修士論文

基本周波数パターン生成過程モデルの指令差分を利用した焦点制御の改良

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

48116417 川口 拓也

目次

図目次

表目次

第1章

序論

1.1　研究背景

音声を用いたマンマシンインターフェースには，そのマシンを見る必要がないというメリットがあるため，主にカーナビゲーションシステムなどの運転中のドライバーをサポートするシステムや，視覚情報を受け取れない障碍者向けのテキスト読み上げシステムとして早くから用いられている．最近では，スマートフォンの操作に音声を用いることができるなど，音声を用いたマンマシンインタフェースが身近なものになりつつある．

このように用途が広がっている音声を用いた入出力機能であるが，人間の音声コミュニケーションにおいて情報の伝達に用いられる韻律情報の，制御の高度化が大きな課題となっている．音声の韻律はアクセント，イントネーション，リズムの 3 要素が主な構成要素であり，その特徴量は音声の基本周波数，話速，パワーである．

韻律情報の制御の高度化により，読み上げ調の合成音声と比べて表情豊かな合成音声が得られると考えられている．特に楽しさ，悲しさ，驚きなどを表現する発話スタイル音声合成や，発話の特定の部分を強調することによる明確な焦点を付与した音声合成は，マンマシンインターフェースへの応用においてわかりやすい情報の提供を可能にできる．

本研究では，韻律情報の制御のうち焦点の付与を行うテキスト音声合成システムについて述べる．発話の焦点は話し手が特に明確に伝えたいと意図した部分に置かれ，焦点を置かない場合よりも効果的な情報伝達を行うことができる．テキスト音声合成システムに取り入れることで，マンマシンインターフェースにおいてユーザーに対して効率的な情報伝達が可能になると考えられる．

1.2　研究目的

現在 Text-to-Speech システムとして盛んに研究されている，隠れマルコフモデル (Hidden Markov Model; HMM) を用いた音声合成方式は，HMM の適応技術などを用いて “柔軟な音声合成” を可能にすることができる一方で，基本周波数パターンで代表される韻律の取り扱いには不向きという問題がある．これは，韻律が単語や句などの長時間に渡る特徴量であるが，HMM 音声合成では制御対象の特徴ベクトルがフレーム単位という，韻律と比べて非常に短い時間での取り扱いを行なっていることに起因する．

この問題に対して，基本周波数パターン生成過程モデル [1] を用いた韻律の制御が提案されている．このモデルは対数で表した基本周波数 (F_0) パターンを 2 種類の指令成分を用いて近似し，各フレームにおける F_0 の値の代わりにこの指令成分を制御対象とすることにより，良好かつ柔軟な韻律制御が実現できる．既に，特に焦点を置かない発声と特定の語句に焦点を置いた発声のパラレルコーパスを用いて，両者の F_0 パターンの指令の差分を二分木で学習し，それを用いて，特に焦点を付与せずに合成した音声の F_0 パターンの指令を変更することによって焦点を置いた音声を合成する手法 [2, 3] が開発されている．さらに，同様の手法で，スタイル変換，声質変換も行われている [4].

しかしながら，その制御は F_0 パターンが中心で，基本的に他の韻律的特徴は操作の対象外とされていた．焦点が付与された場合，休止の位置と長さもそれに応じて変化することが多い．これを踏まえ，休止のあるなしを考慮した手法を開発することを目標とした．また，指令の差分に着目する手法では，パラレルコーパスで指令の 1 対 1 対応を仮定しているが，実際の発声ではこのような明確な対応が得られにくい場合がある．この問題に対しても対策を取ることとした．

第2章

音声合成技術

2.1　はじめに

音声合成とは、専用の機械あるいはソフトウェアを用いて人工的に音声をつくる技術の総称であり、音声を用いた様々なマンマシンインタフェースを実現する上で、音声認識と共に欠かせない技術になっている．

最古の音声合成器は、1791 年にオーストリアの von Kempelen によって作成されたものであると言われている [5]．この音声合成器は機械式のものであり、声帯を模したリードおよび声道を模した共鳴部で構成されている．手動式であるが、共鳴部の形を変えることにより声道を変形させることと同じ効果が得ることができ、母音と子音を分けて発音することができたと言われている．これは、音声合成の先駆けとして非常に素晴らしいものであった．

近年、音声合成は、コンピュータの高性能化に合わせるように、専用の LSI やボードを用いたものから、PC 上で動くソフトウェアや組み込み機器で動くミドルウェアへと利用できる範囲が広がっている [6, 7]．これらの音声合成ソフトウェアはテキスト情報を音声として伝達するため、視覚情報を受け取れない障害者向けのテキスト読み上げシステムや運転中のドライバーをサポートするカーナビ等に用いられている．

前述の通り，音声合成とは人工的に音声をつくる技術のことであるが，テキストから音声が合成されることを明示する場合にはテキスト音声合成 (Text-to-Speech Synthesis; TTS) と呼ばれることもある．音声合成の研究はコンピュータの高性能化に合わせて、ルールベースによる手法からデータを集めて統計的に処理する手法へと発展してきた．人手で調整されたダイフォン [1]などの音素単位を接続するよりも、数時間から数十時間の長さの音声データベースから適切な音声単位を選択する方が合成音声の品質の向上を期待できる．このような手法は単位選択型音声合成方式 [8]、あるいはコーパスベース音声合成方式と呼ばれ、現在広く使われている．コーパスベース方式は高品質な合成音声を得られる手法である一方で，次のような弱点が存在する．

1. 声質，発話スタイルを多様にするほど必要な音声データベースの量が増大する
2. 合成音声の品質にばらつきが生じる

1. は音声データベースに無い声質の音声を合成できないということであり，2. はうまく接続できる音声単位がデータベースにあれば高品質な音声が得られるが，なかった場合には不連続生じて品質が下がってしまうことを表す．

この弱点を克服する手法として，統計的パラメトリック音声合成方式 [9] と呼ばれる方式が注目を集めている．統計的パラメトリック方式は，データに基づいた手法という意味ではコーパスベース方式の一種である．特に，統計モデルとして隠れマルコフモデル (hidden Markov model; HMM 図 2.1) を用いた方式は，効率的な学習アルゴリズムや関連ツールが

[1]ある音素から次の音素の中心までで定義される音素単位

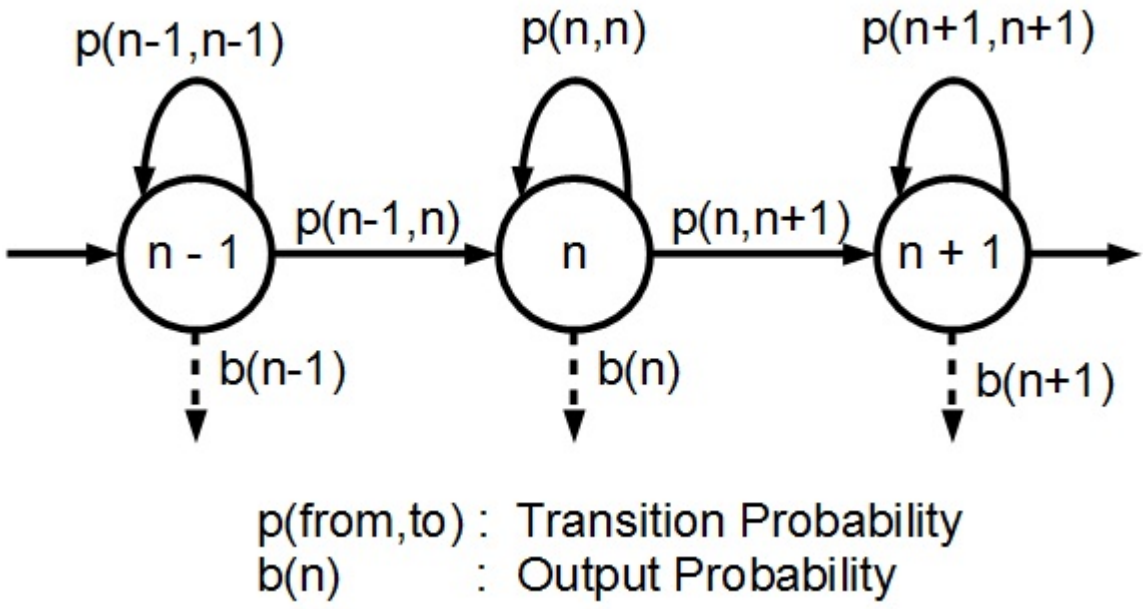


図 2.1: 隠れマルコフモデルの例

利用できることから，統計的パラメトリック方式の代表的な手法になっている．

2.2　隠れマルコフモデル (HMM) による音声合成

図 2.3 に HMM 音声合成の概念図を示した [10]．HMM 音声合成の基本的な考え方はとても簡単に表すことができる．学習用音声データの集合と対応するラベルをそれぞれ O, W、合成音声と対応するラベルをそれぞれ o, w とした時、次の式で表すことができる．

$$\text{学習：} \quad \lambda_{max} = \arg\max_{\lambda} P(O|W, \lambda) \tag{2.1}$$

$$\text{合成：} \quad o_{max} = \arg\max_{o} P(o|w, \lambda_{max}) \tag{2.2}$$

ただし、λ は HMM のモデルパラメータである [11].

HMM のモデル学習は尤度最大化基準に基づいた学習アルゴリズムによる．通常、各 HMM は音素等の音声単位に対応する長さの音声をモデル化する．合成を行う際に、音声波形再合成のために基本周波数 (F_0) に関する情報が必要であるため、学習音声のスペクトルパラメータおよび F_0 パラメータの 2 つのパラメータを結合した特徴ベクトルの列をモデル化する．F_0 パラメータは無声区間において値が存在しないという性質を持つために、特徴ベクトルの F_0 の部分には状態出力確率分布が用いられる [12]．また、動的特徴とよばれる各パラメータ列の時間方向微分に対応するパラメータを付加する．これは HMM が時間方向の相関をモデル化しにくいという点をを補うためである．これに加えて明示的な状態継続長モデルを導入することも行われる [13]．HMM では時系列の時間方向の伸縮変動は状態遷移確率によりモデル化されるが、音声の時間的な構造を精度よくモデル化するには不十分だからである．このようなモデルは Hidden semi–Markov model(HSMM 図 2.2) などと呼ばれる．

また、各音素 HMM が用いるコンテキスト依存モデルは先行・後続音素のみに依存するのではなく、アクセント型、品詞、文長、文内位置などの言語的な情報にも依存している．これはスペクトルパラメータが主として音素コンテキストに影響を受けるのに対して、F_0 パラメータおよび継続長は言語的な情報にも影響を受けるからである．これらのコンテキストの組み合わせから、コンテキスト依存モデルの数が膨大となるため、クラスタリング手法を用いて類似したモデルあるいは状態出力分布を統合することが行われる．その際、スペクトルと F_0 はそれぞれ別のコンテキストに依存するため、状態出力分布のスペクトル部と F_0 部は独立にクラスタリングされる．同様に状態継続長分布に関してもクラスタリングを行う．これにより、スペクトル、F_0、継続長のすべてを一つの確率モデルによってモデル化することができ、合成時に必要な全てのモデルパラメータを同時に自動学習することが可能となっている [14]

一方で音声の合成は、与えられたテキストに対応するラベル列に従って音声単位 HMM を連結することによって得られる HMM から、音声パラメータの列を生成することにより行われる．式 2.2 の通り、音声パラメータの生成は HMM からの出力確率を最大化するように行われる．この際に動的特徴を取り入れることにより、時間的になめらかに変化する

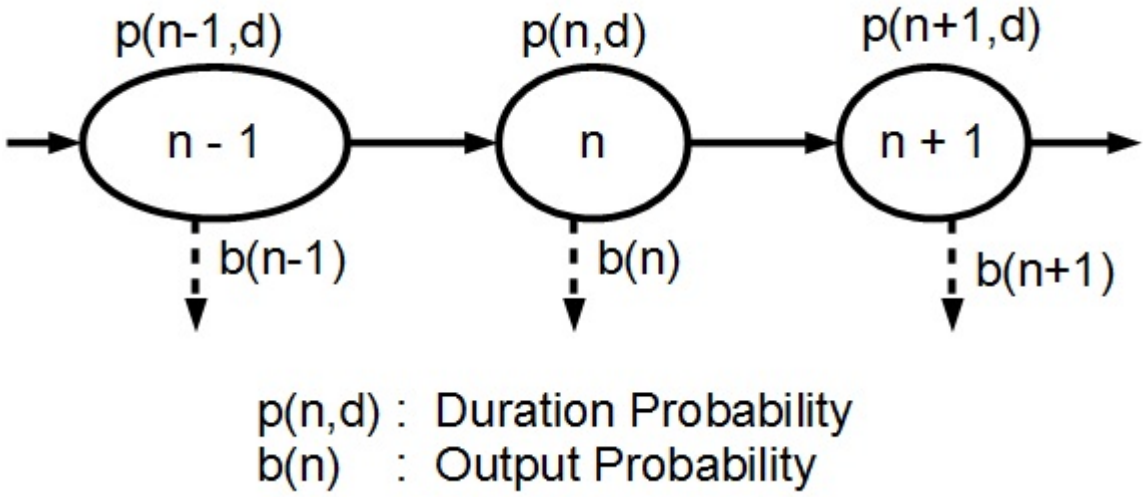


図 2.2: 隠れセミマルコフモデルの例

音声パラメータ列の生成ができる [15]．スペクトルパラメータとしては、合成フィルタとの整合性の良い形式に定義されたメルケプストラムなどが用いられる [16]．また、後述するように、HMM 音声合成は、学習したパラメータに対して何らかの処理を行うことが可能であるため、学習データにない声質やスタイルの音声の合成を行うこともできる．

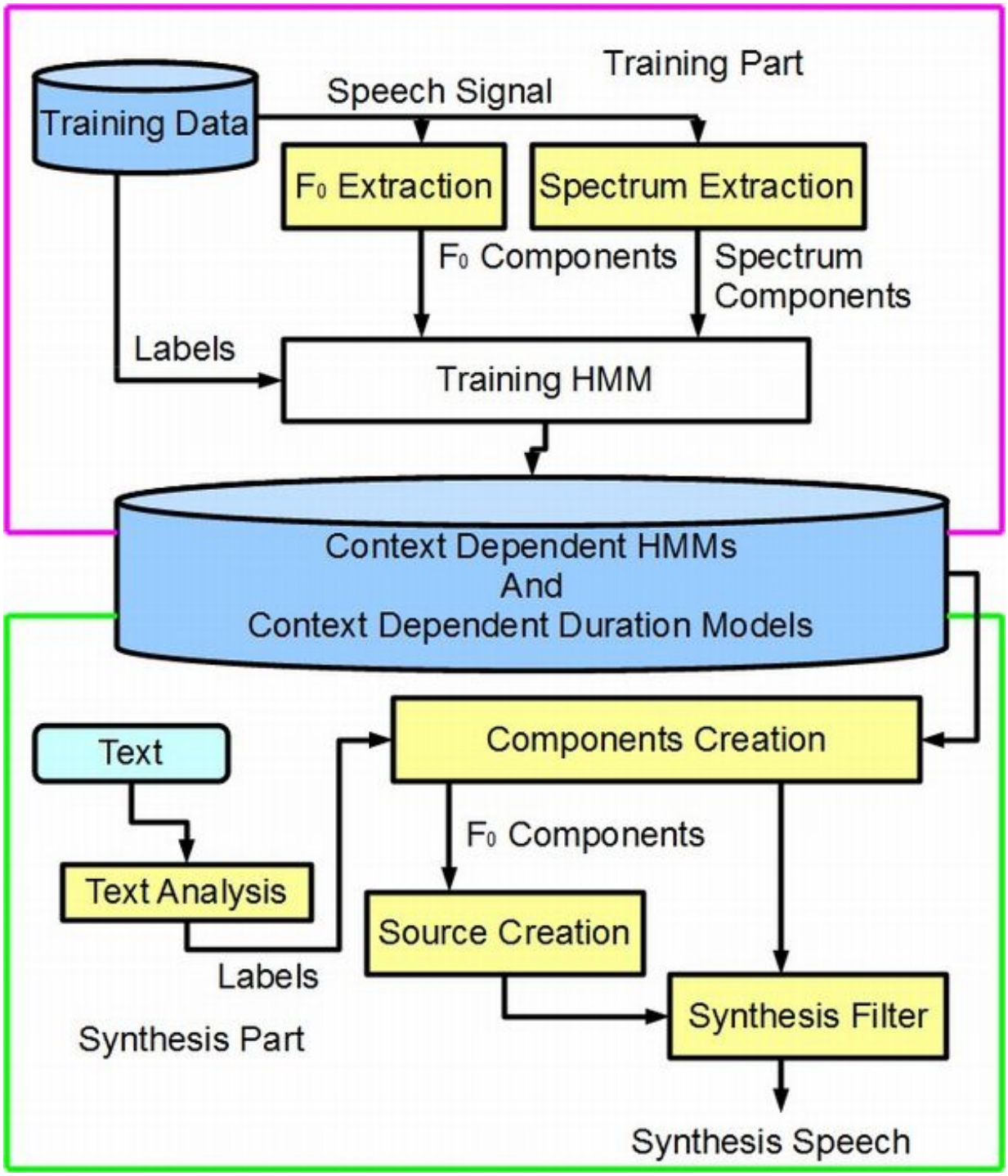


図 2.3: HMM 音声合成の概念図

2.3　HMM 音声合成の応用

HMM 音声合成の特徴として、比較的容易に多様な声質、発話スタイルの音声が得られるという特徴がある．HMM 音声合成と関連した研究として次のようなものがある．

2.3.1　話者適応

音声認識の分野においても HMM が用いられている．通常、多数の話者の音声データを用いて学習した不特定話者 HMM が用いられているが、特定の話者に対して認識精度を向上させるために、話者適応という手法が用いられることがある．これは、少量の特定話者データを用いて、モデルパラメータを当該話者に合わせるよう自動調整することにより、認識精度を向上させる手法である．このような話者適応の手法を音声合成に適応することにより、特定の話者の声質あるいは発話スタイルを真似ることが可能となる [17]．スペクトル、F_0、継続長に関するすべてのパラメータが適応されるため、韻律に関する特徴を含めて適応することができる．

2.3.2　話者補間

HMM 音声合成では、各話者の特徴は HMM のモデルパラメータとして表現されているため、2 つの話者モデルがあったとき、それらのモデルパラメータを適当な方法で補間することにより、2 人の話者の中間的な性質を持ったモデルを得ることができる [18]．複数の話者モデルがあった場合、あるいは、各特定話者モデルを特定の発話スタイルに置き換えた場合も同様である [19]．

2.3.3　固有声手法

前述の話者補間では、元となる話者モデルの数が多ければ多いほど、得られる合成音声の自由度は高くなるが、その一方でそれぞれのモデルにどのような補間係数を設定すればよいかということが、ユーザにとって把握しがたいものとなってくる．このため、主成分分析などの手法を話者モデルのパラメータ集合に適応することにより、全話者空間を少数のパラメータで表そうとするのが、固有声手法である [20]．元来、固有声手法は、音声認識のための話者適応手法であり、数単語程度の極度に少量の適応データに基づいて話者適応を行おうとするものである．音声合成での応用では、次元圧縮された話者空間で重み係数を設定することにより、ユーザは容易に所望の「声」の話者モデルを生成することが可能となる．なお、固有声手法においては、次元圧縮された空間の各軸の意味は明らかではないが、各軸を発話スタイルなどの表現語に対応付ける手法も提案されている [21]．

2.3.4　多言語音声合成

HMM 音声合成システムは、言語に依存する部分が殆ど無いという特徴がある．言語に依存するのは、音素などの音声単位を含むコンテキストの定義と、コンテキストクラスタリングのための質問セットの集合のみであり、ソフトウェアツール自体に言語へ依存した部分は全くない．このため、様々な言語へ容易に適応することができる．

2.4　基本周波数パターン生成過程モデル

HMM 音声合成では，音素などの短い単位でモデル化しているため，長い区間に渡る F_0 の変化を組み込むのが難しい．基本周波数パターン生成過程モデル [1](以下 F_0 モデル) は，この F_0 の変化を表現する手法の一つである．F_0 モデルは喉頭の生理的・物理的特性を基に，声帯振動制御機構を定量的にモデル化している．

F_0 モデルでは対数軸上で表現した F_0 パターンが，次の 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．1 つ目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，フレーズ成分と呼ばれる．2 つ目は個々の単語または単語の連続に付随する局所的な起伏に対応するもので，アクセント成分と呼ばれる．

2.4.1　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，声帯振動の始まりよりおよそ 300ms 前から準備され，やや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き，発話の終わり近くで急峻に下降する成分である．このフレーズ成分を式で表すと，式 2.3 のようになる．

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & (t \geq 0) \\ 0 & (t < 0) \end{cases} \tag{2.3}$$

ここで α はフレーズ指令に対する系の速さを求める定数である．

2.4.2　アクセント成分

アクセント成分は個々の単語または連続した単語に付随するもので，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持ち，低いモーラに移る時にやや先行して下降し始める成分である．このアクセント成分を式で表すと式 2.4 のようになる．

$$G_a(t) = \begin{cases} \min\left[1-(1+\beta t)e^{-\beta t}, \gamma\right] & (t \geq 0) \\ 0 & (t < 0) \end{cases} \tag{2.4}$$

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，γ は式 2.4 中の $\min\left[1-(1+\beta t)exp(-\beta t, \gamma)\right]$ が有限の時間内に一定値に達することを保証する定数である．

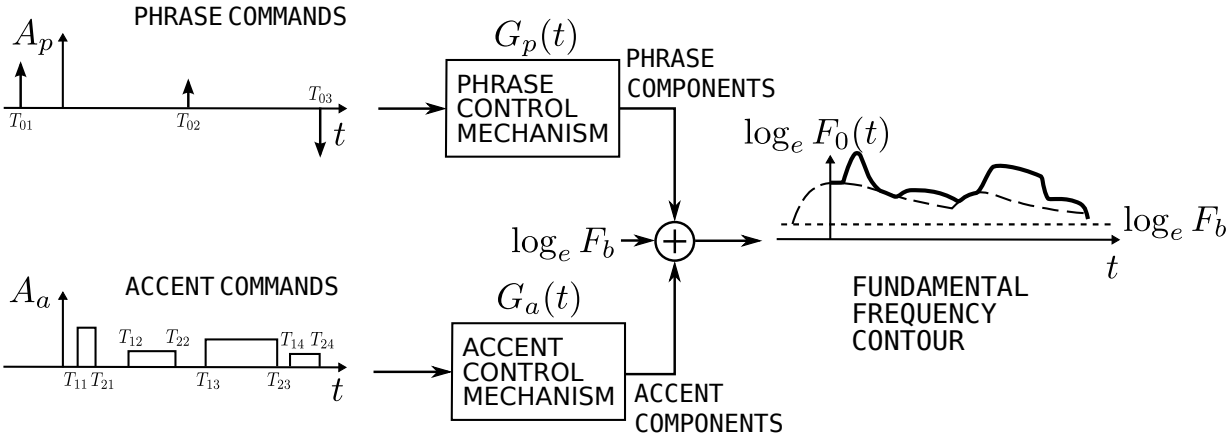


図 2.4: F_0 パターンの例

2.4.3　F_0 パターン生成

F_0 パターンをこの 2 つの成分を用いて表すと式 2.5 のように，各々のパターンが独立であるという仮定のもとで，個々の成分と各話者の発話スタイルに固有な基底周波数との線形和で表すことができる．

$$\begin{aligned}\ln F_0(t) &= \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) \\ &\quad + \sum_{j=1}^{J} A_{aj}\{G_a(t - T_{1j}) - G_a(t - T_{2j})\}\end{aligned} \tag{2.5}$$

ただし，式 2.5 中の変数はそれぞれ次のような意味を持つ．

F_b ： F_0 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_{pi} ： i 番目のフレーズ指令の大きさ
T_{0i} ： 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_{aj} ： j 番目のアクセント指令の大きさ
T_{1j} ： j 番目のアクセント指令の立ち上がり位置
T_{2j} ： j 番目のアクセント指令の立ち下がり位置

図 2.4 は文音声の基本周波数パターンを想定したもので，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続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 F_0 のパターンの様子が生成されている．

また，F_0 モデルの大きな特徴として，基本周波数のパターンを記述するのに必要なパラメータの数が HMM 音声合成と比べて少ないという特徴がある．

第3章

韻律情報の制御

3.1　はじめに

現在 Text-to-Speech システムとして盛んに研究されている音声合成方式として HMM 音声合成がある。HMM 音声合成は前章にて挙げた通り、HMM の適応技術等を用いて話者変換などの “柔軟な音声合成” を行うことができる。しかしながら、HMM 音声合成では制御対象としている特徴ベクトルが、フレーム単位という非常に短い時間を対象としたものであるため、基本周波数 (F_0) パターンで代表される韻律の制御には不向きであるという問題を持っている。これは、韻律は単語、句などといったフレーム単位と比べて非常に長い時間に渡る特徴であることに依るところが大きい。さらに、HMM 音声合成により得られた韻律的特徴と入力テキストとの明確な対応を取ることが難しく、合成音声に追加的な処理を行い “柔軟な音声合成” を行うことが困難という問題もある。

この問題に対して、F_0 モデルを用いた韻律の制御が提案されている。前章にて挙げた通り、F_0 モデルは比較的長時間に渡る F_0 の変化を記述することができ、フレーズ指令、アクセント指令の値を調整することにより韻律の制御を行える。F_0 モデルを用いた韻律制御の手法として、特に焦点を置かない発声と特定の語句に焦点を置いた発声のパラレルコーパスを用いて、両者の F_0 パターンの指令を変更することによって焦点を置いた音声を合成する手法 [2, 3] や、同様にしてスタイル変換、声質変換を行う手法 [4] が開発されている。

3.2　F_0 モデルを用いた焦点制御手法

先行研究 [2, 3] では、図 3.1 のように、同一話者による、特に焦点を置かない発声と特定の名詞句に焦点を置いた発声との F_0 パターンの違いを、指令の差分として捉え、テキストの言語情報などから推定する二分木を少量のコーパスを用いて学習、構築し、そこから得られる差分に基いて、特に焦点を置かない合成音声に対し、その F_0 パターンの指令の大きさを変更して、焦点制御を実現している。その際、差分学習のためのコーパスの話者は、焦点制御を含まない音声合成の音声コーパスの話者と同一である必要がないという特徴を持つ。

差分情報は、同一文で焦点のあるものとないものそれぞれの音声の韻律的特徴量の実測データと、発話文の言語情報、発声時に指定した焦点の位置の情報を用いている。また、指令の大きさはフレーズ指令を先に推定し、推定されたフレーズ指令の大きさを用いてアクセント指令の大きさを推定している。その際に指令の大きさが負の値になった場合、その大きさを 0 としている。

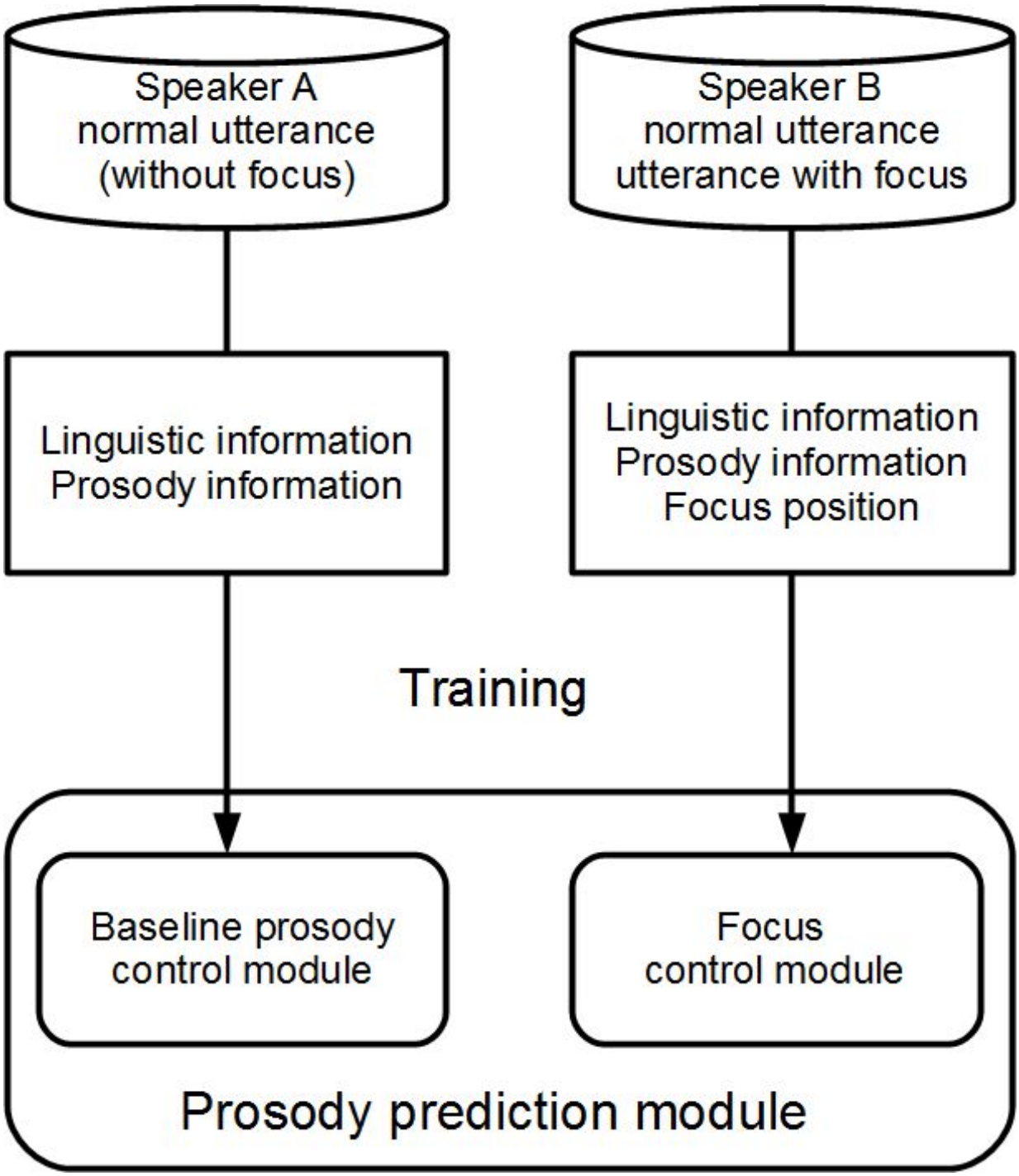


図 3.1: 韻律制御部の構築

第4章

指令差分に基づく焦点制御の改良

4.1　はじめに

従来の F_0 モデルを用いた焦点制御の手法は，F_0 パターンを制御することが中心であり，基本的に他の韻律的特徴は操作の対象外となっていた．しかし，焦点が付与された場合は，休止の位置や長さもそれに応じて変化する．

特定の文章の読み上げにおいて焦点の付与を行うことを考えた場合，焦点の付与に伴って変化する休止は，文章そのものを変更することはなく，もとから存在した休止の消滅あるいは時間の増減，もしくは新しい休止の挿入となる．

表 4.1: CART の質問セット

項目	カテゴリ数
当該文節モーラ数	10
1 つ前文節モーラ数	10
1 つ後文節モーラ数	10
当該文節アクセント型	6
1 つ前文節アクセント型	6
1 つ後文節アクセント型	6
当該文節自立語の品詞	16
1 つ前文節自立語の品詞	16
1 つ後文節自立語の品詞	16
当該文節に焦点が付与されているか	4
焦点付加音声において対応する文節の直前文節が休止かどうか	2
焦点付加音声において対応する文節の直後文節が休止かどうか	2
(アクセント指令のみ) 直前のフレーズ指令の大きさ	(実数値)

4.2　焦点制御への休止区間の組み込み

前述のように，焦点が付与された場合，その直前あるいは直後の休止の様子が変化することが多い．休止は発話の時間構造に密接に関係するため，テキストから，まずこれを推定し，その結果を用いて，指令の推定を行うのが妥当である．ここでは，休止の推定が既に行われたと仮定し，休止の変動を取り入れる焦点制御を行う．先行研究と同様に二分木を用いることとし [2, 3]，その入力項目（質問セット）を表 4.1 のようにした．

本研究においては，焦点を付与した文節の直前，直後に休止があるか否かを入力項目に加えている．休止があるか否かを判断するために自立語の品詞のカテゴリに対し，文頭，休止，文末の 3 種を追加した．また，アクセント指令については，フレーズ成分との関連があるため，直前のフレーズ指令の大きさを入力項目に加えている．

各カテゴリにおける，詳しい内容は以下の通りである．

- モーラ数
  1 モーラから 10 モーラの 10 種
- アクセント型

0 型から 5 型の 6 種

- 品詞
  動詞，名詞，代名詞，形容詞，形状詞，連体詞，副詞，接続詞，感動詞，助詞，助動詞，接頭辞，接尾辞，文頭，休止，文末の 16 種
- 当該文節に焦点があるかどうか
  (当該/直前/直後) 音節に焦点がある，その他の場合である，の 4 種

また，指令の差分に着目した手法では，パラレルコーパスで，指令の 1 対 1 対応を仮定し，指令の大きさの差のみを推定対象としている．だが，実際の発声では，このような対応が得られない場合がある．これに対して，フレーズ指令については，大きさ 0 の指令を仮定することで，対応を取った．アクセント指令については，1 文節につき 1 つのアクセント指令とすることで，対応を取った．

4.3　合成実験と評価

提案手法の有効性を確認するため，音声の合成を行った．予め焦点の付与による休止の有無が決定できたと仮定し、休止の有無に応じた焦点付き合成音声の作成を行った。

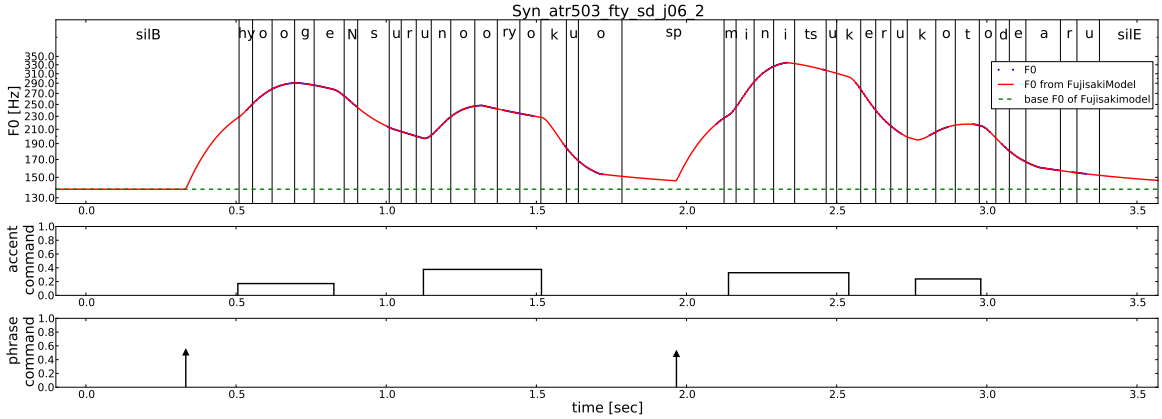
まず，女性話者 1 名が読み上げた 30 文の焦点が付加されてない音声と，同じ文章で焦点が付加された音声とのパラレルコーパスを用い，F_0 パターンの指令の差分を二分木で学習した．次に，HMM 音声合成にて作成した焦点を考慮しない音声に対し，差分を適応して F_0 パターンの修正を行った．

HMM の学習には話者 FTY による ATR 日本語音素バランス文 503 文の A セットから I セットの 450 発声を用いた．HMM の構築に用いた特徴量ベクトルは，25 次までのメルケプストラム係数と対数 F_0，およびそれらの 1 次微小変化分，2 次微小変化分の計 78 次元とした．HMM は単一ガウス分布を 5 状態の left-to-right 型 HMM とした．この HMM を用いて ATR503 文の J セットのうち 10 文を合成し，差分の適応を行った．なお，パラレルコーパスに用いた 30 文はすべて ATR503 文の A セットの文章であるため，HMM の学習に用いた文章と重複がある．

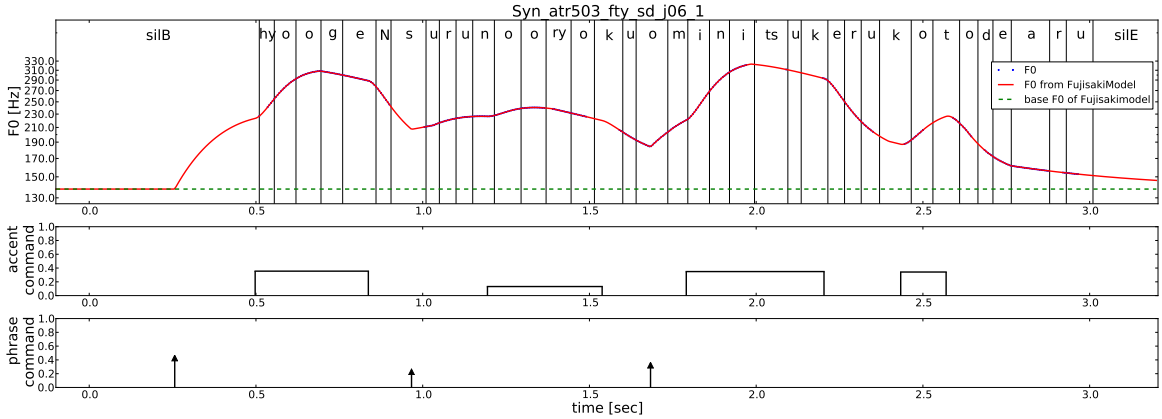
音素アライメントに Julius[1]，HMM の構築に HTS[2]，音声の合成に SPTK[3]を用いた．また二分木の構築には Classification And Regression Trees(CART) を用い，CART の学習に wagon[4]を用いた．また、F_0 モデルのパラメータのうち、α、β、γ についてはそれぞれ $\alpha = 3.0[s^{-1}]$、$\beta = 20[s^{-1}]$、$\gamma = 0.9$ で一定とした。

合成した音声の F_0 パターンの例を図 4.1、図 4.2 に示した．図 4.1 では、どちらも「表現する能力を身につけることである」のうち「身に付ける」の部分に焦点を付加している．また、図 4.2 では、どちらも「平均倍率を下げた形跡がある」のうち、「下げた」の部分に焦点を付加している。休止の付加によって単に発声が引き伸ばされるだけではなく，フレーズ指令，アクセント指令共に異なった適応を受けていることがわかる．

[1]http://julius.sourceforge.jp/
[2]http://hts.sp.nitech.ac.jp/
[3]http://sp-tk.sourceforge.net/
[4]http://www.cstr.ed.ac.uk/projects/speech_tools/

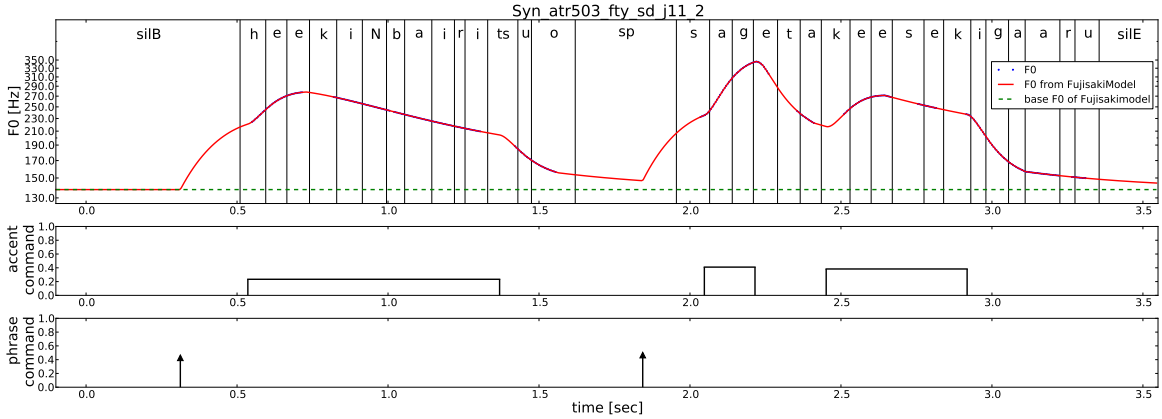


(a) 休止あり

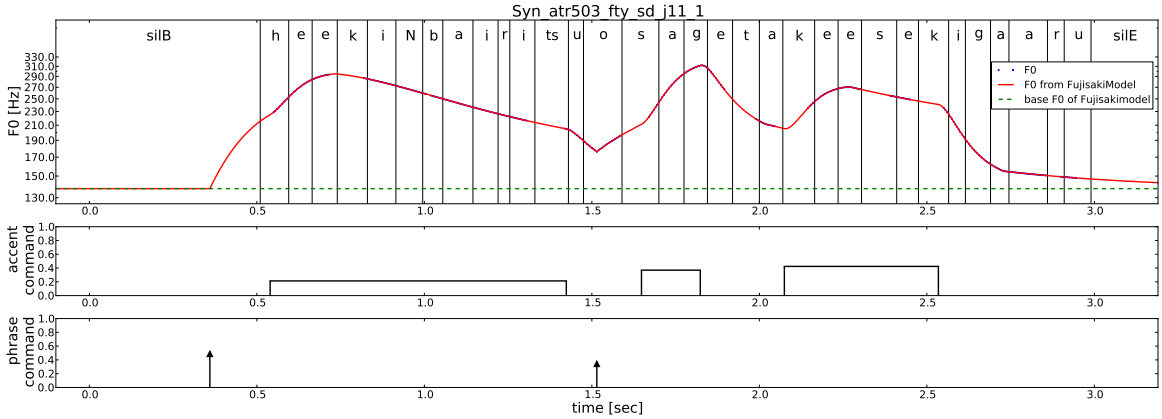


(b) 休止なし

図 4.1: 合成した音声の例 [1](「表現する能力を身につけることである」)



(a) 休止あり



(b) 休止なし

図 4.2: 合成した音声の例 [2](「平均倍率を下げた形跡がある」)

第5章

結論

5.1　まとめ

近年、音声を用いたマンマシンインターフェースは用途が広がっているが、人間の音声コミュニケーションにおいて情報の伝達に用いられる韻律情報の制御の高度化が大きな課題となっている。現在 Text-to-Speech システムとして盛んに研究されている HMM 音声合成方式では、“柔軟な音声合成” を可能にできる一方で F_0 パターンで代表される韻律の取り扱いには不向きという問題がある。この問題は HMM 音声合成がフレーム単位の短い時間で処理を行う為に生じている。

F_0 モデルは F_0 パターンの表現を 2 種類の指令成分を用いて表現するモデルであり、比較的長い時間単位で処理を行えるため、HMM 音声合成と比べて韻律の制御を行いやすい。既に F_0 モデルを用いて韻律の制御を行う手法として、パラレルコーパスより F_0 パターンの指令の差分を学習し、HMM 音声合成にて特に焦点を付与せずに合成した音声の F_0 パターンの指令を変更することにより焦点を付与した音声を合成する手法が開発されている。

しかしながら、その制御の対象は F_0 パターンが中心であり、他の韻律的特徴は操作の対象外とされていた。特に休止の位置や長さは、焦点の付与に応じて変化することが多く、韻律の制御の際に考慮すべき要素となりえる。本論文では休止の有無を考慮した焦点制御の為に、文頭、休止、文末を単語とみなしてパラレルコーパスの差分を学習することを行った。また、指令の差分に着目した手法では、パラレルコーパスにおいて 1 対 1 対応を仮定し、指令の大きさの差のみを推定対象としている。しかし実際の発声では、指令の対応関係が 1 対 1 対応とならずに差分の学習が煩雑になることがあった。この問題に対して、フレーズ指令は大きさ 0 の指令を仮定することにより、アクセント指令は 1 文節に対して 1 つのアクセント指令とすることにより、1 対 1 対応を常に取ることとした。これを踏まえて予め休止の有無が決定できた場合を仮定して、休止の有無を取り入れた焦点付きの音声の合成を行い、休止の有無に応じた合成音声を得た。

5.2　今後の展望

今後の展望としては、デュレーションの取り込みが考えられる。各音素の長さを焦点の有無にあわせて調節することにより、さらなる柔軟性が得られると考えられる。また、休止の有無の判別を自動で行い、焦点を付与する部分を指定するだけで休止を考慮した焦点付き合成音声を生成するシステムへの発展が考えられる。

謝辞

最初に，３年間教官として指導してくださった広瀬啓吉教授及び峯松信明教授に感謝いたします．両教授には研究に関する助言を始め多様な面から支えて頂きました．支援がなかったならばここまで研究を続けることができなかったと思います．次に，技官の高橋登氏，広瀬・峯松研究室秘書の池上恵氏，折茂結実子氏に感謝いたします．両氏には直接・間接問わず多くの支援を頂きました．特に計算機の管理やティーチングアシスタントでの書類の管理など，協力が無ければ順調に進むことはなかったと思います．そして，共に研究室生活を過ごした広瀬・峯松研究室の皆様に感謝いたします．特に同期である橋本浩弥氏には非常に多くの支援を頂きました．同氏の支援が無ければ確実に詰みの状態に陥っていたと思います．最後に，休日に一緒に遊び気分転換の助けとなった友人達，ここまで私を育ててくれ，支えになってくれた両親，親戚の皆様に感謝します．

2013 年 2 月 6 日

川口 拓也

参考文献

[1] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” *J. Acoust. Soc. Jpn.(E)*, 5, 233-242 (1984).

[2] K. Ochi, K. Hirose and N. Minematsu, “Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,*, 4485–4488 (2009–4).

[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. International Conference on Speech Prosody*, 171–174 (2012–5).

[4] 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, Florence, August 28–31, pp.2793-2796 (2011–8).

[5] 古井貞熙, “ディジタル音声処理,” 東海大学出版会 (1985).

[6] 三留幸夫, “音声合成の応用と今後の展望,” 日本音響学会誌, 49, 875–880 (1993).

[7] 籠嶋岳彦, “テキスト音声合成技術実用化の動向,” 日本音響学会誌, 67, 23–27 (2011).

[8] A. Hunt, A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” *Proc. ICASSP* 373–376, (1996).

[9] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Commun.*, 51, 1039–1064 (2009).

[10] 徳田恵一, “統計的パラメトリック音声合成技術の動向,” 日本音響学会誌, 67, 17–22 (2011)

[11] 徳田恵一, “特集号：音声情報処理技術の最先端 (1)HMM による音声認識と音声合成,” 情報処理学会誌, 45, 1005–1011 (2004).

[12] 徳田恵一, 益子貴史, 宮崎昇, 小林隆夫, “多空間上の確率分布に基づいた HMM,” 信学論 (D-II), J83-D-II, 1579–1589 (2000).

[13] H. Zen, K.Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, “Hidden semi–Markov model based speech synthesis,” *IEICE Trans. Inf. Syst.*, E90-D, 825–834 (2007).

[14] 吉村貴克, 徳田圭一, 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,” 信学論 (D-II), J83-D-II, 1879–1589 (2000).

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, “Speech parameteter generation algorithms for HMM–based speech synthesis,” *Proc. ICASSP*, 3, 1315–1318 (2000).

[16] 徳田圭一, 小林隆夫, 深田俊明, 斎藤博徳, 今井聖, “メルケプストラムをパラメータとする音声のスペクトル推定,” 信学論 (A), J74-A, 1240–1248 (1991).

[17] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata and J. Isogai, “Analysis of speaker adaptation algorithms for HMM–based speech synthesis and a constrained SMAPLR adaptation algorithm,” *IEEE Audio Speech Lang. Process,*, 17, 66-83 (2009).

[18] T. Toshimura, K. Tokuda, T. Masuko, T. Kobayashi and K. Kitamura, “Speaker interpolation for HMM–based speech synthesis system,” *J. Acoust. Soc. Jpn. (E)*, 21, 199–206 (2000).

[19] M. Tachibana, J. Yamaguchi, T. Masuko and T. Kobayashi, “Speech synthesis with various emotional expressions and speeking styles by style interpolation and morphing,” *IEICE Trans. Inf. Syst.*, E88-D, 2484–2491 (2005).

[20] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, “Eigenvoices for HMM–based speech synthesis,” *Proc. ICSLP*, 1269–1272 (2002).

[21] K. Miyanaga, T.Masuko and T. Kobayashi, “A style control technique for HMM–based speech synthesis,” *Proc. INTERSPEECH*, Vol. II, 1437–1440 (2004).

発表文献

[1] 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, 3–P–34b, (2013–3). (To Appear)

[2] 川口拓也, “格フレーム辞書を用いた単語置換に基づく音声認識用言語モデルの高精度化,” 東京大学工学部電子情報工学科卒業論文, (2011–2).