修士論文

韻律の観点からの HMM音声合成の高度化

2013 年 2 月 6 日

指導教員　広瀬 啓吉 教授

東京大学大学院 情報理工学系研究科
電子情報学専攻

48-116438　橋本 浩弥

内容梗概

音声は人間がコミュニケーションを図る上で、重要な伝達手段の一つである。よって、機械を用いて自動的に音声を認識・合成することが可能になれば、人間と機械が音声で対話をすることによって意思疎通を図れるようになり、より人間と機械の距離を縮めることができる。そして、音声の認識・合成技術の発展によって、対話システムによる様々な案内サービスや学習支援などの提供が可能になり、社会インフラの向上が見込まれ、我々の生活がより便利になることが期待される。さらに、ユーザが思い描く通りの音声を任意のテキストから自動的に創りだすことが出来れば、映画やテレビ番組などのコンテンツを従来に比べて容易に作成することができるようになり、コンテンツの幅が広がることも期待できる。

そのようなテキストを入力とし、対応する音声を自動的に合成するシステムは、テキスト読み上げ音声システム（Text-To-Speech; TTS) と呼ばれ、世界中で研究されてきた。しかし、実用レベルの品質には長いこと達しなかった。その原因の一つとして、音声の多様性にある。現在人間が利用している音声は、言語と密接な結びつきがあるが、文字と音声は必ずしも対応していないことが知られている。例えば、これは日本語の例になるが、「粘板岩（ねんばんがん)」は、3 回の「ん」があるが、これは音韻論の観点からは全て異なる音であることが知られている。また、音声の自然性に関係する重要な要素の一つであるアクセントやイントネーションは、音素単位[1]ではなく、単語や文全体にわたって表れる特徴であるため、適切にモデル化することが困難であった。TTS システムを実現するためには、このような複雑な音声の特徴をテキストから適切に再現する必要がある。さらに、テキストから読み上げられる音声は、話者、話者の発話スタイル、感情、強調、意図など、様々な要因によって変化するものであるため、柔軟な音声合成システムを実現するためには、ユーザが直感的に意図した通りの音声を合成できるようなシステムが必要不可欠である。

音声合成システムは、このような多くの課題に直面してきたが、近年、計算機の目覚ましい発達、大規模コーパスの整備、及び数理統計手法の発展などを背景として、急速に発展してきている。実際、最近のモバイル端末において、Apple の「Siri」[2]や NTT docomo の「しゃべってコンシェル」[3]など、音声合成を利用したサービスが搭載されるようになり、音声をインターフェースに利用したシステムが注目されるようになってきた。だが、現在に至っても未だにその性能や品質は十分であるとは言えない。現状のシステムでは、ユーザが自由に合成音声を制御することが困難である。また、合成された音声は、音質が十分

[1]語の意味を区別する音声の最小単位
[2]http://www.apple.com/jp/ios/siri/
[3]http://www.nttdocomo.co.jp/service/information/shabette_concier/

ではなく、特に韻律に不自然なところがあり、人間が実際に発生した音声とは大きな隔たりがある。この韻律は、アクセントやイントネーションなど、音の強弱・長短・高低などによって表現される言葉のリズムであり、音声の自然性や発話意図などに関係する重要な要素である。

そこで本研究では、主に韻律に注目して音声合成システムの改善に取り組む。まず初めに、音の高低による韻律表現に注目する。音の高さは一般にピッチと呼ばれているが、工学的には基本周波数という特徴量がおよそそれに対応していることが知られている。しかし、基本周波数は安定して抽出することが困難であり、また、音の高低による韻律的特徴は、フレーム単位ではなく単語やそれより長い単位で表れるため、容易に取り扱うことが困難であった。ここで、基本周波数の時系列パターンを表現するモデルとして、基本周波数パターン生成過程モデルというモデルがある。このモデルは生理的・物理的特性に基づいており、少数のパラメータで基本周波数パターンを表現することができる。そして、このモデルパラメータは言語情報と対応がよくとれることが知られている。ところが、基本周波数パターンからモデルパラメータを自動で抽出することが困難であるという問題があった。そこで、音声合成システムの 1 つである HMM 音声合成で用いられているコンテキストラベルを利用することにより、モデルパラメータの抽出性能を既存の手法と比較して、大幅に改善する手法を提案した。また、提案した手法を利用することにより、音声合成の品質の改善だけではなく、焦点制御などが従来に比べて容易に実現できることを示した。

ここで、提案手法に利用しているコンテキストラベルとは、音声合成において、テキストには直接表れない韻律などの特徴を表現するために、音素以外に様々な情報を加えられたラベルのことを指す。しかしこのラベルは、次のような問題がある。ラベルに用いられているアクセント句は、定義に曖昧性がある上、話者や話者の発話速度、発話スタイルに依存してその長さが変化するため、自動抽出が困難である。本来、TTS を目的としたラベルはテキストから推定される情報のみを用いなければならない。また、多様な音声を実現するために付加情報を加える場合は、ユーザが直感的に操作可能なものである必要がある。さらに従来のラベルでは、アクセント句の位置番号などの絶対的な情報が用いられているが、これでは任意の長さの文章を合成するために理論上ラベルの種類が無限に必要であり、また、一部の発話構造が異なるだけで、文全体のラベルが変化してしまうという問題がある。そこで従来のラベルの問題点を改善するために、コンテキストラベルの改良をした。具体的には、アクセント句の代わりに定義の曖昧性が少なく話者性に依存しない文節を利用し、位置番号などの絶対的な情報ではなく、前後の単語や文節などの相対的な情報を用いることにした。提案したラベルを用いることにより、合成音声の品質が改善されることを聴取実験により確認した。また、このラベルは、合成音声の品質が改善されるだけではなく、従来に比べてテキストから容易に、かつ安定して抽出することが可能になるという利点がある。

本論文で提案する基本周波数パターン生成過程モデルのモデルパラメータの自動抽出を高精度化する手法や、コンテキストラベルは、自然で多様な合成音声システムを実現するための重要なステップである。

目次

図目次

表目次

第1章

序論

1.1　研究の背景

テキストから音声を人工的に合成する、テキスト読み上げシステム（Text-To-Speech; TTS) が実用段階に近づいてきた。現在利用されている音声合成は、主に音声波形の素片を繋いでいく、波形接続型のシステムが主流であるが、近年、統計的手法を用いた手法が注目されている。その代表例として、隠れマルコフモデル（Hidden Markov Model; HMM）に基づく音声合成システムがある [1]。このモデルでは、音声分析再合成技術を用いることによって、音声の波形を直接取り扱うのではなく、特徴量ベースで取り扱い、学習用音声コーパスから HMM を学習する。そして合成時は、テキストを入力とし、HMM から生成された特徴量から音声を合成する。ここで、構築した HMM に適応や変換をかけることにより、従来の波形接続方式と比べて比較的容易に、話者、あるいは感情、発話スタイル等を柔軟に制御可能であることが知られている [2, 3]。

一方で、現状の HMM 音声合成は、その品質が波形接続に比べて低いという問題がある。その原因は様々な要因があるが、ここでは韻律に注目して 2 つの問題点を指摘する。1 つ目は、HMM 音声合成では、特徴量をフレーム単位で扱うため、より長時間にまたがって表れる韻律的特徴のモデル化が困難であるという問題である [4]。ここで、音声の基本周波数パターンを表現するモデルとして、基本周波数パターン生成過程モデルがある [5]。このモデルは生理的・物理的根拠に基づいており、少数のパラメータで基本周波数パターンをよく記述することができ、言語情報とよく対応が取れるという特徴がある。そのため、このモデルパラメータを用いることにより、焦点制御やスタイル制御などが容易に実現する手法が既に提案されている [6, 7]。しかし、基本周波数パターンからモデルパラメータを抽出することは逆問題であり、解析的に解くことができないという問題がある。モデルパラメータの自動抽出の代表的な先行研究としては、成澤による手法や [8]、言語情報を用いて改善を加えた古山の手法がある [9, 10]。しかしこの手法では、基本周波数パターンにおいて基本周波数が観測されない無声区間を 3 次スプラインを用いて補間しているが、これは長い無声区間等で本来存在しないピークを生じさせ、誤った抽出をしてしまうことがある。そのため、自動抽出性能が十分ではなく、上記で述べたスタイル制御や焦点制御に関する研究における生成過程モデルのパラメータは手動抽出されている。2 つ目は、HMM 音声合成で用いられているコンテキストラベルに関する問題である。この問題は、音声認識システムと対比させながら述べていく。HMM を用いた手法は元々、音声認識において発達してきた手法であり、基本的な要素は共通している。音声認識においては、その目的から主に音韻性のみが注目されるため、該当音素に対し、前後の音素で区別したトライフォン (triphone) が HMM の単位として主に用いられているが[1]、音声合成においては、テキストには直接表れない韻律などの特徴を表現するために、音素に加えて様々な情報を加えたラベル（コンテキストラベル）が用いられている。しかし、ラベルの種類を多くしていくと、存在しうるラベルの組み合わせが爆発的に増加していくため、ほとんどのラベルにおいて、データスパースネスの問題が発生する。これに対して、ベイジアンネットワークを用い、ラベル同士の因果関係を見出すことによって、重要なラベルを取捨選択する手法や [11]、

[1] 前後 2 つまで考慮したクインフォン (quinphone) も広く用いられている

従来のアクセント型に基づくコンテキストに代わり、音声の基本周波数を音素ごとに量子化したものをコンテキストとして用いる手法 [12] などが提案されているが、あまり効果を上げていない。音声認識では、様々な話者による多様な音声から話者性を取り除き、発話内容を決定する問題であるため、統計的機械学習と非常に相性が良いが、音声合成はその逆問題であるため、自然で多様な音声を実現するためには、音声の特徴を適切に捉えた、品質の良いラベルが重要であると考えられる。そしてそのラベルの種類はなるべく少数であり、テキストから容易かつ安定に推定されるものである必要がある。しかし、従来の日本語音声用ラベルには、次のような問題がある。韻律を担う重要な要素であるアクセント（イントネーション）を表現するために用いられているアクセント句は、定義に曖昧性があり、テキストだけではなく、話者や話者の発話速度、発話スタイルに依存するため、自動推定することが困難である。また、文の長さや文中における絶対的な位置情報がラベルとして用いられているが、これでは任意の長さの文を生成可能にするために、非常に多くのラベルの種類を必要とする上、文の一部分のみ発話構造が変化した場合に文全体のラベルが変化してしまう。

1.2　研究の目的

以上の観点から、韻律に注目して HMM 音声合成の改善に取り組む。具体的には、基本周波数パターン生成過程モデルのモデルパラメータ自動抽出性能を高精度化することによって、焦点付与などの柔軟な音声合成を可能にするシステムを実現する。また、HMM 音声合成で用いられているコンテキストラベルを改良することによって、安定してテキストから自動推定することが可能なラベルになり、より利便性の高い音声合成システムを目指す。

1.3　本論文の構成

本論文は全 6 章から構成される。まず第 2 章では、音声合成システムとその周辺技術について述べる。第 3 章では、基本周波数パターン生成過程モデルパラメータの自動抽出に関して、代表的な先行研究と、提案手法についてその詳細を述べる。そして、モデルパラメータ抽出性能の比較実験により、提案手法の有効性を確認する。第 4 章では、前章で提案した生成過程モデルパラメータの自動抽出手法を利用することにより、HMM 音声合成において、韻律の改善、学習コーパスの整備、焦点制御など、様々な有効活用が容易に実現できることを示す。第 5 章では、HMM 音声合成において従来用いられてきたコンテキストラベルの問題点を指摘してから、それを改善するラベルを提案する。そして、その有効性を聴取実験によって確認する。第 6 章で本論文をまとめ、今後の展望について述べる。

第2章

音声合成に関する諸研究

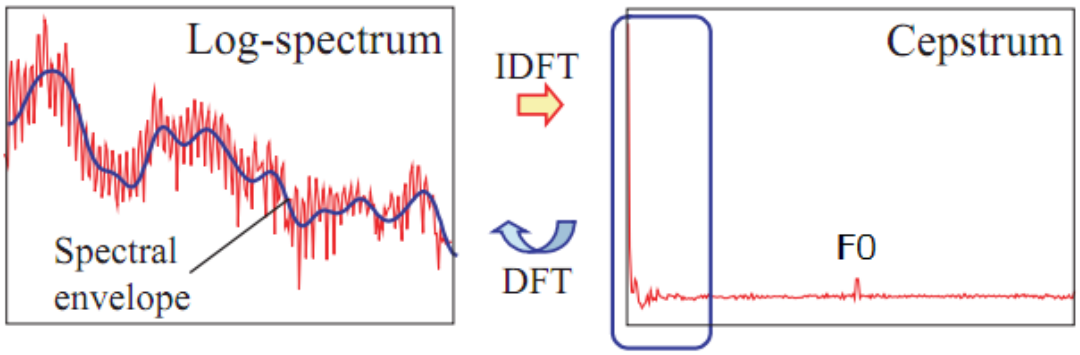


図 2.1: 音声のスペクトルと基本周波数

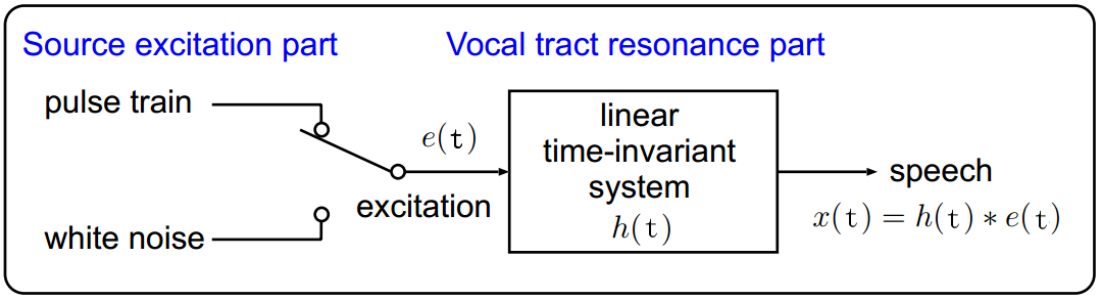


図 2.2: 音声の生成過程と特徴量

2.1　はじめに

本章では、まず音声の生成過程から説明し、そこから導き出される音声の特徴量について説明する。次に、その特徴量の 1 つである基本周波数の時系列パターンをモデル化した基本周波数パターン生成過程モデルについて説明する。最後に、TTS システムの代表的な手法の 1 つである HMM 音声合成について、その要素技術を述べる。

2.2　音声の生成過程と特徴量

2.2.1　音声の生成過程

音声は、声帯の振動あるいは声道のせばめでの乱流で生成した波形が、声道の形状の特性によって周波数的に加工されたものということができる。声道の形状から決まる伝達特性は、個々の音がどの音素に対応するかといった音声の音韻性に深く関与する。これに対して、音源の波形、特に声帯の振動による有声音源の波形とその繰り返しの特徴は、アクセント、イントネーション、リズムといった音声の韻律性に関与している。音韻性と韻律性を比較すると、前者が音素程度の比較的狭い範囲に、後者が単語あるいはそれ以上の広い範囲と関係するため、その観点から、それぞれ分節的特徴、超分節的特徴とよばれる。線形分離等価回路モデル (Source-Filter Model) では、音声信号を $x(t)$ とすると、$x(t)$ は声

帯振動の駆動音源波 $e(t)$ と声道のインパルス応答波 $h(t)$ の畳み込みで表されると考える。$x(t)$、$e(t)$、$h(t)$ それぞれのフーリエ変換を $X(z)$、$E(z)$、$H(z)$ とし、その両辺の対数をとると、次式のように時間領域で畳み込まれた音源信号の成分と声道伝達特性の成分が、和の形で表されることになる。

$$\ln|X(z)| = \ln|E(z)| + \ln|H(z)| \tag{2.1}$$

音声波形から一部分を切り出し、窓関数をかけて、短時間フーリエ変換をして、その対数振幅をプロットしたのが図 2.1 の左側である。この図において、櫛状の成分（赤線のぎざぎざ部分）が声帯音源によるものである。この例のように有声のときは周期性があり、その基本周期（間隔）の逆数を基本周波数という。基本周波数は声の高さにおよそ対応する特徴量である。音声を合成するとき、声帯振動による音源波は、有声のときは、パルス波で、無声のときは、ガウス性雑音（白色雑音）で近似される。一方、包絡成分 (青線）が声道伝達特性によるものであり、スペクトル包絡 ($\ln|H(z)|$) と呼ばれる。このスペクトル包絡が、「あ」などの音を特徴付ける音韻性におよそ対応している。これらを踏まえて、線形分離等価回路モデルによる音声合成の概念図を示したのが図 2.2 である。

ここで、人間の聴覚特性を考慮した上で、スペクトル包絡を効率的に表現する特徴量として、音声合成では、メル一般化ケプストラム（MGC; Mel-generalized cepstral coefficients）などが一般に用いらている。

2.2.2　基本周波数

音声の基本周波数は先に述べたように、およそ声の高さに対応し、イントネーション、ストレス、アクセント、リズムなどを表現し、発話スタイルや感情にも深く関係する重要な特徴量である。基本周波数を抽出する手法は、自己相関関数により波形の周期性を検出する方法 [13] やケフレンシー軸（図 2.1 の右側）でピークを探索することにより抽出する手法 [14] が代表的な手法であるが、YIN らによる手法 [15]、瞬時周波数振幅スペクトルに基づく手法 [16]、高 SNR 下における波形から高速に基本周波数を抽出する手法 [17] など、様々な手法が提案されている。しかし、今日に至るまでディファクトスタンダードと呼べるような手法は存在しないのである。なぜ基本周波数を安定して抽出することが困難であるのかというと、音声は常に周期も振幅も連続的に変化するため、厳密な意味での基本周期は存在せず、音声波形のどの時間区間を周期として採用すれば良いのか自明ではないからである。さらに、音声は声帯の振動を伴う有声（Voice）と振動を伴わない無声（UnVoice)があり、その判別を行うこと（V\UV Detection）が困難であるという問題がある。また、声帯の振動は常に安定したものではなく、さまざまな揺らぎが生じることも抽出を困難にする大きな要因となっている。

一方で自然な音声を合成するためには、すべての基本周波数を正確に再現する必要があるわけではないことが知られている。例えば、人間の聴覚上、マイクロプロソディと呼ばれる微細な変動は、情報伝達に大きな影響を与えるものではないとされている。そして、基本周波数の時系列パターンは、単語あるいはそれ以上の広い範囲に渡って表れる特徴が重要であるため、そのような基本周波数パターンの特徴を適切に捉えたモデル化が重要で

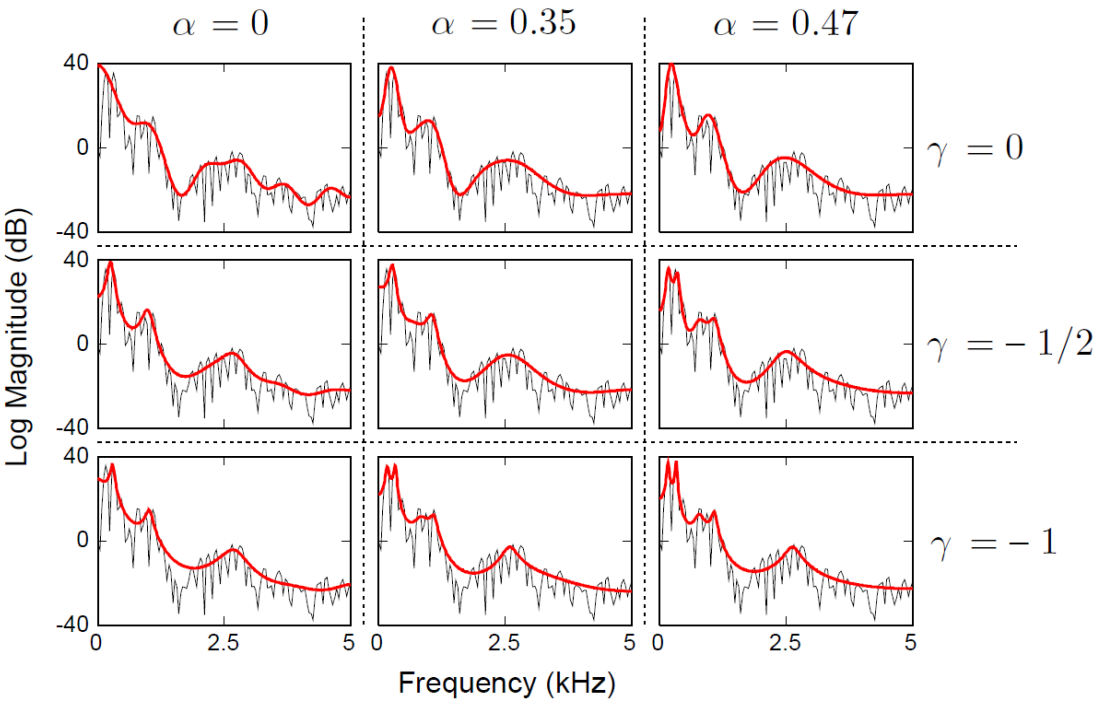


図 2.3: メル一般化ケプストラムによるスペクトル包絡の特性 (文献 [22] より引用)

ある。また、モデル化することによって、少数のパラメータで韻律を制御できるようになるというメリットがある。基本周波数パターンのモデルは、Penta Model や [18]、中国語（北京語）などの声調言語のための Tone nucleus model [19]、Tonal Tilt Model [20] などがあるが、藤崎が提案した基本周波数パターン生成過程モデルは [5]、少数のパラメータで基本周波数パターンを記述することができ、言語情報とよく対応が取れることが知られている。

2.2.3　メル一般化ケプストラム

音声の音韻性を表現する代表的な特徴量として、音声認識では MFCC(Mel-frequency cepstral coefficients) が広く利用されているが、MFCC は、スペクトルのフィルタバンク出力から得ている特徴量であるため、元のスペクトルを再現するという目的には適していない。そこで、音声合成では声道特性の伝達関数として、全極型モデルと極零型モデルが考案されている。全極型は、

$$H(z) = \frac{1}{1 - \sum_{m=0}^{M} c_\gamma(m) z^{-m}} \tag{2.2}$$

で表され、極零型は、

$$H(z) = \exp \sum_{m=0}^{M} c_\gamma(m) z^{-m} \tag{2.3}$$

で表される。全極型は、一般的に比較的少ない次数で、音韻性の再現に重要とされるスペクトルのピーク（フォルマント）をよく捉えることができるが、解の安定性が保証されていないという問題がある。一方極零型は、安定的に解を求めることができるが、その解から得られる曲線は滑らかであり、スペクトル包絡の特徴を十分に再現するためには、一般的に次数を比較的多くする必要があるという問題がある。

そこで、その両方の特性を併せ持つメル一般化ケプストラム（MGC; Mel-generalized cepstral coefficients）が考案された。MGC は次式で表される。

$$H(z) = \left(1 + \gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_{\alpha}^{-m}\right)^{\frac{1}{\gamma}} \tag{2.4}$$

$$z_{\alpha}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{2.5}$$

これは、$\gamma = -1$ のとき、全極型 (式 (2.2)) と等価になり、$\gamma = 0$ のとき、極零型 (式 (2.3)) と等価になる。ここで、式 (2.5) はオールパスフィルタであり、人間の聴覚特性を考慮したメル尺度化に対応している。メル尺度とは、音高の知覚的尺度であり、メル尺度の差が同じであれば、人間が知覚する音高の差が同じであることを意味する。メル周波数 f_{mel} は周波数 f に対して、

$$f_{mel} = 1127.01048 \ln\left(1 + \frac{f}{700}\right) \tag{2.6}$$

のように周波数ウォーピングを施すことで得られる。これは人間の周波数分解能が、およそ対数軸上で等間隔であり、低域ほど分解能が高く、高域ほど分解能が低いことを意味している。この式とほぼ等価な変換が式 (2.5) であり、16 kHz サンプリングでは、およそ $\alpha = 0.42$ が対応しているとされる。

MGC によって生成されたスペクトル包絡の例を図 2.3 に示す。α が大きいほど低域のスペクトルがよく再現されている。また、$\gamma = -1$ のときはスペクトルのピークが強調されており、$\gamma = 0$ のときはスペクトル包絡が滑らかであることがわかる。そして、MGC は MGLSA フィルタによって [21]、効率的に音声合成が可能であることが知られている。

2.3　基本周波数パターン生成過程モデル

2.3.1　基本周波数パターン生成過程モデルの概要

基本周波数パターン生成過程モデルとは、喉頭の生理的・物理的特性に基づいて、声帯振動制御機構を定量的にモデル化したものである。

このモデルは、対数軸状で表現した基本周波数パターンが、次に述べる 2 種類の成分と話者の発話スタイルに固有な値の和として表されるとしている。その 1 つは、句頭から句末に向かう緩やかな下降に対応するもので、これをフレーズ成分と呼ぶ。2 つ目は、個々の単語または単語の連鎖に付属する局所的な起伏に対応するもので、これをアクセント成分と呼ぶ。図 2.4 はこのモデルの概念図であり、文音声の基本周波数パターンを想定した

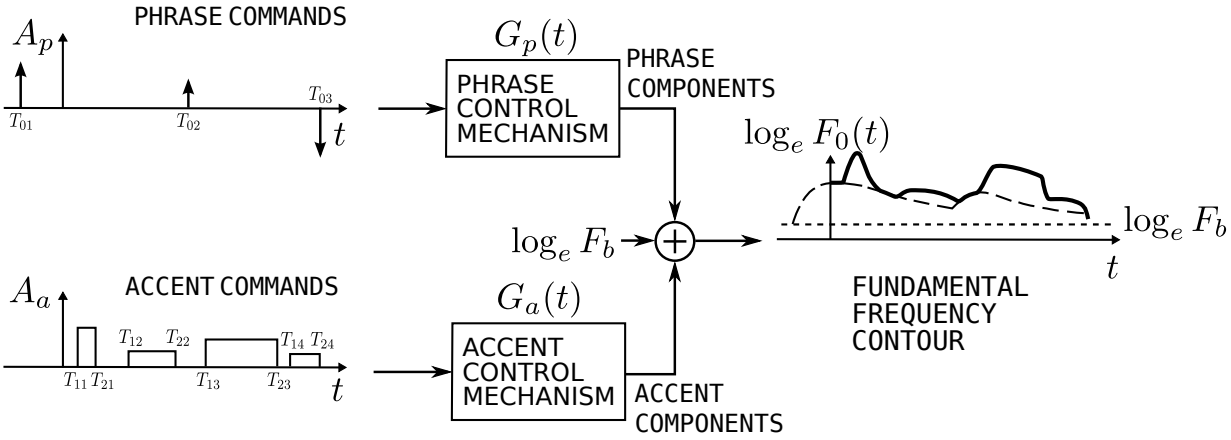


図 2.4: 基本周波数パターン生成過程モデル

ものである。入力となる 2 種類の指令のうち、フレーズ成分の指令はインパルスとして、アクセント指令は方形波として表され、個々の単語または単語連鎖毎に生起してアクセント成分を生じさせている。最後に、この 2 種類の成分と基底周波数が足し合わされて、声帯振動の対数基本周波数パターンが表されている。ここで、時刻 t における基本周波数を $F_0(t)$ とおくと、次式のように表される。

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{p,i} G_p(t - T_{0,i}) + \sum_{j=1}^{J} A_{a,j} \{G_a(t - T_{1,j}) - G_a(t - T_{2,j})\} \qquad (2.7)$$

ただし、式中の変数はそれぞれ次のような意味を持つ。

F_b	:	基本周波数パターンの基底値 (基底周波数)
I	:	文中のフレーズ指令の数
$A_{p,i}$	:	i 番目のフレーズ指令の大きさ
$T_{0,i}$	:	i 番目のフレーズ指令が生起する時点
J	:	文中のアクセント指令の数
$A_{a,j}$	:	j 番目のアクセント指令の大きさ
$T_{1,j}$	:	j 番目のアクセント指令の立ち上がり位置
$T_{2,j}$	:	j 番目のアクセント指令の立ち下がり位置

このモデルでは少数のパラメータで基本周波数パターンをよく表しているが、次に述べる様々な要因により、基本周波数パターンだけからパラメータを推定することが困難であるという問題がある。実測される基本周波数パターンは、式 (2.7) に示す連続曲線ではなく、無声化による基本周波数の欠落、声帯振動の不規則性や基本周波数抽出アルゴリズムの不完全性に起因する誤った基本周波数の抽出がされるためである。

2.3.2　フレーズ成分

フレーズ成分は、単独発話では 1 つであるが、文の発話では複数個存在し得るものであり、声帯振動のはじまりより約 300〜400 ms 前から準備され始め、上昇しながら最大値に達した後、緩やかに下降して一定の値まで漸近していく成分である。この成分は質量とばね定数を持つ 2 次の力学系が瞬間的な外力 (インパルス外力) を受けた場合の運動で表現される。そこで、フレーズ成分を質量、ばね定数、摩擦抵抗を持った力学系のインパルス応答を用いて近似し、かつこの仮想的な力学系が線型性を持つ臨界制動系であると仮定すると、フレーズ成分に相当する $G_p(t)$ は次式で表される。

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t) & (t \geq 0) \\ 0 & (t < 0) \end{cases} \tag{2.8}$$

ここで α はフレーズ指令に対する系の速さを定める係数であり、日本人の話者の平均的な値として、経験的に 3.0 を用いると良いことが知られている [23]。

2.3.3　アクセント成分

基本周波数パターンを構成しているもう 1 つの成分であるアクセント成分は、個々の単語又は連続した単語に付随するもので、アクセントが高いモーラの発音にやや先行して上昇し始め一定の値に漸近し、そのまま高いモーラが続く間は高い値を保ち、低いモーラに移るときにやや先行して下降し始める成分である。この成分は、質量とバネ定数を持った 2 次の力学系が一定時間続くステップ的な外力を受けた時の運動で表現される。そのため、このアクセント成分 $G_a(t)$ は次式で表される。

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma] & (t \geq 0) \\ 0 & (t < 0) \end{cases} \tag{2.9}$$

ここで β はアクセント成分の立ち上がりの速さを定める係数であり、平均的な値として 20.0 が用いられる [23]。また、上式の $\min[1 - (1 + \beta t)\exp(-\beta t), \gamma]$ は、実際の基本周波数パターンにおいて $G_a(t)$ が有限の時間内に上限値 γ に達することを表すためである。この γ の値は通常 0.9 が用いられる [23]。

2.3.4　基本周波数パターン生成過程モデルの生理学的、物理学的根拠

生成過程モデルではフレーズ成分、アクセント成分を分離して線形な力学系を仮定することで、基本周波数パターンを定量的に扱っているが、これは物理的、生理的に妥当であることが藤崎らによって示されている [24]。以下にその概要を示す。

i)　声帯の長さの変化と対数基本周波数の変化との関係

声帯の長さは、咽頭を動きを制御することで変化している。図 2.5 に喉頭の概略図を示

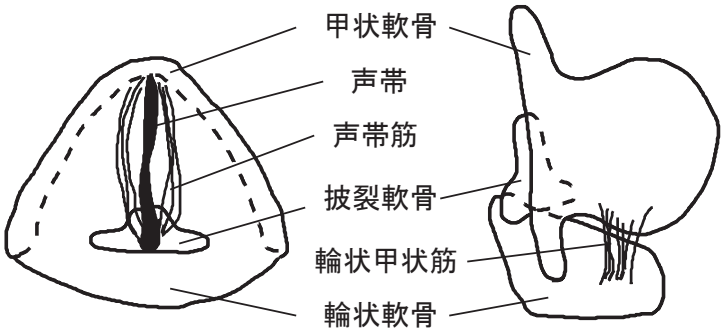


図 2.5: 咽頭の概略図 (左は上から見た図、右は横から見た図)

す。声帯は前後に引っ張られると薄い膜状になるが、その際左右の披裂軟骨を近づけることによって声門を狭め、さらに気流がある程度以上の強さになるとベルヌーイ効果で振動をし始め、声門が開閉する。声帯のすぐ脇にある声帯筋に力を入れて引っ張るとき、そのばね定数は張力にほぼ比例することが実験的に知られている [25, 31]。ここで T を張力、l を声帯の長さ、a を $T = 0$ におけるばね定数、b を比例定数とすると

$$\frac{dT}{dl} = a + bT \tag{2.10}$$

であるからこれを解いて

$$T = \left(T_0 + \frac{a}{b}\right) \exp\left(\, b\,(l - l_0)\right) - \frac{a}{b} \tag{2.11}$$

となる。ただし、T_0 は声帯にあらかじめ加えた張力、l_0 は $T = T_0$ のときの声帯の長さである。このとき $T_0 \gg a/b$ であるとすると

$$T \simeq T_0 \exp(bx) \tag{2.12}$$

となる。ただし x は l_0 を基準としたときの声帯の伸びである。一般的に弾性膜の固有振動数は張力の平方根に比例するので、弾性膜の振動の基本周波数は

$$F_0 = c_0 \sqrt{\frac{T}{\rho}} \tag{2.13}$$

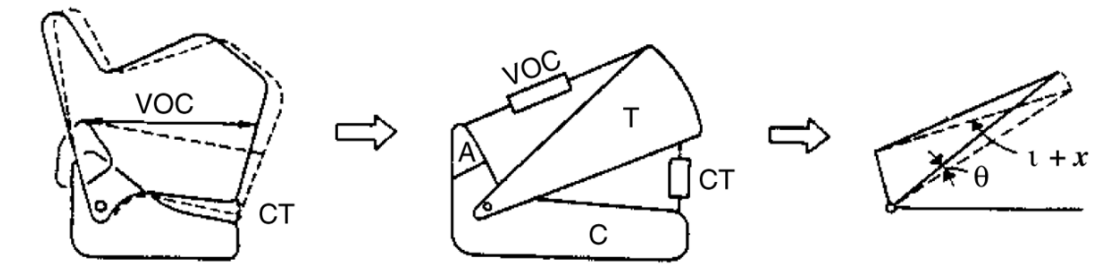
と与えられる。ただし ρ は弾性膜の単位面積あたりの密度、c_0 は弾性膜の大きさに反比例する定数である。したがって式 (2.12),(2.13) から

$$\ln F_0 = \ln F_b + \frac{b}{2} x \tag{2.14}$$

となる。ただし $F_b = \sqrt{T_o/\rho}$ である。また厳密には式 (2.14) の第 1 項は x によって多少変化するが、$\ln F_0$ の値は主に第 2 項によって決まる。以上から x が時間的に変化する場合を考えると

$$\ln F_0(t) = \ln F_b + \frac{b}{2} x(t) \tag{2.15}$$

となり、これは基本周波数の対数が声帯の長さに比例して変化することを示している。



VOC : Vocalis M.
CT　: Criothyroid M.

T: Thyroid
C: Cricoid
A: Arytenoid

ι: Length of vocalis
x: Elongation of vocalis
θ: Angular dispacement of thyroid

図 2.6: 咽頭の構造とその力学系としての取り扱い

ii)　喉頭の調節機構

声の高さの調節に関与する喉頭の構造を模式的に示すと、図 2.6 のようになる。輪状軟骨を基準に考えると、甲状軟骨は声帯筋と輪状甲状筋によって互いに逆の方向へ引っ張られて平衡状態を保っているが、輪状甲状筋の張力が少し強まると、甲状軟骨を前へ倒そうとする。回転するとき、甲状軟骨は 1 つの質量と見なされ、声帯筋と輪状甲状筋はばねと見なすことができる。甲状軟骨と輪状軟骨をつなぐ輪状甲状関節には、運動の自由度が 2 つあり、輪状甲状筋には、甲状軟骨を平行移動させる斜部と回転移動させる直部の 2 つの部分がある。この 2 つの運動を図 2.6 に倣って模式的に考えると、回転運動方程式が求まる。つまり甲状軟骨が平行移動する場合はインパルス関数入力に対する線形 2 次系の応答になり、甲状軟骨が回転する場合はステップ関数入力に対する線形 2 次系の応答になる。すなわち、前者からフレーズ成分が、後者からアクセント成分が生じていると見なすことができる。結局のところ甲状軟骨の平行移動と回転とが声帯の伸びを決定し、それぞれが 2 次系の動きをするので、対数基本周波数パターンは、それらの 2 つの成分の和として反映される。すなわち平行移動による声帯の伸びを $x_1(t)$、回転による声帯の伸びを $x_2(2)$ とすると、式 (2.15) は

$$\ln F_0(t) = \ln F_b + \frac{b}{2}(x_1(t) + x_2(t)) \tag{2.16}$$

となる。以上の観点から基本周波数パターンは、フレーズ成分とアクセント成分の和として表すことができ、従って基本周波数モデルが生理学的、物理学的根拠に基づくモデルであることが示された。

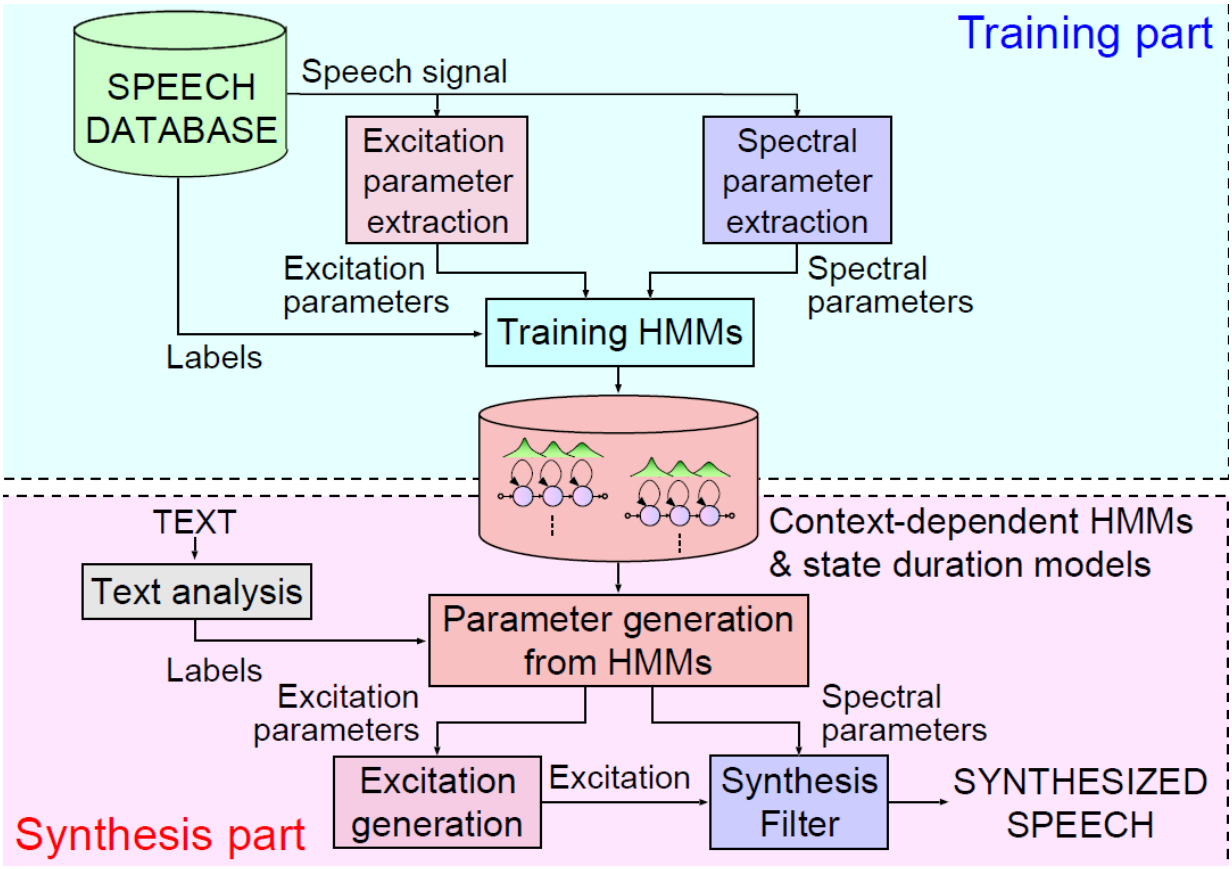


図 2.7: HMM 音声合成 (文献 [22] より引用)

2.4　HMM 音声合成

2.4.1　HMM 音声合成の概要

HMM 音声合成は、TTS(Text-to-Speech) システムの 1 種であり、音声を特徴量で表現する音声分析合成技術と、数理統計モデルの 1 つである隠れマルコフモデル (Hidden Markov Model; HMM) を土台としている。

HMM 音声合成システムの概要を図 2.7 に示す。まず、学習用音声からスペクトル包絡特徴量と基本周波数を抽出し、それぞれの特徴量を連結したベクトルデータとテキストから抽出したラベルデータを用いて HMM を学習する。合成時は、入力テキストから抽出されたラベルに対して HMM から尤度最大化基準によって生成された特徴量を用いて音声を合成する。

HMM 音声合成は、音声を特徴量で取り扱うため、柔軟性に優れ、話者適応などの様々な手法が提案されているが、現状合成音声の品質は十分ではなく、様々な改良が盛んに研究されている。

2.4.2　HMM 音声合成の定式化

TTS は次式のように表現することができる。

$$\tilde{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p(\boldsymbol{x}|\boldsymbol{X}, w, W) \tag{2.17}$$

ここで、$\boldsymbol{x}$ が合成音声、$\boldsymbol{X}$ が学習用音声、w が合成する音声のテキスト、W が学習用音声のテキストである。つまり、学習用音声とそのテキスト及び、入力されたテキストから最も最適な合成音声 $\tilde{\boldsymbol{x}}$ を出力するタスクであると言える。そして次式のように展開する。

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{X}, w, W) &= \int p(\boldsymbol{x}, \boldsymbol{\lambda}|\boldsymbol{X}, w, W)\mathrm{d}\boldsymbol{\lambda} \\ &\simeq \int p(\boldsymbol{x}|w, \boldsymbol{\lambda}) p(\boldsymbol{\lambda}|\boldsymbol{X}, W)\mathrm{d}\boldsymbol{\lambda} \\ &\simeq p(\boldsymbol{x}|w, \hat{\boldsymbol{\lambda}}) \end{aligned} \tag{2.18}$$

ただし、

$$\hat{\boldsymbol{\lambda}} = \underset{\hat{\boldsymbol{\lambda}}}{\operatorname{argmax}}\, p(\boldsymbol{\lambda}|\boldsymbol{X}, W) \tag{2.19}$$

である。このようにモデルパラメータセット $\boldsymbol{\lambda}$ を導入し、確率の加法定理 (sum rule) と乗法定理 (product rule) を用い、積分値はパラメータの最大値で近似することによってこのような式変形ができる。そして、

$$\begin{aligned} \hat{\boldsymbol{\lambda}} &= \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, p(\boldsymbol{\lambda}|\boldsymbol{X}, W) \\ &= \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, \frac{p(\boldsymbol{X}|W, \boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\boldsymbol{X})} \\ &= \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, p(\boldsymbol{X}|W, \boldsymbol{\lambda})p(\boldsymbol{\lambda}) \\ &\simeq \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, p(\boldsymbol{X}|W, \boldsymbol{\lambda}) \end{aligned} \tag{2.20}$$

となる。ここではベイズの定理を用いて式変形している。また、最後の近似は事前分布が一様分布であることを仮定したものであり、ML 推定 (Maximum Likelihood Estimation) と呼ばれる。この事前分布に Gauss-Wishart 分布などの分布を導入すれば、MAP 推定（Maximum a Posteriori Estimation）となる。そしてこの式をさらに変形すると、

$$\begin{aligned} \hat{\boldsymbol{\lambda}} &= \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, p(\boldsymbol{X}|W, \boldsymbol{\lambda}) \\ &= \underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \int p(\boldsymbol{X}, \boldsymbol{O}|W, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{O} \\ &\simeq \underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \int p(\boldsymbol{X}|\boldsymbol{O}) p(\boldsymbol{O}|W, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{O} \\ &\simeq \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{O}}|W, \boldsymbol{\lambda}) \end{aligned} \tag{2.21}$$

ただし、

$$\hat{\boldsymbol{O}} = \underset{\boldsymbol{O}}{\operatorname{argmax}}\, p(\boldsymbol{X}|\boldsymbol{O}) \tag{2.22}$$

である。ここで導入したモデルパラメータセット $\boldsymbol{O}$ は、学習用音声の特徴量と考えることができる。最後に、式 (2.18) についても同様に式変形をする。

$$\begin{aligned}
\tilde{\boldsymbol{x}} &= \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p(\boldsymbol{x}|\boldsymbol{X}, w, W) \\
&\simeq \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p(\boldsymbol{x}|w, \hat{\boldsymbol{\lambda}}) \\
&= \underset{\boldsymbol{x}}{\operatorname{argmax}} \int p(\boldsymbol{x}, \boldsymbol{o}|w, \hat{\boldsymbol{\lambda}}) \mathrm{d}\boldsymbol{o} \\
&\simeq \underset{\boldsymbol{x}}{\operatorname{argmax}} \int p(\boldsymbol{x}|\boldsymbol{o}) p(\boldsymbol{o}|w, \hat{\boldsymbol{\lambda}}) \mathrm{d}\boldsymbol{o} \\
&\simeq \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p(\boldsymbol{x}|\hat{\boldsymbol{o}})
\end{aligned} \tag{2.23}$$

ただし、

$$\hat{\boldsymbol{o}} = \underset{\boldsymbol{o}}{\operatorname{argmax}}\, p(\boldsymbol{o}|w, \boldsymbol{\lambda}) \tag{2.24}$$

である。ここで導入したモデルパラメータセット $\boldsymbol{o}$ は、合成音声の特徴量と考えることができる。

以上の式展開によって得られた式をまとめると以下のようになる。

$$\hat{\boldsymbol{O}} \simeq \underset{\boldsymbol{O}}{\operatorname{argmax}}\, p(\boldsymbol{X}|\boldsymbol{O}) \tag{2.25}$$

$$\hat{\boldsymbol{\lambda}} \simeq \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{O}}|W, \boldsymbol{\lambda}) \tag{2.26}$$

$$\hat{\boldsymbol{o}} \simeq \underset{\boldsymbol{o}}{\operatorname{argmax}}\, p(\boldsymbol{o}|w, \boldsymbol{\lambda}) \tag{2.27}$$

$$\hat{\boldsymbol{x}} \simeq \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p(\boldsymbol{x}|\hat{\boldsymbol{o}}) \tag{2.28}$$

これら 4 つの式の内、上 2 つはそれぞれ学習用音声からの特徴量の抽出、その特徴量から音響モデルの学習と解釈できる。そして、下 2 つはそれぞれモデルからのパラメータ生成、生成したパラメータからの音声合成と解釈できる。

ここで音響モデルとして、隠れマルコフモデル (HMM) を仮定する手法が HMM 音声合成である。

2.4.3　隠れマルコフモデル (HMM)

隠れマルコフモデル (Hidden Markov Model; HMM) は、不確定な時系列のデータをモデル化するための有効な統計的手法である。出力シンボル系列が与えられても状態遷移系列は一意に決まらず、観測できるのはシンボル系列だけであることから hidden(隠れ) マルコ

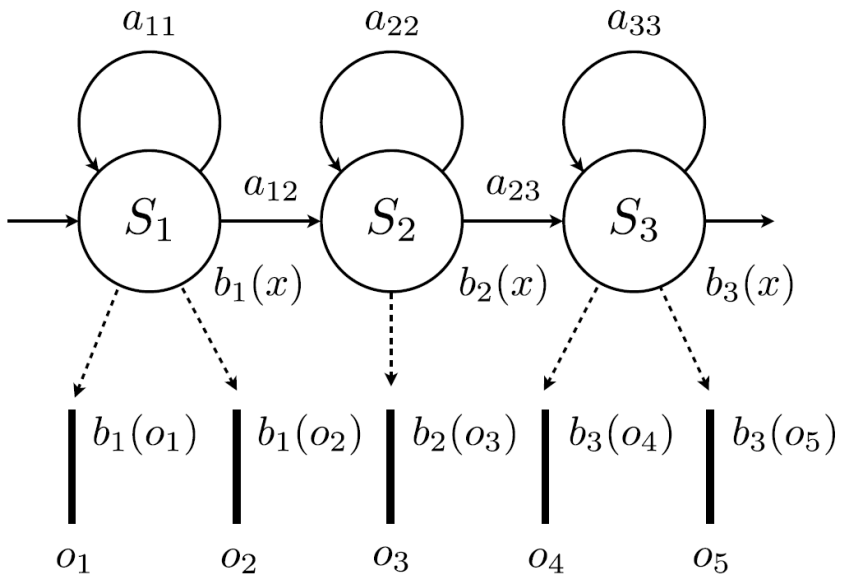


図 2.8: 隠れマルコフモデル (HMM)

フモデルと呼ばれる。図 2.8 に示すのは典型的な left-to-right 型の HMM である。a_{ij} は状態 S_i から状態 S_j への遷移確率を表し、$b_i(x)$ は状態 S_i における観測シンボル x の生起確率分布を表す。HMM 音声合成では、観測シンボル o_i が各フレームにおける音声の特徴量をまとめたベクトルになる。そして、生起確率分布 $b_i(x)$ には多くの場合、ガウス分布に基づくものが用いられる。HMM は音声認識の分野ですでに広く使われている手法である。

ここで、入力 λ が与えられたときに、出力シンボル系列 $\boldsymbol{o} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T\}$ が観測される確率は、forward-backward アルゴリズムによって効率的に解くことができる。前向き確率を $\alpha_t(\cdot)$ とし、後向き確率を $\beta_t(\cdot)$ とし、次式のように定義する。

$$\alpha_0(j) = \begin{cases} 1 & j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.29}$$

$$\begin{aligned} \alpha_t(j) &= P(\boldsymbol{o}_1, \cdots, \boldsymbol{o}_t, q_t = j|\lambda) \\ &= \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(\boldsymbol{o}_t) \quad \left(\begin{array}{l} t = 1, 2, \cdots, T \\ 1 \leq j \leq N \end{array} \right) \end{aligned} \tag{2.30}$$

$$\beta_{T+1}(i) = \begin{cases} 1 & i = N \\ 0 & \text{otherwise} \end{cases} \tag{2.31}$$

$$\begin{aligned} \beta_t(i) &= P(\boldsymbol{o}_{t+1}, \cdots, \boldsymbol{o}_T, |q_t = i, \lambda) \\ &= \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j) \quad \left(\begin{array}{l} t = T-1, \cdots, 1 \\ 1 \leq i \leq N \end{array} \right) \end{aligned} \tag{2.32}$$

ここで、$q_t = j$ は時刻 t に j 番目のステートにいる状態のことを指す。すると、入力 λ が与

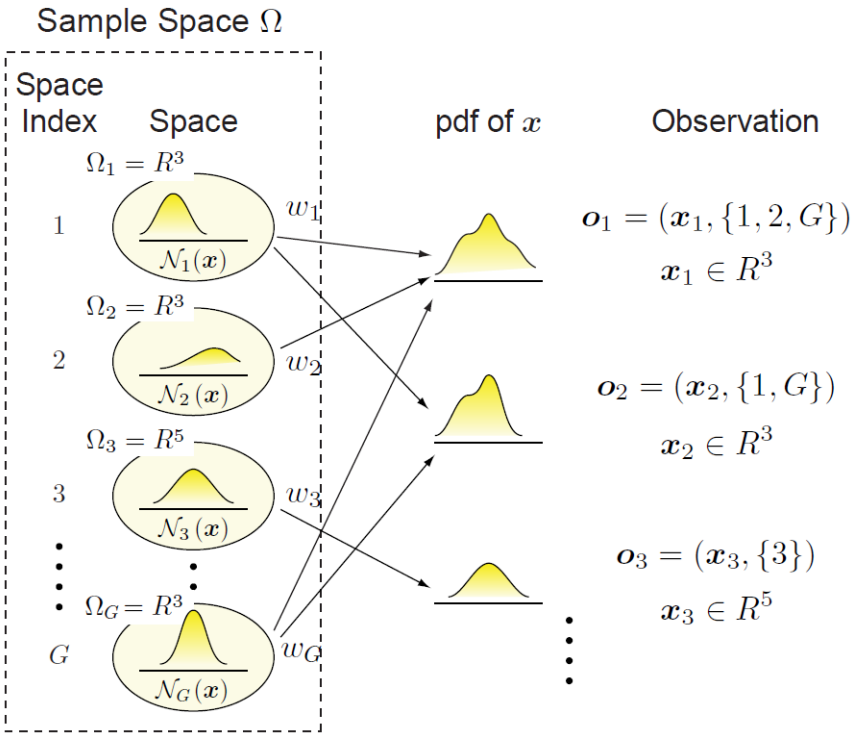


図 2.9: 多空間確率分布 (文献 [22] より引用)

えられたときに、出力シンボル系列 $\boldsymbol{o}$ が観測される確率は次式によって求まる。

$$\begin{aligned} P(\boldsymbol{o}|\lambda) &= \sum_{i=1}^{N} P(\boldsymbol{o}, q_t = i|\lambda) \\ &= \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \end{aligned} \tag{2.33}$$

また、学習データからモデルパラメータ $\theta = \{a_{ij}, b_i(\boldsymbol{x})\}$ を求めることは、Baum-Welch アルゴリズムによって局所解を求めることができる。

2.4.4　多空間確率分布 HMM

F0 は無声区間では定義されない（値を持たない）ため、通常の HMM では F0 を取り扱うことができない。そこで多空間確率分布という概念を導入する。

多空間確率分布について図 2.9 を例に説明する。G 個の空間 $\Omega_1, \Omega_2, \cdots, \Omega_G$ からなる標本空間 Ω を考える。

$$\Omega = \bigcup_{g=1}^{G} \Omega_g \tag{2.34}$$

各空間 Ω_g は n_g 次元の実空間 R^{n_g} とする。G 個の空間の次元 n_g は、互いに異なる値でも一部が同じ値でも良い。各空間 Ω_g の確率を w_g とし、各空間のもつ確率密度関数を $p^{(g)}(\boldsymbol{x}),\ \boldsymbol{x} \in$

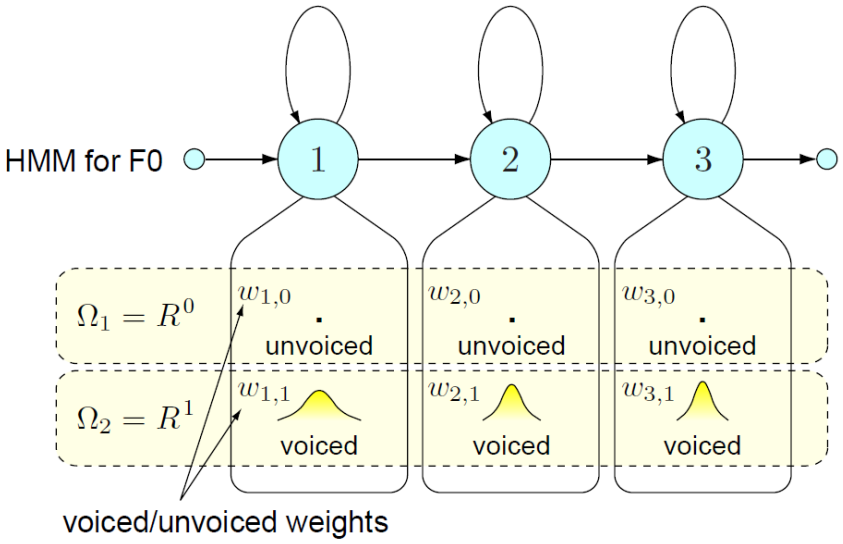


図 2.10: 多空間確率分布 HMM(文献 [22] より引用)

R^{n_g} とする。ただし、

$$\sum_{g=1}^{G} w_g = 1 \tag{2.35}$$

かつ、

$$\int p^{(g)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1 \tag{2.36}$$

であるとする。また、$n_g = 0$ のときは、$p^{(g)}(\cdot) = 1$ と定義する。そして、空間インデックスの集合 X とベクトル $\boldsymbol{x}$ から、$\boldsymbol{o}$ が観測されると考える。つまり、

$$\boldsymbol{o} = (X, \boldsymbol{x}) \tag{2.37}$$

である。ただし、X に含まれる空間インデックスが表す空間は、全て同じ次元でなければならない。このとき、$\boldsymbol{o}$ の観測確率は次式で表される。

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g p^{(g)}(V(\boldsymbol{o})) \tag{2.38}$$

ただし、

$$S(\boldsymbol{o}) = X, \quad V(\boldsymbol{o}) = \boldsymbol{x} \tag{2.39}$$

である。

この多空間確率分布を HMM に導入したモデルが多空間確率分布 HMM(MSD-HMM) である [26]。これにより、図 2.10 のように、F0 を出力する分布と無声であることを示す（値を出力しない 0 次元の）分布を同時に内包することができる。

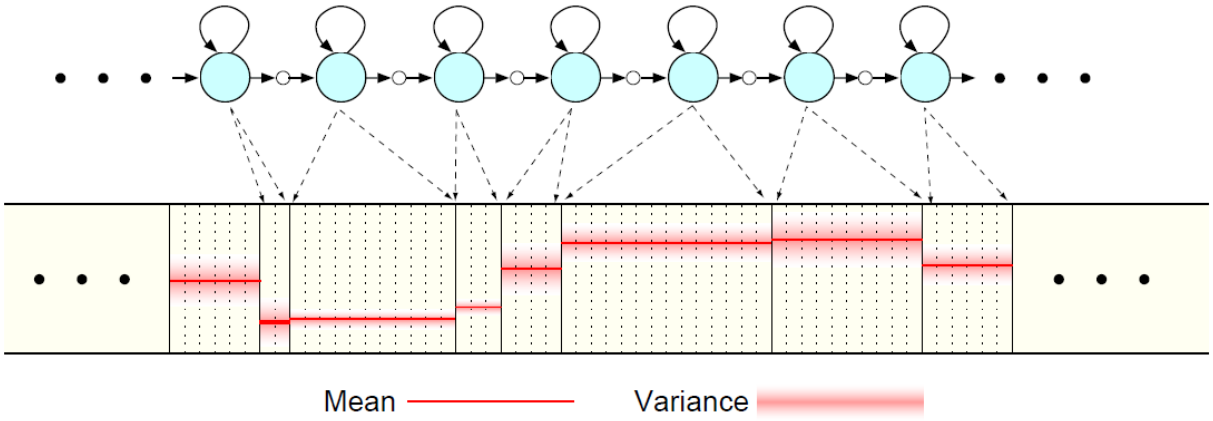


図 2.11: 動的特徴量を用いていない HMM の例 (文献 [22] より引用)

2.4.5　動的特徴量とマルチストリーム化

HMM 音声合成では、合成時に尤度最大化基準でパラメータを生成するため、各 HMM から出力されるパラメータは、基本的に各 HMM の生起確率分布であるガウス分布の平均値となる。しかし、このままでは図 2.11 に示されるように、出力されるパラメータは同じ HMM に所属する限り、常に一定の値になってしまう。

そこで、この問題を解決するために動的特徴量が用いられている。次式で示されるように、HMM の観測シンボルは、静的特徴量だけでなく、その 1 次微分、2 次微分も特徴量として加えている。

$$\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta^2\boldsymbol{c}_t^\top]^\top \tag{2.40}$$

$$\Delta\boldsymbol{c}_t = \frac{1}{2}(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1}) \tag{2.41}$$

$$\Delta^2\boldsymbol{c}_t = \boldsymbol{c}_{t-1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t+1} \tag{2.42}$$

これによって、パラメータ生成時に動的特徴量を制約に加えることにより、滑らかなパラメータ生成を可能にしている。

ここで、音韻的特徴と韻律的特徴の時間的対応が取れるように、MGC と F0 は同一のベクトルに纏められているが、これらの独立性が比較的高いことが知られている。さらに、F0 の有声/無声の判定は空間選択の重みの大小だけによって決定されるため、F0 の動的特徴量はロバスト性が低い。そのため、マルチストリーム化によってそれぞれ重みを付けて独立の分布として取り扱う。

以上より、HMM 音声合成で用いられる特徴量ベクトルの構成は図 2.12 のようになる。ここで、$\boldsymbol{c}_t$ が MGC 等のスペクトル包絡特徴量であり、p_t が F0 である。また、観測ベクトル $\boldsymbol{o}_t$ は次式によって求められる。

$$b_j(\boldsymbol{o}_t) = \prod_{s=1}^{S} (b_j^{(s)}(\boldsymbol{o}_t^{(s)}))^{w_s} \tag{2.43}$$

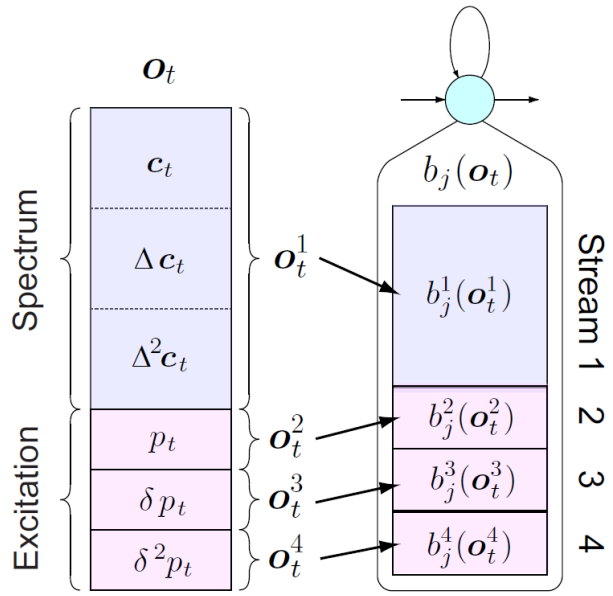


図 2.12: 特徴量ベクトルの構成 (文献 [22] より引用)

$$p_L - p_C + p_R/A : a_{C1}\_a_{C2}/B : b_{L1}\_b_{L2}\_b_{L3} - b_{C1}\_b_{C2}\_b_{C3} + b_{R1}\_b_{R2}\_b_{R3}$$
$$/C : c_{L1}\_c_{L2}\_c_{L3}\_c_{L4} - c_{C1}\_c_{C2}\_c_{C3}\_c_{C4}\_c_{C5} + c_{R1}\_c_{R2}\_c_{R3}\_c_{R4}$$
$$/D : d_{L1} - d_{C1}\_d_{C2} + d_{R1}/E : e$$

図 2.13: コンテキストラベルの構成

2.4.6　コンテキスト依存モデル

TTS は、テキストから音韻的特徴と韻律的特徴の両方を同時に再現しなければならないため、HMM の単位は、図 2.13 のように非常に複雑なラベルが用いられている。ラベルの内容については第 5 章にて詳しく扱うので、ここではラベルのクラスタリングについて述べる。図 2.13 に示されるラベルの種類は、各要素ごとの種類の積になるため非常に大きい数となる。このままでは、学習データのサンプル数よりもラベルの種類の方が多くなり、データスパーネスの問題が発生してしまうため、決定木によるクラスタリングが行われる。HMM のクラスタ S が質問 q によってクラスタ S_{q+} とクラスタ S_{q-} に分割される際に、その分割を適応するか否かを決定する基準となるのが次式で示される最小記述長 (MDL) 基準である [27]。

$$\Delta q = \frac{1}{2}\left\{\Gamma(S_{q+})\log|\Sigma_{S_{q+}}| + \Gamma(S_{q-})\log|\Sigma_{S_{q-}}| - \Gamma(S)\log|\Sigma_S|\right\} + K\log|\Gamma(S_0)| \quad (2.44)$$

ここで、$\Gamma(S)$ はクラスタ S に含まれる学習データの量、Σ は各クラスタの共分散行列、K は特徴量ベクトルの次元数、S_0 は決定木のルートクラスタである.

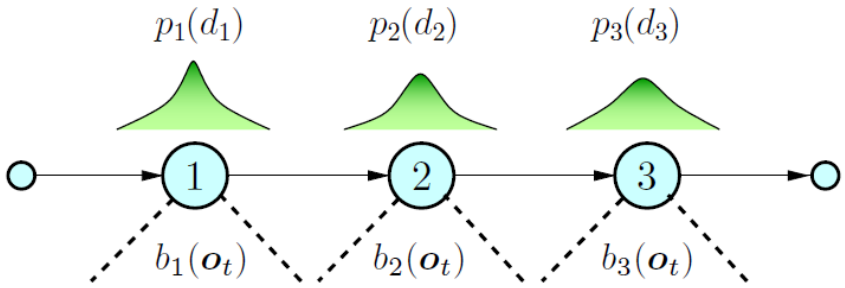


図 2.14: 隠れセミマルコフモデル (HSMM)　(文献 [22] より引用)

2.4.7　状態継続長モデル

HMM は音声認識で幅広く用いられてきたモデルであるが、このままでは継続長のモデル化に問題が生じる。なぜなら、遷移確率は必ず 1 以下であるため、尤度最大化基準でそのままパラメータを推定すると、各ステートを必ず 1 回しか通らないためである。そのため、初期の HMM 音声合成では、学習時に各ステートで留まる回数をカウントしておき、別途継続長をモデル化するという事が行われていたが、明示的に継続長分布を含んだ隠れセミマルコフモデル (Hidden Semi-Markov Model, HSMM) を利用する手法が、全らによって提案された [28]。

ここで、入力 λ' が与えられたときに、出力シンボル系列 $\boldsymbol{o} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T\}$ が観測される確率は、従来の HMM と同様に forward-backward アルゴリズムによって効率的に解くことができる。前向き確率を $\alpha'_t(\cdot)$ とし、後向き確率を $\beta'_t(\cdot)$ とし、次式のように定義する。

$$\alpha'_0(j) = \begin{cases} 1 & j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.45}$$

$$\begin{aligned} \alpha'_t(j) &= P(\boldsymbol{o}_1, \cdots, \boldsymbol{o}_t, q_t = j | q_{t+1} \neq j, \lambda') \\ &= \sum_{d=1}^{t} \sum_{\substack{i=1 \\ i \neq j}}^{N} \alpha'_{t-1}(i) a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\boldsymbol{o}_s) \quad \left(\begin{array}{l} t = 1, 2, \cdots, T \\ 1 \leq j \leq N \end{array} \right) \end{aligned} \tag{2.46}$$

$$\beta'_{T+1}(i) = \begin{cases} 1 & i = N \\ 0 & \text{otherwise} \end{cases} \tag{2.47}$$

$$\begin{aligned} \beta'_t(i) &= P(\boldsymbol{o}_{t+1}, \cdots, \boldsymbol{o}_T, | q_t = i, q_{t+1} \neq i, \lambda') \\ &= \sum_{d=1}^{T-t} \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij} p_j(d) \prod_{s=t+1}^{t+d} b_j(\boldsymbol{o}_s) \beta'_{t+d}(j) \quad \left(\begin{array}{l} t = T-1, \cdots, 1 \\ 1 \leq i \leq N \end{array} \right) \end{aligned} \tag{2.48}$$

ここで、$q_t = j$ は時刻 t に j 番目のステートにいる状態のことを指す。すると、入力 λ' が与えられたときに、出力シンボル系列 $\boldsymbol{o}$ が観測される確率は次式によって求まる。

$$P(\boldsymbol{o}|\lambda') = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{d=1}^{t} \alpha'_{t-d}(i) a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\boldsymbol{o}_s) \beta'_t(j) \tag{2.49}$$

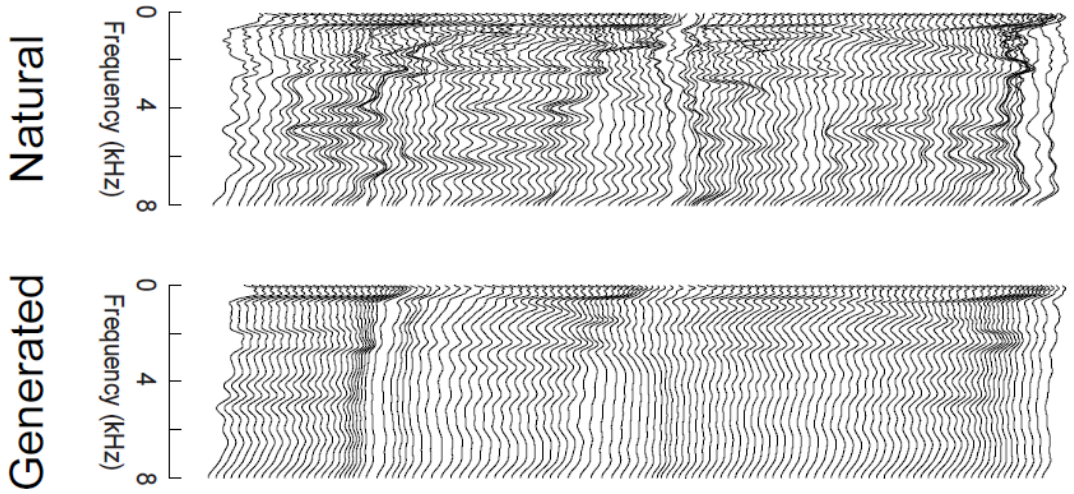


図 2.15: 合成音声における過剰なスペクトルの平滑化 (文献 [22] より引用)

また、学習データからモデルパラメータを求めることも、従来の HMM と同様に Baum-Welch アルゴリズムによって局所解を求めることができる。

この HSMM は状態継続長用に別途モデル化する必要がない上、従来の HMM における話者適応等の技術をそのまま利用することができ、その上、継続長も含めて適応することが容易に可能になるという利点がある。

2.4.8　系列内変動 (GV)

HMM のような何らかの統計手法を利用して、尤度最大化基準でパラメータを生成すると、学習データの「平均値」が出力されることなる。しかし、このまま得られたパラメータで音声を合成すると、なまった音声になることが知られている。HMM から生成されたスペクトルの例を図 2.15 に示す。自然音声に比べて、微細な構造が失われていることがわかる。そして、このような過剰な平滑化により音声の肉声感が失われてしまうという問題がある。

そこで、発話全体にわたる分散を制約条件に加えることにより、過剰な平滑化を防ぐ手法を戸田らが提案した [29]。これを Gloval Variance(GV) という。学習に用いる特徴量ベクトルの次元数を D とし、発話全体の総フレーム数を I とすると、特徴量ベクトル系列 $\boldsymbol{c}$ の系列内変動 $v(\boldsymbol{c})$ は、以下の式で表される。

$$v(\boldsymbol{c}) = [v(1), v(2), \cdots, v(D)] \tag{2.50}$$

$$v(d) = \frac{1}{I}\sum_{i=1}^{I}(c_{i,d} - \bar{c_d})^2 \tag{2.51}$$

$$\bar{c_d} = \frac{1}{I}\sum_{i=1}^{I} c_{i,d} \tag{2.52}$$

この系列内変動 $v(\boldsymbol{c})$ を利用したパラメータ生成法は次節で述べる。

2.4.9　音響特徴量生成アルゴリズム

テキストから音声の特徴量の生成は、次式によって定式化される。

$$p(\boldsymbol{o}|w,(\hat{\lambda})) = \sum_{q} p(\boldsymbol{o}|\boldsymbol{q},\hat{\lambda})p(\boldsymbol{q}|w,\hat{\lambda}) \tag{2.53}$$

ここで、2.4.2 で利用した手法と同様の近似式を用いることにより、

$$\hat{\boldsymbol{q}} = \underset{\boldsymbol{q}}{\operatorname{argmax}}\, p(\boldsymbol{q}|w,\hat{\lambda}) \tag{2.54}$$

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, p(\boldsymbol{o}|\hat{\boldsymbol{q}},\hat{\lambda}) \tag{2.55}$$

となる。ただし、合成時に必要なパラメータは静的特徴量なので、合成時は $\boldsymbol{o}$ ではなく、$\boldsymbol{c}$ について解くことに注意[1]する。これは、状態系列（継続長）を先に決定し、それを用いてパラメータ生成することを意味している。そして式 (2.55) は、$\boldsymbol{c}$ で偏微分することにより解析的に解くことができ、以下の式で音声の特徴量系列を得ることができる。

$$\boldsymbol{W}^{\top}\Sigma_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\boldsymbol{c} = \boldsymbol{W}^{\top}\Sigma_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{\mu}_{\hat{\boldsymbol{q}}} \tag{2.56}$$

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c} \tag{2.57}$$

一方、前節で解説した GV を導入したパラメータ生成は次式で表される。

$$\underset{\boldsymbol{c}}{\operatorname{argmax}}\, p(\boldsymbol{o}|\hat{\boldsymbol{q}},\hat{\lambda})^{\omega} p(v(\boldsymbol{c})|\hat{\lambda}_v) \tag{2.58}$$

これは解析的に解くことができないため、勾配法やニュートン法により逐次アルゴリズムを利用してパラメータを生成する。

2.5　まとめ

本章では、今日の音声工学における音声合成のための要素技術について述べた。この章で取り扱った基本周波数パターン生成過程モデルと HMM 音声合成は、本研究の核となる重要な技術である。

[1] この学習時と生成時で最適化するパラメータが異なるミスマッチを解消する手法として、全らが Trajectry HMM という手法を提案している [30]

第3章

基本周波数パターン生成過程モデルパラメータの自動抽出の高精度化

3.1　はじめに

本章では、前章で述べた基本周波数パターン生成過程モデルについて、そのモデルパラメータを音声から自動抽出する手法について述べる。

先行研究としては、Left-to-Right AbS の処理を用いる手法 [32]、ローパスフィルタを用いる方法 [33,34]、F0 パターンに対する LPC 分析を用いる手法 [35]、ハイパスフィルタを用いる方法 [36] などがあるが、ここではまず初めに代表的な例として、成澤、古山らによる 3 次スプラインを用いた手法について解説する [8–10]。成澤の手法は F0 パターンから直接パラメータを推定する手法であり、古山の手法は、成澤の手法に対して、言語情報を用いて修正したものとなっている。次に、その手法の問題点を指摘してから、新しいモデルパラメータの自動抽出手法を提案する。そして、モデルパラメータの抽出性能の比較実験により、提案手法の有効性を示す。

3.2　従来手法

成澤の手法は、前処理、初期値抽出、最適化の 3 段階からなる。前処理では、F0 パターンの平滑化と F0 が存在しない無声区間を補間するために、3 次スプラインを用いる。初期値抽出では、3 次スプラインの微分係数を利用することにより、初期値を決定する。そして、最適化では、AbS（Analysis-by-Synthesis）処理によってパラメータを微修正する。

3.2.1　前処理

抽出された基本周波数パターンに対して、無声区間 $[t_{i+1}, t_{j-1}]$ は、次式のような 3 次曲線近似によって補間される。

$$\ln F_0(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \tag{3.1}$$

ここで、係数 $[a_0, a_1, a_2, a_3]$ は次式を満たす。

$$\begin{cases} \ln F_0(t_i) &= a_0 + a_1 t_i + a_2 t_i^2 + a_3 t_i^3 \\ G_0(t_i) &= a_1 + 2a_2 t_i + 3a_3 t_i^2 \\ \ln F_0(t_j) &= a_0 + a_1 t_j + a_2 t_j^2 + a_3 t_j^3 \\ G_0(t_j) &= a_1 + 2a_2 t_j + 3a_3 t_j^2 \end{cases} \tag{3.2}$$

$G_0(t)$ は $\ln F_0(t)$ の勾配である。そして、最後に F0 パターン全体に対し、区分的三次曲線近似による平滑化を行う。係数の計算方法は次のようなものである。まず、有声区間の先頭区分においては次式により係数を求める。

$$\begin{cases} \Sigma_{i=1}^{N_1} \ln F_0(t_i) &= N_1 a_0 + \Sigma_{i=1}^{N_1} a_1 t_i + \Sigma_{i=1}^{N_1} a_2 t_i^2 + \Sigma_{i=1}^{N_1} a_3 t_i^3 \\ \Sigma_{i=1}^{N_1} \ln F_0(t_i) t_i &= \Sigma_{i=1}^{N_1} a_0 t_i + \Sigma_{i=1}^{N_1} a_1 t_i^2 + \Sigma_{i=1}^{N_1} a_2 t_i^3 + \Sigma_{i=1}^{N_1} a_3 t_i^4 \\ \Sigma_{i=1}^{N_1} \ln F_0(t_i) t_i^2 &= \Sigma_{i=1}^{N_1} a_0 t_i^2 + \Sigma_{i=1}^{N_1} a_1 t_i^3 + \Sigma_{i=1}^{N_1} a_2 t_i^4 + \Sigma_{i=1}^{N_1} a_3 t_i^5 \\ \Sigma_{i=1}^{N_1} \ln F_0(t_i) t_i^3 &= \Sigma_{i=1}^{N_1} a_0 t_i^3 + \Sigma_{i=1}^{N_1} a_1 t_i^4 + \Sigma_{i=1}^{N_1} a_2 t_i^5 + \Sigma_{i=1}^{N_1} a_3 t_i^6 \end{cases} \tag{3.3}$$

N_1 は有声区間の先頭 200[ms] 中に含まれるフレームの総数であり、前式より求めた係数は有声区間の先頭 150[ms] を平滑化するのに用いられる。それ以降の区分については、次式から求めた係数を用いて、150[ms] ごとに平滑化を行う。

$$\begin{cases} \ln F_0(t_j) & = a_0 + a_1 t_j + a_2 t_j^2 + a_3 t_j^3 \\ G_0(t_j) & = a_1 + 2a_2 t_j + 3a_3 t_j^2 \\ \Sigma_{i=j}^{j+N_2} \ln F_0(t_i) & = N_2 a_0 + \Sigma_{i=j}^{j+N_2} a_1 t_i + \Sigma_{i=j}^{j+N_2} a_2 t_i^2 + \Sigma_{i=j}^{j+N_2} a_3 t_i^3 \\ \Sigma_{i=j}^{j+N_2} \ln F_0(t_i) t_i & = \Sigma_{i=j}^{j+N_2} a_0 t_i + \Sigma_{i=j}^{j+N_2} a_1 t_i^2 + \Sigma_{i=j}^{j+N_2} a_2 t_i^3 + \Sigma_{i=j}^{j+N_2} a_3 t_i^4 \end{cases} \tag{3.4}$$

ただし、j は前区分の直後にあるフレームの添字であり、N_2 は 150[ms] 中に含まれるフレームの総数である。

このようにして得られた F0 パターンは至るところで連続かつ微分可能であり、これを以降平滑化 F0 パターンと呼ぶ。

3.2.2　初期値抽出

i) アクセント指令の初期値抽出
前節で求めた 3 次曲線の 1 次微分の正の極値と負の極値を 1 つの組として、1 つのアクセント指令とする。
アクセント指令の大きさは、正の極値の時刻を $t_{1,j}$ とすると、次式となる。

$$A_{a,j} = \frac{e}{\beta} G_0(t_{2,j}) \tag{3.5}$$

同様にして、負の極値についても大きさを求め、正と負の両方の極値から求めた指令の大きさの平均値を初期値とする。ただし、片方の極値しか存在しない場合は、式 (3.5) の値をそのまま初期値とする。
アクセント指令の初期位置は、極大値、極小値の位置より $1/\beta$ 前の時点をそれぞれ立ち上がり位置、立ち下がり位置とする。ただし、発話の先頭で極小値が先に検出された場合は、1 型のアクセント句であると考え、立ち上がり位置は立ち下がり位置より平均 1 モーラ長（1/7s）前の時点とし、また、発話の終わりで極小値が検出されなかった場合は、0 型のアクセント句であると考え、発話の終わりの位置を立ち下がり位置とする。

ii) 第 1 フレーズ指令の初期値抽出
平滑化 F0 パターンからアクセント成分を差し引いた残差パターン $R_0(t)$ から、第 1 フレーズ指令を推定する。ここで、各発話ごとの平滑化 F0 パターンの開始時点を T_s とし、j 番目 ($j = 1, 2, \cdots, J$) のアクセント指令の立ち上がり位置を $T_{1,j}$、立ち下がり位置を $T_{2,j}$ とし、$R_0(t)$ が最大値となる時刻を $T_{R_0\max}$ とする。このとき、第 1 フレーズ指令の生起位置は、$T_{R_0\max}$ より $1/\alpha$ 前の時点とする。また、第 1 フレーズ指令の大きさは次式から求める。

$$A_{p,1} = \frac{e}{\alpha} \{\ln R_0(T_{R_0\max}) - \ln F_b\} \tag{3.6}$$

そして、区間 $[T_s - 0.5, T_{2,1}]$ における $R_0(t)$ に対して、フレーズ指令のみパラメータの逐次処理をして、部分的に最適化する。

iii) 第 2 フレーズ指令以降の初期値抽出

第 2 以降のフレーズ指令は残差 F0 パターンの積分値を利用して、一定値以上の閾値に達したとき、指令を抽出する。

これまでに推定された i 個のフレーズ指令のパラメータの初期値から求められるフレーズ成分を $R_0(t)$ から差し引いた残差パターンを $R_i(t)$ とする。区間 $[T_{2,j}, T_{2,J}]$ にわたって $R_i(t)$ の積分値を求め、その値がある閾値を超えた場合、$R_i(t)$ の積分値が増加し始める地点を $i+1$ 番目のフレーズ指令の生起位置 $T_{0,i+1}$ とする。ただし、$T_{0,i+1}$ が $j+1$ 番目のアクセント指令の内部に存在する場合は、$T_{2,j}$ と $T_{1,j+1}$ の間で平滑化 F0 パターンの 1 次微分係数が負から正へと変化する時点を $T_{0,i+1}$ とし、この時点が存在しないならば、$(T_{2,j} + T_{1,j+1})/2$ を $T_{0,i+1}$ とする。そして、時刻 $t_{i\max} = T_{0,i+1} + 1/\alpha$ における $R_i(t_{i\max})$ を $R_{i\max}$ とし、次式によりフレーズ指令の大きさを求める。

$$A_{p,i+1} = \frac{e}{\alpha}\{\ln R_{i\max} - \ln F_b\} \tag{3.7}$$

そして、区間 $[T_{2,j}, t]$ における $R_i(t)$ に対して、フレーズ指令のみパラメータの逐次処理をして、部分最適化する。以降、$R_i(t)$ の積分値をリセットし、i の値を 1 増やして一連の作業を繰り返す。

3.2.3　最適化

前節までに求めたモデルパラメータを逐次処理により最適化する。各分析区間 $[t_i, t_j]$ において、実測の F0 パターンを $F_0(t)$ とし、推定した指令から生成された F0 パターンを $f_0(t)$ とし、次式が最小となるようにパラメータを微小変化させる。

$$\frac{2}{j-i+1}\sum_{n=i}^{j}\ln\frac{F_0(t_n)}{f_0(t_n)} \tag{3.8}$$

i)　言語情報を用いた指令の修正

古山らは、成澤の手法におけるモデルパラメータの初期値を、言語情報から修正する手法と [9]、事前に用意したコーパスから決定する手法 [10] を提案している。

1 つ目の手法について述べる。この手法は、前節で述べた手法から抽出されたモデルパラメータの初期値を、言語情報と照らし合わせて修正する。ここでいう言語情報とは、コーパス（男性アナウンサ 1 名による NHK ラジオ第 1 放送の朗読番組「私の本棚」の朗読音声、1 回分、15 分間、85 文）から手動で抽出したモデルパラメータと、テキストから抽出したアクセント型や、係り受けの情報との対応関係を指す。そして、次に述べる 6 つのルールに従って修正する。

1) 自動推定したアクセント指令について、立上り・立下りの位置が言語情報による位置から ±1 モーラの範囲でずれているものに関しては、言語情報に対応した位置に移動させる。

2) 言語情報によるアクセントの立上り・立下りが存在する部分にアクセント指令の立上りも立下りも存在せず、かつ、その部分が文の最終アクセント句でない場合は、新たにアクセント指令を立てる。アクセント指令の大きさは、(アクセント指令全体の大きさの平均) − (アクセント指令の大きさの標準偏差) から算出する。
3) 言語情報によるアクセントの立上り・立下りが存在しないところにアクセント指令が存在する場合は、そのアクセント指令を除去する。
4) ICRLB 境界[1]、またはショートポーズ[2]の存在する文節境界以外の文節境界の直後に、ある閾値より大きく、かつある閾値より継続時間が長いアクセント指令が存在する場合 (ここでは、アクセント指令の大きさが 0.37 より大きく、かつ、継続時間が 250ms より長いか、または、大きさが 0.6 以上、かつ、継続時間が 150ms 以上である場合とした)、そのアクセント指令の直前にフレーズ指令を新しく立てる。
5) あるアクセント句 A の開始時刻と次のアクセント句 B の開始時刻の中間の時刻 C から、アクセント句 A の継続時間の 30%だけ先行する時刻を D とおく。また、時刻 C からみてアクセント句 A の継続時間の 10%だけ後ろの時刻を E とおく。もし、時刻 D と時刻 E の間にフレーズ指令が存在する場合は、これを除去する。すなわち、i 番目のフレーズ指令の立上り時間を $P_i(t)$ とおき、i 番目のフレーズ指令の直後にある文節の開始時刻を $B_j(t)$、直前にある文節の開始時刻を $B_{j-1}(t)$ とおくとき、

$$\frac{4B_{j-1}(t) + B_j(t)}{5} < P_i(t) < \frac{2B_{j-1}(t) + 3B_j(t)}{5} \tag{3.9}$$

であるならば、i 番目のフレーズ指令を除去する。
6) ICRLB 境界に対応する文節境界、またはショートポーズのある文節境界にフレーズ指令が存在しない場合は、新たにフレーズ指令を立てる。

2 つ目の手法について述べる。この手法では、コーパス（ATR 日本語音声データベース [38] から男性話者 MHT が発生した音声 503 文）から手動抽出したモデルパラメータを様々な言語情報を利用して学習した、決定木の一種である CART(Classification And Regression Trees) [37] から初期値を決定する。

3.3　従来手法の問題点

成澤による手法では、基本周波数パターンにおいて、基本周波数が観測されない無声区間を 3 次スプラインを用いて補間しているが、これは長い無声区間等で本来存在しないピークを生じさせ、誤った抽出をしてしまうことがある。そしてこれは、古山らの手法により、言語情報から正しい位置に初期値を修正しても、結局最適化の段階でモデルパラメータが誤ったものになりうることを意味する。また、古山らの手法では 1 人の話者から構築したコーパスを元にして得られた知見を根拠にしているが、それが一般の話者に適用可能なのか、汎用性に疑問が残る。

[1]ICRLB(Immediate Constituent with a Recursively Left-Branching structure: 最大左枝分かれ句) とは、文の構文木において、右枝分かれ境界で前後を区切られ、かつ左枝分かれ境界のみを含む単語連鎖のこと
[2]句読点等による発話休止区間

3.4　提案手法

従来手法の問題点を踏まえて、F0 が存在しない区間（無声区間）を補間せずにモデルパラメータを抽出する手法を提案する。また、観測された F0 の内、母音区間のみを利用することで microprosody 等の影響を受けにくいよりロバストな手法を目指す。

提案手法によるモデルパラメータの抽出は、成澤らの手法と同様に、前処理、初期値推定、最適化の 3 段階からなる。提案手法に必要な情報は、時間情報を含んだ音素ラベル、アクセント句とアクセント型が記されたラベルである。

3.4.1　前処理

モデルパラメータの抽出に不要と考えられる F0 の除去を行う。ある 1 つの母音区間内で、前後の音素における最大値、最小値を超える値が存在する場合、その母音区間内の F0 の分散が 0.1 以下になるまで、中央値から最も外れている F0 を除外する。その後、母音以外のすべての F0 を除去する。これにより、スプライン補間による本来存在しない F0 に影響を受けないだけだなく、母音部だけをみることにより、microprosody 等の影響を小さくすることができる。

3.4.2　初期値推定

基本周波数パターンについて、0.3 [sec] 以上の休止区間までの区間を 1 つの呼気段落とし、それぞれ呼気段落単位で独立にモデルパラメータの推定を行う。HMM 音声合成で用いられているラベル情報から、アクセント句、アクセント型の情報を抜き出し、アクセント句につき 1 つのアクセント指令があるとする。そして、アクセント句の直前にフレーズ指令の挿入判定を行う。フレーズ指令を挿入判定は、アクセント句内の各モーラを HL（High, Low）の 2 値で表し、L に該当する音素の最小値における F0 の値と、それ以前のフレーズ指令が該当位置において生成する F0 の値を比較し、最小値の F0 が上回った場合、フレーズ指令を立てる。ここで、アクセント型が LHHLL のような場合、左側の L と右側の L の両方について検討し、小さい方を採用する。ただし、呼気段落最初のフレーズ指令が生成するピークは、アクセント区内の F0 の最大値の半分以上でなければならないものとし、フレーズ指令直後のアクセント指令は 0.2 以上の大きさがなければならないものとする。アクセント指令の大きさはフレーズ指令の大きさを推定した後、アクセント句の F0 の最大値をとる地点を基準に推定する。フレーズ指令の開始時刻は、呼気段落初めの指令は、第 1 モーラの母音部開始時刻 0.3 [sec] 前とし、それ以外の指令の開始時刻は、該当アクセント句の開始時刻とする。フレーズ指令は、呼気段落の終了後の無音区間部内で打ち切るとする。アクセント指令の開始時刻は、1 型のアクセント句のときは、第 1 モーラの母音部開始時刻 0.15 [sec] 前とし、それ以外は、第 2 モーラの母音部開始時刻 0.1 [sec] 前とする。アクセント指令の終了時刻は、0 型のアクセント句のときは、アクセント句終了時刻の 0.1 [sec] 前とし、それ以外は、アクセント核に該当する音素の母音部開始時刻 0.05 [sec] 前とする。ただし、該当音素が促音のときは 1 モーラ後ろにずれるものとする。

フレーズ指令は必ずしも係り受けの構造に従うわけではないため、アクセント句とアクセント句の間にのみ存在するという、必ず成り立つ制約のみを用い、F0 パターンから直接推定している。また、アクセント指令の初期値に依存することなく、大きさを求めている。

3.4.3　最適化

最急降下法により、モデルパラメータの最適化を行う。最急降下法は目的関数を最小化するアルゴリズムの 1 つであり、パラメータを初期値から逐次更新していくことにより最適なパラメータを探索する。パラメータの更新式は次式であらわされる。

$$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\Theta}^{(k)} - \epsilon \nabla E \tag{3.10}$$

$$E = ||\boldsymbol{F} - \tilde{\boldsymbol{F}}||^2 \tag{3.11}$$

ここで $\boldsymbol{\Theta}$ は最適化するパラメータ群を並べたベクトルであり、k は反復回数、ϵ は学習係数、E は目的関数である。そして $\boldsymbol{F}$ は呼気段落の開始時刻から終了時刻までにおける、F0 パターンの時系列ベクトルであり、$\tilde{\boldsymbol{F}}$ はモデルから生成された F0 パターンの時系列ベクトルであり、目的関数 E はその 2 乗誤差である。また、i 番目のフレーズ指令が生成する F0 パターンの時系列ベクトルを $\boldsymbol{p}_i$ とし、その 1 次微分を $\boldsymbol{p}'_i$ とすると、フレーズ指令の大きさと開始時刻についての E の偏微分はそれぞれ、

$$\frac{\partial E}{\partial A_{p,i}} = 2\langle \boldsymbol{p}_i, \tilde{\boldsymbol{F}} - \boldsymbol{F} \rangle \tag{3.12}$$

$$\frac{\partial E}{\partial T_{p_s,i}} = 2A_{p,i}\langle \boldsymbol{p}'_i, \tilde{\boldsymbol{F}} - \boldsymbol{F} \rangle \tag{3.13}$$

($\langle\cdot\rangle$ 内積をあらわす)

となり、他のパラメータについても同様にして求まるため、E の勾配は解析的に求めることができる。尚、$\boldsymbol{F}, \tilde{\boldsymbol{F}}, \boldsymbol{p}, \boldsymbol{p}'$ 等の時系列ベクトルは、各時刻において F0 が観測されない無声のときは、すべて 0 として扱う。パラメータの最適化は前処理を施した基本周波数パターンに対して行う。最適化するパラメータは各指令の大きさと位置であり、基底周波数 F_b に関しては、既知とした。最急降下法の探索範囲は、フレーズ指令の大きさ ±0.1、フレーズ指令の開始時刻 ±0.1[sec]、アクセント指令の開始時刻 ±0.1[sec]、アクセント指令の終了時刻 ±0.1[sec] とした。また、アクセント指令の終了時刻と開始時刻の差が 0.05[sec] 以上となるようにし、隣り合うアクセント指令は、小さい方の指令が大き方の指令に重ならないようにし、フレーズ指令の開始時刻は、対応するアクセント句のアクセント指令の開始時刻以前かつ、直前のアクセント指令の終了時刻以降になるようにした。そして、大きさが 0 以下になった指令を除去した。

3.5　評価実験

提案手法の生成過程モデルパラメータの抽出性能を評価するために、成澤による手法を従来手法として比較した [8]。

表 3.1: 生成過程モデルパラメータ自動抽出性能の実験条件

サンプリング周波数	16 kHz
フレーム周期	5 msec
基底周波数	60 Hz(既知)
音声コーパス	ATR 日本語音声データベース
話者	男性 1 名（MHT）
評価文数	503 文

表 3.2: モデルパラメータの抽出性能

	conventional		proposed	
	recall	precision	recall	precision
phrase command	93.3	**88.8**	**96.7**	80.4
accent command (onset)	83.4	94.0	**91.6**	**97.9**
accent command (reset)	78.5	88.5	**89.2**	**95.4**

3.5.1　実験条件

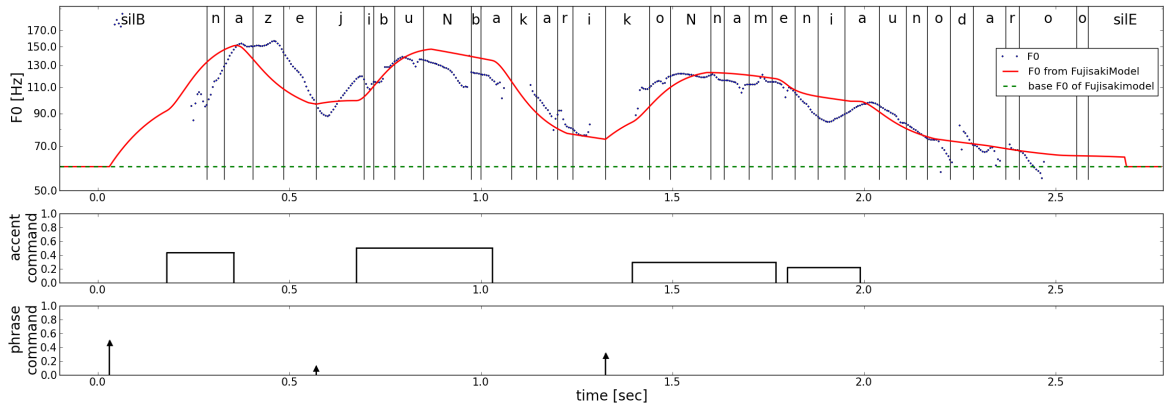
それぞれの手法で抽出したモデルパラメータと正解のモデルパラメータから、次式で定義される再現率（recall) と適合率 (precision) を求め、比較した。

$$recall = \frac{\text{正解数}}{\text{手動抽出数}} \times 100 \quad [\%] \tag{3.14}$$

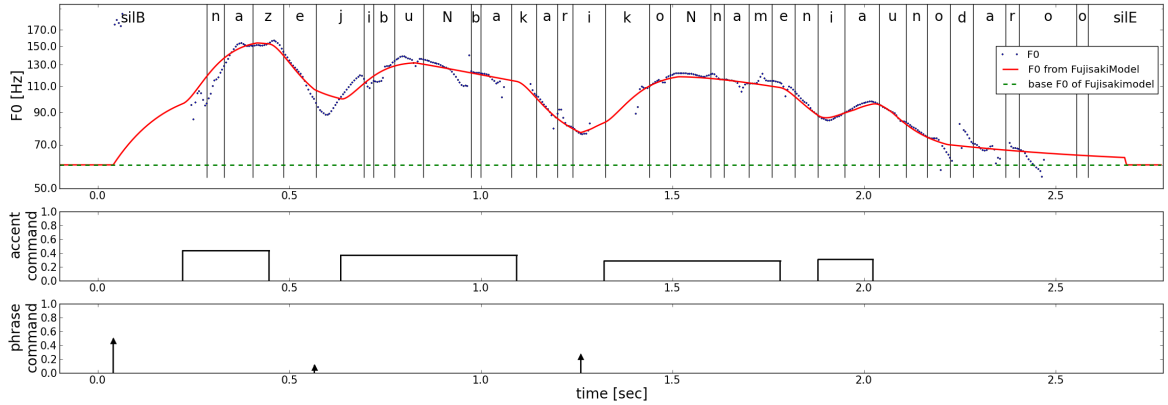
$$precision = \frac{\text{正解数}}{\text{自動抽出数}} \times 100 \quad [\%] \tag{3.15}$$

ここで正解とは、手動で抽出した指令の該当するモーラから前後 1 モーラ以内に指令があるものを指す。そして、フレーズ指令の生起位置、アクセント指令の立ち上がり位置、立ち下がり位置についてそれぞれ再現率と適合率を求めた。ただし、指令の大きさが 0.1 以下のものについては、F0 パターンの再現性において重要ではないため、ここでは除外した。

実験条件を表 3.1 に示す。音声データは ATR 日本語音声データベース [38] の B セットの中から、話者 MHT を選択した。両手法において、F0 は STRAIGHT [39] を用い、フレーム周期 5 [msec]、最小値 60 [Hz]、最大値 200 [Hz] で抽出し、基底周波数 F_b は 60 [Hz] とした。提案手法において、音素ラベルの時間情報は Julius[3]を用いて得て、アクセント情報は、HTS-2.1[4] のデモスクリプトに付属しているラベルを利用した。ただし、sp（ショートポーズ）ラベルは音声に合わせて修正した。最急降下法における学習係数の値は 0.002 とした。



(a) 初期値抽出



(b) 最適化後

図 3.1: 提案手法によるモデルパラメータの抽出例（文：なぜ、自分ばかりこんな目に合うのだろう。）

3.5.2　結果

結果を表 3.2 に示す。フレーズ指令の再現率、アクセント指令について、高い性能を示していることがわかる。提案手法のフレーズ指令の適合率が低いのは、正解データではアクセント指令のみで F0 パターンを表現しているところを、提案手法ではフレーズ指令とアクセント指令の両方で F0 パターンを表現している箇所があるためである。しかし、この様な場合は F0 パターンの再現性という観点からは問題ないと考えられる。また、モデルパラメータの抽出に関して、あるべきところに指令が存在しないのは、最適化のプロセスから外れ、韻律の自然性に大きく影響するため、再現率の方が適合率に比べて重要であるといえる。アクセント指令が従来手法に比べて大きく改善している理由として、提案手法では、F0 パターンの無声区間を補間せずに扱っているため、存在しない F0 への誤った

[3] Julius, http://julius.sourceforge.jp/
[4] HTS, http://hts.sp.nitech.ac.jp/

抽出や最適化がされにくくなっているためであると考えられる。尚、古山の手法は再現することが困難であるため、直接の比較は行なっていないが、文献による数値と比較すると、提案手法の方が有効であると考えられる。

3.5.3　考察

提案手法による自動抽出の 1 例を図 3.1 に示す。母音を中心とした聴覚的に重要であると考えられる F0 部分を適切に捉えられていることがわかる。提案手法では、F0 パターンそのものに対して最適化をしているのではなく、前処理を施した F0 パターンに対して最適化をしているため、元の F0 パターンの再現性という観点からのディストーションは最適ではない。しかし、F0 パターンのモデル化の目的は、実測の F0 パターンを忠実に再現することではなく、聴覚的に自然な韻律パターンを少数のパラメータで実現することである。この観点において、提案手法のように F0 を選別するか、聴覚特性を考慮して適切に重み付けすることが重要である。

既知の課題としては、撥音「ん」について扱いが挙げられる。撥音は母音に比べると F0 が不安定な傾向にあるため、今回の手法では他の子音と同様に無視してしまっているが、モデルパラメータの抽出において無視できない場合がある。そのため適切な前処理を施してから取り入れる必要がある。また、アクセント指令は、アクセント句と 1 対 1 に対応していると仮定しているが、これは必ずしも成り立たない。さらに、アクセント句は第 6 章で述べるような問題点がある。よって、さらなるモデルパラメータの自動抽出を高精度化するためには、より小さい単位である文節や形態素単位でアクセント指令を抽出する必要がある。

3.6　まとめ

本章ではまず、基本周波数パターン生成過程モデルパラメータを自動抽出する先行研究を紹介した。先行研究の問題点として、無声区間における F0 パターンの補間によって、本来存在しない F0 による悪影響があった。そこで、F0 パターンの補間を必要としないモデルパラメータを自動抽出する新しい手法を提案した。そして、評価実験により提案手法の有効性が確認された。

第4章

HMM音声合成における基本周波数パターン生成過程モデルの応用

4.1　はじめに

本章では、前章で提案した生成過程モデルのモデルパラメータ自動抽出手法を用いることにより、HMM 音声合成において様々な有効活用が容易に実現できることを示す。ここでは、学習コーパスの改善、学習コーパスの整備、焦点制御の 3 つの手法を紹介する。

4.2　HMM 音声合成における学習への利用

4.2.1　背景

音声分析再合成技術から抽出された F0 パターンは、マイクロプロソディ等の要因により、微細な変動を含んでいる。これは、F0 の動的特徴量 (Δ, Δ^2) に影響するため、韻律の自然性に重要な F0 パターンのグローバルな特徴に悪影響を及ぼす。そこで、HMM 音声合成における学習データの F0 パターンを基本周波数パターン生成過程モデルで平滑化することにより、韻律の自然性を改善する手法を提案する。

4.2.2　提案手法

学習用音声から抽出された F0 パターンについて、生成過程モデルパラメータを抽出する。そして元の観測された F0 パターンを、モデルパラメータから生成した F0 パターンに置き換える。これにより、ラベルとよく対応のとれた自然な F0 パターンになり、より自然な合成音声が実現されることが期待される。

4.2.3　実験

学習データの F0 パターンを、前章で提案した手法によって自動抽出したモデルパラメータから再生成した F0 パターンに置き換えて HMM を学習し、合成した音声を提案手法とし、学習データの F0 をそのまま用いて HMM を学習し、合成した音声を従来手法とする。ただし、モデルパラメータから再生成した F0 パターンについて、元の F0 パターンが無声区間である部分については、同様にその区間を無声区間であるとする。提案手法の有効性を聴取実験によって確認する。

i)　実験条件

実験条件を表 4.1 に示す。音声データは前章と同じ、ATR 日本語音声データベース [38] から話者 MHT を選び、全 503 文のうち、サブセット A から I までの 450 文で HMM を学習し、サブセット J の 53 文を合成して評価した。音声の分析は STRAIGHT を用いて [39]、F0、スペクトル包絡特徴量、非周期性指標を抽出した。分析条件は、フレーム周期 5 [msec] であり、F0 の探索範囲は最小値 60 [Hz]、最大値 200 [Hz] である。HMM に用いた特徴量は、0 から 39 次元までのメルケプストラムと 0-1、1-2、2-4、4-6、6-8 [kHz] の 5 帯域の平均非周期性指標、対数 F0、およびそれらの Δ、Δ^2 を含めた 138 次元のベクトルとした。メル

表 4.1: 生成過程モデルによる学習コーパスの改善の実験条件

音声コーパス	ATR 日本語音声データベース
話者	男性 1 名（MHT）
サンプリング周波数	16 kHz
フレーム周期	5 msec
特徴量	一般化メルケプストラム 40 次元 平均非周期性指標 5 次元 対数基本周波数 1 次元 および、それぞれの Δ, Δ^2 パラメータ
HMM	5 状態 left-to-right
学習文数	450 文（A-I セット）
評価文数	53 文（J セット）

ケプストラムと平均非周期性指標は、スペクトル包絡特徴量と非周期性指標からそれぞれ SPTK [1] を用いて求めた。HMM は HTS-2.1 を用いて構築した。状態継続長分布を明示的に含んだ 5 状態 left-to-right HSMM を用い、各状態の出力は単一の対角共分散ガウス分布とし、決定木によるコンテキストクラスタリングを行い、木の停止基準には MDL 基準を用いた。

従来手法による合成音声と提案手法による合成音声のどちらがより自然であるかを 8 人の被験者が主観評価した。評価は 5 段階であり、提案手法の方が明らかに良いと評価されたときを 2 とし、提案手法の方が良いと評価されたときを 1 とし、どちらともいえないときを 0、従来手法の方が良いと評価されたときを -1 とし、従来手法の方が明らかに良いと評価されたときを -2 とした。

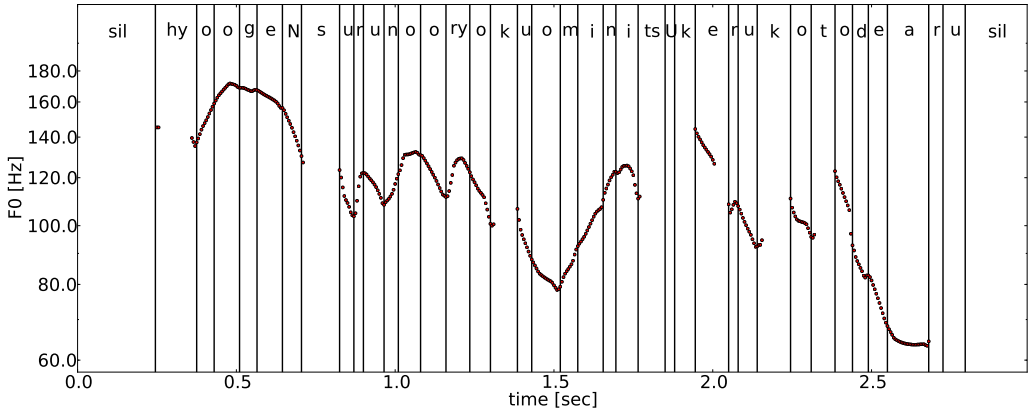
ii)　結果

有意水準 5%として、全文における評価平均の信頼区間は 0.074 ± 0.069 であり、わずかではあるが、提案手法のほうが優れている結果となった。特に評価の高かった 1 例を図 4.1 に示す。「能力を」の部分について、不自然な F0 の変動が改善されていることがわかる。

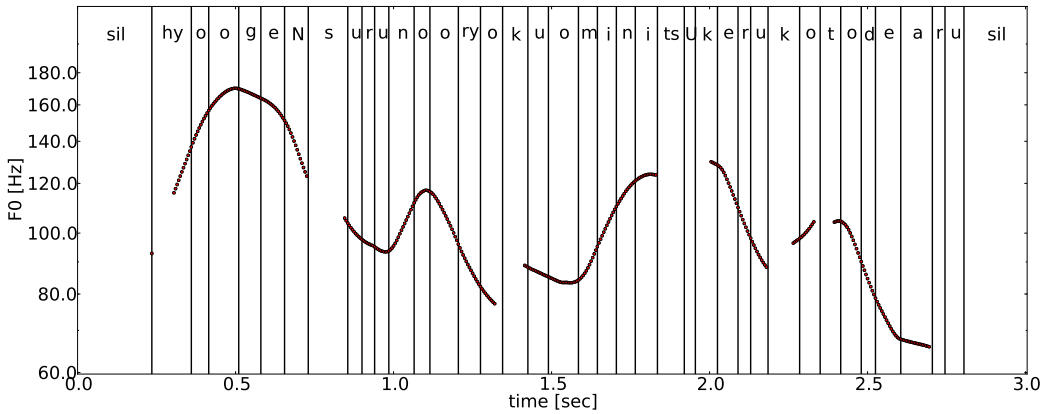
iii)　考察

提案手法の方が評価が悪かった文は、学習時にアクセント指令の抽出に失敗している箇所が悪影響を与え、アクセントが正しく再現されなかったためだと考えられる。また、有声無声の判定が悪化しているケースがあった。そのため、有声/無声のエラーを改善するために Wang らの手法を導入した [40]。これは、学習データの F0 を無声区間も含めて生成過程モデルから生成された F0 に置き換え、合成時に無声子音に関してのみ無声として取り扱うという手法であり、中国語の音声において成果を上げている。そこで、先の提案手法

[1] SPTK, http://sp-tk.sourceforge.net/



(a) 従来手法



(b) 提案手法

図 4.1: 生成過程モデルによって改善された F0 パターンと従来の F0 パターン（文：表現する能力を身につけることである。）

との比較実験を行った。しかし、6 人の被験者による聴取実験の結果は -0.286 ± 0.117 となり、良い結果は得られなかった。尚、有声/無声のエラーを改善する手法としては他に、Yu らによる手法 [41] などが提案されており、今後検討を進めていく必要がある。

4.3　HMM 音声合成学習コーパスの評価と選択

4.3.1　背景

HMM 音声合成では、音声分析再合成技術が用いられているが、分析の不安定性等から学習に悪影響を与えるサンプルが含まれることがある。そのため、F0 の観点から学習データの選別を試みる。ここで F0 に注目する理由を述べる。F0 は 2.2.2 で述べた通り、倍ピッチや半ピッチなどの抽出ミスや、マイクロプロソディ等の現象により、安定して抽出する

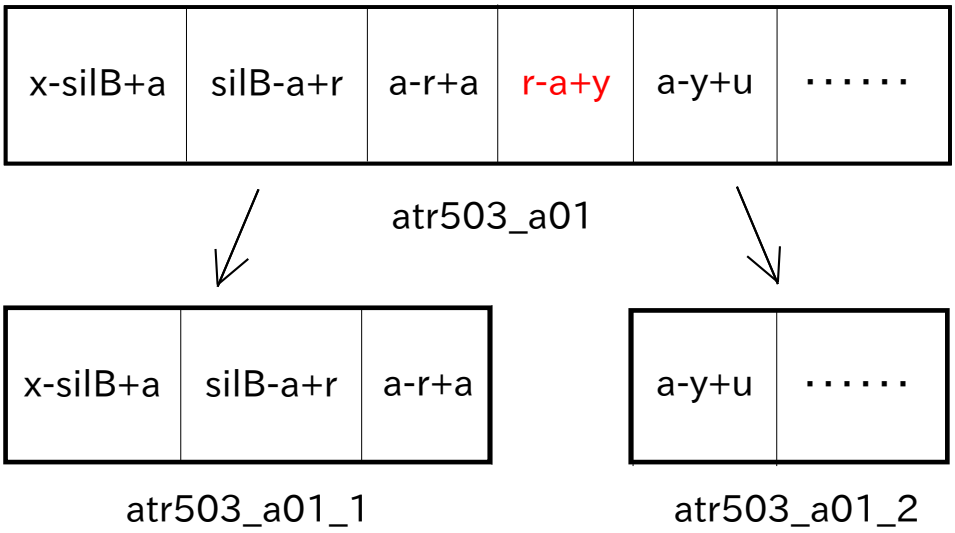


図 4.2: 学習データの除外・分割例（音素「r-a+y」を除外している）

ことが困難な特徴量である。また、スペクトル包絡は F0 を元に抽出するため、F0 の抽出ミスはスペクトル成分にも影響を及ぼすからである。

そこで、生成過程モデルから生成された F0 パターンと大きくかけ離れた F0 は、特徴量の抽出が失敗していると考え、学習コーパスから除外する手法を提案する。

4.3.2　提案手法

学習用音声について、抽出された F0 パターンと生成過程モデルから再生成された F0 パターンとの F0 差分を各フレームごとに計算する。この F0 差分を元にして、学習に不必要な音素区間を図 4.2 のように除外する。F0 が抽出された各音素区間ごとに定義される F0 差分の最大値が、全音素区間の F0 差分最大値の上位 5%,10%,30%に含まれる場合、その音素区間を除外して、学習データを分割する。残った学習データを用いて HMM を学習する。

4.3.3　実験

提案手法の有効性を主観評価実験により確認する。

i)　実験条件

実験条件を表 4.2 に示す。音声データは ATR 日本語音声データベース [38] の B セットの中から、男性話者 MMI と女性話者 FTY を選択した。各話者について、全 503 文のうち、サブセット A から I までの 450 文で HMM を学習し、サブセット J の 53 文を合成した。音声の分析は STRAIGHT を用いて [39]、F0、スペクトル包絡特徴量、非周期性指標を抽出した。フレーム周期は 5 [msec]、F0 は、女性話者 FTY は最小値 150 [Hz]、最大値 500 [Hz] で、男性話者 MMI は最小値 80 [Hz]、最大値 250 [Hz] でそれぞれ抽出した。HMM に用いた

表 4.2: 学習コーパスの評価・選択の実験条件

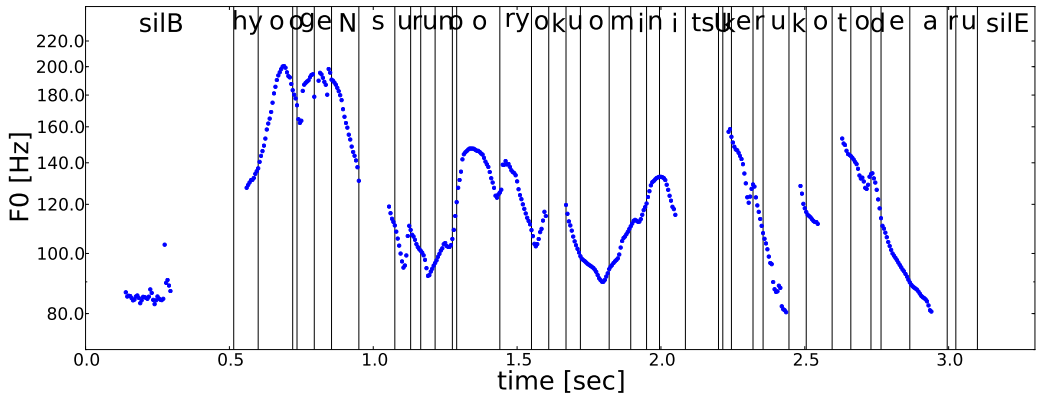
音声コーパス	ATR 日本語音声データベース
話者	男性 1 名（MMI）、女性 1 名（FTY）
サンプリング周波数	16 kHz
フレーム周期	5 msec
特徴量	一般化メルケプストラム 40 次元 平均非周期性指標 5 次元 対数基本周波数 1 次元 および、それぞれの Δ, Δ^2 パラメータ
HMM	5 状態 left-to-right
学習文数	450 文（A-I セット）
評価文数	20 文 $\times$ 3（J セット 53 文からランダムに 20 文選択、3 手法）

表 4.3: 学習コーパスの評価・選択の主観評価実験の結果

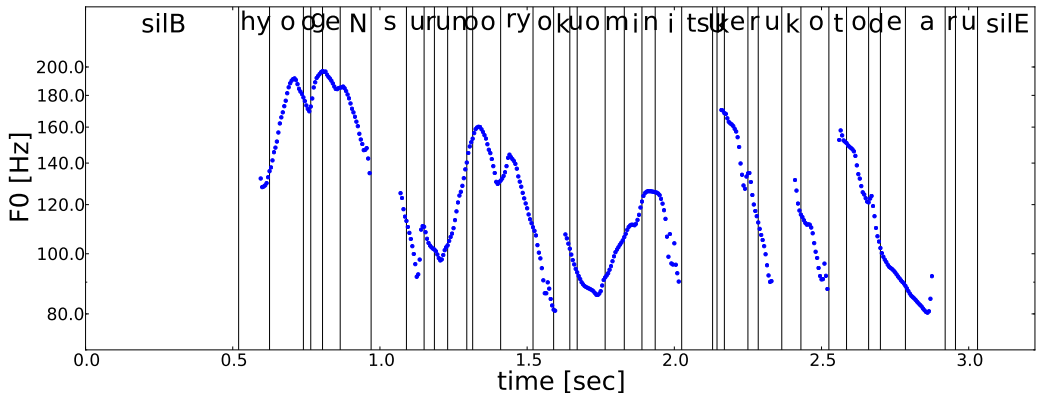
話者	除外音素の割合	スコア (95 %の信頼区間)
MMI	5%	0.300 ± 0.148
	10%	0.139 ± 0.147
	15%	-0.378 ± 0.162
FTY	5%	-0.100 ± 0.152
	10%	-0.075 ± 0.158
	15%	-0.381 ± 0.146

特徴量は、0 から 39 次元までのメルケプストラムと 0-1、1-2、2-4、4-6、6-8 [kHz] の 5 帯域の平均非周期性指標、対数 F0、およびそれらの Δ、Δ^2 を含めた 138 次元のベクトルとした。メルケプストラムと平均非周期性指標は、スペクトル包絡特徴量と非周期性指標からそれぞれ SPTK を用いて求めた。HMM は HTS-2.1 を用いて構築した。状態継続長分布を明示的に含んだ 5 状態 left-to-right HSMM を用い、各状態の出力は単一の対角共分散ガウス分布とし、決定木によるコンテキストクラスタリングを行い、木の停止基準には MDL 基準を用いた。

従来手法による合成音声と提案手法による合成音声のどちらがより自然であるかを、9 人の被験者が主観評価した。話者 1 人につき、J セットの 53 文からランダムで 20 文を選び、20 文に対し 3 段階の除外手法による計 60 文を、被験者が主観評価した。評価は 5 段階とし、提案手法の方が優れているときを 2、提案手法の方がやや優れているときを 1、どちらともいえないときを 0、従来手法の方がやや優れているときを -1、従来手法の方が優れているときを -2 とスコアを付けた。



(a) 従来手法



(b) 提案手法

図 4.3: 学習コーパスの評価・選択によって改善された F0 パターンと従来の F0 パターン（文：表現する能力を身につけることである。）

ii)　結果

結果を表 4.3 に示す。話者 MMI で除外音素が 5%の場合には提案手法の優位性が示された。除外音素の増加につれて結果が悪くなる傾向が見られるが、HMM の学習量減少による音声の品質低下なので妥当な結果と言える。女性話者 FTY についてはどの場合も結果が悪かったが、FTY の音声は分析で得た F0 パターンの乱れが少なく、F0 差分が小さい傾向にあったため、問題のある音素区間の除去の効果より HMM の学習量の減少による悪影響が表れたと考えられる。図 4.3 では、F0 パターンが改善された例を示す。「表現」の部分など全体的に F0 の乱れが収まっている点が確認される。

今回の手法では各音素区間で F0 差分の最大値を元に除外しているが、この手法では除外する音素のバランスが取れていないので、音素ごとに F0 差分が大きい区間を除外するといった手法が考えられる。また、話者の違いにより、除外条件が自動的に最適化されるような手法を検討していく必要がある。

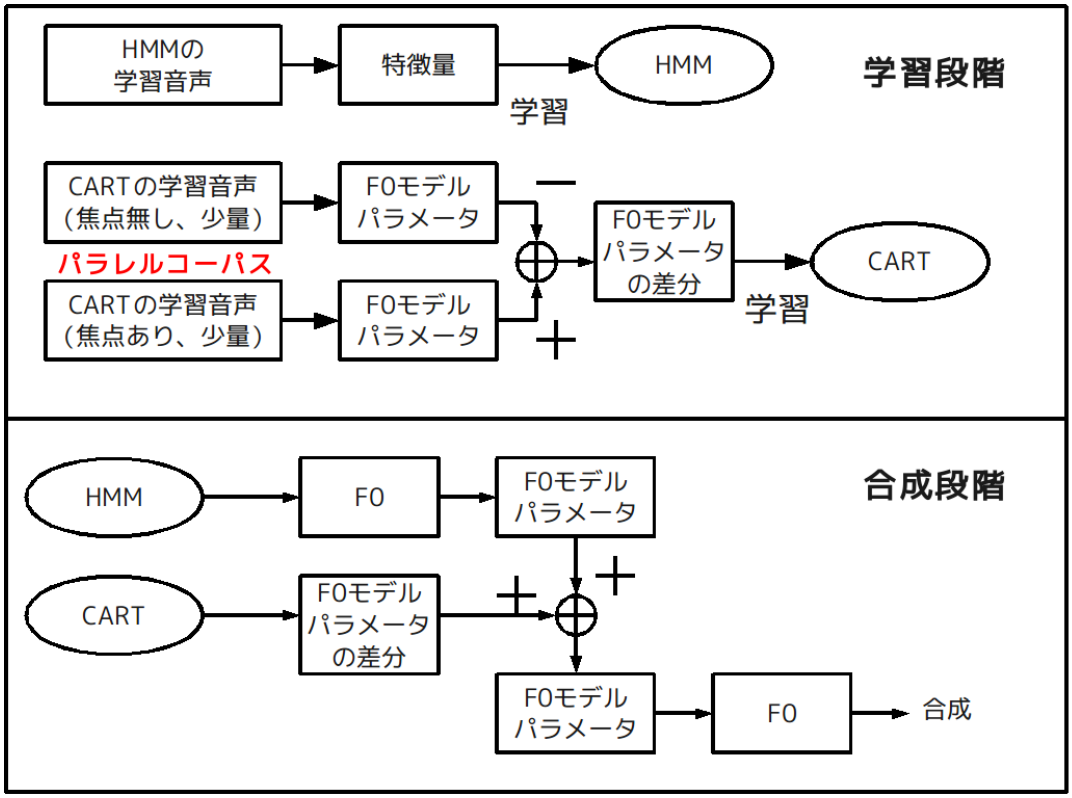


図 4.4: 焦点制御システムの概要

4.4　HMM 音声合成における焦点制御

4.4.1　背景

生成過程モデルを利用した焦点制御の先行研究としては越智らによる手法がある [6]。しかし、焦点を付与していないベースライン音声合成として、生成過程モデルパラメータを直接 2 分木で推定しているため、2 分木の学習には、生成過程モデルパラメータがラベル付けされた音声コーパスが相当量必要であり、作成には時間を要していた。また、F0 パターンが HMM とは別個に生成されるため、スペクトルなどの他の特徴量とミスマッチが生じるという問題もある。

そこで、HMM から生成された F0 パターンから生成過程モデルパラメータを抽出し、生成過程モデルの差分を適用することにより焦点付与する手法を提案する。

4.4.2　提案手法

提案手法の概要を図 4.4 に示す。焦点を付与しない音声と焦点を付与した音声の生成過程モデルパラメータの差分は、決定木の一種である CART(Classification And RegressionTrees) を用いて推定する。

CART の学習方法は、次のようなものである。まず、1 人の話者が同一文について、焦

表 4.4: CART の説明変数

当該指令の大きさ
当該指令の 1 つ前の指令の大きさ
当該指令のあるアクセント句と焦点を付加するアクセント句までのアクセント句の数
当該指令のあるアクセント句のモーラ数
当該指令のあるアクセント句の 1 つ前のアクセント句のモーラ数
当該指令のあるアクセント句の 1 つ後のアクセント句のモーラ数
当該指令のあるアクセント句のアクセント型
当該指令のあるアクセント句の 1 つ前のアクセント句のアクセント型
当該指令のあるアクセント句の 1 つ後のアクセント句のアクセント型
当該指令のあるアクセント句の前にショートポーズがあるかどうか

点を特に置かないで読み上げた音声と、あらかじめ指定した箇所の文節に焦点を置いて読み上げた音声とのパラレルコーパスを用意する。次に、焦点付与した音声と付与していない音声から得られた生成過程モデルパラメータの差分で、表 4.4 に示されるラベルデータと共に CART を構築する。

焦点付与音声は、次のようにして得る。まず、HMM 音声合成で生成された（焦点が付与されていない）F0 パターンから生成過程モデルパラメータを抽出する。次に、この F0 モデルパラメータに、CART から推定される差分を加え、F0 パターンを再生成する。そして、この生成された F0 パターンと HMM 音声合成で生成された他の特徴量を用いて音声合成を行い、焦点付与された合成音声を得る。

尚、一連の流れにおいて、対応する指令が存在しない場合は全て大きさ 0 の指令があるとしている。この手法は、HMM を学習するための話者（最終的に合成される音声の話者）と、焦点を付与するシステムを構築するためのパラレルコーパスの話者が、必ずしも同一人物である必要がないという特徴がある。

4.4.3　実験

提案手法によって合成された焦点付き音声が、意図した箇所に焦点を知覚できるか調査する知覚実験を行った。また、焦点を付与することによる音声の音質劣化の程度を調べるため、自然性評価実験を行った。

i)　実験条件

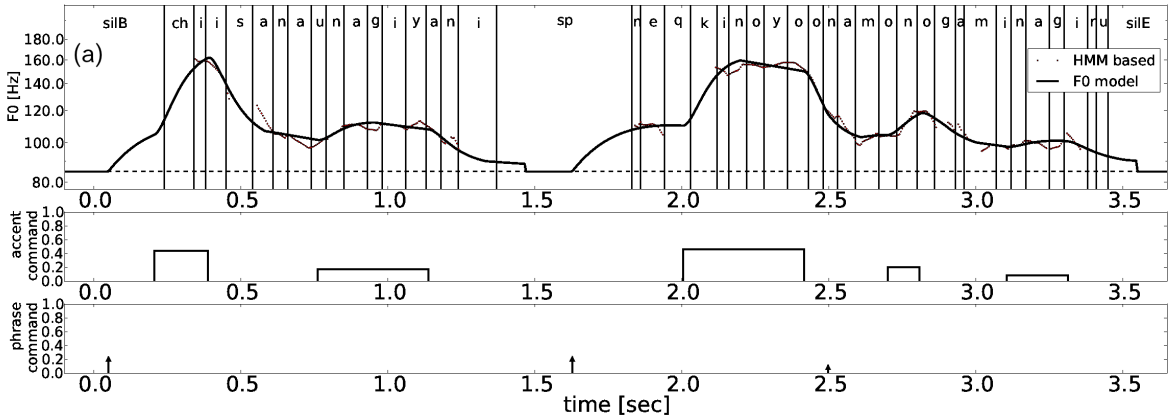
実験条件を表 4.5 に示す。HMM の学習用音声は ATR 日本語音声データベース [38] から話者 MHT を選び、全 503 文のうち、サブセット A から I までの 450 文で HMM を学習し、サブセット J の 53 文を合成して評価した。音声の分析は STRAIGHT [39] を用いて、F0、スペクトル包絡特徴量、非周期性指標を抽出した。分析条件は、フレーム周期 5 [msec] である。HMM に用いた特徴量は、0 から 39 次元までのメルケプストラムと 0-1、1-2、2-4、

表 4.5: 焦点付与の実験条件

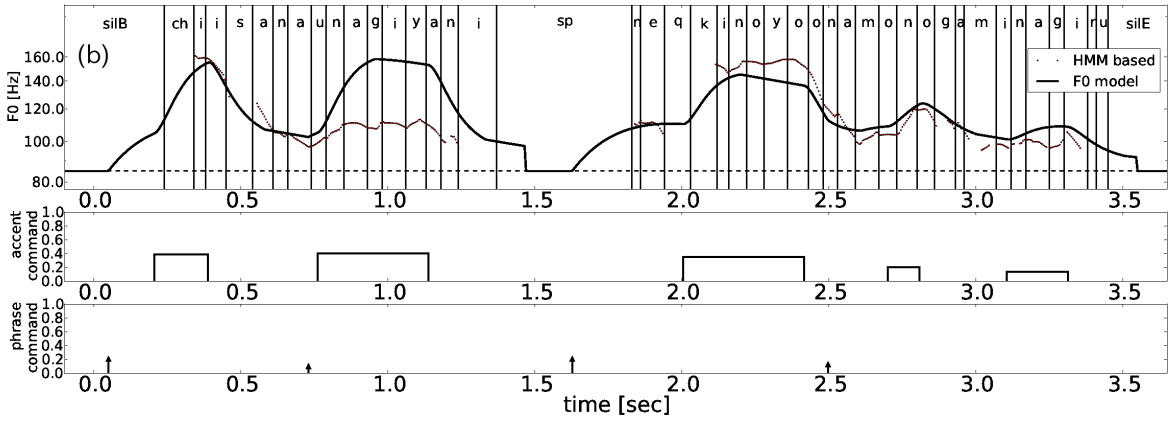
音声コーパス	ATR 日本語音声データベース
合成音声の話者	男性 1 名（MHT）
焦点音声コーパスの話者	女性 1 名
サンプリング周波数	16 kHz
フレーム周期	5 msec
特徴量	一般化メルケプストラム 40 次元 平均非周期性指標 5 次元 対数基本周波数 1 次元 および、それぞれの Δ, Δ^2 パラメータ
HMM	5 状態 left-to-right
HMM の学習文数	450 文（A-I セット）
CART の学習分数	50 文 232 発話（A セット）
評価文数（焦点の知覚実験）	18 文（J セット 53 文からランダムに選択）
評価文数（自然性の評価実験）	20 文（J セット 53 文からランダムに選択）

4-6、6-8 [kHz] の 5 帯域の平均非周期性指標、対数 F0、およびそれらの Δ、Δ^2 を含めた 138 次元のベクトルとした。メルケプストラムと平均非周期性指標は、スペクトル包絡特徴量と非周期性指標からそれぞれ SPTK を用いて求めた。HMM は HTS-2.1 を用いて構築した。状態継続長分布を明示的に含んだ 5 状態 left-to-right HSMM を用い、各状態の出力は単一の対角共分散ガウス分布とし、決定木によるコンテキストクラスタリングを行い、木の停止基準には MDL 基準を用いた。CART の学習には、女性話者 1 名が ATR503 文の A セット 50 文を読み上げた、焦点あり音声 182 発声、焦点なし音声 50 発声の、計 232 発声を用いた。CART の学習には Edinburgh 大学が提供している wagon[2]を用いた。F0 パターンから生成過程モデルパラメータを抽出する方法は、CART の学習用音声については人手で行い、テスト時の HMM 合成音声については、前章の提案手法による自動抽出である。

焦点知覚実験では、被験者に焦点を付与しない合成音声を聞かせ、直後に同じ文でテスト音声を聞かせた。そして、テスト音声に焦点があると判断したら、焦点がある部分にマークさせ、焦点がないと判断したら、焦点なしにマークさせた。被験者は日本語を母語とする 11 名である。テスト音声は ATR503 文のうち、J セット 53 文からランダムに 18 文選び、そのうち 16 文に焦点を置き、残る 2 文については焦点を置かなかった。自然性評価実験では、焦点を付加した音声 10 文と焦点を付加しない音声 10 文の、計 20 文をランダムに聞かせ、音声の自然性を 5 段階で主観評価させた（5 が最も自然であり、1 が最も不自然であるとした）。この 20 文も J セット 53 文からランダムに選んだ。



(a) 焦点なし



(b)「鰻屋に」の部分に焦点を付加

図 4.5: HMM から生成された F0 パターンと焦点付加後の生成過程モデルによる F0 パターンの例（文：小さな鰻屋に熱気のようなものがみなぎる。）

ii)　結果と考察

焦点知覚実験では 77%の確率で、正しく焦点がついていると判断された。また、自然性評価実験では、焦点なしの音声の自然性の平均が 3.25(標準偏差 0.93) に対して、焦点ありの音声の自然性の平均が 3.07(標準偏差 0.96) であり、焦点付与による明確な音質の劣化は見られなかった。

提案手法で合成した音声の F0 パターンの 1 例を図 4.5 に示す。焦点を置いた 1.0 秒付近の「鰻屋に」の部分の F0 が大きくなっていることが確認できる。

焦点知覚実験で、合成音声に焦点をほぼ適正に付加できたことが示されたが、F0 パターンを変化させても、焦点ではなく発話スタイルの変化と捉えてしまうことがあった。焦点を付けたと容易に知覚できる音声を合成するためには、F0 以外にも、パワーや継続長などを同時に変化させていく必要があると考えられる。

[2] http://www.cstr.ed.ac.uk/projects/speech_tools/

4.5　まとめ

本章では、基本周波数パターン生成過程モデルを利用することによって、HMM 音声合成において、学習コーパスの改善、学習コーパスの整備、焦点制御する手法を提案した。そして、提案手法の有効性が評価実験により確認された。これは、生成過程モデルの自動抽出性能が実応用に使える水準になったことを意味している。

また、これらの手法は互いに組み合わせることが可能である。例えば、学習コーパスの内、不適切なサンプルをあらかじめ除外した上で、生成過程モデルで平滑化された F0 パターンと、元の F0 パターンとの差分を同時に学習することによって、韻律のグローバルな特徴を適切に再現した上で、モデルでは表現されていない部分も含めた高品質な音声合成が可能になることが期待される。また、各種実験の結果で述べたように、依然として解決しなければならない課題が残っている。今後、これらのことを検討していく必要がある。

第5章

HMM音声合成における
コンテキストラベルの改良

5.1　はじめに

本章では、第 2 章で述べた HMM 音声合成システムで使われているコンテキストラベルについて述べる。代表的な HMM 音声合成システムである HTS[1] で用いられている日本語音声用ラベルについて述べてから、その問題点を指摘し、それを改善するラベルを提案する。そして、提案手法の有効性を聴取実験によって確認する。

5.2　従来のコンテキストラベル

従来用いられてきたコンテキストラベルを表 5.1 に示す。韻律に関するラベルはアクセント句単位で定義されていることがわかる。一般に、句頭においてピッチの上昇を伴う場合をアクセント句境界があるとし、ピッチが下降する直前のモーラをアクセント核と呼ぶ。アクセント核は、アクセント句につき高々 1 個のアクセント核があるとすることが多いが、ピッチの上昇を伴わない（少ない）場合、副次アクセントとして定義されることがある。しかし、ピッチの上昇を伴うか、伴わないかは明確に区別できるものではないという問題がある。アクセント句境界は主にテキストのみから推定されることが一般的であるが、本来、話者の発話速度、発話スタイルによって変化するものである。そのため、学習データにおいては、テキストだけではなく、音声の基本周波数も利用して自動推定する研究も提案されてはいるが [42], 現状では手動で抽出することが多いのが実情である。また、従来用いられているラベルにあるアクセント句の位置は、ある同じテキストを読み上げた 2 つの音声について、一部分だけアクセント句の長さが異なる場合、その後続部分が同じ発話構造をもっていたとしても、ラベルが異なったものになってしまうという問題がある。そして、アクセント句や呼気段落は、発話によっては文の長さと同様に明確な上限がないため、任意の文章を生成可能にするためには非常に多くのラベル数を必要とし、可能なラベルの組み合わせが爆発的に増加するという問題がある。

5.3　提案手法によるコンテキストラベル

前節で指摘した問題点を踏まえて、設計方針としては、発話スタイルによって長さが変わってしまう情報や、絶対的な位置情報（呼気段落におけるアクセント句の位置や、文中における呼気段落の位置）を用いず、可能な限り相対的な（直前直後の）情報を用いることにより、1 文の長さにラベルの種類が依存しないようにする。

提案手法によるコンテキストラベルを表 5.2 に示す。提案手法の特徴として、次のようなものが挙げられる。

- アクセント句の代わりに、文節を用いている。

文節は、アクセント句に比べて、話者性に依存せず、言語情報のみから一意に決定されるものであるため、曖昧性が少ないという利点がある。文節境界は、名詞連続の場合を除い

[1] HTS, http://hts.sp.nitech.ac.jp/

表 5.1: 従来手法によるコンテキストラベル

先行音素
当該音素
後続音素
アクセント句内モーラ位置 (単位: モーラ)
アクセント型とモーラ位置との差 (単位: モーラ)
先行品詞 ID
先行品詞の活用形 ID
先行品詞の活用型 ID
当該品詞 ID
当該品詞の活用形 ID
当該品詞の活用型 ID
後続品詞 ID
後続品詞の活用形 ID
後続品詞の活用型 ID
先行アクセント句の長さ (単位: モーラ)
先行アクセント句のアクセント型
先行アクセント句と当該アクセント句の接続強度
先行アクセント句と当該アクセント句間のポーズの有無
当該アクセント句の長さ (単位: モーラ)
当該アクセント句のアクセント型
先行アクセント句と後続アクセント句の接続強度
当該呼気段落でのアクセント句の位置
疑問文かそうでないか
後続アクセント句の長さ (単位: モーラ)
後続アクセント句のアクセント型
後続アクセント句と当該アクセント句の接続強度
後続アクセント句と当該アクセント句間のポーズの有無
先行呼気段落の長さ (単位: モーラ)
当該呼気段落の長さ (単位: モーラ)
文中での当該呼気段落の位置
後続呼気段落の長さ (単位: モーラ)
文の長さ (単位: モーラ)

表 5.2: 提案手法によるコンテキストラベル

先行音素
当該音素
後続音素
先行モーラのアクセント （0:Low, 1:High）
当該モーラのアクセント （0:Low, 1:High）
後続モーラのアクセント （0:Low, 1:High）
単語内における位置の正順 (単位: モーラ)
単語内におけるモーラ位置の逆順 (単位: モーラ)
文節内におけるモーラ位置の正順 (単位: モーラ)
文節内におけるモーラ位置の逆順 (単位: モーラ)
先行単語のモーラ数
当該単語のモーラ数
後続単語のモーラ数
先行文節のモーラ数
当該文節のモーラ数
後続文節のモーラ数
先行単語の品詞 ID1
当該単語の品詞 ID1
後続単語の品詞 ID1
先行文節における自立語の品詞 ID1
当該文節における自立語の品詞 ID1
後続文節における自立語の品詞 ID1
先行単語の品詞 ID2
当該単語の品詞 ID2
後続単語の品詞 ID2
先行文節における自立語の品詞 ID2
当該文節における自立語の品詞 ID2
後続文節における自立語の品詞 ID2
単独で 1 モーラの母音であるか (0:No, 1:Yes)
当該モーラが長母音を含むか (0:No, 1:Yes)

て、ほぼ正確に自動推定することができる。ただし、「…、という…」のようなケースは読点直前と読点直後の「と」を含めて 1 つの文節とされることが多いが、それでは 1 つの文節句中に休止が入ってしまうため、ここでは、読点は必ず文節句境界があるとし、直後の「と」は自立語を持たない単独の文節句であるとして取り扱う。

- 文節を基本単位としては最長の単位とすることにより、その長さを高々20 モーラとすることができる（名詞連続を除く）。
- 単語や、文節において、文や呼気段落における位置情報を用いるのでなく、直前直後の相対的な情報を用いる。

これにより、ラベルが文の長さに依存しないため、ラベルの数を従来に比べて大幅に抑制することができる。また、今回は直前直後の情報のみを用いているが、音声認識で用いられている quinphone と同様に、学習データの量が十分にあれば、前後 2 つまで考慮しても良いと考えられる。

- アクセントを高低の 2 値のみで表現している。

アクセント句を用いていないため、アクセント型の代わりに、アクセントを H(High) と L(Low) の 2 値で表現している。副次アクセントは通常のアクセントと区別せず、その単語にアクセントがあるものとしている (1 型を除いて、1 モーラ目が L、2 モーラ目が H とする)。副次アクセントは、「ある」、「とき」などの付属語としての役割が強い語句で多くみられるが、これは、これらの語句の文中での役割が自立語と比較して小さいため、明確なアクセントとして表現されないためと考えられる。実際、強調が置かれた時にはアクセントが明確に現れる。そして、このラベルではアクセント句境界を推定する必要がなく、単語アクセントのみを推定すれば良いことを示している。例えば従来は、「東京」と「大学」がそれぞれ単語単独では 0 型のアクセントであるが、それが「東京大学」になるとき 5 型のアクセントになるとされる。しかし、これは「大学」が 1 型のアクセントに変化したと考えることもできる。このように考えることにより、「東京」にのみ強調を置くことが容易になるというメリットがある。また、0 型が連続する場合、途中の L が消失しているように聴こえることが多いが、これは、アクセント結合によるアクセントの変化ではなく、副次アクセントと同様に、アクセントが明確に現れていないだけであると考えられる。実際、ゆっくりと明瞭に読み上げる時には、アクセントが明確に表れることからも、アクセント結合とは異なる現象であると考えられる。このように考えることで、多くのケースにおいて曖昧性をなくすことができる。残る問題として、「強ければ」のように 1 型でも 2 型でも良い場合は、その可能な候補をテキストから推定し、発話速度等を加味して決定するようなシステムが必要であると考えられる。また、”形容詞”+”名詞”はアクセント結合をしてもしなくても良い場合が多く、同様にどちらにおいても対応可能にする必要がある。無論、3 単語以上の名詞連続は大きな課題である。

- 単母音、長母音を明示化している。

「…のお客…」と「能力」は初めの 3 音素が/noo/であり、同じ音素列になってしまうた

表 5.3: コンテキストラベルの改良の実験条件

音声コーパス	ATR 日本語音声データベース
話者	男性 1 名（MMI）、女性 1 名（FTY）
サンプリング周波数	16 kHz
フレーム周期	5 msec
特徴量	一般化メルケプストラム 40 次元 平均非周期性指標 5 次元 対数基本周波数 1 次元 および、それぞれの Δ, Δ^2 パラメータ
HMM	5 状態 left-to-right
学習文数	450 文（A-I セット）
評価文数	20 文（J セット 53 文からランダムに選択）

め、これを明示的に区別するラベルを加えている。尚、今回は音素ラベルに長母音を含んだものを用いていないため、”長母音を含むか”といったラベルを加えているが、勿論、音素ラベルに長母音を加えることも可能である。

その他、品詞 ID1 とは、”動詞”,”名詞”,”形容詞”,”形状詞”,”連体詞”,”副詞”,”接続詞”,”代名詞”, ”感動詞”,”助詞”,”助動詞”, ”接頭辞”,”接尾辞”,”文頭”,”休止”,”文末”であり、品詞 ID2 とは、”自立可能”,”非自立可能”,”一般”,”普通名詞”,”数詞”,”固有名詞”,”名詞的”, ”動詞的”,”形容詞的”,”形状詞的”, ”格助詞”,”準体助詞”,”副助詞”,”接続助詞”,”係助詞”,”終助詞”,”助動詞語幹”,”タリ”,”フィラー”である。これらは、Unidic[2]に基づくものであり、品詞 ID1 については、”文頭”,”休止”,”文末”を品詞として追加している。文頭、文末、文中の休止区間（ショートポーズ）を品詞扱いしておくことにより、単語や文節単位でみたときに、前後に休止があるのかどうかという情報が組み込まれている。

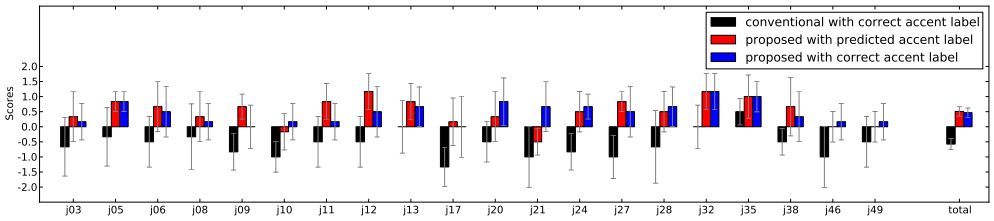
5.4　実験

従来手法によるラベルと提案手法によるラベルのそれぞれを用いて HMM を学習し、音声を合成した。そして、主観評価実験により音声の自然性を比較した。

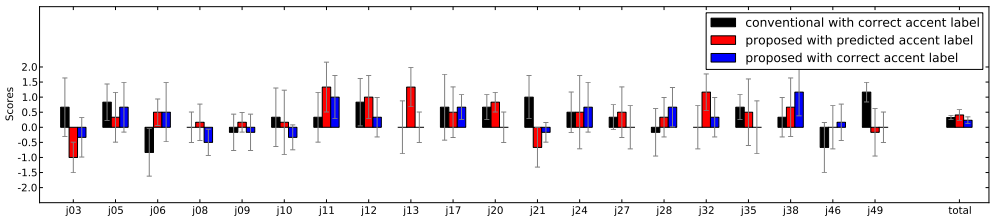
5.4.1　実験条件

実験条件を表 5.3 に示す。音声データは ATR 日本語音声データベース [38] の B セットの中から、男性話者 MMI と女性話者 FTY を選択した。各話者について、全 503 文のうち、サブセット A から I までの 450 文で HMM を学習し、サブセット J の 53 文を合成した。音声の分析は STRAIGHT を用いて [39]、F0、スペクトル包絡特徴量、非周期性指標を抽出した。フレーム周期は 5 [msec]、F0 は、女性話者 FTY は最小値 120 [Hz]、最大値 400 [Hz]

[2]Unidic, http://www.tokuteicorpus.jp/dist/



(a) FTY



(b) MMI

図 5.1: コンテキストラベルの改良の主観評価実験の結果

で、男性話者 MMI は最小値 60 [Hz]、最大値 250 [Hz] でそれぞれ抽出した。HMM に用いた特徴量は、0 から 39 次元までのメルケプストラムと 0-1、1-2、2-4、4-6、6-8 [kHz] の 5 帯域の平均非周期性指標、対数 F0、およびそれらの Δ、Δ^2 を含めた 138 次元のベクトルとした。メルケプストラムと平均非周期性指標は、スペクトル包絡特徴量と非周期性指標からそれぞれ SPTK を用いて求めた。HMM は HTS-2.1 を用いて構築した。状態継続長分布を明示的に含んだ 5 状態 left-to-right HSMM を用い、各状態の出力は単一の対角共分散ガウス分布とし、決定木によるコンテキストクラスタリングを行い、木の停止基準には MDL 基準を用いた。

従来手法によるコンテキストラベルは、手動抽出されたものを用いた。提案手法において、アクセントに関するラベルは鈴木らの手法によって自動推定したラベルと [43]、手動抽出されたラベルの 2 種類を用意した。形態素解析は Mecab[3]を用いているが、読み誤り、及びそれに起因すると思われるアクセント誤りについては手動で修正している。合成された音声 53 文の内、無作為に 20 文選び、それぞれについて、従来手法による合成音声と、提案手法による 2 種類の合成音声の合計 3 種類、全体で 60 文の音声を用意した。そして、音声の自然性を 6 人の被験者が主観評価した。評価は 5 段階であり、明らかに品質が良いと評価されたときを 2 とし、品質が良いと評価されたときを 1 とし、どちらともいえないと評価されときを 0 とし、品質が悪いと評価されたときを -1 とし、明らかに品質が悪いと評価されたときを -2 とした。

[3]Mecab, https://code.google.com/p/mecab/

5.4.2　結果

結果を図 5.1 に示す。それぞれのバーは、被験者の平均値とその 95%信頼区間を表示している。横軸は J セット 53 文から無作為に抽出された文番号を示しており、最後は 20 文全体のスコアの平均である。20 文全体でのスコアは、それぞれ、FTY の従来手法は -0.59 ± 0.18、提案手法（推定されたアクセントラベル）は 0.50 ± 0.16、提案手法（手動抽出したアクセントラベル）は 0.47 ± 0.15 であり、MMI の従来手法は 0.33 ± 0.07、提案手法（推定されたアクセントラベル）は 0.41 ± 0.16、提案手法（手動抽出したアクセントラベル）は 0.23 ± 0.11 であった。

これらをまとめると、MMI については有意な差がでなかったが、FTY については提案手法の方が有意に優れている結果となった。傾向としては、提案手法の方が音質が優れている傾向にあるが、一部の文においてイントネーションがうまく再現されず、大きく評価を落としていた。

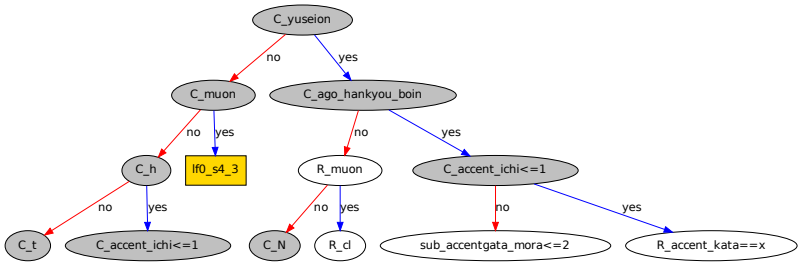
5.4.3　考察

このような結果になった原因を調査するため、学習時に構築された決定木を調べた。その一部が図 5.2、図 5.3 である。これはそれぞれ、対数基本周波数 (LF0) とメル一般化ケプストラム（MGC）の決定木の一部である。この決定木は全て HMM5 状態の内、3 状態目である。LF0 に関して提案手法は、音韻に関する質問が優先され、韻律に関する質問が従来手法に比べて若干少ない傾向があった。そのため、一部のイントネーションが十分に再現されなかったと考えられる。一方 MGC については、従来手法と提案手法に比べて大きく傾向が異なることはなかった。ただし FTY に関しては、従来手法における MGC に関する木のみ、他の木に比べて一回り小さく、これが音質劣化の原因であると考えられる。

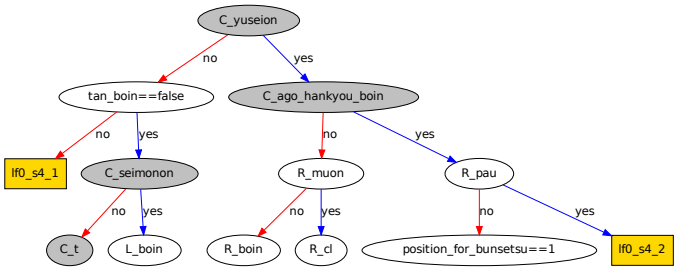
5.5　まとめ

本章では、HMM 音声合成におけるコンテキストラベルの改良を提案した。そして、その有効性を聴取実験により確認した。さらに、提案したラベルは文の長さに依存していないため、任意の長さの文に対して、安定して音声を合成ができることが期待される。今回は、学習と評価に用いたコーパスが ATR 日本語音声データベースであるため、比較的短い文しかなかったが、より一文が長いコーパスを用いることにより、提案手法の有効性が期待される。また、定義が不明確で、自動抽出が困難なアクセント句を必要としないため、ラベルの作成コストを削減できることが期待される。

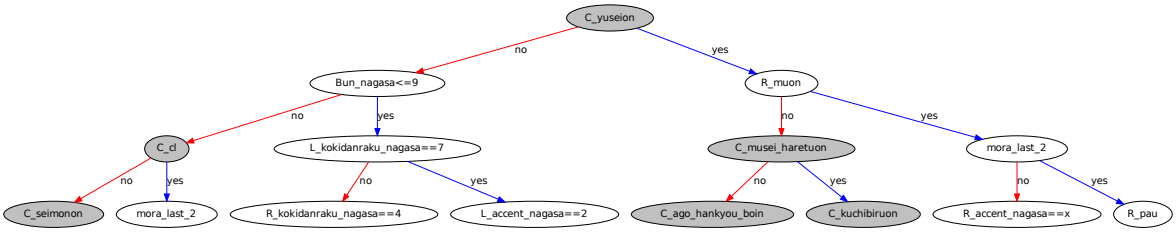
多様な音声合成を実現するためには、より多くのパラメータを必要とするため、スパース性は避けて通れない問題である。そのため、音声の性質を適切に捉えた質の良いラベルは必要不可欠であり、また、そのラベルは安定して自動抽出可能なものであるか、ユーザが直感的に操作可能なものでなければならない。今回提案したラベルを更に改良し、発話スタイルや感情などの様々な音声を高品質かつ、ユーザが直感的に操作可能なシステムを目指していくことが、今後の検討課題である。



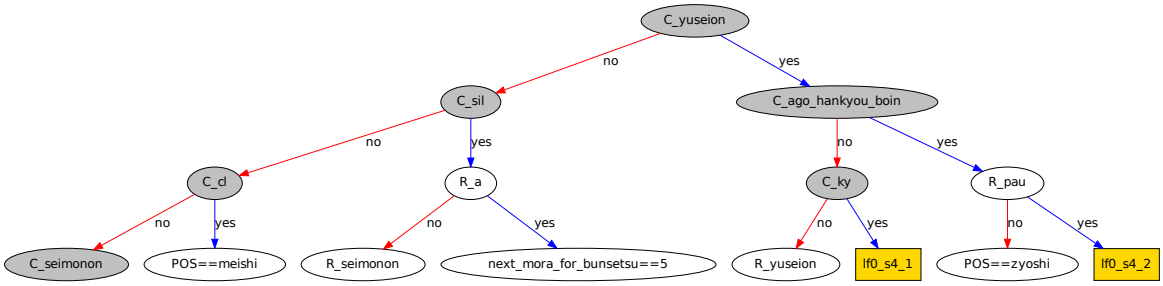
(a) FTY, 従来手法



(b) FTY, 提案手法（推定されたアクセントラベル）

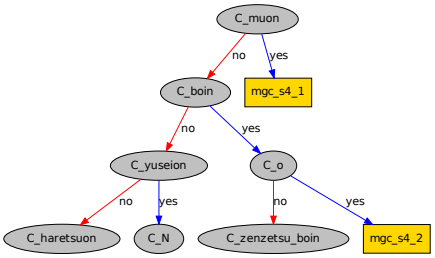


(c) MMI, 従来手法

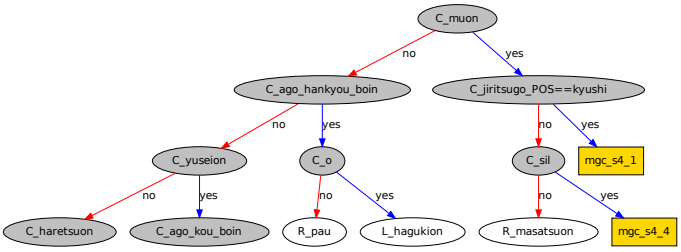


(d) MMI, 提案手法（推定されたアクセントラベル）

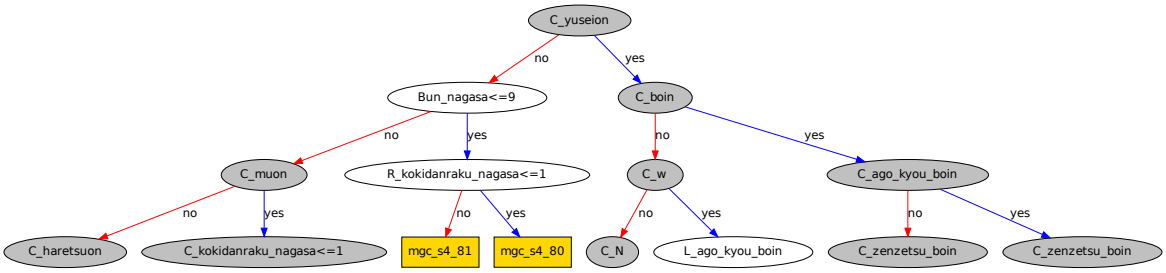
図 5.2: 対数基本周波数における決定木の例



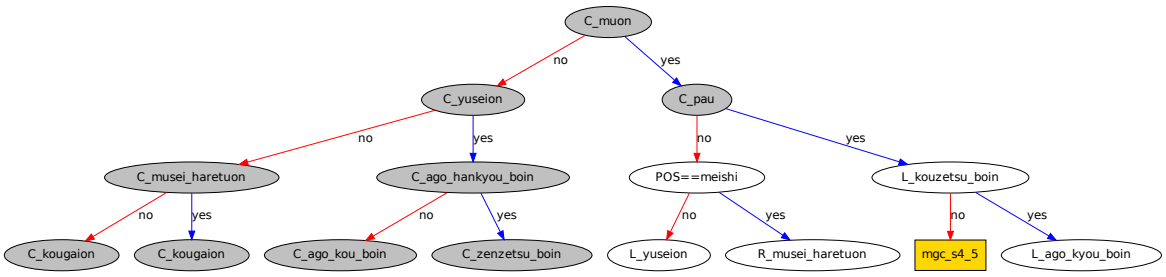
(a) FTY, 従来手法



(b) FTY, 提案手法（推定されたアクセントラベル）



(c) MMI, 従来手法



(d) MMI, 提案手法（推定されたアクセントラベル）

図 5.3: 一般化メルケプストラムにおける決定木の例

第6章

結論

6.1　本論文のまとめ

本論文では、韻律の観点から HMM 音声合成の高精度化を試みた。主な成果は、基本周波数パターン生成過程モデルにおけるモデルパラメータの自動抽出の高精度化と、HMM 音声合成で用いられているコンテキストラベルの改良である。

基本周波数パターン生成過程モデルは、生理的・物理的特性に基づいており、少数のパラメータでイントネーションの元となる基本周波数パターンを良く記述することができるため、焦点制御を始めとする様々な有効活用が期待できるが、精度良くモデルパラメータを自動抽出することが困難であるという問題があった。そこで、HMM 音声合成に用いられているコンテキストラベルを利用することにより、高精度なモデルパラメータを自動抽出する新しい手法を提案した。提案手法の抽出性能が従来手法に比べて大幅に高いことを比較実験により示した。そして、提案手法によって自動抽出された生成過程モデルのパラメータを利用することにより、HMM 音声合成において、学習コーパスの改善、学習コーパスの整備、焦点制御する手法を提案し、様々な有効活用ができることを示した。

また、HMM 音声合成に用いられていたコンテキストラベルは、定義に曖昧性があり、テキストからの自動抽出が困難なラベルが用いられていた。さらに、位置番号等の絶対的な情報が用いられているが、これは任意の長さの文を生成可能にするために非常に多くのラベルの種類を必要とする上、文の一部分のみ発話構造が変化した場合に文全体のラベルが変化してしまうという問題が生じていた。そこで、文節という単位に注目し、また相対的な情報を用いることによりこの問題を改善するラベルを提案した。主観評価実験により提案手法の有効性が確認された。また、提案したコンテキストラベルは従来に比べてテキストからの推定が容易になるというメリットがある。

6.2　今後の展望

本論文で提案した基本周波数パターン生成過程モデルのモデルパラメータの自動抽出を高精度化する手法は、従来のコンテキストラベルを利用しているため、これをもう 1 つの提案である改良されたコンテキストラベルを用いることにより、さらなる抽出性能の高精度化が期待できる。そして、ユーザが直感的に操作することができ、発話スタイルや感情などの多様な音声を高品質で合成することが可能なシステムを実現するためには、ラベルのさらなる改良を始めとして、様々な角度から改良を進めていく必要がある。

謝辞

3 年間の研究生活にわたって、常日頃からご指導、ご鞭撻を承りました指導教員の広瀬啓吉教授に、ここで感謝の意を記します。峯松信明教授も、大変お世話になり深謝いたします。そして、日頃の研究活動を支えてくださった高橋登技官、秘書の池上恵氏、折茂結実子氏にも厚く御礼を申し上げます。システム情報学第一研究室の齋藤大輔助教、及び博士課程の鈴木雅之氏には、私が卒論生の頃から研究について、研究方法や論文添削などの様々な指導をしていただき、本当に有難うございました。また、同期では特に柏木陽佑氏によく研究の相談にのってもらい、深く感謝しています。川口拓也氏、甲斐常伸氏、加藤集平氏は、3 年間共に同じ研究室を過ごし、楽しい日々を送らせていただきました。池島純氏、水上智之氏、槇佑馬氏、中村新芽氏は、卒業研究をお手伝いさせていただき、私にとっても大変貴重な経験となりました。広瀬・峯松研究室の方々は、皆優しい方達であり、素晴らしい研究室であると確信しています。最後に、自分を支えてくれた友人と家族に感謝します。

2013 年 2 月 6 日
橋本 浩弥

参考文献

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” Proc. EUROSPEECH, pp. 2523–2526, 1997.

[2] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis,” IEICE Trans. Inf. & Syst., vol. E88-D, no. 3, pp. 503–509, 2005.

[3] T. Nose, Y. Kato, and T. Kobayashi, “A speaker adaptation technique for MRHSMM-based style control of synthetic speech,” Proc. ICASSP, pp. 833–836, 2007.

[4] K. Hirose, “Speech prosody in spoken language technologies,” Journal of Signal Processing, vol. 12, no. 1, pp. 7–16, 2008.

[5] H. Fujiaski, and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” J. Acoust. Soc. Japan (E), vol. 5, no.4, pp. 233–242, 1984.

[6] K. Ochi, K. Hirose, and N. Minematsu, “Control of prosodic focus in corpus-based generation of funda- mental frequency based on the generation process model,” Proc. INTERSPEECH, pp. 1216, 2008.

[7] 見原 隆介, 越智 景子, 広瀬 啓吉, 峯松 信明, “基本周波数パターン生成過程モデルに基づくコーパスベース韻律生成における発話スタイル制御”, 日本音響学会春季講演論文集, 1-Q-24(d), pp. 379–382, 2011.

[8] S. Narusawa, N. Minematsu, K. Hirose and H. Fujisaki, “A Method for Automatic Extraction of Model Parameters from Fundamental Frequency Contours of Speech,” Proceedings of ICASSP, vol.1 pp. 509–512, 2002.

[9] K. Hirose, Y. Furuyama, S. Narusawa, N. Minematsu, and H. Fujisaki, “Use of linguistic information for automatic extraction of F0 contour generation process model parameters,” Proc. Oriental COCOSDA, pp. 38–45, 2003.

[10] K. Hirose, Y. Furuyama, and N. Minematsu, “Corpus-based extraction of F0 contour generation process model parameters,” Proc. INTERSPEECH, pp. 3257–3260, 2005.

[11] H. Lu, and S. King, “Bayesian Networks to nd relevant context features for HMM-based speech synthesis,” Proc. INTERSPEECH, 2012.

[12] 大木 康次郎, 能勢 隆, 小林 隆夫, “F0 量子化に基づく韻律コンテキストを用いた HMM 音声合成”, 電子情報通信学会技術研究報告. SP, 音声, vol.109, no. 356, pp. 141–146, 2009.

[13] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, “Average magnitude difference function pitch extractor,” IEEE Trans. Acoust. Speech Signal Process., vol. ASSP-22, no. 5, pp. 353–362, 1974.

[14] A.M. Noll, “Short-time spectrum and “cepstrum” techniques for vocal pitch detection,” J. Acoust. Soc. Am., vol. 36, no. 2, pp. 269–302, 1964.

[15] A. Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music” J. Acoust. Soc. Am., vol. 111, no. 4, pp. 1917–1930, 2002.

[16] T. Tanaka, T. Kobayashi, D. Arifianto, and T. Masuko, “Fundamental frequency estimation based on instantaneous frequency amplitude spectrum,” Proc. ICASSP. vol. 1, pp. 329–332, 2002.

[17] 森勢 将雅, 河原 英紀, 西浦 敬信, “基本波検出に基づく高 SNR の音声を対象とした高速な F0 推定法,” 電子情報通信学会論文誌. D, 情報・システム, vol. 93, no. 2, pp. 109–117, 2010.

[18] Y. Xu, “Transmitting tone and intonation simultaneously—the parallel encoding and target approximation (PENTA) Model,” Proceedings of international symposium on tonal aspects of languages: with emphasis on tone languages, pp. 215–220, 2004.

[19] J. Zhang, and K. Hirose, ”Tone nucleus modeling for Chinese lexical tone recognition,” Speech Communication, Vol. 42, No. 3-4, pp. 447–466, 2004.

[20] P. Tailor, “Analysis and synthesis of intonation using the tilt model,” The Journal of the acoustical society of America, vol. 107, pp. 1697–1714, 2000.

[21] 小林 隆夫, “メル一般化対数スペクトル近似 (MGLSA) フィルタ”, 電子通信学会論文誌 A, vol. 70, no. 3, pp. 471–480, 1985.

[22] K. Tokuda, and H. Zen, “Fundamentals and recent advances in HMM-based speech synthesis,” Proc. INTERSPEECH tutorial, 2009.

[23] H. Fujisaki, “From information to intonation,” Proceedings of 1993 International Symposium on Spoken Dialogue, pp. 7–18, 1993.

[24] H. Fujisaki, “Modelling the process of fundamental frequency contour generation”, Speech perception, production and linguistic structure, pp. 313–328, 1992.

[25] F. Buchthal and E. Kaiser, “Factors determining tension development in skeletal muscles,” Acta Physiologica Scandinavica, vol. 8, pp. 38–74, 1944.

[26] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling,” Proc. ICASSP, pp. 229–232, 1999.

[27] K. Sinoda, and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” J. Acoust. Soc. Jpn. (E), vol. 21, no. 2, pp. 79–86, 2000.

[28] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” IEICE Trans. Inf. & Syst., vol. E90-D, no. 4, pp. 825–834, 2007.

[29] T. Toda and K. Tokuda, “Speech parameter genaration algorithm considering global variance for HMM-based speech synthesis,” Proc. INTERSPEECH, pp. 2801–2804, 2005.

[30] H. Zen, K. Tokuda, and T. Kitamura. “Reformulating the hmm as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences,” Computer Speech & Language, vol. 21, no. 1, pp. 153–173, 2007.

[31] W. Sandow, ”A theory of active state mechanisms in isometric muscular contraction,” Science, vol. 127, pp. 760–762, 1958.

[32] E. Geoffrois, “A pitch contour analysis guided by prosodic event detection,” Proc. EUROSPEECH, vol. 2, pp. 793–796, 1993.

[33] V. Ström, “Detection of accents, phrase boundaries and sentence modality in German with prosodic features,” Proc. EUROSPEECH, pp. 2039–2041, 1995.

[34] A. Sakurai, and K. Hirose, “Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours,” Proc. Int. Conf. on Spoken Language Processing, pp. 817–820, 1996.

[35] J. Mersdorf, A. Rinscheid, M. Brüggen and K.U. Schmidt, “Coding of large intonational units by linear prediction,” ESCA Workshop on Intonation: theory, pp. 18–20, 1997.

[36] H. Mixdorff, Y. Hu, and G. Chen, “Towards the automatic extraction of Fujisaki model parameters for Mandarin,” Proc. INTERSPEECH, pp. 873–876, 2003.

[37] L. Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone, “Classification and regression trees,” Wadsworth, Pacific Grove, California, 1984.

[38] A. Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp. 357–363, 1990.

[39] H. Kawahara, I. Matsuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.

[40] M. Wang, M. Wen, K. Hirose, and N. Minematsu, “Improved generation of fundamental frequency in HMM-based speech synthesis using generation process model,” Proc. INTERSPEECH, pp. 2166–2169, 2010.

[41] K. Yu and S. Young, “Continuous F0 modeling for HMM based statistical parametric speech synthesis,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1071–1079, 2011.

[42] 山本 麻美, 趙 國, 山下 洋一, “言語情報と F0 情報を利用したアクセント句境界の自動定”, 電子情報通信学会技術研究報告，SP2010–109, pp. 37–42, 2011.

[43] 鈴木 雅之, 黒岩 龍, 印南 圭佑, 小林 俊平, 清水 信哉, 峯松 信明, 広瀬 啓吉, “CRF を用いた日本語東京方言のアクセント結合自動推定”, 日本音響学会秋季講演論文集, 2-2-12, pp. 299–302, 2012.

発表文献

国際会議論文

[1] K. Hirose, H. Hashimoto, J. Ikeshima, and N. Minematsu, “Use of Generation Process Model for Synthesizing Fundamental Frequency Contours in HMM-based Speech Synthesis,” Proceedings of the 11th International Conference on Signal Processing (ICSP), 2012.

[2] H. Hashimoto, K. Hirose, and N. Minematsu, “Improved Automatic Extraction of Generation Process Model Commands and Its use for Generating Fundamental Frequency Contours for Training HMM-based Speech Synthesis,” Proc. INTERSPEECH, 2012.

[3] K. Hirose, H. Hashimoto, J. Ikeshima, and N. Minematsu, “Fundamental Frequency Contour Reshaping in HMM-based Speech Synthesis and Realization of Prosodic Focus Using Generation Process Model,” Proc. Speech Prosody, 6th International Conference, SS1-3, 2012.

[4] K. Hirose, T. Matsuda, H. Hashimoto, and N. Minematsu, “REPRESENTING FUNDAMENTAL FREQUENCY CONTOURS GENERATED BY HMM-BASED SPEECH SYNTHESIS USING GENERATION PROCESS MODEL,” Proceedings of the IEEE International Workshop on MACHINE LEARNING FOR SIGNAL PROCESSING(MLSP), 2011.

[5] K. Hirose, K. Ochi, R. Mihara, H. Hashimoto, D. Saito, and N. Minematsu, “Adaptation of Prosody in Speech Synthesis by Changing Command Values of the Generation Process Model of Fundamental Frequency,” Proc. INTERSPEECH, 2011.

国内研究会論文

[6] 橋本 浩弥，広瀬 啓吉，峯松 信明，“日本語 HMM 音声合成のコンテキストラベルの改良”，電子情報通信学会技術報告， SP2012-103， pp. 31–36, 2013.

国内全国大会論文

[7] 橋本 浩弥，鈴木 雅之，広瀬 啓吉，峯松 信明，“日本語 HMM 音声合成のコンテキストラベルの改良”，日本音響学会春季講演論文集, 2013-3，(発表予定)

[8] 川口 拓也，橋本 浩弥，広瀬 啓吉，峯松 信明，“基本周波数パターン生成過程モデルの指令差分に基づく焦点制御の改良”，日本音響学会春季講演論文集, 2013-3，(発表予定)

[9] 水上 智之，橋本 浩弥，広瀬 啓吉，峯松 信明，“基本周波数パターン生成過程モデルによる HMM 音声合成学習コーパスの評価と選択”, 日本音響学会春季講演論文集, 2013-3，(発表予定)

[10] 槇 佑馬，鈴木 雅之，橋本 浩弥，峯松 信明，広瀬 啓吉 “点予測を用いたアクセント結合自動推定”，日本音響学会春季講演論文集, 2013-3，(発表予定)

[11] 中村 新芽，鈴木 雅之，峯松 信明，橋本 浩弥，広瀬 啓吉，中川 千恵子，中村 則子，平野 宏子，田川 恭識 “日本語音声教育のための韻律読み上げ Web チュータの開発と評価”，日本音響学会春季講演論文集, 2013-3，(発表予定)

[12] 橋本 浩弥，広瀬 啓吉，峯松 信明，“音声合成のための言語情報を利用した基本周波数パターン生成過程モデルパラメータの自動抽出の高精度化”，日本音響学会春季講演論文集, pp. 441–444, 2012-3.

[13] 池島 純，橋本 浩弥，広瀬 啓吉，峯松 信明，“基本周波数パターン生成過程モデルを用いた音声合成の焦点制御の検討”, 日本音響学会春季講演論文集, pp. 445–448, 2012-3.

[14] 橋本 浩弥，齋藤 大輔，峯松 信明，広瀬 啓吉，“基本周波数パターン生成過程モデルを用いた声質変換の高精度化に関する検討”，日本音響学会春季講演論文集, pp. 413–416, 2011-3.

学位論文

[15] 橋本 浩弥，“基本周波数パターン生成過程モデルを利用した声質変換の高精度化に関する検討”，東京大学工学部電子工学科卒業論文, 2011.