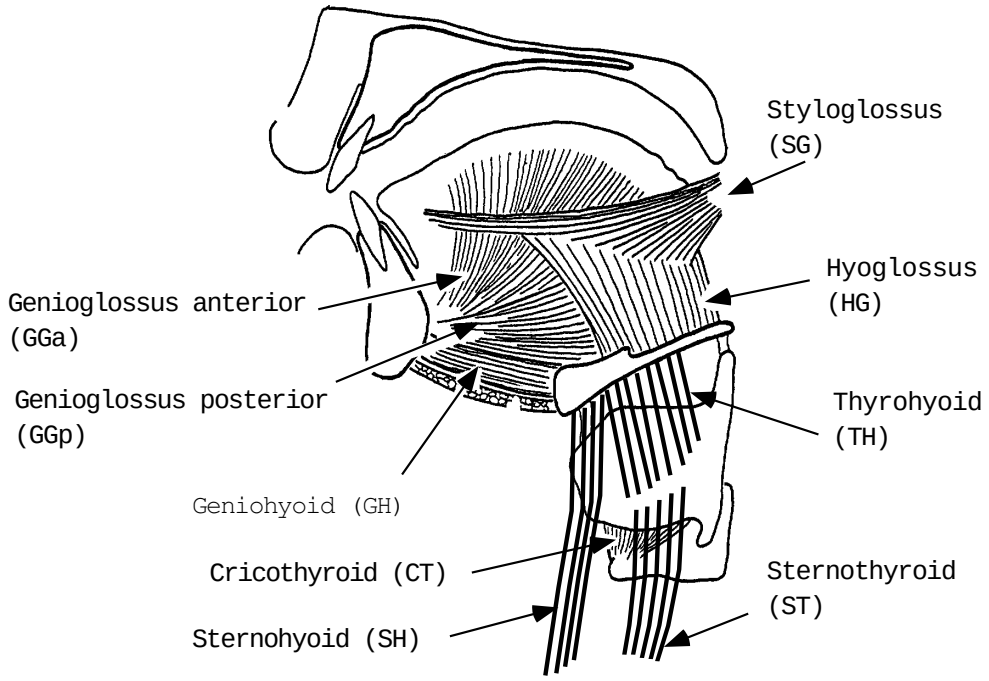


図 5.3: Schematic drawing of the extrinsic tongue muscles (surrounded) and the laryngeal muscles.

である。舌を 31 個の 3 角形要素で近似し，平面応力状態として計算した。舌の変形には，オトガイ舌筋 (GGa, GGp, GGm)，茎突舌筋 (SG)，舌骨舌筋 (HG) の 5 つの外舌筋が使用される。これらの外舌筋は，周囲の硬性器官とも接続されており，それぞれの位置計算にも用いられる。これらの筋は，筋束の終端となる複数の節点に作用する収縮力発生要素としてモデル化した。図 5.4 に本モデルの筋の走行を示す。図 5.4 に示された筋のうち，太線が解剖学的に基本となる筋の走行を示し，細線は個人データに適応させるために追加した要素である。これらの筋の活動量及び，舌骨・下顎骨の移動による舌の接続点の位置変化を入力として，舌の変形及び舌が舌骨，下顎骨の接続点に与える力が計算される。

(2) の硬性器官の位置変化はマススプリングモデルを用いて計算した。本モデルでモデル化した硬性器官は，下顎骨，舌骨，甲状軟骨，輪状軟骨，披裂軟骨，口蓋 (palate)，頸椎，胸骨 (sternum) である。これらの硬性器官は，筋，靱帯，関節などによって互いに接続されている。これらのうち，下顎骨，舌骨，甲状軟骨，輪状軟骨，披裂軟骨は移動可能な硬性器官である。筋の活動量が与えられると，筋，

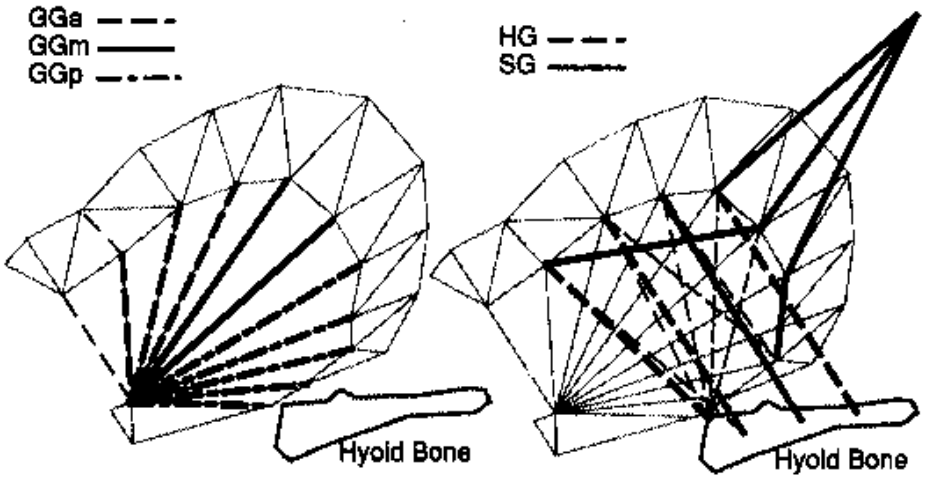


図 5.4: Implementation of the extrinsic tongue muscles in the finite element tongue model. The thick lines show standard muscle arrangements, and the thin lines indicate ad hoc adjustments required for the subject.

靱帯，関節などとの力の授受が計算され，最終的にすべての力が釣り合う位置にそれぞれの器官が移動する。また，下顎骨・舌骨は舌の有限要素法モデルとも接続されており，舌からも力の作用を受ける。以下に，硬性器官に力を与える各要素について述べる。

すべての筋は収縮力発生要素と弾性要素を並列に持つバネとしてモデル化した。更に，筋は空間的な広がり (筋束の幅) を持つため，一つの筋を複数の要素の並列接続により表現した。靱帯は収縮力発生要素を持たないことをのぞき筋と同様である。関節については，過度の回転と移動を制限するための弾性要素よりモデル化した。下顎の関節（側頭下顎関節）は回転と移動の二つの運動要素を持つことが知られており [60][61]，関節の回転に伴い下顎の関節突起が関節窩の湾曲に沿って移動すると考えられている。本モデルにおいても，関節の回転と関節窩の湾曲に沿った移動を取り入れ，下顎のモデル化を行っている。

輪状軟骨は，甲状軟骨と関節を介して接続しており，また気管とも結合している。更に輪状軟骨の後板は食道入口部の組織を隔てて頸椎に面している。従って，輪状軟骨は，喉頭筋やこれらの周囲構造からの力を受けて，主として頸椎の形状に沿う上下運動が生じる。5.1 節に述べたように，頸椎は輪状軟骨の高さで前方に湾曲しているために，喉頭の上下運動によって輪状軟骨の回転が引き起こされる

と考えられる。本モデルでは擬似的に輪状軟骨後板の上下 2 か所と頸椎の間に頸椎と垂直な方向に力を発生するバネを接続することによって，頸椎の湾曲と輪状軟骨の回転の関係をモデル化した。有限要素法とマススプリングモデルの結合は，舌と喉頭との力の授受を表現するために，以下の方法により行った。有限要素法で計算された舌の舌骨・下顎骨との接続点に掛る力は，マススプリングモデルの入力となり，反対に，マススプリングモデルで計算された舌骨・下顎骨の位置変化は，有限要素法の入力となる。従って，有限要素法モデルとマススプリングモデルの計算を交互に繰り返すことにより，マススプリングモデルで計算される舌の接続点に加わる力と，有限要素法で計算される舌の接続点に加わる力が釣り合う位置の検索を行った。両者の力が釣り合ったときの位置形状が調音モデルの最終状態となる。

5.2.2 音声の合成

音声の合成は Sondhi と Schroeter による調音パラメータ駆動の合成方式 [62] に基づいている。この合成法では 2.3.4 節にて示した音響管の電気回路網モデルを用いて計算された声道伝達関数と，2.3.3 節に示した石坂と Flanagan による 2 質量モデル [34] を用いて生成された音源波形によって音声の合成が行われる。

前節で述べたように，本研究の「調音モデル部」は 2 次元モデルである。従って，声道断面から得られた声道横断長を断面積関数に変換し，声道伝達関数の計算を行った。声道横断長から断面積関数への変換は以下の式を用いて計算した。

$$\begin{aligned} Area(d,x) &= \alpha(d)x^{\beta(d)}, & x \le X(d) \\ Area(d,x) &= \alpha(d)x^{\beta(d)} + \gamma(d)x, & x > X(d) \end{aligned} \tag{5.1}$$

ここで，$Area(d,x)$ は口唇からの距離が d で声道横断長が x のときの声道断面積である。$\alpha(d), \beta(d), \gamma(d)$ は口唇からの距離 d によって変化する係数である。$\alpha(d), \beta(d)$ は従来の $\alpha - \beta$[63] と同様の係数であるが，本モデルでは上下の歯列間の空間に対して新しいパラメータ $\gamma(d)x$ を導入した。$X(d)$ は舌が上顎の歯列から離れるときの声道横断長であり，x が $X(d)$ より大きくなると，$\gamma(d)x$ が $Area(d,x)$ の計算に加えられる。この項により，開口に伴い舌が上顎の歯列から離れるときに声道断面積が不連続に増加する現象を表現できる。また，本モデルでは口唇の変形が考

慮されていない。つまり，声道の正中断面のみを対象としているため，口唇の丸めを再現することができない。したがって，/u/，/o/など丸めを伴う母音の生成に際しては，口唇部にMR画像から抽出した口唇開口部の面積をそのまま用いた。

音源部におけるF0の変化は，調音モデルの声帯長に基づいて，2質量モデルのパラメータである質量とスティフネスを変化させることによって制御した。声帯長には，調音モデルの最終状態における声帯筋 (vocalis) の長さを用いた。

二つの質量要素の単位長当たりの質量は，声帯長の変化により声帯が均一に伸縮すると仮定すると，声帯長の変化に逆比例すると考えられる。この関係を以下の式で表した。

$$\frac{m_i}{m_{i0}} = \frac{l_{g0}}{l_g}, \quad i = 1, 2 \tag{5.2}$$

ここで，$m_i (i = 1, 2)$ は声帯長が l_g のときの質量，m_{i0} と l_{g0} は質量及び声帯長の初期値である。また声帯組織のスティフネスは，伸長の少ない領域と伸長の多い領域で上昇率の異なる非線形の関数である。[34],[64]。従って，本モデルでは，声帯長の変化に対する各質量要素のスティフネスの変化は3次の多項式で近似した。

$$\frac{k_i}{k_{i0}} = a\left(\frac{l_g}{l_{g0}}\right)^3 + b\left(\frac{l_g}{l_{g0}}\right)^2 + c\left(\frac{l_g}{l_{g0}}\right) + d, \quad i = 1, 2 \tag{5.3}$$

ここで，a, b, c, d は各項の係数，k_i と k_{i0} はスティフネス及びその初期値である。

5.3 モデル形状の構築とパラメータの推定

調音モデルを構築し音声を生成するために必要な資料は，発話器官の基本形状，モデルの駆動に用いる筋収縮力パタン，音声合成に用いる声道と音源のパラメータ，などである。これらの資料を得るために，種々の実験に基づいて計測と推定を行った。また，1名の男性被験者のデータのみを使用したのは，個人の特性をなるべく忠実に再現することを目的としたためである。

5.3.1 発話器官の形状計測と筋電信号計測

調音モデルを構成するためには，発話器官の形状，筋の付着位置などの形態的データと，駆動源となる筋収縮力パタン，筋スティフネスなどの生理的データが必

要となる。本研究では，MRI を用いて発話器官と声道形状の計測を行い，更に筋電信号の計測を行った。二つの実験に使用した発話タスクは，日本 5 母音/aiueo/と母音/a/の下降音階であり，被験者は 43 歳の男性である。これらの計測資料と解剖学的知識に基づいて，一人の個人を対象としたモデルを構築し，筋電信号から筋収縮力を求めるためのパラメータを推定した。これら資料は，音声合成パラメータの推定，及び，モデルの評価実験にも使用した。

MRI 実験により発話器官の形状と声道輪郭，断面積関数の計測を行った。5 母音の計測には，Spin Echo 法により 3mm 間隔で 7 スライスの矢状断面を撮像した。この中から正中矢状断面を選んで，器官形状を抽出した。更に，5mm 間隔で声道全体の冠状断面を撮像し，声道断面積の計測を行った。これらの 5 母音の計測においては，矢状方向の撮像で 2min，冠状方向の撮像に 4min の撮像時間を要した。その間，被験者は声道形状を保持しながら持続発生と息つぎを繰り返した。母音/a/の下降音階の測定には，Field Echo 法を使用して，10mm スライス厚を持つ正中矢状断面の撮像を行った。1 回の撮像時間は 4s であり，各ピッチごとに 1 回の撮像を行った。これらの画像より，舌の形状，喉頭軟骨の位置，声帯長などの抽出を行った。なお，この母音の下降音階の計測では，約 1.5 オクターブの範囲で F0 を変化させた。また，喉頭位置及び声道形状，F0 との相互関係のモデル化を容易にする目的で，被験者は F0 の下降において咽頭の上下運動を強調した発声を行った。

筋電信号の計測には，同様の 5 母音と下降音階の発話タスクを使用し，2 回に分けて実験を行った。1 回目の実験には，X 線マイクロビーム装置を併用し，発話器官の位置と形状をモニタした。被験筋は，オトガイ舌筋後部 (GGp)，茎突舌筋 (SG)，舌骨舌筋 (HG)，オトガイ舌骨筋 (GH)，顎二腹筋 (ABD) であり，すべて針金電極を用いて筋電信号を記録した。2 回目の実験は，筋電計側のみを単独で行った。被験筋は，オトガイ舌筋前部 (GGa)，胸骨甲状筋 (SH)，輪状甲状筋 (CT) である。GCa に対しては，口腔底に設置した表面電極，その他の筋については，針金電極を用いて測定した。なお，この 2 回目の実験では，1 回目の実験際に録音した音声を被験者に聞かせて，できるだけ同じように発声するように指示した。

5.3.2 計測結果と音声合成パラメータの推定

図 5.5 は 5 母音における器官形状と声道輪郭のトレース図である。

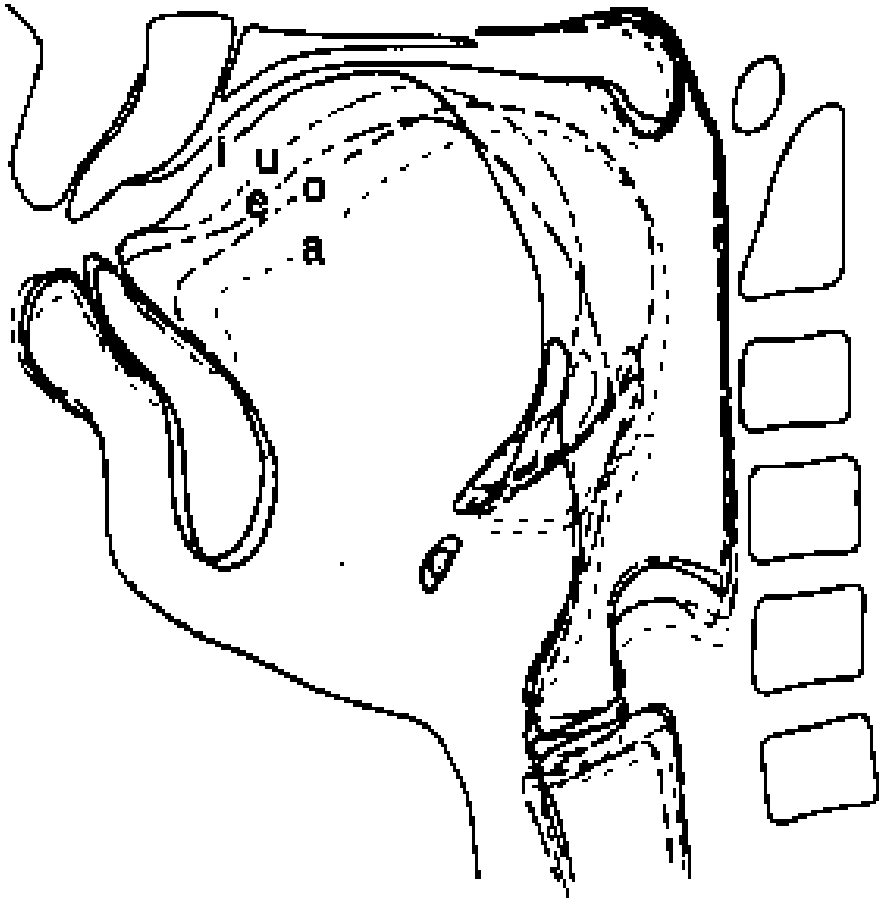


図 5.5: Tracings of mid-sagittal MR images of Japanese five vowels.

MRI では歯列が撮像されないが，図中には安静時の画像から歯の形状を推定して示してある。この図から，母音/u/の舌形状が/i/に近く，また，母音/a/の舌の位置が/o/より後方にある，などの被験者固有の特徴があることが分かる。図 5.6 は母音/a/の下降音階における MR 画像のトレース図である。図中には舌，下顎 (mandible)，舌骨 (hyoid bone)，甲状軟骨 (thyroid cartilage)，披破軟骨 (arytenoid cartilage)，輪状軟骨 (cricoid cartilage)，頸椎 (cervical spine) の位置変化を示してある。この図より，母音/a/を下降音階で発声した場合，以下に挙げるような現象が起こっていることが分かる。1) 喉頭の下降が生じ，声道長が増大する。2) 舌骨が舌と共に後下方へ移動する。3) 狭めの位置が軟口蓋の付近から咽頭腔まで移動

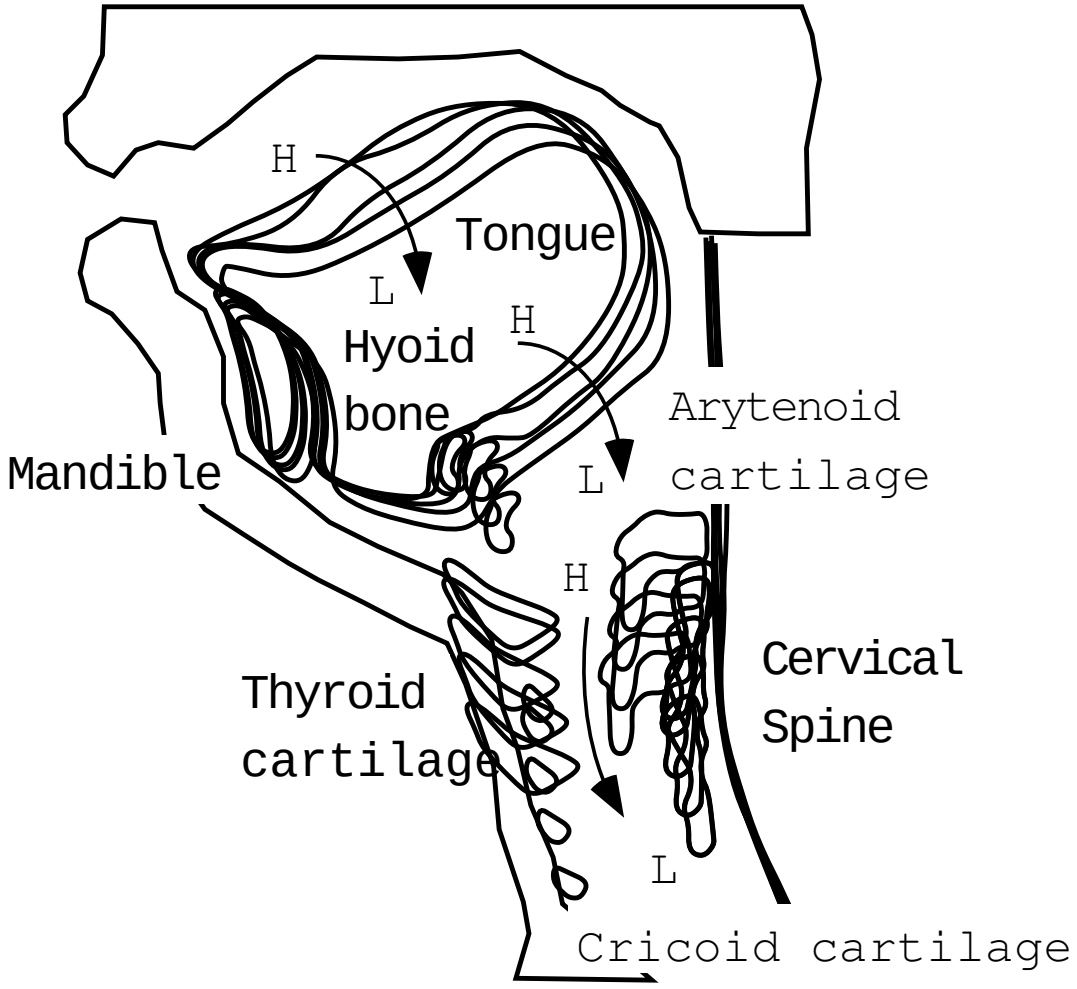


図 5.6: Tracings of mid-sagittal MR images during sustained vowel /a/ superimposed for five F0 levels.

する。4) 前室の体積は増大し後室の体積は減少する。

図 5.7 に 5 母音を発生したときの筋電活動を，図 5.8 に母音/a/の下降音階を発

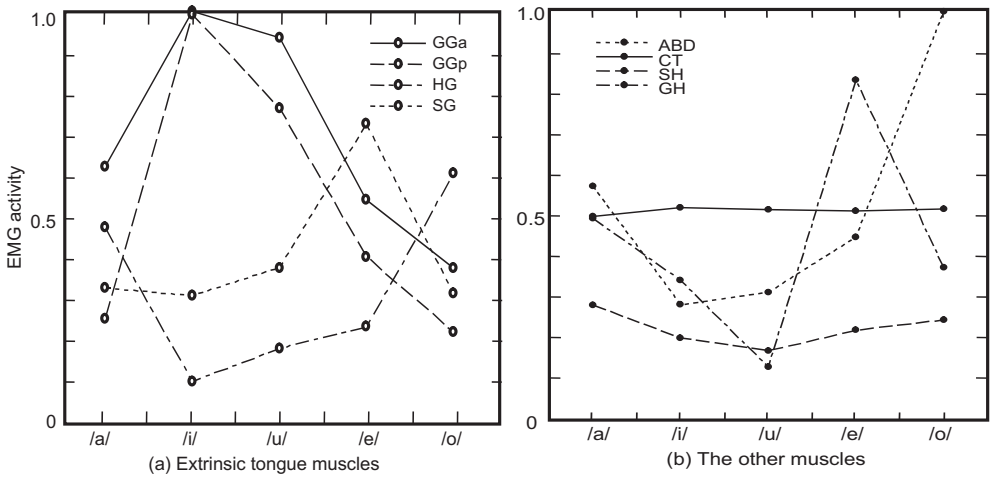


図 5.7: Normalized EMG activites of (a) the extrinsic tongue muscles and (b) other muscles during sustained production of Japanese five vowels with constant F0 levels. EMG values are normalized by the maxima for the two experimental tasks.

声したときの筋電活動を示す。

なお，筋電信号は全波整流した後 200ms の時間窓で平滑化し，各発話の定常区間における平均値を計算し，更に各筋ごとに最大値で正規化を行った値である。

図 5.8 より同じ母音/a/を発声しているにも関わらず，外舌筋など喉頭以外の筋が大きく変化していることが分かる。F0 の下降と共に活動が上昇している筋 (HG, SG) は舌を後方（背側）へ引くための筋であるが，同時に舌骨も後方へ引く効果を持つ。その結果，甲状軟骨を後方へ回転させ F0 下降を促進させる作用が示唆される。また，F0 の高い領域で活動している筋 (GGp, GH) は，舌骨を前方 (腹側) に引き，甲状軟骨を前方に回転させ F0 を上昇させると考えられる。SH は舌骨を下方 (尾側) へ引く作用を持ち，その結果として喉頭の下降を引き起こすと考えられる。以上の結果は，5.1 節に示したメカニズムを支持している。以上の結果をもとに，筋活動から収縮力を求めるための変換係数，及び筋のスティフネス値を推定した。これらのパラメータ値は，実測された筋電信号を入力として計算されたモデルの形状が，同じ発話タスクを行ったときに撮像した MR 画像の正中矢状断

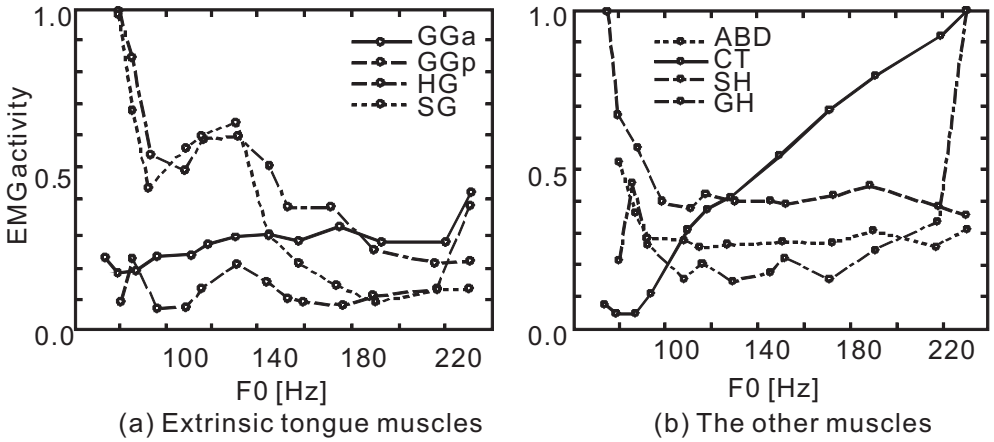


図 5.8: Changes of EMG activities of (a) the extrinsic tongue muscles and (b) other muscles during sustained vowel /a/ with a descending F0 scale. The EMG values are plotted in the same normalized axis.

面画像に重なるように調整することによって決定した。また，本モデルで導入されているにも関わらず今回測定できなかった筋の活動量については，MR 画像の形状の変化を参考にして推定した。

音声の合成に必要なパラメータは，声道横断長から断面積関数へ変換するための式 (5.1) の係数及び，音源の生成に用いる 2 質量モデルのパラメータである。式 (5.1) の係数は，5 母音発声時に撮像した声道の冠状断面画像に基づいて推定した。5mm 間隔で撮像された冠状断面は 1mm ごとのスライス間補間を行い，3 次元に再構成した。得られた 3 次元画像より声道の中心線に垂直な断面を求め，声道断面積及び声道横断長を測定した。図 5.9 に測定結果を示す。この結果を式 (5.1) に当てはめ，声道断面積の推定誤差が最小になるように各係数を求めた。X(d) は，すべての位置 (d) で 1.2cm として計算した。図 5.10 にパラメータ α, β, γ の推定結果を示す。

2 質量モデルのパラメータは二つの質量要素の質量及びスティフネスを除いて石坂らの基本的なモデルと同じ値を用いた。声帯長から質量及びスティフネスを求める式 (5.2) と式 (5.3) の係数は，母音の下降音階の MRI 実験より得られた F0 と声帯長の関係に従って推定した。各係数は，式 (2)，式 (3) を用いてモデルにより生成される F0 と声帯長の関係が，推定データに一致するように最適化した。ただし，MR 画像データの解像度が十分でなく，解剖学上の声帯長を正確に測定する

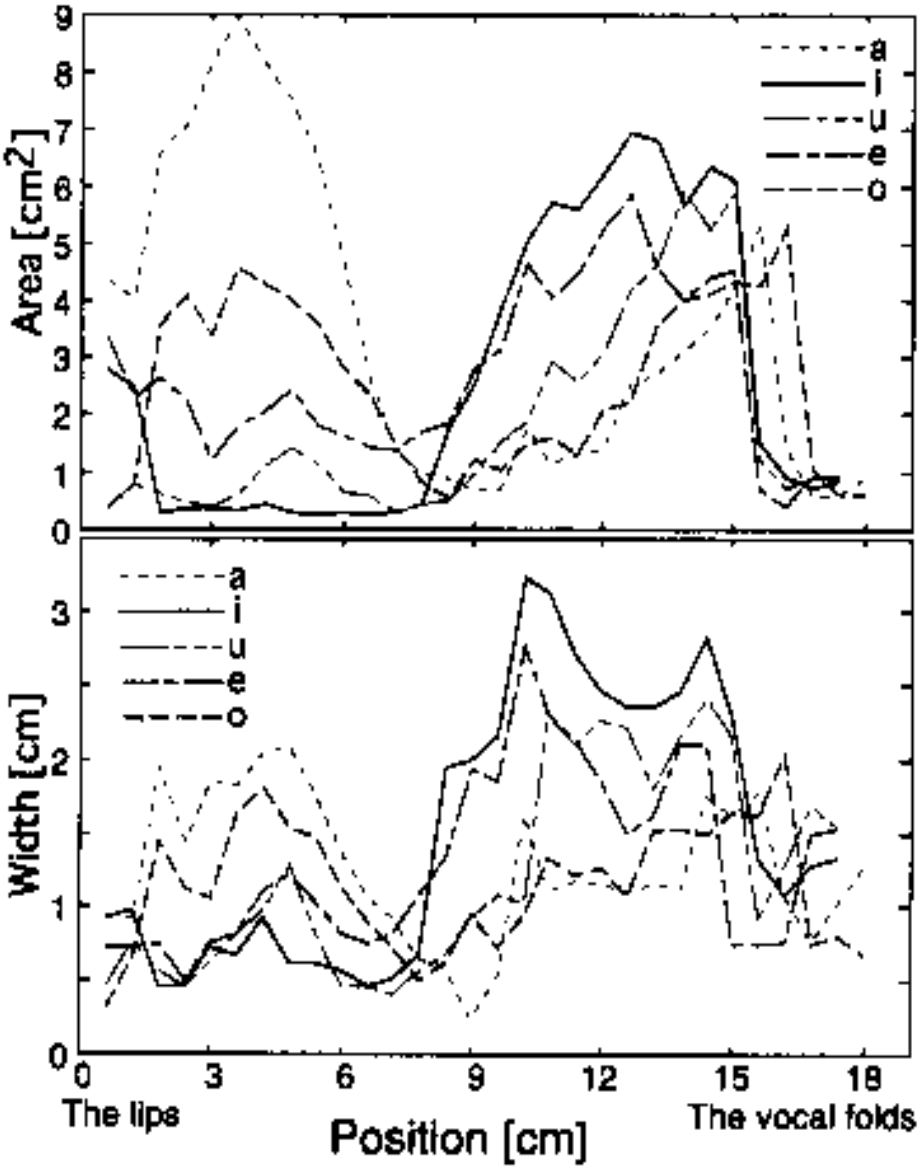


図 5.9: The vocal tract area functions (upper) and the mid-sagittal vocal tract widths (lower) for Japanese five vowels, measured from reconstructed 3D MRI data.

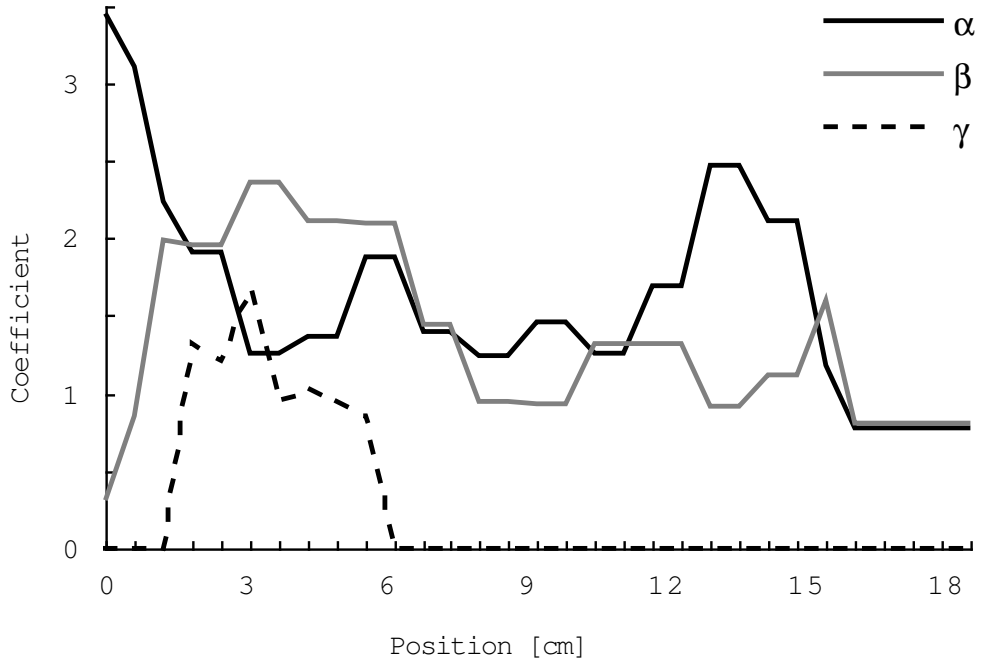


図 5.10: Estimated three coefficients (α, β, γ) for the conversion from mid-sagittal vocal tract shape to the area function.

ことは困難であった。MR 画像上において計測した声帯長は，甲状軟骨の前下端付近の点と披裂軟骨の前下端付近の点との距離を示すもので，声帯長の真値とは比例関係にあると見なせる値である。図 5.11 に推定した係数により計算された声帯長の変化に対する質量及びスティフネスの変化を，また図 5.12 にそれらのパラメータを用いて生成した音源波形の F0 と声帯長の関係及びパラメータ推定に用いた測定値を示す。図 5.12 よりモデルにおける声帯長と F0 の関係が実測値とほぼ一致することが分かる。

5.4 モデルの評価実験

5.4.1 5 母音と固有ピッチの生成

本論文に示す調音モデルは母音生成時の発話器官全体の位置形状を再現するモデルである。このモデルの動作を評価するために，はじめに筋電信号を入力としてモデルから 5 母音を生成する実験を行い，合成音のホルマント周波数を実音声と比較

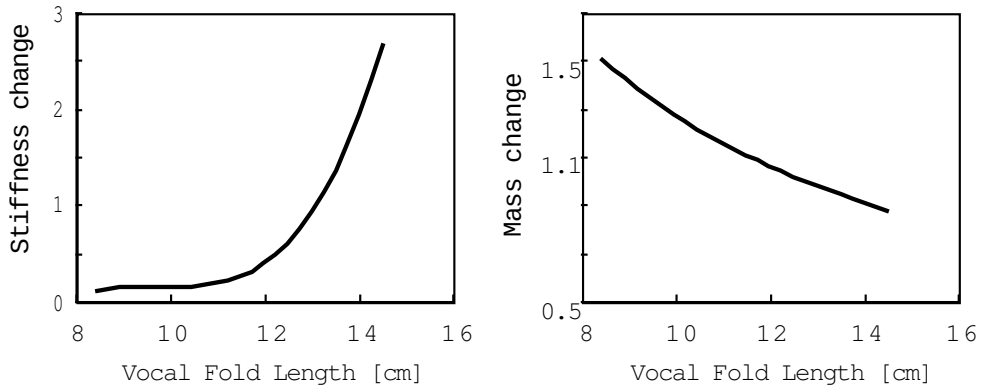


図 5.11: Estimated mapping functions to obtain the mass and stiffness parameters of the two-mass model from the change in vocal fold length.

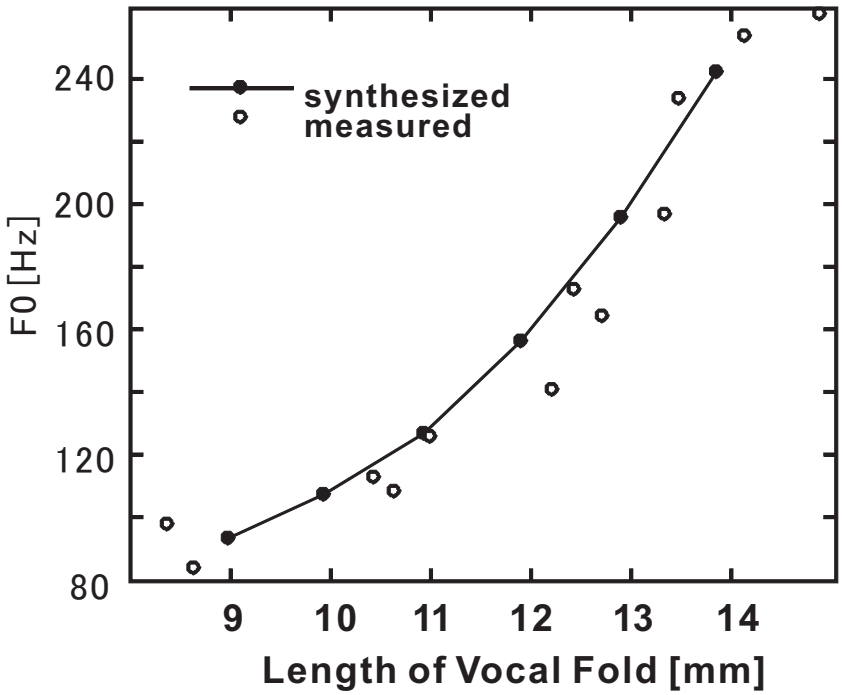


図 5.12: The relationship between F0 and vocal fold length derived from MRI measurement and the synthesis by the model. The open circles indicate the measured data, and the solid line shows the result from vowel sysnthesis by the model. The measure for vocal fold length corresponds to the distance between the landmarks on the arytenoid and thyroid cartilages, which is proportional to the anatomical length.

した。次に，母音の固有ピッチの現象を再現することを目的として，母音の調音動作における F0 の変化を調べた。これには，調音動作を強調した母音/a/，/i/，/u/を生成し，調音動作の F0 に及ぼす影響を検討した。

方法

X 線マイクロビーム [65] と同時に計測した 5 母音発声時の筋活動 (図 5.7) をモデルの入力として 5 母音を生成し，合成された音声と実測音声とのホルマント周波数の比較を行った。

筋電計測と MRI 実験では，ともに音声の測定が行われたが，MRI 測定時の音声は騒音のために分析ができないので，筋電計測実験で測定した音声を比較に用いた。実測音声のホルマント周波数の分析は，F0 の影響を除くため，1 ピッチより短いデータから正確に求めることができる MCLP[66] を用いた。音声のサンプリング周波数は 10kHz，LPC の次数 12，分析区間長は 24 サンプルであり，各母音の定常な部分の中心付近 50ms に含まれる区間を用いて求めた。母音/a/，/i/，/u/の調音動作の強調には，各母音の調音に関する主働筋である外舌筋を 30%増大させる操作を行った。/a/では SG，HG の活動を，/i/では GGa，GGp の活動を，/u/では SG，GGp の活動を増大させた。また，/u/の口唇の開きも調音動作の強調のために 20%減少させた CT 筋はすべての発声で一定とした。

結果及び考察

図 5.13 に合成音声のホルマントと，実音声のホルマントの比較を示す。

F1，F2，F3 の合成音声と実音声との誤差は，平均で 10.7%であった。/a/と/o/の合成音声は F2 が若干高く母音が中性化する傾向にあり，/i/と/u/では反対に調音が強調される傾向が見られる。しかし，他の母音の領域に移動するほどの誤差ではない。よって，本モデルは筋電信号を入力として実測音声に近い 5 母音を生成することができることが分かる。個人性に関しては，合成音声のホルマントは全体的に低く，成人男性の傾向 [67] を示している。しかし，合成音声を聴取したときの筆者らの主観的な評価では，個人を識別できるほどのホルマント周波数の一致は得られていない。

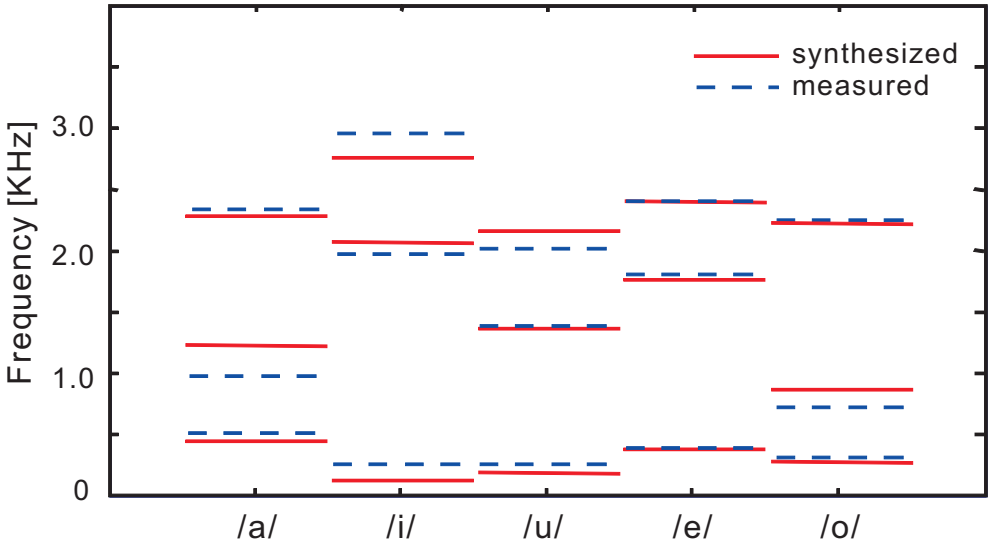


図 5.13: Lower three formant frequencies of the measured and synthesized Japanese five vowels. The solid and dashed lines indicate synthesized and measured values, respectively.

ホルマント周波数が正確に一致しない理由として複数の要因が考えられる。舌モデルの 2 次元化などの調音モデル作成上の単純化や，筋電計測と MRI 実験における調音状態の相異などがまず考えられる。それ以外に，MR 画像からの声道断面形状の抽出にも問題があると考えられる。本研究で用いた 3 次元画像は 5mm 間隔で撮像した冠状断面から 1mm 間隔になるように補間し再構成している。そのため，冠状断面と垂直な断面となる咽頭部の声道断面では前後方向の分解能が低下し，切り出された画像の輪郭は不明瞭なものとなる。このことから，式 (5.1) の推定に用いた声道断面形状も抽出誤差を含んでいると考えられる。このため，咽頭部の声道断面が実際より大きく推定され，図 5.13 に示すような誤差が生じたのではないかと思われる。

強調母音の生成結果を表 5.2 に示す。強調母音は，それぞれの母音のホルマントパターンが互いに分離する方向に移動しており，音響的にも強調されることが示されている。合成音声の F0 は，/a/で低く，/i/，/u/で高い。強調された母音では，この傾向が顕著になっている。従って調音を強調する動作はホルマント周波数に現れた母音特徴ばかりでなく，母音の固有ピッチの傾向をも強調する結果が得られている。また，母音を強調したときの舌骨の位置は，狭母音で前方へ，広

表 5.2: Acoustic results from a simulation of vowel enhancement.

Utterance	F0[Hz]	F1[Hz]	F2[Hz]	F3[Hz]	Vocal fold length [cm]
/a/	138.7	494	1,212	2,304	1.170
/a/enhanced	135.7	496	1,132	2,400	1.165
/i/	143.6	222	2,105	2,709	1.197
/i/enhanced	148.4	213	2,140	2,716	1.207
/u/	140.6	236	1,489	2,202	1.192
/u/enhanced	141.6	187	1,444	2,208	1.200

母音で後方に移動している。これらの結果は,「舌と喉頭との舌骨を介する接続により母音の固有ピッチが生じる」という理論 (tongue-pull theory)[59] と一致すると考えられる。

5.4.2 F0 下降に伴う母音のホルマント変化の生成

F0 調節による声道形状の変化及びホルマント周波数の変化を再現することを目的として，母音/a/の下降音階の音声を生成した。生成された声道形状と MR 画像，及び合成音声と実音声とを比較することにより，モデルの評価を行った。

方法

X 線マイクロビームと同時に計測した母音/a/の下降音階発声時の筋活動 (図 5.8) を入力として，母音/a/の下降音階を生成した。実音声のホルマント周波数の分析は前項と同様に MCLP を用いた。

結果及び考察

モデルにより生成された声道形状の変化を図 5.14 に，ホルマント周波数 (F1, F2 ,F3) と F0 の変化を図 5.15 に示す。ただし，最も高い F0 を発声したときの測定音声のホルマントは，声帯の閉鎖区間が短いため，安定に抽出することができなかっ

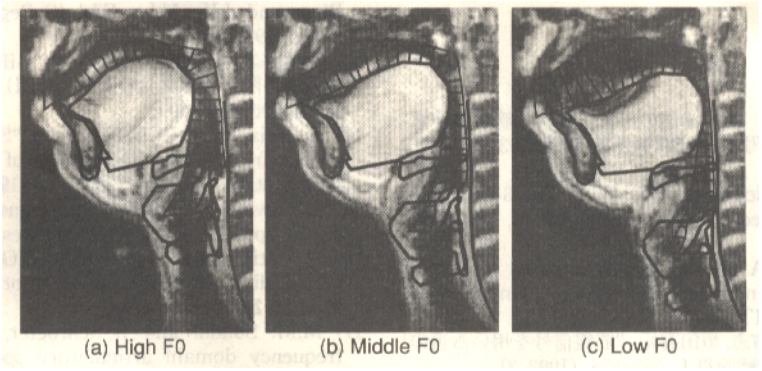


図 5.14: Examples of the model profile showing the effect of F0 change on the articulatory and laryngeal configuration.

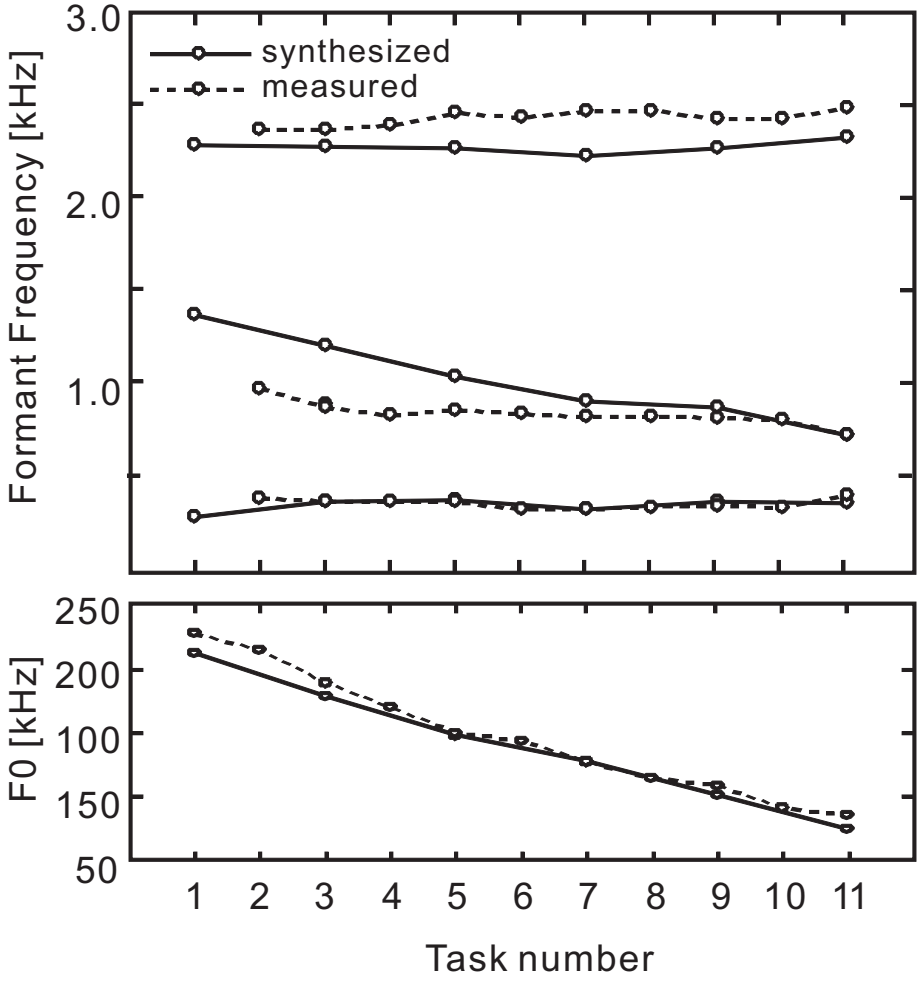


図 5.15: Changes in formant frequencies for the measured and synthesized vowels during sustained vowel /a/ with a descending F0 scale.

た。従って，図 5.15 にはその値が欠損している。図 5.14 の背景はそれぞれに対応する F0 を発声したときに撮像した MR 画像である。

図 5.14 では，異なる F0 においてモデルと MR 画像における声道形状がほぼ一致しており，F0 の下降に伴う舌の後方移動と喉頭の下降が再現されていることが分かる。図 5.15 から，この被験者の F0 下降時のホルマント周波数変化は，F2 の下降 F3 の上昇などの傾向が見られることが分かる。本モデルの結果においても，各ホルマントの遷移の傾向は正しく再現されている。しかし，F2 の変移の傾斜はかなり急であり，測定値と異なっている。この原因は，MR 画像を用いて喉頭の上下動に関するパラメータを調節したにも関わらず，筋電計測実験における音声を比較に用いたためではないかと考えられる。更に，このような F0 を大きく変化させるような発声を行う場合，F0 の非常に高い領域及び低い領域では舌の過緊張が起こり，通常の母音発話とは異なる舌の変形が生ずる可能性がある。従って，式 (5.1) のような声道横断長と声道断面積を直線的に対応づける式に対して，持続母音発声時の声道形状から推定したパラメータをあてはめた結果，極端な声道の変形を忠実に再現することができなかったのではないかと思われる。

5.5　本章のまとめ

本章では，前章で明らかにした F0 調節機構を有し，舌骨を介した喉頭と舌との力の授受を考慮した発話器官の２次元正中矢状断面のモデルを提案した。筋活動量を入力とした音声合成実験により，F0 調節が声道形状に影響を与える現象，および声道形状制御が F0 に影響を与える現象を再現できることを確認した。このことから，前章で明らかにした F0 調節機構により F0 調節と声道形状制御との間に相互作用が生じること，および本モデルにてその相互作用を実現できることが明らかになった。また，本モデルではパラメータを特定話者に最適化することで，個人性を有する音声の合成を試みた。しかし，日本語５母音の合成音声と実音声の第 1-3 フォルマント周波数の平均誤差は 10.7%となり，個人を識別できる程度の一致は得られなかった。このことから，実用化を行うには音質の改良が必要であることがわかった。

本モデルでは可能な限り計測されたデータを用いてパラメータを設定し，計測

が困難なパラメータについては解剖学的知識を基に手動で調節を行うことでモデルの構築を行った。しかし，本モデルの構築に使用したデータは多岐にわたり，中には計測が容易でないデータも含まれる。よって，全てのデータを収集することは被験者に大変な負担を強いるものであった。異なる話者から少しずつデータを収集することも考えられるが，発話の方法には個人差があり，異なる話者から得られたデータから構築されたモデルには信頼性に問題が生じる。そのため，本研究では１名の話者のモデルを用いて検証を行った。よって，F0 調節に関連する発話機構の解明は行うことはできたが，F0 変化によるフォルマント周波数の変化量などについて一般的な傾向として定量的に評価を行うことはできなかった。今後，機会があれば複数の話者のモデルを構築し，これらについても明らかにしていきたい。

第 6章

3次元MRI動画を用いたフィルタ(声道形状)モデルの高精度化による音声の高品質化

前章では，喉頭による F0 調節と舌による調音動作との間に相互作用が存在し，発話器官の音声生成モデルを用いることでそれらを表現できることを示した。しかし，前章のモデルを用いて生成された音声は，定量的にはフォルマント周波数の誤差が 10.7 %と大きく，実用化という観点からは不十分な音質であった。このことから音質が本方式の実用化へのボトルネックになると考えられる。本章では，この実用化へのボトルネックを解消するために行った合成音声の高品質化について述べる。前章のモデルでは，MRI により計測された 2 次元正中矢状断面の声道形状の再現を目標にモデルの構築を行った。しかし，音声の合成には声道の 3 次元形状より得られる声道断面積関数が必要である。また，声道の 2 次元矢状断面から声道断面積関数への変換方法については，未だ確立された方法が存在しない。よって，2 次元から 3 次元への変換時に大きな誤差が生じた可能性が疑われる。前章のモデルを含めて，これまで多くの音声生成モデルにおいて 2 次元正中矢状断面のモデルが用いられてきた。その理由は，3 次元の声道形状を考えるとモデルが複雑になることに加え，発話動作時の 3 次元声道形状を測定することが非常に困難であったことが挙げられる。これに対し，近年 3 次元 MRI 動画撮像法が開発された。本章では，この 3 次元 MRI 動画データを用いてフィルタ (声道形状) のモ

デルを高精度化することにより実用可能な音声の合成を目指す。発話器官の動作を再現する力学的モデルの 3 次元への拡張は党らによって進められており [12]，舌の有限要素法の改良などにより高精度に発話器官の動作を再現することが可能になってきている。しかし，音質については，個人性を再現できるまでには至っていない。よって，本章では計算コストも考慮し実用化へのアプローチとして，発話時の声道断面積関数を忠実に再現する声道断面積モデルを提案する。

ただし，本章にて提案するモデルにおける韻律には直接入力された F0 が使用される。よって，これまで明らかにしてきた F0 調節と声道形状制御との相互作用を再現する機構を含むものではない。これは，本章の目的がフィルタ部について注目し，３次元声道形状データを用いて高精度に声道断面積を再現することにより音質の改善が可能であることを確認することであるためである。評価実験では同じ単語を同じ様に発話した時の F0 と声道形状を用いて合成を行うため，相互作用を含んだ自然な音声が合成することができる。しかし，自然性の高い自由な文を生成するには，相互作用を再現するために，党らのような 3 次元の発話器官モデルに拡張するか，発話器官モデルのシミュレーション結果に基づいて得られた知見を基に声道断面積モデルと F0 を同時に制御する機構を新たに設ける必要がある。現状では，計算コストなどを考慮すると，実用化には後者が有望であると考えている。よって，本モデルではこれらを可能にするために，発話器官の位置と声道断面積モデルのパラメータとの変換が容易になるように声道内の物理的な位置とモデルのパラメータとの位置関係の情報が失われないことを考慮してモデルを作成した。本モデルを用いて幾つかの単語を合成し，実音声と比較することで，音質の評価を行った。

6.1　本方式の概要

本方式の概要を図-6.1 に示す。

本方式では，近年開発された 3 次元 MRI 動画から得られた声道断面積データを用いて，ソース (声道形状) 部の高精度化を行うことで，高品質な音声の合成を行う。音声合成への応用を行うには，3 次元 MRI から計測された声道断面積関数データを少ないパラメータにモデル化する必要がある。従来より幾つかの声道形状を

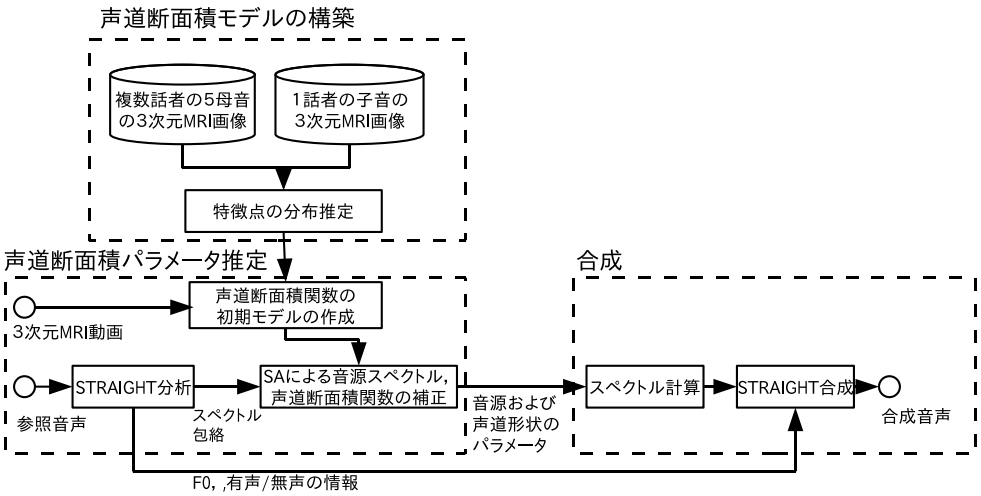


図 6.1: 本方式の概要

SA はシミュレーテッドアニーリングを示す。

表現するモデルが提案されているが，それらは 2 次元正中矢状断面の声道形状を幾何学的にモデル化したものが多い。3 次元 MRI 動画より得られた声道は複雑な形状をしており，実形状に忠実に声道形状を表現するにはそれらのモデルは適しているとは言えない。本方式では，はじめに 3 次元 MRI 動画より得られた声道断面積関数を精度良く表現できる声道断面積モデルの構築を行う。モデルには，接続面の位置と面積をパラメータとする円錐台の連続体を用いる。声道形状から声道伝達関数への計算に円錐台の連続体から声道伝達関数を計算する手法 [35][36] を用いることにより，単位長の短い円筒管に展開する必要が無く，計算コストが少ないという利点がある。本論文では，この円錐台の接続箇所を声道形状を特徴付ける為に必要な点として特徴点と呼ぶこととする。

特徴点のパラメータは，F0 調節時の声道形状の変形などを実現するためにパラメータの変換などの処理を行うことを考慮すると，音素や話者によらず発話器官の位置と対応づけられる同じ基準で決定されることが望ましい。よって，あらかじめ複数話者の５母音および１話者の複数の子音の声道形状を解剖学的な基準で４つに分割し，それぞれの部分について全ての声道形状を表現するために最低必要な特徴点を求め，各特徴点の声道中心線の上の位置に対する分布を推定する。声道断面積関数から声道断面積モデルのパラメータへの変換はこの特徴点の分布を

用いて行う。

3 次元 MRI 動画を得るためには数百回の繰り返し発話を，仰臥位の姿勢，頭部の固定という特殊な環境で行う必要がある [68]。さらに撮像時には撮像装置から大きな騒音が発生する。その結果，繰り返し発話においてタイミングの揺れや調音の歪みが生じ、声道形状の測定誤差および発話の不自然性を生じる。また，声道形状のモデル化によっても誤差が生じる。これらの誤差を補償するために，3 次元 MRI 動画から得られた声道断面積モデルを初期値として，同じ単語を発話した参照音声のスペクトルと合成音声のスペクトルが等しくなるように，シミュレーテッドアニーリング (以下 SA 法とする) を用いて声道断面積モデルのパラメータ補正を行う。

合成部では，推定された声道断面積モデルのパラメータと音源スペクトルのパラメータ，および参照音声の STRAIGHT 分析結果より得られた F0 と帯域毎の有声/無声の比率を用いて STRAIGHT の音声合成方式に従い音声の合成を行う。

6.2 声道形状の計測

6.2.1 MRI による声道形状計測

本方式では，特徴点の分布計算および声道断面積推定において計測された声道形状を使用する。声道形状の測定には，3 次元 MRI 動画を用いた [42]。また，声道断面積関数の計測には竹本らの手法 [69] を用いた。以下、本研究で用いた声道断面積データについて示す。

6.2.2 定常母音

定常母音発声時の声道形状は特異点分布の計算に用いる。定常母音の声道断面積関数には，北村らが報告した 5 母音 11 名の実験データ [70] のうち 8 名のデータを使用した。これらのデータは，仰臥位で約 3 分間母音発声を持続しながら撮像した声道形状に，予め口腔造影剤を用いて撮像された歯列画像を補充し [46] 声道断面積を測定した結果である。表-6.1 に撮像条件を示す。

表 6.1: 定常母音の撮像条件

発話者	成人 8 名 (うち女性１ 名)
発話タスク	定常発話 (日本語５母音)
撮像条件	エコー時間 (TE): 11 ms 繰り返し時間 (TR): 3000 ms スライス方向: 矢状 スライス厚: 2 mm スライス間隔: 2 mm スライス数: 41 or 51 枚 撮像領域: 256 x 256 mm 分解能: 512 x 512 pixels

6.2.3 5 母音の連続発話

5 母音連続発話時の声道断面積関数には，竹本らの日本語 5 母音の連続発話を撮像した実験データ [42] を用いた。このデータは，1 名の話者が日本語 5 母音/aiueo/を連続発話した際の 3 次元の動画像に対し歯列画像を補充し声道断面積を測定したものである。撮像は仰臥位で発話を促すノイズバースト信号に合わせて，128 x 5 回，計 640 回の繰り返し発話により撮像されている。表-6.2 に撮像条件を示す。

6.2.4 子音を含む単語発話

子音を含む単語発声時の声道断面積は，母音の連続発話と同様の手法 [42] を用いて撮像した 3 次元 MRI 動画より測定した。歯列画像についても同様の手法により補充した。表-6.3 に撮像条件を示す。

MRI 動画撮像のフレームレートについては発話者への負担を減らすため 20 fps と遅くなっている。音声の合成を行うには調音動作の速い子音部ではフレーム間隔 50 msec は長すぎる。よって，3 次元 MRI 動画から得られた声道断面積モデルの初期値のパラメータをスプライン補間を用いて時間方向に再サンプリングを施し 10 msec 間隔のデータとしてパラメータの推定に用いた。

表 6.2: 日本語 5 母音連続発話の撮像条件

発話者	成人男性 1 名
発話タスク	日本語 5 母音/aiueo/の連続発話
撮像条件	エコー時間 (TE): 4.0 ms
	繰り返し時間 (TR): 1900 ms
	スライス方向: 矢状
	スライス厚: 6 mm
	スライス間隔: 4 mm
	スライス数: 20 枚
	撮像領域: 256 x 256 mm
	分解能: 256 x 256 pixels
	フレーム数: 56 フレーム
	フレームレート: 30 fps

6.3 声道モデル

6.3.1 声道断面積モデルの構成

図-6.2 に，6.2.3 節で示した方法により測定した 1 名の話者が連続母音「あいうえお」を発話した時の声道断面積の変化の例を示す。図には，連続母音の発話時に撮像した約 40 フレームの声道形状から見やすさを考慮して類似した結果を間引き，7 フレームの結果のみを記載した。この図より声道形状が，破線の位置を接続面とする円錐台の連続体として表現できることが推測される。主声道については，この接続面の位置を声道形状の特徴点とし，特徴点を接続面とする円錐台の連続体として声道断面積関数をモデル化する。

断面積関数の各特徴点は口腔・咽頭腔の形態的特徴を反映しており，健常な話者においては同様の位置に特徴点が存在すると推測される。複数の話者，音素に共通な特徴点の決定方法については，次節に示す。

口蓋扁桃（扁桃腺）については話者毎に大きさが大きく異なり，話者によっては存在を明確に判断できない場合がある。本研究では，顕著な口蓋扁桃を有する

表 6.3: 子音を含む単語の撮像条件

発話者	成人男性 1 名
発話タスク	単語発話 (散る, 詐欺, シェフ, 逆, 除夜, 柿, 拒否, 島, 祖母, 手話, ペケ)
撮像条件	エコー時間 (TE): 3 ms 繰り返し時間 (TR): 950 ms スライス方向: 矢状 スライス厚: 4 mm スライス間隔: 4 mm スライス数: 24 枚 撮像領域: 256 x 256 mm 分解能: 256 x 256 pixels フレーム数: 18 フレーム フレームレート: 20 fps

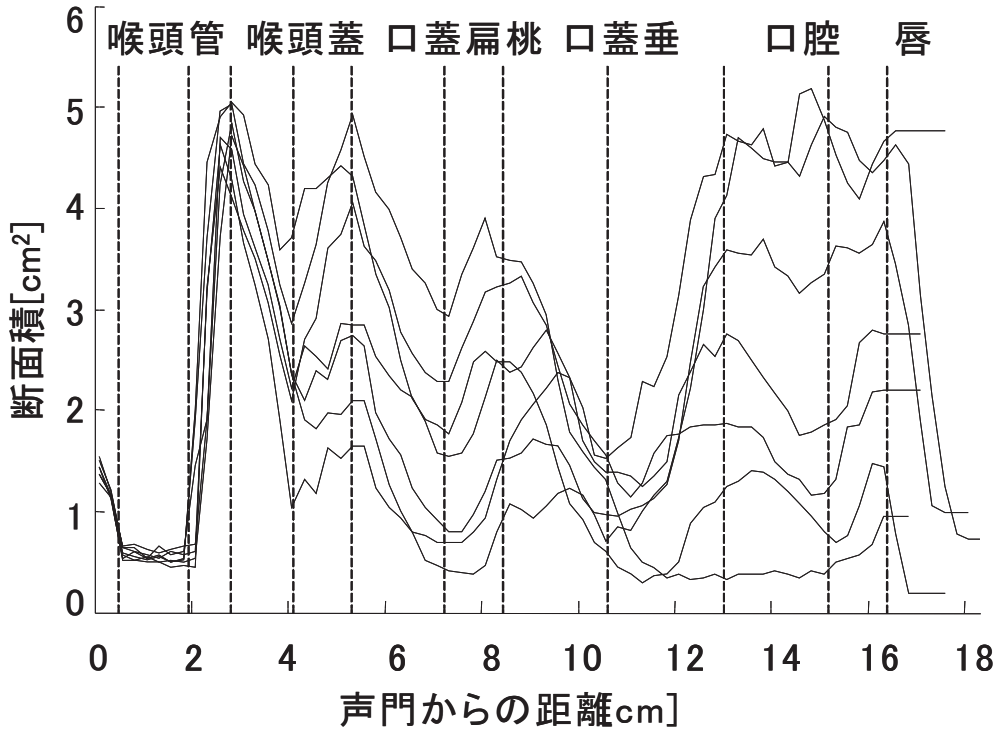


図 6.2:　母音連鎖発話時の声道断面積の変化

連続発話「あいうえお」を撮像した 3 次元 MRI 動画より抽出した声道断面積関数を重ねて表示

発話者を特徴点の分布推定実験に含めることにより，口蓋扁桃の形状を表現できるモデルを構築した。

声道形状から声道伝達関数への変換では，主声道，鼻腔以外にも梨状窩，副鼻腔も大きな影響を与えることが知られている [71][72]。本モデルでは主声道以外に梨状窩，鼻腔，副鼻腔をモデル化した。梨状窩は左右ともに等間隔の 4 つの円錐台の連続体でモデル化した。鼻腔は左右の鼻道の断面積を加算し 1 本の音響管で近似した。副鼻腔は鼻腔との接続部の円柱と副鼻腔と体積のみが等しく長さの十分に短い円柱から成るヘルムホルツ共鳴器に類した 2 管モデルで表現した。梨状窩と主声道との接続点は喉頭腔と中咽頭腔の境界から約 2.5 mm 声帯側の位置，鼻腔と主声道との接続点には中咽頭と口腔の境界を用いた。ただし，喉頭腔と中咽頭腔の境界には，ラベリングを容易にするため，喉頭腔と中咽頭腔の接続部近傍の口唇側で最も声道断面積が広くなる位置を用いた。

特徴点の決定方法

複数の話者の 5 母音を表現できる特徴点の決定方法を以下に示す。はじめに，声道伝達関数を正確に表現するために最低限必要な特徴点の数と位置を，複数の話者の 5 母音の声道断面積関数から求める。その結果より，声道中心線軸上の位置に対する特徴点の出現頻度を求める。つぎに，各々の特徴点の出現頻度分布が正規分布に従うと仮定し，G 個の特徴点の分布を式-6.1 に示す混合正規分布のパラメータとして推定する。パラメータ推定には EM アルゴリズム [73] を用いた。

$$\phi(x) = \sum_{k=1}^{G} \pi_k N_k(x; \mu_k, \sigma_k) \tag{6.1}$$

ここで，$N_k(x; \mu_k, \sigma_k)$ は，第 k 番目の特徴点の出現確率を表す平均 μ_k 分散 σ_k の正規分布の確率密度関数である。また，π_k は第 k 番目の特徴点の出現する比率を表し，

$$\pi_k \geq 0, \sum_{k=1}^{G} \pi_k = 1 \tag{6.2}$$

を満たす。混合正規分布の次数の決定には赤池情報量基準 (AIC)[74] を用いた。ここで得られた混合正規分布の次数が必要な特徴点の数となる。混合正規分布のパラメータを用いて声道断面積関数から特徴点を決定する方法は 6.3.3 節に示す。

各声道断面積関数における必要最低限の特徴点の位置と数は，以下の手順で求めた。まず、3 次元 MRI データより求めた声道断面積関数上の全ての計測点（以下データ点とする）を初期値とする。次に，1 点ずつデータ点を削除し，削除されたデータ点を前後のデータ点を用いて補間した時の伝達関数と初期値の伝達関数との誤差が最も少なくなるデータ点の位置を探索する。そして、その時の誤差が所定の閾値を超えない最小のデータ点数となる組み合わせを決定する。この結果が必要最低限の特徴点となる。データ点を N 点に削減した場合の最適な特徴点の位置を求めるには，N 点の位置の全ての組み合わせについて伝達関数を計算し，誤差が最小となる組み合わせを探索する必要がある。本研究では，$N+1$ 点の結果から N 点での最適位置の近似解を以下の手順で求めた。

1. $N+1$ 点の中の i 番目の点を削除する。
2. $i+1$ 番目の点を $i-1$ 番目の点の位置から $i+2$ 番目の点の位置の範囲で変化させ，伝達関数の誤差が最も少なくなる位置を探索する。つぎに，探索の対象とする点を $i+2$ 番目から $N+1$ 番目まで，および，$i-1$ 番目から 1 番目まで順次置き換えて探索を繰り返す。
3. 各点の位置が安定するまで 2 を繰り返す。
4. 削除する点 i を 1 から $N+1$ まで置き換え 1,2,3 を繰り返し，最終的に計算された伝達関数の誤差が最も少なかったデータ点の組み合わせを N 点の最適な位置とする。

ここで，伝達関数の誤差の評価には以下の式を用いた。

$$D = \sum_{k=1}^{K} \left(\log S_i(f_k) - \log S_d(f_k)\right)^2 \tag{6.3}$$

$S_i(f_k)$ は計測された声道断面積から計算された周波数伝達関数の初期値，$S_d(f_k)$ はサンプル点を削除した声道断面積から計算された周波数伝達関数，f_k は対数領域で等間隔に分布する周波数を示す。ただし，データ点数の削除を終了するための評価値には，音質との対応が判断しやすく，閾値が決定しやすいパラメータとして，初期値の伝達関数に対する第 1 から第 4 フォルマント周波数の相対誤差の最大値を用いた。終了判定の閾値は 5 %とした。

6.3.2 特徴点の分布推定

母音

特徴点の分布推定は，MR 画像を参考に，喉頭腔，中咽頭腔，口腔，口唇部に分けて行った。口唇部については口腔との境界，先端の 2 点とした。使用したデータは 6.2.2 節に示した声道断面積関数のデータである。

図-6.3 に各区間に与えるデータ点の数と声道伝達関数の歪みの関係を示す。発話者 8 名の 5 母音の結果を重ねて表示した。図-6.3(a) が喉頭腔，図-6.3(b) が中咽頭腔，図-6.3(c) が口腔の結果である。図に示す歪みは、最小の数のデータ点の組を決定するために用いる第 1 〜第 4 フォルマント周波数の相対誤差の最大値である。破線は閾値 (5 %) を示す。本結果より全ての母音について，喉頭腔，中咽頭腔は 1 から 2 点，口腔部は 2 から 4 点程度で表現できることがわかる。各部を表現するために必要なデータ点数の 8 話者 5 母音における平均値はそれぞれ，喉頭腔 1.4 個，中咽頭腔 1.5 個，口腔 2.5 個であった。

全ての発話者，全ての母音の特徴点の出現位置を，喉頭腔，中咽頭腔，口腔毎に声帯側の境界を 0，口唇側の境界が 1 になるように正規化し，出現頻度を表示した結果を図-6.4 に示す。図中の実線は EM アルゴリズムを用いて混合正規分布 (式-6.1) を当てはめた結果である。

表-6.4 に混合正規分布の次数 G(特徴点の数) を変化させたときの各部位ごとの AIC の値を示す。AIC が最小となるときの次数 G を各部ごとに必要な特徴点の数とした。喉頭腔は 3 次，中咽頭腔は 3 次，口腔は 4 次である。図-6.4 に示された正規分布は，この結果により決定された次数 G を用いて推定した結果である。

表 6.4: 特徴点の数 (次数 G) を変化させたときの各部の AIC

次数 G	1	2	3	4	5
喉頭腔	-63.57	-64.03	-71.35	-66.95	-63.81
中咽頭腔	-5.26	-18.91	-20.00	-18.84	-14.99
口腔	20.08	-3.21	-6.75	-6.96	-5.31

表-6.4 より口腔に関しては特徴点が 4 次の時と 3 次の時との AIC の差が少な

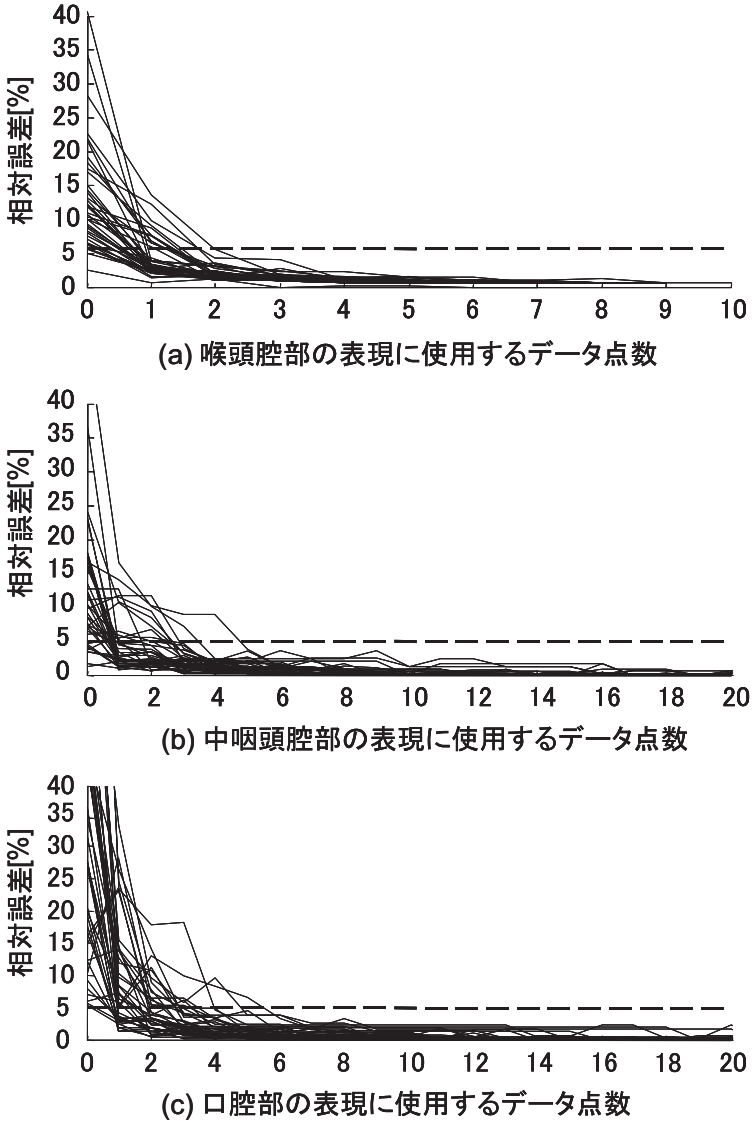


図 6.3:　データ点数と伝達関数の劣化歪の関係

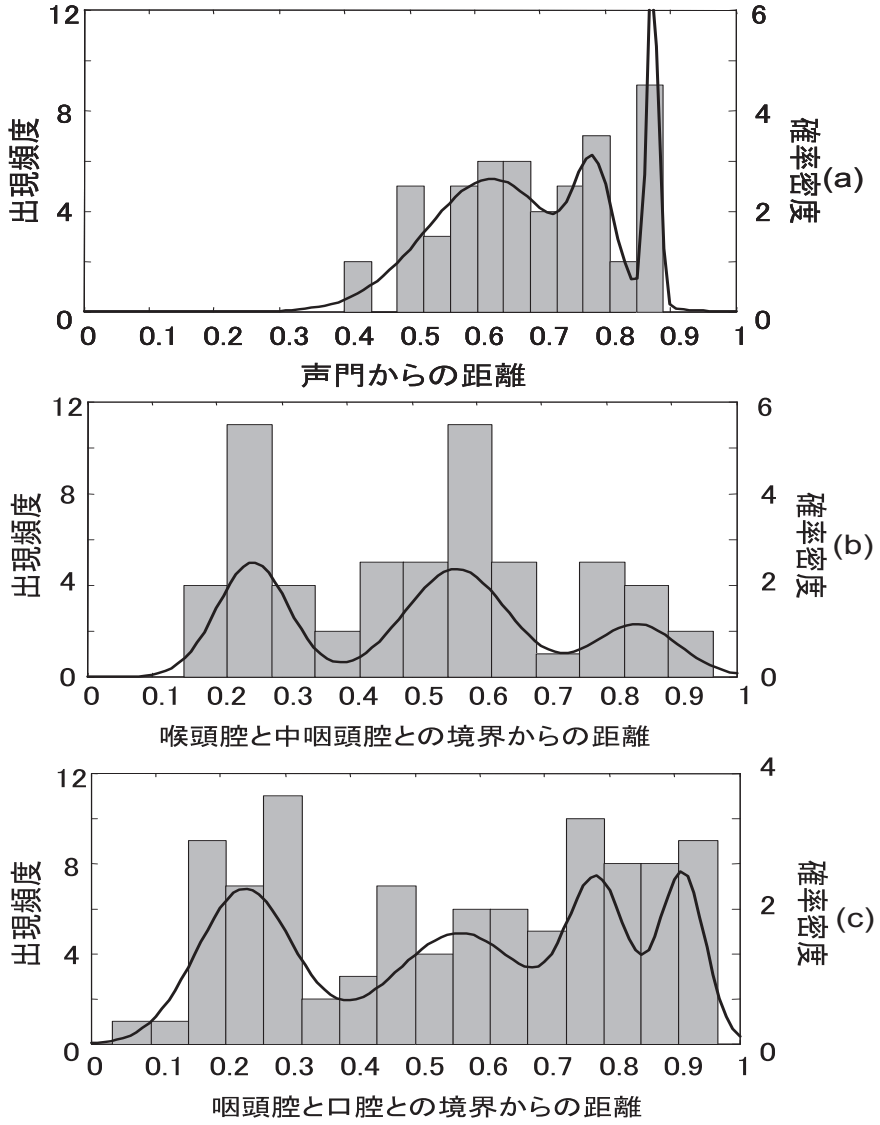


図 6.4: 特徴点の分布（a) 喉頭腔 (b) 中咽頭腔 (c) 口腔
実線は特徴点の位置の分布を混合正規分布としたときの確率密度関数

い。検証の為に，全ての発話者，全ての母音について，3 次の場合と 4 次の場合に得られた特徴点の分布を用いて 6.3.3 節に示す方法に従い声道断面積モデルを作成し，モデル化によるフォルマント周波数の相対誤差を計算した。表-6.5 に結果を示す。表-6.5 から，次数 3 の場合に F4 の誤差の最大値が非常に大きくなってい

表 6.5: 口腔の特徴点の数 (次数 G) とフォルマント周波数の相対誤差

次数 G		F1	F2	F3	F4
3	平均 (%)	2.3	1.6	1.0	1.2
	最大 (%)	8.1	7.8	4.2	15.4
4	平均 (%)	1.8	1.8	1.0	0.9
	最大 (%)	8.1	8.4	8.4	3.4

る。母音の場合は F4 も音質に大きく影響することを考慮して次数 4 の結果を採用した。表-6.6 に推定した次数 4 の混合正規分布 (式 6.1) のパラメータを示す。

表 6.6: 特徴点の分布を混合正規分布で近似した時の式-6.1 のパラメータ

	喉頭腔			中咽頭腔			口腔			
k	1	2	3	1	2	3	1	2	3	4
π_k	0.67	0.17	0.16	0.34	0.47	0.19	0.32	0.32	0.20	0.16
μ_k	0.63	0.78	0.87	0.26	0.57	0.85	0.24	0.57	0.78	0.91
σ_k	0.10	0.03	0.01	0.06	0.08	0.07	0.07	0.10	0.05	0.04

子音

母音と比較して子音は種類が非常に多く，多くの発話者に対して全ての子音の 3 次元 MRI 動画を撮像することは困難である。また，破裂子音や摩擦子音などの子音では，母音と比較し声道伝達関数の誤差が音質へ与える影響が少ない。以上のことから，話者 1 名の子音の声道断面積関数を用いて 6.3.2 節で得られた母音の声道断面積モデルを元に子音を表現できる声道断面積モデルへの拡張を行った。

口唇部については，口唇に調音点をもつ子音を表現できないため，接続部，先端に中間点を追加し 3 点とした。その他の部分については，実測された声道断面積関数から求めた声道伝達関数と母音の特徴点の分布を用いてモデル化した声道断面積関数から計算された声道伝達関数を比較し，モデル化による誤差が大きい場合は誤差を減少させる位置に特徴点の追加を行った。

子音への拡張に用いた声道断面積関数は，6.2.4 節で測定された単語に含まれる全ての子音の中心フレームでのデータである。モデル化前の計測された声道断面積から求めたスペクトルと，6.3.3 節に示す方法によりモデル化された後のスペクトルとの差から，モデル化により生じる誤差を計算した。結果を表-6.7 の上段（追加無しと記載）に示す。表中の F1〜F4 は，音源が声帯に存在する場合の第 1〜4 フォルマント周波数の相対誤差である。表中の歪みは音源が狭めの口唇側にある場合の対数スペクトルの歪みを示す。この結果より，声道断面積モデル (追加無し) では歪みが大きくなる音素があることがわかる。それらは，口蓋に狭めがある音素であった。子音の発声では母音と比較し極端な狭めがあり，その位置や形状の精度がスペクトルに大きく関与するため，狭めの部分の精度不足が，この誤差の原因ではないかと考えられる。よって，声道断面積が最も小さくなる位置（調音点）に特徴点を 1 点追加した。その結果を表-6.7 の下段に示す。この結果より，母音から得られた声道断面積モデルに口唇部と子音調音点に特徴点を加えることで子音についても精度の高い声道断面積モデルが構築できることがわかる。

表 6.7: 子音をモデル化した時のフォルマント周波数の相対誤差とスペクトル歪

		F1	F2	F3	F4	スペクトル歪
追加	平均 (%)	3.5	3.2	2.8	2.8	3.7 dB
無し	最大 (%)	20.0	14.1	9.6	10.5	9.7 dB
追加	平均 (%)	1.8	2.7	1.9	2.1	3.0 dB
有り	最大 (%)	6.9	9.2	10.4	11.4	5.3 dB

6.3.3 声道断面積関数から特徴点への変換

声道断面積関数から声道断面積モデルへのモデル化は，混合正規分布のパラメータを用いて，特徴点の位置と面積を求めることにより行う。以下に手順を示す。

1. 声道断面積関数および 3 次元 MRI の画像を参考にし，口唇，口腔，中咽頭腔，喉頭腔の境界を目視にてラベリングし，その位置および面積を初期モデルに加える。

2. 口腔，中咽頭腔の区間内で最も面積が小さくなる位置を調音点として初期モデルに加える。

3. 中咽頭腔と喉頭腔との境界から 2.5mm 声帯側を梨状窩の接続点として初期モデルに加える。

4. 表-6.6 に示す特徴点を初期モデルに加える。追加する特徴点の位置は，隣合う特徴点間を直線補間した声道断面積関数モデルと実測された声道断面積関数との差の自乗和が最小となるように組み合わせ探索により決定する。なお、この最適化は表-6.6 より求めた各々の特徴点の出現確率が他の特徴点の出現確率より高くなる範囲で行う。

6.4 声道断面積パラメータの補正

6.4.1 母音

高品質な音声を合成できる声道断面積モデルのパラメータを得るため，声道断面積モデルから計算された声道伝達関数が実音声の STRAIGHT 分析により得られたスペクトル包絡と一致するように SA 法を用いてパラメータの補正を行う。STRAIGHT により得られるスペクトル包絡には音源のスペクトルが含まれているため，スペクトルの比較を行うには音源のモデルが必要となる。声帯音源のモデルは数多く提案されている [32][75][34][76]。それらの多くは時間領域の音源波形モデルである。本研究では，周波数領域で誤差を評価しパラメータの補正を行う。よって，パラメータの変化とスペクトルの変化の対応が直接的な周波数領域のモデル

の方が望ましい。ただし，スペクトル形状の自由度が高すぎるモデルでは，声道形状の補正が正しく行えない。よって，以下に示す声帯音源のスペクトルモデル $V_s(f)$ を使用する。

$$Sf_1(f) = \frac{s_1 s_1^*}{(s-s_1)(s-s_1^*)} \tag{6.4}$$

$$Sf_2(f) = \frac{s_2^2}{(s-s_2)^2} \tag{6.5}$$

$$F_c(f) = 1 - \frac{\alpha}{1+e^{\frac{-\beta(f-f_c)}{f_c}}} \tag{6.6}$$

$$V_s(f) = (Sf_1(f) + Sf_2(f))F_c(f) \tag{6.7}$$

ここで，$s = j2\pi f, s_1 = 2\pi(jF_1 + B_1), s_2 = 2\pi B_2$ である。図-6.5 に 1 例を示す。

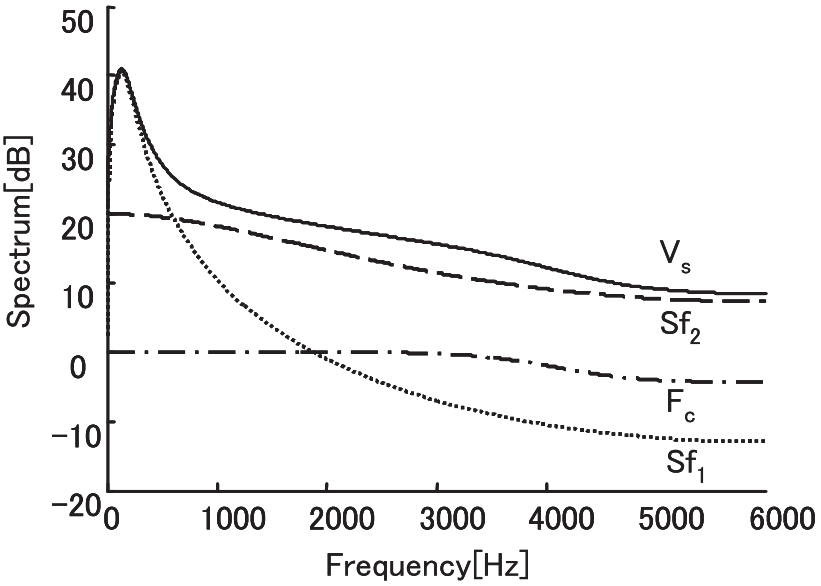


図 6.5: 提案方式の音源スペクトル V_s

式 (6.4) は音源スペクトル包絡の第 1 ピークの形状を表現している。F_1 はピークの位置，B_1 はピークのバンド幅を表す。式 (6.5) は約 1 kHz 以上の領域のスペクトルの傾斜を表す。式 (6.6) は高域部の補正を示している。α, β, f_c は，補正する周波数帯域およびゲインを調整するパラメータである。本モデルでは高品質な音声を合成するため 6 kHz 程度の帯域が必要があると考えている。しかし，高域部では 1 次元の音響管への近似が成立しない場合がある。よって，式 (6.6) によって高域部における複雑な形状に起因するスペクトルのピークやゼロが生じた場合の補

正を行う。

音源波形のスペクトル形状のパラメータについても，声道パラメータと同時に補正を行う。表-6.8 に母音のスペクトル包絡の計算に使用するパラメータを示す。主声道における「その他」とは，口腔と中咽頭腔の境界，中咽頭腔と喉頭腔の境界，喉頭腔の終端，および梨状窩の接続部を示す。

表 6.8: 母音合成のパラメータ

周期性音源	F_1, B_1, B_2, α	SA 有
	β, f_c	SA 無
主声道	口唇　位置，断面積（3 次）	SA 有
	口腔　位置，断面積（4 次）	SA 有
	中咽頭腔　位置，断面積（3 次）	SA 有
	喉頭腔　位置，断面積（3 次）	SA 有
	調音点　位置，断面積（1 次）	SA 有
	その他　位置，断面積（4 次）	SA 有
梨状窩	位置，断面積（4 次 x 2 本）	SA 有
鼻腔	位置，断面積（40 次）	SA 無
	接続部の面積	SA 有
副鼻腔	管の長さ，断面積（2 次 x 4 個）	SA 無

表-6.8 において「SA 有」は SA 法を用いて補正の行われるパラメータを示す。また，音源パラメータにおいて，$\beta = 10, f_c = 4k$Hz とした。鼻腔，副鼻腔は MRI から測定された値を固定値として用いた。

パラメータ補正にて使用する評価関数を以下に示す。

$$
\begin{aligned}
E \;=\; & W_{spc}\frac{1}{K}\sum_{k=1}^{K}\left(\frac{S_{ac}(f_k)-S_{sy}(f_k)}{S_{so}(f_k)}\right)^2 + \\
& W_{fm}\frac{1}{P}\sum_{p=1}^{P}\left(\frac{F_{ac}(p)-F_{sy}(p)}{F_{ac}(p)}\right)^2 + \\
& W_{area}\sum_{n=1}^{N}\left(\frac{A(n)-A_i(n)}{A_i(n)}\right)^2\frac{L_i(n-1)+L_i(n)}{2L_{all}} + \\
& W_{len}\sum_{n=1}^{N}\left(\frac{L(n)-L_i(n)}{L_{all}}\right)^2
\end{aligned}
\tag{6.8}
$$

式 (6.8) において S_{ac} は参照音声のスペクトル，S_{sy} は合成音声のスペクトル，S_{so} は音源の微分波形のスペクトルを示す。f_k は対数領域で等間隔に分布する周波数を示す。$F_{ac}(p)$ は参照音声の p 番目のフォルマント周波数，$F_{sy}(p)$ は合成音声の p 番目のフォルマント周波数を示す。$A(n)$ は n 番目の特徴点の声道断面積，$A_i(n)$ は n 番目の特徴点の声道断面積の初期値を示す。$L(n)$ は n 番目と $n+1$ 番目の特徴点間の長さ,$L_i(n)$ は n 番目と $n+1$ 番目の特徴点間の長さの初期値を示す。L_{all} は声道長を示す。$W_{spc}, W_{fm}, W_{area}, W_{len}$ は各項に乗ずる重み係数を示す。ここで，第 1 項はスペクトルの歪みを表す項である。フォルマント周波数の誤差は音質に与える影響が大きいと考えられるため、これを修正するために第 2 項を追加した。第 3 項と第 4 項は声道形状の初期値からの変形に対するペナルティを表す。第 3 項は声道断面積の変形に対するペナルティであり，前後の特徴点までの長さの平均を重みとして乗じた。第 4 項は各特徴点間の長さの変形に対するペナルティである。フォルマントは 4 次まで ($P = 4$)，変形のペナルティは主声道と梨状窩 ($N = 26$) を用いた。梨状窩のパラメータは，他のパラメータを固定した状態で，以下の評価式を用いて補正した。

$$E = \sum_{k=1}^{K} (\log S_{ac}(f_k) - \log S_{sy}(f_k))^2 \tag{6.9}$$

ここで，f_k は線形周波数軸上に等間隔に分布する周波数を示す。

一般に音声から声道形状を求める逆問題の場合は解が 1 対多になるため，推定される声道形状に制約を設けたり，時間変化に対して滑らかに変化するように制約を設けることにより解を一意に定める。本手法では，解の初期値として MRI から計測された声道形状を用いることで，初期値に最も近い局所解が真の最適解であると仮定し解を探索する。よって，真の最適解からかけ離れた解が得られることは無いと考えられる。しかし，最適解付近にも複数の局所解が存在する可能性があるため，声道パラメータの初期値からの変形率に 2 種類の制約を設けた。初期値から大きく離れないようするための変形率の大きさに対する制約と，隣合うパラメータ間で変形率が滑らかに変化するための隣接するパラメータの変形率の差に対する制約である。この 2 種類の制約により現実的な声道形状を保ちながら解の探索を行った。また，パラメータ補正とパラメータの時間方向への平滑化を交互に数回繰り返し行うことにより時間方向に対しても滑らかに変化するパラメー

タを求めた。

6.4.2 子音

子音のパラメータについても母音と同様に声道パラメータの補正を行う。子音の音源には，声帯音源以外に声道内の狭めの位置に生じる雑音源がある。雑音源には，任意の周波数を境に高域，低域で異なるスペクトル傾斜を持つ雑音源を用いた [77]。狭めよりわずかに口唇側の位置に双極子音源を挿入したときの口唇までの伝達関数に雑音源のスペクトルを畳み込むことにより，狭めの雑音源により生じる波形のスペクトルを計算した。狭めの位置の雑音源から口唇までの伝達関数の計算では，鼻腔への分岐を無視し，声帯は閉鎖端として計算した。また，声帯音源と狭めによる雑音源の両方を有する子音のスペクトルは，両音源から生成されるスペクトルを合わせることにより求めた。

子音のスペクトル包絡計算には表-6.8 に加え表-6.9 のパラメータを用いた。

表 6.9: 子音合成のパラメータ（追加分）

非周期性音源	音源挿入位置	SA 無
	低域の傾斜	SA 有
	高域の傾斜	SA 有
	高域と低域の境界周波数	SA 有
	ゲイン	SA 有

表-6.9 中の音源挿入位置は測定された声道断面積関数より直接決定した。それ以外は全て SA 法によりパラメータの補正を行なった。

パラメータ補正にて使用する評価関数を式 (6.10) に示す。

$$
\begin{aligned}
E \;=\; & W_{spc}\frac{1}{K}\sum_{k=1}^{K}(\log S_{ac}(f_k)-\log S_{sy}(f_k))^2 + \\
& W_{area}\sum_{n=1}^{N}(\frac{A(n)-A_i(n)}{A_i(n)})^2\frac{L_i(n-1)+L_i(n)}{2L_{all}} + \\
& W_{len}\sum_{n=1}^{N}(\frac{L(n)-L_i(n)}{L_{all}})^2 \\
& S_{sy}(f_k)=\max(S_v(f_k),S_c(f_k))
\end{aligned}
\tag{6.10}
$$

式 (6.10) において S_v は声帯に音源がある場合の有声音声に対するスペクトルである。S_c は狭めに音源がある場合のスペクトルである。その他の変数は式 (6.8) と同じである。母音の場合と比較し，子音の場合はスペクトルのピークの位置だけで音韻が特徴づけられるものではない。よって，零点についてもピークと同様に正確に補正するために，スペクトルの歪みの計算には対数スペクトルを用いた。声帯音源の有無および狭めでの雑音源の有無は参照音声波形の周期性/非周期性の比率より判断した。

6.5 音声合成

声帯に音源を持つ有声音声は，モデルより計算されたスペクトル包絡と参照音声から得られた F0, 周期性/非周期性の比率を用いて STRAIGHT により合成した。狭めの雑音源から生成される子音の音声は，白色雑音に狭めの雑音源から口唇までのスペクトルを畳み込むことにより合成した。声帯と狭めの両方に音源を持つ子音の音声は，それぞれの音源により合成された波形を加算することにより合成した。

本方式の評価をするために，音声の合成実験を行った。話者は成人男性 1 名，発話は母音「あいうえお」および子音を含む単語「あざ」,「かき」である。参照音声は，MRI 撮像後に同じパイロット信号を聞きながら防音室内で仰臥位の状態での発話を収録した。収録では約 10 回程度連続して発話し，その中で安定して発話されていると判断された中から任意に 1 つ選択した。音声のサンプリング周波数は 12 kHz を用いた。鼻腔および副鼻腔については，同発話者を用いた過去の研究の測定結果を用いた [71]。梨状窩の形状は，音韻間で変化が少ないことが報告されている [70]。よって，全ての音韻に対して平均的な同一の形状を初期値として用いた。

6.5.1 母音の合成実験

母音「あいうえお」の合成結果を図-6.6 に示す。

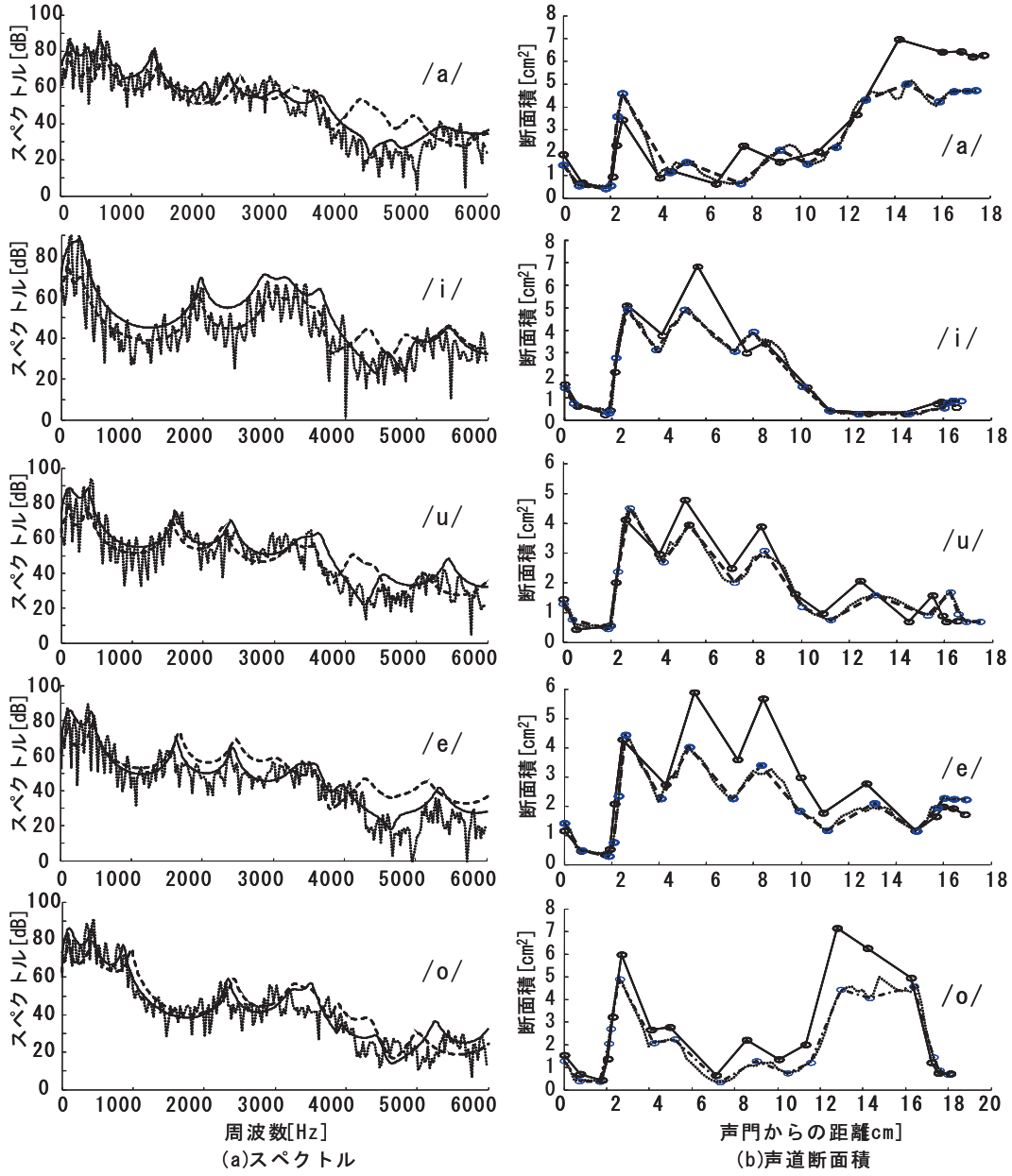


図 6.6: 連続母音「あいうえお」における母音スペクトルと声道断面積モデル
上から順に，各母音の中心部における結果を示す。(a) 点線は参照音声の FFT 結果，破線は MRI から抽出した初期モデルから計算したスペクトル，実線は補正された声道断面積モデルから計算したスペクトルを示す。(b) 点線は MRI から計測した声道断面積関数，破線は声道断面積の初期モデル，実線は補正された声道断面積モデルを示す。

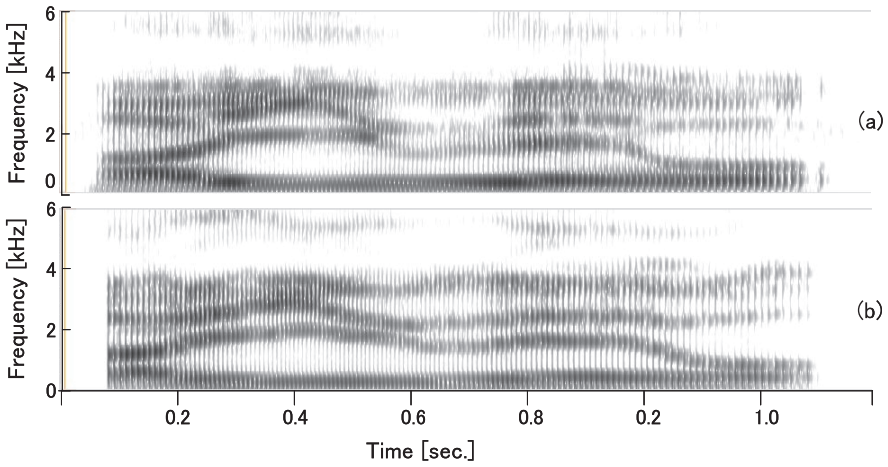


図 6.7: スペクトログラム「あいうえお」(a) 元音声 (b) 合成音声

表 6.10: スペクトル誤差と変形量の平均値

	スペクトル		変形量	
	SA 前	SA 後	面積	長さ
あいうえお	$7.45dB$	$3.94dB$	$0.64cm^2$	$0.16cm$
かき	$8.60dB$	$3.43dB$	$0.77cm^2$	$0.15cm$
あざ	$6.54dB$	$3.63dB$	$0.97cm^2$	$0.16cm$

図-6.6 の/a/は「あ」の母音中心，/i/は「い」の母音中心，/u/は「う」の母音中心，/e/は「え」の母音中心，/o/は「お」の母音中心での結果を示す。図-6.6 (a) は参照音声のスペクトルおよび声道断面積モデルから計算したスペクトルを示す。点線は参照音声の FFT 結果，破線は MRI から抽出した声道断面積関数の初期モデルから計算したスペクトル，実線はパラメータ補正後の声道断面積モデルから計算したスペクトルを示す。各母音について 4 kHz 以下のスペクトルを比べると、参照音声の FFT 結果から推測されるフォルマント周波数が初期モデルから推定したフォルマント周波数と比較的近いことがわかる。このことから声道断面積モデルが MRI の測定結果から適切に抽出されていることがわかる。この発話者の場合，4 〜 5 kHz に梨状窩による零点が存在する。梨状窩の形状の測定精度，開口端補正等の精度が向上すれば初期モデルのスペクトルが参照音声のスペクトルにより近づくことが予想される。パラメータ補正後は，全ての帯域で目標値に近いスペクトルが得られており，補正が正しく行われていることが推測される。図-6.6 (a)-/a/ の声道断面積モデルから計算されたスペクトルでは，500 Hz, 2500 Hz 付近に零点が見られる。これは，母音の/a/の発声時に鼻咽腔の開口が生じて鼻腔との結合により生じた零点である。図-6.6 (b) に声道形状の比較図を示す。図の点線は MRI から測定された声道断面積関数，破線は測定された声道断面積関数から抽出した声道モデルの初期値，実線は補正された声道形状を示す。補正された声道形状と測定した声道断面積関数の比較から,「あ」の口腔部,「い」の中咽頭部など断面積の大きい部分では測定した形状より断面積が大きく補正される場合が多い。これは MRI 撮像時は繰り返し発話であるため通常発話と比較して動作が抑えられ，口腔が自然な発話より小さくなっていたためではないかと考えられる。図-6.7 に参照音声および合成音声のスペクトログラムの結果を示す。また，表-6.10 にスペクトル誤差および特徴点の面積と特徴点間の長さの変形量の平均値を示す。ここで，SA 前のスペクトル誤差とは声道断面積関数よりえられた声道断面積関数の初期モデルに対して音源のパラメータのみ SA 法にて補正した結果である。スペクトログラムの結果より，全体的に，各フレーム間が滑らかに接続されており，各フレーム間で大きく離れた局所解に陥ることなく，声道パラメータの推定が正しく行われていることがわかる。また，補正の変形量も少ないことから実声道形状に近い形状が得られていることが推測される。ただし，母音間の遷移部分では，合成音声の

スペクトログラムの方が参照音声と比較しなだらかに変化している。これは，局所解をさけるために加えた声道形状の変形に対する制約および時間方向へのパラメータの平滑化の影響であると考えられる。遷移部分の推定を改善するには，動作の変化が大きなフレームではMRIの測定誤差も大きくなることから，計測された声道形状の変動に応じて，変形に対する制約や平滑化のパラメータを変化させるなどの対策が必要であろう。

6.5.2 子音を含む単語の合成実験

図-6.8の/k/は無声破裂音（「かき」の先頭の/k/）の結果，/z/は音源が2つある有声摩擦子音/z/の結果を示す。

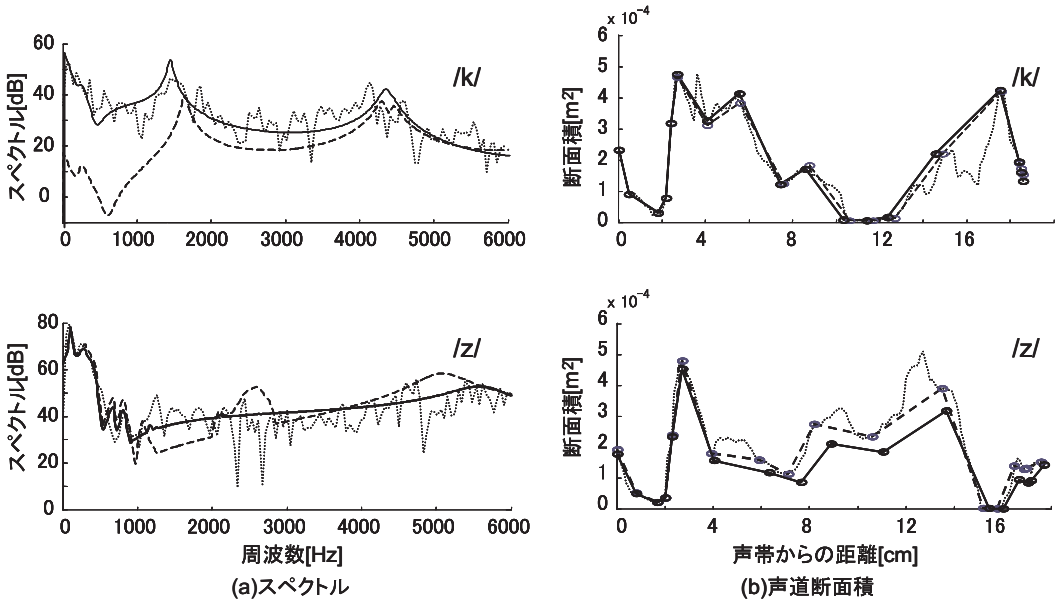


図 6.8: 子音を含む単語「かき」「あざ」における子音スペクトルと声道断面積モデル

/k/は「か」の子音の中心，/z/は「ざ」の子音の中心部における結果を示す。(a) 点線は参照音声のFFT結果，破線はMRIから抽出した初期モデルから計算したスペクトル，実線は補正された声道断面積モデルから計算したスペクトルを示す。(b) 点線はMRIから計測した声道断面積関数，破線は声道断面積の初期モデル，実線は補正された声道断面積モデルを示す。

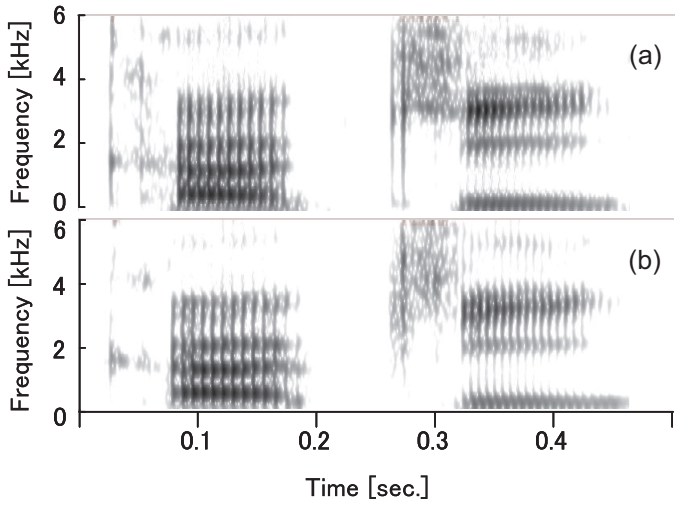


図 6.9: スペクトログラム「かき」(a) 元音声 (b) 合成音声

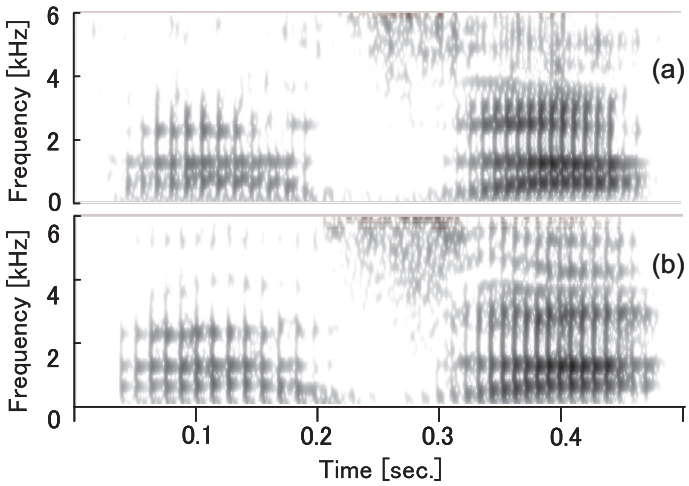


図 6.10: スペクトログラム「あざ」(a) 元音声 (b) 合成音声

両者のスペクトログラムを比較すると、子音部分においても実測データから生成された声道モデルより合成した音声は参照音声によく似ていることがわかる。このことから，子音についても初期モデルが適切に抽出されていることがわかる。/k/,/z/の子音のように狭めのある子音の場合，狭めより声帯側の形状はスペクトルのピークにほとんど影響しないため，パラメータ補正において解が一意に定まらない。よって，狭めによる音源が存在する場合は式 (6.10) の W_{area}, W_{len} を大きくすることで MRI から得られた声道形状からの誤差に重みが置かれるようにした。その結果，/k/,/z/ともに声道形状の推定結果は初期値に良く似た形状となっている。表-6.11 に本実験で用いた重み係数を示す。図-6.9,6.10 のスペクトログラムの

表 6.11: 重み係数

	W_{spc}	W_{fm}	W_{area}	W_{len}
母音	1.0	4.0	0.1	0.1
子音	0.1	-	1.0	1.0

比較結果より,「かき」「あざ」ともに元音声に類似した結果が得られており，本方式の有効性が確認できる。また，表-6.10 より,「あいうえお」と同様に大きな変形なく，実声道形状に近い形状が推定されていることがわかる。

本方式では 6kHz までの周波数帯域を対象としているため音声のスペクトル全域を再現するものではなく、摩擦子音などではパワーの集中する高域まで十分に表現できない。しかし，12 kHz でサンプリングされた音声を実際に聞けばわかるように，帯域制限された音声を用いても，個人性，感情を含めて，十分に表現可能であり，利用可能な市場は存在すると考えている。ただし，エンターテイメントなどの分野への展開を考えると，広帯域への拡張は今後の重要な課題であろう。

6.6 本章のまとめ

本章では，フィルタ (声道形状) のモデルについて，これまで行ってきた２次元断面上での模倣を３次元に拡張し，３次元 MRI 動画と実音声を用いて声道形状の推定を高精度化することにより高品質な音声の合成を可能にした音声分析合成方

式を提案した。声道形状のモデルには，複数話者の母音および１話者の複数の子音の MRI データを基に構築した全ての声道形状を同じ次数のパラメータで表現できる声道断面積モデルを用いた。このモデルは，F0 調節時の声道形状の変化などの発話機構の生理学的知識に基づくパラメータの変換に対応するため，解剖学的な位置の基準を考慮し構築を行った。分析合成時には，３次元 MRI 動画から声道断面積モデルの初期値となるパラメータを決定し，実音声を用いて最適化を行うことでパラメータを推定した。本方式を用いた幾つかの単語の合成実験を行った結果，元音声と合成音声とのスペクトル歪みの平均は約 3.67 dB となり，人間の音声生成メカニズムに基づく音声合成方式を用いても，３次元の声道形状を用いて最適化を行うことで，高品質な音声を合成することができることをが明らかとなった。

第 7章

あとがき

本論文では，人間の音声生成メカニズムに基づく音声合成方式を用いたテキスト音声合成システムの開発を目標として行ってきた研究についてまとめた。

音声生成メカニズムに基づく音声合成方式により，自然性の高い音声の合成を行うには，音声生成の生理機構を解明することが重要な課題の１つである。本研究では，はじめに，内咽頭筋以外の活動による F0 調節の生理機構の解明を行った。複数の話者が下降音階で定常母音を発声した時の正中矢状断面を MRI を用いて撮像し喉頭周囲構造の位置変化を分析した。その結果、輪状軟骨と甲状軟骨だけでなく頸椎、舌骨などの器官も F0 調節に大きな役割を果たしていることが確認された。その中でも頸椎の自然湾曲による輪状軟骨の回転は従来の研究で指摘されていなかった新しい知見である。この機構は、喉頭の上下運動を輪状甲状調節の回転に変換する生理機構である。この機構により、従来から問題とされてきた F0 下降時の舌骨下筋の活動、喉頭の下降傾向を説明することができる。よってこれらは，CT の弛緩と共に F0 下降の主要因であると考えられる。また、今回得られた結果は，従来の報告にあるような甲状輪状関節の水平移動の成分や CT 以外の内喉頭筋の F0 への影響なども推測することができるものであった。

このような喉頭の上下動を含めた F0 調節機構を考慮すると，F0 調節と声道形状制御との間に相互作用が生じることが予想される。本研究ではつぎに，この F0 調節機構を実装し，舌と喉頭とを含む全ての発話器官間相互の力の授受を考慮することにより，F0 調節と声道形状制御との相互作用を実現する発話器官の生理学的モデルを構築した。このモデルを用いて，日本語５母音および調音を強調した

ときの５母音を生成し調音動作のF0に及ぼす影響を調べた。その結果，調音の強調動作により各母音のフォルマントが分離するとともに，母音の固有ピッチも強調される結果が得られた。これは「母音の固有ピッチが舌と喉頭との舌骨を介する接続により生じる」という理論を支持する結果であった。また，F0を下降させた時の音声を生成し，F0調節の調音に与える影響を調べた。その結果，F0調節により声道形状がMRIにより計測された形状と同様に変化する結果が得られた。これらのことから，F0調節と声道形状制御との相互作用が存在し，本モデルを用いてその相互作用を表現することが可能であることが確認された。この相互作用の実現により，F0変化による音声の周波数特性の変化を含む音声をモデルを用いて生成することが可能となった。よってこの機構を合成システムに実装することにより，テキスト音声合成システムのデータベースに異なるF0の同じ音素を複数蓄積する必要がなくなり，自然性が高くコンパクトな合成音声システムの開発が可能になると考えられる。しかし，本モデルを用いて生成した音声は，定性的には相互作用を含む実音声と同じフォルマント変化の傾向を示したが，定量的には日本語５母音の合成音声と実音声の第1-3フォルマント周波数の平均誤差が10.7%と多くの誤差を含み実用化には不十分な結果であった。このことから音質の劣化が本方式の実用化へのボトルネックになることが予想された。音質劣化の原因としては，モデル化の際に行った様々な単純化や計測データに含まれていた誤差などが考えられる。特に，このモデルのパラメータはMRIの矢状断面の２次元声道形状の再現を目標に最適化を行ったものであるが，音声合成には３次元形状が必要であることから，2 次元データから3次元データに変換する際に生じた誤差が主原因ではないかと考えられた。

よって最後に，本方式の実用化へのボトルネックを解消するために、声道フィルタ部についてこれまで行ってきた２次元断面上での声道形状の模倣を３次元声道形状を模倣するモデルに拡張することで，合成音声の高品質化をを目指した。本研究では，３次元MRI動画から得られた声道形状のパラメータを実音声を用いて最適化することで高品質な音声の合成を可能にする声道断面積モデルによる分析合成方式を提案した。あらかじめ，複数話者の母音および１話者の複数の子音のMRIデータを基に全ての声道形状を表現できる汎用の声道断面積モデルを作成した。この声道断面積モデルはF0変化時の声道形状の変化などの発話機構の生理学

的知識に基づくパラメータの変換に対応するため，解剖学的な位置の基準を考慮し作成を行った。分析合成に用いるパラメータは，汎用のモデルを基準に３次元MRI 動画から初期のパラメータを推定し，実音声を用いて最適化を行うことで推定した。本方式を用いて幾つかの単語を合成した結果，元音声と合成音声とのスペクトルの平均誤差は約 3.67dB となった。この結果より，人間の音声生成メカニズムに基づく音声合成方式を用いても高品質な音声を合成することができることが明らかとなった。

本研究の最終的な目的は，人間の音声生成メカニズムに基づく音声合成方式によるテキスト音声合成システムの実用化である。残念ながら，システムの完成には至らなかった。今後の課題を以下に示す。

1. 先に明らかにした F0 調節の生理機構を声道断面積モデルによる分析合成方式に実装できるように適切なモデル化を行う。そして，F0 の違いによる声道形状の変化を含む音声を収録することなく合成できることを確認する。
2. 多くの単語について声道断面積モデルの分析を行い，音素を追加することで，任意の音声を合成できるシステムを構築を行う。また，前後の音素環境の違いによる調音結合をモデル化し全ての音素環境を含む音声を収録することなく合成できることを確認する。
3. 個人性や感情音声の生理的要因を分析しモデル化を行う。

これらを行うことで，高品質で，声質，感情など多様な音声を表現でき，かつ，コンパクトなテキスト音声合成システムの開発を行いたいと考えている。

謝辞

本研究を行なうに当たり, 終始御指導を賜わった党 建武 教授，本多 清志 博士に深謝致します。

また, 日頃から有益な御助言をいただき, 多面に渡って励ましていただいた甲南大 北村 達也 准教授，ATR 竹本 浩典 博士，ATR 正木 信夫 博士に感謝致します。そして，MRI の撮像実験にご協力頂いた ATR BAIC 島田 育廣 博士，藤本 一郎 氏に感謝致します。

最後に, 本論文をまとめるに当たってご助言をいただいた北陸先端大 赤木 正人 教授，小谷 一孔 准教授，徳田 功 准教授，および，本論文の作成に御協力いただいた党研究室 藤田 覚 氏，ATR 人間情報研究所 第４研究室および人間情報科学研究所 BPI プロジェクトに所属されておりました研究員の皆様に厚く御礼申し上げます。

ありがとうございました。

参考文献

[1] 広川智久, 箱田和男, 中津良平, “波形接続型規則合成法における波形選択法,” 信学技報, SP89-114, Jan. 1989.

[2] ニック・キャンベル, アラン・ブラック, “CHATR : 自然音声波形接続型任意音声合成システム,” 信学技報, SP96-7, May. 1996.

[3] J. Schroeter and M. Sondhi. “Techniques for estimating vocal-tract shapes from the speech signal,” IEEE Trans. Speech and Audio Processing, 2, 133-150, (1994).

[4] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokamik, P. Rubin and E. Saltzman, “Accurate recovery of articulator positions from acoustics: New conclusions based on human data,” J. Acoust. Soc. Am. 100(3), 1819-1834, (1996).

[5] 白井克彦, 誉田雅彰, “音声波からの調音パラメータの推定,” 電子情報通信学会論文誌 (A), J61-A, 5, 409-416, (1978).

[6] 鈴木紳, 岡留剛, 誉田雅彰, “音響調音対コードブックを用いた音声からの調音運動の逆推定”, 電子情報通信学会論文誌 (A), J85-A, 8, 840-846, (2002).

[7] 後藤正三, 三輪譲二, “調音音響変換 A-b-S 法を用いた VCV 音声の動的声道形推定,” 信学技報, SP2002-175, 35-40, (2003)

[8] Y. Katsuki, “The function of the phonatory muscles,” Jpn. J. Physiol. 1, 29-36 (1950).

[9] A. Sonninen, “Is the length of the vocal cords the same at all different level of singing ?,” Acta Otolaryngol. Suppl., 118, 219-231 (1954).

[10] 新美成二, “日本語・英語・中国語におけるアクセントの生成の生理学的比較,” 講座 日本語と日本語教育, 杉藤美代子編（明治書院, 東京, 1990）, p.332.

[11] 杉藤美代子, “日本語アクセントの研究,”（三省堂, 東京, 1982）, p.211.

[12] J. Dang, K. Honda, “Construction and control of a physiological articulatory model,” J.Acoust. Soc. Am. 115, 853-870 (2004)

[13] 垣田有紀, 藤村　靖, “米語の母音における筋収縮からフォルマントへの写像,” 音響学会音声研資 S83-100, 791-798 (1984).

[14] R. Wilhelms-tricarico and C. Wu,“ The 3-D-tongue FEM model revised, ” J.Acoust. Soc. Am. 94, 1764-1764 (1993).

[15] J.S. Perkell, ”A physiologically-oriented model of tongue activity in speech production, ” Unpublished Ph. D. Thesis, MIT (1974).

[16] 草川直樹, 本多清志, 垣田有紀,“筋電信号を用いる舌の 2 次元モデル, ”音講論集 I, 273-274 (1992.3).

[17] Y. Kakita and S. Hiki, “A study of laryngeal control for voice pitch based on anatomical model, ” Proc. Speech Commun. Semin. Stockholm, 45-54 (1974).

[18] K. Honda, T. Kurita, Y. Kakita and S. Maeda, “Physiology of the lips and modeling of lip gestures,” J. Phonet. 23, 243-254 (1995).

[19] 比企静雄,“母音の調音器官モデル, ”音声情報処理, 比企静雄編 (東京大学出版会, 東京, 1973), p.95.

[20] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” Speech Communication, 9, 453-467 (1990).

[21] Hisashi Kawai, Tomoki Toda, Jinfu Ni, Minoru Tsuzaki, and Keiichi Tokuda, “Ximera: A New TTS from ATR Based on Corpus-Based Technologies,” ISCA 5th Speech Synthesis Workshop, pp. 179-184, 2004.

[22] 板倉文忠, 斎藤収三, “PARCOR 形音声分析合成方式とその応用,” 日本音響学会誌, vol.27, No.2, pp.114, Feb 1971.

[23] 今井聖, “対数振幅近似 (LMA) フィルタ,” 電子情報通信学会論文誌 (A), J63-A, 12, pp.886-893 (1981)

[24] 菅村昇, 板倉文忠, “線スペクトル対（LSP）音声分析合成方式による音声情報圧縮,” 電子通信情報学会誌, J64-A, [8] pp. 599-606 (1981-8)

[25] H. Kawahara, I. Masuda-Kasuse and Alain de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a reptitive structure in sounds,” Speech Communication, 27, 187-207 (1999).

[26] K. N. Stevens and A. S. House, “Development of a Quantitative Description of Vowel Articulation,” J.Acoust. Soc. Am. 27, 484-493 (1955).

[27] C. H. Coker, “A model of articulatory dynamics and control,” Proc. IEEE, 64, pp.452-460 (1976)

[28] P. Mermelstein, “Determination of the vocal tract shape from the measured formant frequencies,” J.Acoust. Soc. Am. 41, 1283-1294 (1967).

[29] P. Rubin, E. Saltzman, L. Goldstein, R. McGowan, M. Tiede and C. Browman, “Casy and extensions to the task-dynamic model,” 4th Speech Production Seminar, Grenoble, France. (125-128).

[30] 益子 貴史, 徳田 恵一, 小林 隆夫, 今井 聖, “動的特徴を用いた HMM に基づく音声合成,” 電子情報通信学会論文誌 (D-II), J79-D-II, 12, 2184-2190, (1996).

[31] 吉村 貴克, 徳田 恵一, 益子 貴史, 小林 隆夫, 北村 正, “HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化,” 電子情報通信学会論文誌 (D-II), vol.J83-D-II, 12, 2099-2107, (2000).

[32] A. Rosenberg, “Effect of glottal pulse shape on the quality of natural vowels,” J. Acoust. Soc. Am. 49, 583-589 (1971).

[33] G. Fant, J. Liljencrants, Q. Lin, "A Four Parameter Model of Glottal Flow," In STL-QPSR 4, Sweden, 1-13 (1985).

[34] K. Ishizaka and J.L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the cocal cords," Bell Syst. Tech. J.51, 1233-1268 (1972).

[35] R. Causse, J. Kergomard and X. Lurton, "Input impedance of brass musical instruments: Comparison between experiment and numerical models," J. Acoust. Soc. Am., 75, 241-254 (1984).

[36] S. Adachi and M. Yamada, "An acoustical study of sound production in biphonic singing, Xoomij," J. Acoust. Soc. Am., 105, 2920-2932 (1999).

[37] J.L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd Ed., (Springer-Verlag, New York, 1972) Chap. 3, pp. 36-38.

[38] J.S. Perkell, Physiology of Speech Production Results and Implications of a Quantitative Cineradiographic Study MIT Press (1969).

[39] S. Kiritani, K. Itoh and O. Fujimura, "Tongue-pellet tracking by a computer-controlled x-ray microbeam system," J. Acoust. Soc. Am. 57, 1516-1520 (1975).

[40] R.D. Naddler, J.H. Abbs and O. Fujimura, "Speech movement research using the new x-ray microbeam system," Proc. the 11th International Congress of Phonetic Sciences, Tallin, Estonia, 1, 221-224 (1987).

[41] J. Perkell, M. Cohen, M. Svirsky, M. Matthies, I. Garabieta and M. Jackson, "Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements," J. Acoust. Soc. Am., 92, 3078-3096 (1992).

[42] H. Takemoto, K. Honda, S. Masaki, Y. Shimada and I. Fujimoto, "Measurement of temporal changes in vocal tract area function during a continuous vowel sequence using a 3D cine-MRI technique," Proc. 6th Int. Semin. Speech Production, Sydney, pp.284-289, (2003).

[43] C. Yang, H. Kasuya, S. Kanou and S. Satou, “An accurate method to measure the shape and length of the vocal tract for the five Japanese vowels by MRI,” Jpn. J. Logop. Phoniatr., 35, 317-321 (1994).

[44] O. Engwall, “Using linguopalatal contact patterns to tune a 3D tongue model,” Proc. Eurospeech 2001, Vol. 2, pp. 1475-1478 (2001)

[45] M. Matsumura, T. Niikawa, Y. Matsushige, K. Shimizu, Y. Hashimoto and T. Morita, “Measurement of 3D shapes of vocal tract, dental crown, and nasal cavity using MRI,” Tech. Rep. IEICE, MBE93-131, pp. 41-48 (1994).

[46] H. Takemoto, T. Kitamura, Nishimoto and K. Honda, “A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions,” Acoust. Sci. & Tech. 25, 6, 468-474 (2004).

[47] S. Masaki, M.K. Tiede, K. Honda, et al. “MRI-based speech production study using a synchronized sampling method,” J Acoust Soc Jpn(E), 20(5), 375-379, (1999).

[48] 島田，藤本，竹本，高野，正木，本多，武尾，”Synchronized Sampling Method(SSM) を利用した 4D-MRI,” 日本放射線技術学会雑誌第５８巻第１２号, pp.1592-1598, (2002.12)

[49] 藤崎博也, “人間と音声,” 音の科学, 難波精一郎編（朝倉書店, 東京, 1988）, p. 66.

[50] 本多清志, “音声基本周波数の上昇における輪状甲状筋の部位別の機能,” 音講論集 1-2-7, 115-116 (1988.10).

[51] 本多清志, “音声基本周波数の調節に関与する内外喉頭筋の機能,” 音講論集 1-6-1, 157-158 (1987.10).

[52] K. Honda, “Relationship between pitch control and vowel articulation,” in Vocal Fold Physiology, D.M. Bless and J.H. Abbs, Eds. (Collage-Hill Press, SanDiego, CA, 1983), pp.286-299.

[53] 葉山杉夫, “ヒトの発声器官の起源,” 講座 進化 4 形態学からみた進化, 柴谷篤弘, 長野敬, 養老孟司編 (東京大学出版会, 東京, 1991), p.173.

[54] K. Honda and O. Fujimura, “Intrinsic vowel F0 and phrase-final F0 lowering: phonological vs. biological explanations,” in Vocal Fold Physiology, J. Gauffin and B. Hammarberg, Eds. (Singular Publishing Group, Inc., San Diego, CA, 1991), pp.149-157.

[55] A. Sonninen, “The external frame function in the control of pitch in the human voice,” Ann. N.Y. Acad. Sci. 155, 68-89 (1968).

[56] D. J. Broad, “Phonation” in Normal Aspects of Speech, Hearing, and Language, F. D. Minifie, T.J. Hixon and F. Williams, Eds. (Prentice-Hall, Inc., Englewood Cliffs, NJ, 1973), pp.127-167.

[57] G. Fant, “Acoustic Theory of Speech Production,” (Mouton, The Hague, 1960).

[58] I. Lehiste and G.E. Peterson, “Some basic consideration in the analysis of intonation,” J. Acoust. Soc. Am. 33, 419-423 (1961).

[59] K. Honda, “Relationship between pitch control and vowel articulation,” in Vocal Fold Physiology, D.M. Bless and J.H. Abbs, Eds. (College-Hill Press, San Diego, 1983), P.286.

[60] J. Edwards, “Rotation and translation of the jaw during speech,” J. Speech Hear. Res. 33, 550-562 (1990).

[61] E. Vatikiotis-Bateson and D.J. Ostry, “An analysis of the dimensionality of jaw motion in speech,” J. Phonet. 23, 101-117 (1995).

[62] M.M. Sondhi and J. Schroeter, “A hybrid time-frequency domain articulatory speech synthesizer,” IEEE Trans. Acoust. Speech Signal Process. 35, 955-967 (1987).

[63] J.M. Heinz and K.N. Stevens, “On the relations between lateral cineradiographs, area function, and acoustic spectra of speech,” Proc. 5th Int. Congr. Acoust. Liege, A44 (1965).

[64] Y. Kakita, M. Hirano and K. Ohmaru, “Physical properties of the vocal fold tissue: measurements on excised larynges, ”in Vocal Fold Physiology (Univ. Tokyo Press, Tokyo, 1980), P.377.

[65] J.R. Westbury, “The significance and measurement of head position during speech production experiments using the X-ray microbeam system, ”J. Acoust. Soc. Am. 89, 1782-1791 (1991).

[66] 陸金林, 村上秀紀, 粕谷英樹,“複数閉鎖区間を用いた声道伝達関数の推定, ”信学論 J 73-A, 1011-1014 (1990).

[67] 粕谷英樹, 鈴木久喜, 城戸健一,“年齢, 性別による日本語５母音のピッチ周波数とホルマント周波数の変化, ”音響学会誌 24, 355-364 (1968).

[68] 北村達也，竹本浩典，本多清志，島田育廣，藤本一郎，赤土裕子，正木信夫，黒田輝，奥内昇，千田道雄, “座位および仰臥位における声道形状の相違 -開放型 MRI 装置を用いた観測-,”信学技報, SP2004-29, pp. 1-6, (2004)

[69] H. Takemoto, K. Honda, S. Masaki, Y. Shimada and I. Fujimoto, “Measurement of temporal changes in vocal tract area function from 3D cine-MRI data,” J. Acoust. Soc. Am., 119(2), 1037-1049 (2006).

[70] T. Kitamura, K. Honda and H. Takemoto, “Individual variation of the hypopharyngeal cavities and its acoustic effects,” Acoust. Sci. & Tech. 26, 1, 16-26 (2005).

[71] J. Dang, K. Honda and H. Suzuki, “Morphological and acoustical analysis of the nasal and the paranasal cavities,” J. Acoust. Soc. Am., 96, 2088-2100 (1994).

[72] J. Dang, and K. Honda, “Acoustic characteristics of the piriform fossa in models and humans,” J. Acoust. Soc. Am., 101, 456-465 (1997).

[73] たとえば，中山聖一, 確率モデルによる音声認識 (電子情報通信学会, 1988) pp.53-55.

[74] たとえば，坂元慶行, 石黒真木夫, 北川源四郎, 情報量統計学 (共立出版, 1983) pp.42-64.

[75] H. Fujisaki and M. Ljungqvist, “Proposal and evaluation of models for the glottal source waveform,” Proc. IEEE ICASSP 86, Tokyo, 31.2, pp.1605-1608 (1986).

[76] 石塚 正明, 粕谷 英樹, “音声合成用全零型有声音源モデル,” 日本音響学会誌, 50, 5, 361-368, (1994).

[77] S. Narayanen and A. Alwan, “Noise source models for fricative consonants,” IEEE Trans. Speech and Audio Processing, 8(3), 328-344, (2000).

本研究に関する発表論文

学術論文

[1] 平井，党，本多 (1995): “舌と喉頭との相互作用を考慮した発話器官の生理学モデル”，日本音響学会誌　51, 12, 918-928

[2] 平井，本多，藤本，島田 (1994):”F0 調節の生理機構に関する磁気共鳴画像（MRI）の分析” 日本音響学会誌　50, 4, 296-304

[3] 平井，竹本，本多，党 (2008) ”3 次元 MRI 動画と実音声を用いた声道断面積モデルのパラメータ推定”, 日本音響学会誌 64, 4 掲載予定

[4] Honda, K，Hirai, H. Masaki, S. and Shimada, Y. (1999)“ Role of Vertical Larynx Movement and Cervical Lordosis in F0 Control ”Language and speech 42(4) 401-411

学会発表 (査読付き)

[5] Honda, K. Hirai, H. and Dang, J. (1994,9):”A physiological model of speech organs and the implications of the tongue-larynx interaction,” Proc. ICSLP 94, 175-178,Yokohama.

学会発表

[6] Hirai, H, Dang J., and Honda K. (1994,6):”Production of speech from a physiological model of speech organs,” J. Acoust. Soc. Am., 95, p.2823.

[7] 平井，竹本，本多 (2004,3)“ 声道モデルを用いた音声の分析合成方式 ”日本音響学会講演論文，pp.283-284

[8] 平井，竹本，本多 (2004,9)“ 声道モデルを用いた子音の合成 ”日本音響学会講演論文，pp .253-254

[9] 平井，橋本，大西 (2001,9)“ STRAIGHT を用いたテキスト音声合成の開発と評価 ”日本音響学会講演論文，pp.369-370

[10] 平井，党，本多 (1994,3) ”発話器官モデルによる音声の生成 ” 日本音響学会講演論文, pp.669-670.

[11] 平井 (1995),“F0 変化に伴う母音のフォルマント周波数の遷移,” 信学技報, SP94-102, pp.29-36

[12] 平井，本多，藤本，島田 (1993,3)”F0 変化と喉頭軟骨の回転に関する MRI 画像の分析 ”日本音響学会講演論文，pp.199-200.