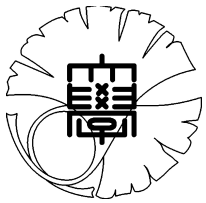
修士論文

音声合成のための
文節単位での感情の程度を考慮した
統計的韻律制御



2006 年 1 月 30 日

指導教官　広瀬 啓吉 教授

東京大学大学院新領域創成科学研究科
基盤情報学専攻　46301

浅野 泰史

内容梗概

音声の韻律的特徴は、文字言語にも含まれる語義・統語・意味などの言語情報のみならず、文字言語には含まれない話者の意図や態度などに関するパラ言語情報、話者の個体差、性別、年齢、感情などに関する非言語情報をも含んでいる。主に超分節的特徴である韻律的特徴によって表される感情の表現の仕方は話者や状況によって様々であり、韻律的特徴量と感情との関係は線形的なものではない。そのため、人手による感情の制御規則の解明や構築は、非常に困難な作業となる。

これらをふまえ、本研究では、韻律的特徴量の中でも基本周波数パターンに焦点を置き、コーパスから統計的手法によって、その制御規則を構築するという手法を取った。また、テキストを入力とした感情音声を合成するために、基本周波数パターン生成課程モデルのパラメータを推定する際に韻律語境界を推定する統計モデル、アクセント核位置の決定規則を用いるという先行研究の手法を採用した。

本稿ではまず、「平静・怒り・喜び・悲しみ」の 4 感情の F_0 モデルパラメータの推定を従来手法により行い、その際に作成された学習木を利用して 2 種類の実験を行った。この実験により、感情の種類と韻律的特徴量の関係を調べた。

次に、文節ごとに記録した発話者の感情の程度の情報を用いた F_0 パターンの作成とその評価を行った。本研究で提案する手法によって、従来手法と比べ より正解に近い F_0 モデルパラメータの推定に成功した。

最後に、推定された F_0 パターンを用いて感情音声を合成し、主観的評価実験を行った。実施した 3 種類の実験により、提案手法による合成感情音声の方が、より明確に元の音声の感情表現を再現できたことが確かめられた。

目次

図目次

表目次

第1章

序論

1.1　本研究の背景

近年、計算機技術の著しい発展と、インターネットをはじめとする情報ネットワークの広汎な拡大・普及に伴い、従来とは比較にならないほど高性能の計算機が一般消費者にも入手可能となり、処理できるメディア (静止画・動画・音声など) や機能も実に多彩なものとなった。しかし、このように情報処理環境が豊かになっていく一方で、情報機器の高機能化のためにますます操作が複雑化していく事が、初心者や高齢者にとって導入のバリアとなっている事も否定できない。

そこで今、誰にでも使用できて親しみやすいインターフェースが求められており、その中で人間らしい、心のこもったコミュニケーションを可能にする手段として、音声が注目を集めている。

音声に関する従来の研究では、主として文字表記と音声との対応という観点から行われてきたため、特に分節的特徴[1] である音韻に重きがおかれていた。しかし、自然会話音声には、文字言語にも含まれる言語情報の他にも話者の意図や態度、感情など様々な情報が含まれており、この言語情報以外の情報の表現を実現することも、人間と機械のコミュニケーションをより円滑にするシステムの構築のためには、欠かすことができない。これらは、主に超分節的特徴である韻律によって表され、それを利用するための研究が現在望まれている。

1.1.1　音声に含まれる情報

文字情報では、文脈などによって言語情報以外の情報が伝達されるが、音声言語では、文字言語に比べ、態度や感情といった情報の比重が増す。音声言語により表現され、伝達される情報は、必ずしも明確に境界を引くものではないが、主に

- 言語情報 - linguistic information
- パラ言語情報 - para-linguistic information
- 非言語情報 - non-linguistic information

の 3 つに大別することができる [1]。ここで、言語情報とは辞書・統語・意味・談話のレベルで文字言語によって陽に表現されるか、あるいは文字言語による表記およびその前後の文脈から容易に、一義的に導出し得るものをさす。それに対し、発話の仕方によって伝わる態度・感情・話者の状態など、記号情報以外で伝わる情報のうち、話し手が意図的に制御できるものをパラ言語情報、話し手が意識的に制御できないものを非言語情報と呼ぶ。対話において、言いたいことが文字情報として完全でなくても意図が伝わり、同じ発話内容でも異なった感情が伝わるのは、パラ言語情報・非言語情報も同時に解釈しているためである。

[1] 音素などといった単語や分を構成する個々の音を規定する特徴

1.1.2　感情音声を表す特徴量

一方、音声を工学的に扱うためには、パラ言語情報・非言語情報を表す特徴量を整理する必要がある。

図 1.1 に示すように、音声は音源で生成された波形が声道伝達特性によって周波数領域で加工されたものと言うことができる。分節的特徴である音韻は主に声道伝達特性によって表されるが、超分節的特長である韻律は、主に有声音源特性によって表され、主に

- 声帯振動の基本周波数
- 音素の継続時間長
- 音源の強度

の３つに区分して扱われる。

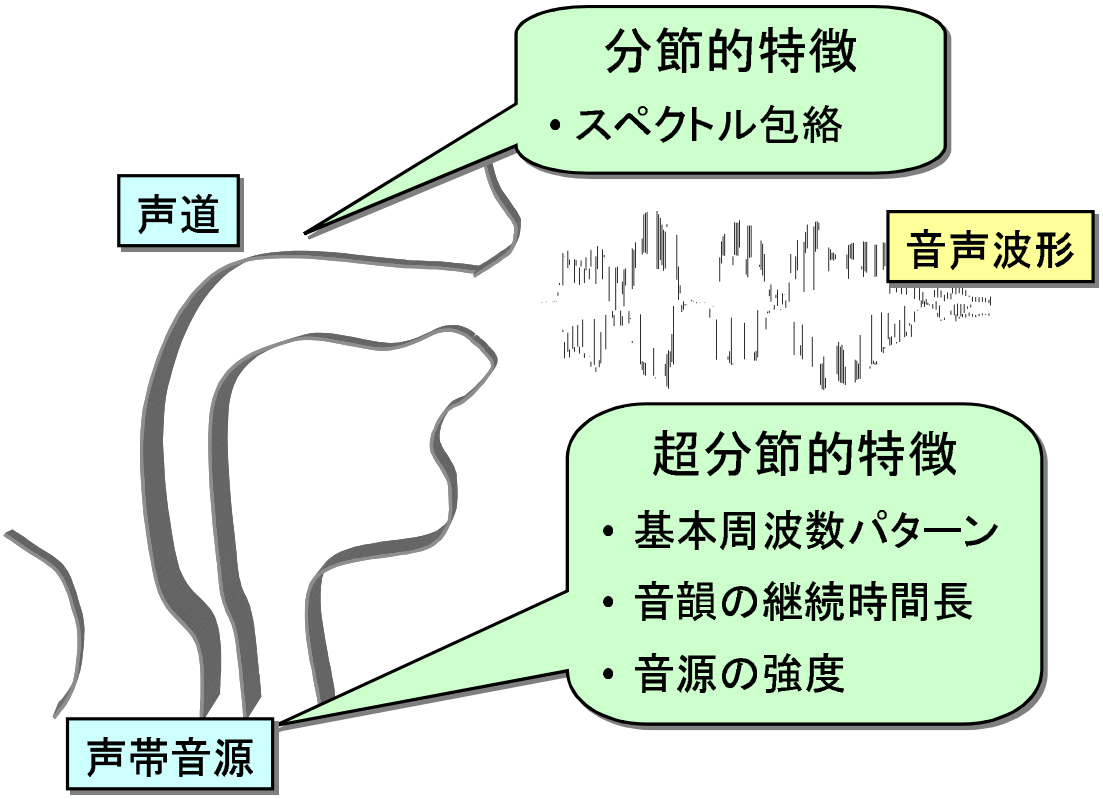


図 1.1: 分節的特徴と超分節的特徴

韻律現象は、分節的特徴に比べて相対的に広い時間領域にわたって生じる現象であり、文字に表現されない音声言語特有のものであるという特徴がある。そして、感性情報は主にこの韻律的特徴によって伝えられる。

1.1.3　感情音声における韻律的特徴

感性情報は主に韻律的特徴によって伝えられることから、これらと韻律的特徴との関係を調べる研究は従来から行われている。

これらの研究では、各種の感情を含ませて収録した単語や文に対して、基本周波数パターン・発話速度などの韻律情報の変化を調べている [2][3][4]。各感情に共通なパターン、特有のパターンが認識されているが、限られた話者・発話を資料としたものであるのが現状である。

重永ら [5] は、感情が平静からの「ずれ」であることに着目し、特徴量も平静からの「ずれ」により表現している。このように正規化した特徴量を用いることによって、最高７０%程度の判別率を得ているが、これも実験に用いている資料数が少ないため汎用的とは言えない。

川波ら [6] は、感性情報をその有無からのみ捉えるのではなく、その強さを考慮した規則化を行うべきであるという観点のもとで、それらと韻律的特徴との関係の定量的な調査を行っている。

武田ら [7] は、各感情の中でも特徴量の捉えやすい「怒り」を取り上げ、怒りの度合いを平常・軽い怒り・怒り・激怒の４段階の感情で表わすことを試みている。怒りの度合いが強くなるにつれて、発話速度が速くなり、基本周波数が高くなる。しかし、激怒になると、基本周波数は上昇するが、発話速度は遅くなるという結果が得られている。

このように、感情音声と韻律的特徴の相関関係については、先行研究により示されているが、その規則は非常に複雑で、その規則はまだ十分に解明されてはいない。実際に、川波ら [6] は、感情の程度の増加に伴い、それを表す韻律的特徴が顕著になる訳ではなく、複数の韻律的な制御方法の中から、様々な方法が適宜取捨選択されて用いられているということを示唆している。

1.1.4　感情音声の合成

現在、良く用いられている音声合成の方式は、次の３種類に分類できる。

1. 波形接続方式
2. 分析合成方式
3. 規則合成方式

番号順にコーパスの蓄積量が少なくなり、柔軟性が高くなるが、音質は劣化する。波形接続方式は、人が発声した音声波形を、そのままあるいは波形符号化して蓄積しておき、必要に応じて繋ぎ合せて出力するものである。分析合成方式は、人が発声した音声波形を分析してパラメータに変換された形で蓄積しておき、それを繋ぎ合せて音声合成器を駆動し、音声を作り出すものである。規則合成方式は、文字列あるいは音素記号列から、音声学的規則に基づいて、音声を作り出すものである。

1.1.3 で述べたように、感情音声に関する規則は非常に複雑であり、その規則を完全に記述するのは難しい。そのため、感情音声合成に関しては波形接続方式、あるいは分析合成

方式が主流であり、既に、それらを用いた研究は数多く行われている。

飯田ら [8] は波形接続方式による感情音声合成を提案している。また、朗読音声の合成において現在盛んに研究されている HMM の枠組みを感情音声に適用し、感情音声合成を実現しようとする試みもある [9][10]。しかし、韻律の制御については未だ不十分な部分が多く、韻律現象と深く結びつく感情音声では特に、その制御法の開発が重要な課題である。

1.1.5　強調の程度の情報と韻律的特徴量の定量的分析

より高い精度で感情音声を扱うことを目指す場合、感情特有の現象を意識した枠組が必要となると考えられる。例えば、感情を込めて文を読み上げる時、1 つの文の中でも特に感情を込めている箇所と平静に近い箇所がある。この感情の程度情報を利用することで、より人間らしい感情表現が実現できるのではないかと予想される。

大野、藤崎らは [11] において音声の強調部分が韻律的特徴の上にどのように表現するかを明らかにするために、F_0 パターンと発話速度に注目し、強調の有無/箇所の違いに関して、F_0 パラメータと局所発話速度比により特徴を定量化し、分析を行った。著者らは分析結果から、フレーズ指令に関しては顕著な特徴の差は見られないが、アクセント指令に関しては強調箇所において指令の大きさが強調箇所において 2 倍以上増大する、指令の持続時間が長くなり、後続の韻律語とアクセント結合が生じ、その他の部分では強調の影響を特に受けない、という傾向を見出している。

また、杉山、藤崎 [12] らは、音源強度にも F_0 パターン生成過程モデルに類似した音源強度変化生成過程モデルを仮定し、音声における強調が、F_0 パターン、局所発話速度、音源強度の 3 つの韻律的特徴の上にどのように表現されているかを系統的、定量的に分析する手法について報告し、また、強調部分に対しては正の音源強度変化指令が生起し、強調しない他の部分では負の指令が生起していることを示した。

立花、西村 [13] は、既存の読み上げ調大規模コーパスをベースとして利用することで、新しい発話スタイルのコーパスを用意する労力を最小化することを目指し、ベースシステムが推定した韻律 (F_0、継続時間長) の出力を修正することで、強調音声の合成を行う手法を提案した。著者らは、大規模コーパスに基づいて構築したベースシステムが推定する韻律を、別の話者の強調音声の小規模コーパスから学習した変化率によって修正し、その韻律を目標としてベースシステムから素片選択を行うことで強調音声の合成を行った。合成音声の強調の有無に関する聴取による主観評価実験によって、強調として認識可能な韻律が著者らの手法で合成可能であることが示されている。

1.2　本研究の目的

このような背景から、本研究では、感情音声における韻律的特徴量の適切な制御規則の構築を目指し、より人間の発声に近い感情を表現した音声の合成を目指す。

韻律を制御する手法としては、人手による発見的なものと、統計的な手法を用いたものがあるが、人手による発見的な手法では、非常に手間がかかる上、感情のような不安定で

それ自体評価の難しいものを扱う場合には、非常に困難である。そのため、本研究では、統計的手法を用いて韻律制御規則の構築を目指す。

統計的手法を用いる場合には、適切にその自由度を減らさないと生成されるパラメータに破綻をきたす恐れがある。先行研究 [14] [15] では、基本周波数生成過程モデル [16](以下、F_0 モデル) に基づいた、感情音声の基本周波数パターン (以下、F_0 パターン) の制御規則の構築と推定の精度の向上を図り、コーパス作成の高精度化や統計モデルへの入力項目の検討などを行ない、実際にテキストから感情音声を合成するシステムを構築している。

本研究では、この枠組みを利用して、発声者の感情の程度の情報を統計モデルの入力項目に追加することによって、より元の音声に近い感情表現を実現するシステムを構築する。また、作成された合成音声の評価を 3 種類の聴取実験によって行い、統計的韻律制御の精度を検証する。

1.3　本論文の構成

本論文では、まず第 2 章で、感情音声合成の枠組みと必要な技術要素について説明する。第 3 章では、F_0 モデルパラメータの推定の枠組みとその際用いる手法について説明する。第 4 章では、感情の種類が推定される F_0 モデルパラメータに対して与える影響について分析を行う。第 5 章では、文節ごとに記録された発声者の感情の程度情報を用いた F_0 パターンの作成評価を行う。第 6 章では、5 章で構築された F_0 パターン作成の枠組みを利用して実際に感情音声を合成し、生成した音声を元に主観的評価実験を行う。最後に第 7 章で結論を述べる。

第2章

感情音声合成に用いる要素技術

2.1　はじめに

任意の漢字仮名混じり文から音声を合成する日本語 Text-to-Speech(以下、TTS) システムの各処理では、適切な制御規則に基づいて入力パラメータを出力パラメータに変換する。

本研究では、TTS システムにおける各処理での制御規則に統計的手法を用いている。そこで、本章では TTS システムの概観について述べた後、それぞれの処理で用いる統計モデルについて述べる。また、F_0 パターンの生成は、F_0 モデルに基づいて行っているが、その原理についても本章で述べる。

2.2　日本語テキスト音声合成システム

分析合成方式による TTS システムの概観を図 2.1 に示す。

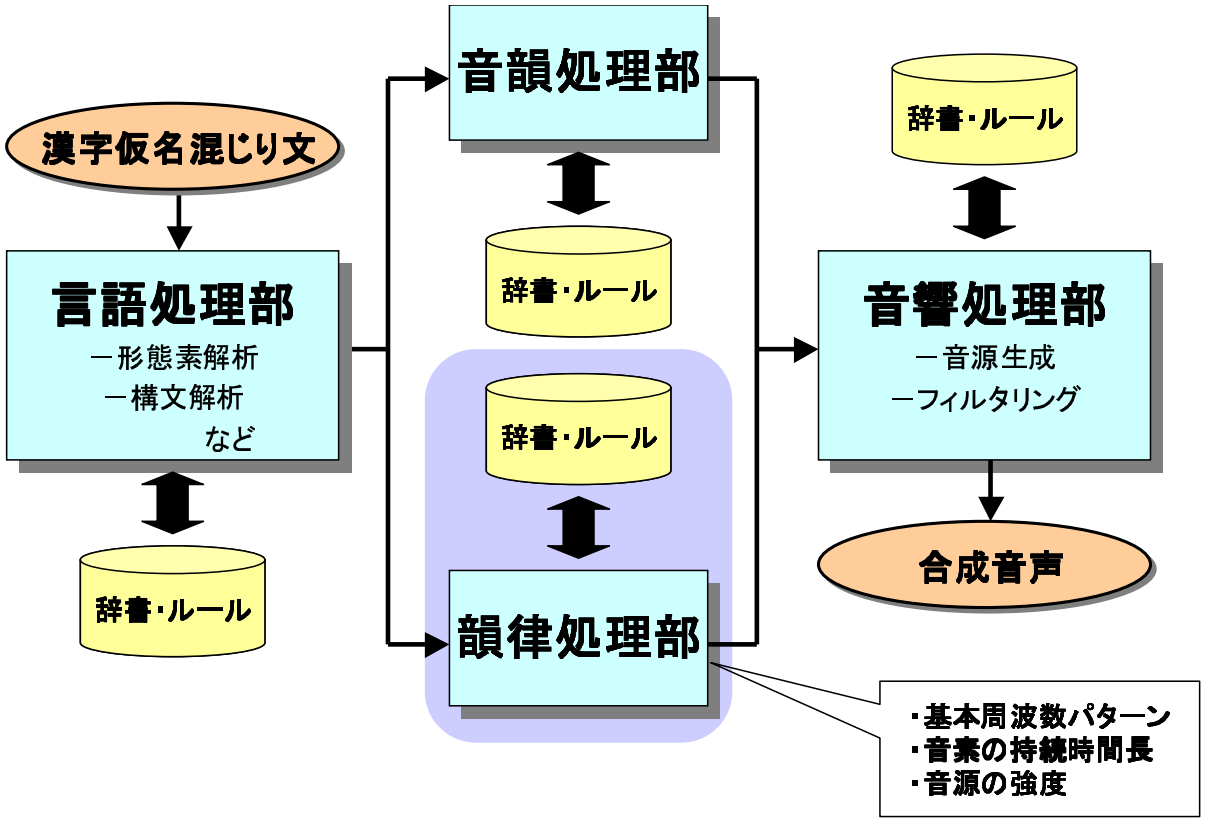


図 2.1: 日本語テキスト音声合成システムの概観

TTS システムは、大きく 4 つの処理部に分けられる。まず、漢字仮名混じり文が入力されると、言語処理部において、形態素解析・構文解析などのテキスト解析が行われ、品詞情報・統語情報などの言語情報を得る。得られた情報をもとに、韻律処理部では F_0 パター

ンなどの韻律的特徴を表す特徴量を決定する一方、音韻処理部では、ケプストラム[1] などの音韻パラメータからスペクトル包絡を得る。音響処理部では、韻律処理部で生成された韻律的特徴量から音源を生成し、音韻処理部で生成されたスペクトル包絡をフィルタリングして、音声波形を出力する。

本研究では、日本語 TTS システムにおける韻律処理部を対象とし、日本語においては、韻律情報の中でも特に重要な F_0 パターンについて扱う。韻律処理部が行うことは、言語情報・パラ言語情報・非言語情報を考慮して、あらかじめ構築されている韻律生成規則を元に、実際に言語情報が入力された時に所望の韻律を生成することである。

2.3　統計モデル

2.3.1　隠れマルコフモデル (HMM)

音韻処理部における、スペクトル包絡生成の手段として、音声認識の分野で良く用いられてきた隠れマルコフモデル (Hidden Markov Model: HMM) を応用する手法が広く用いられている。

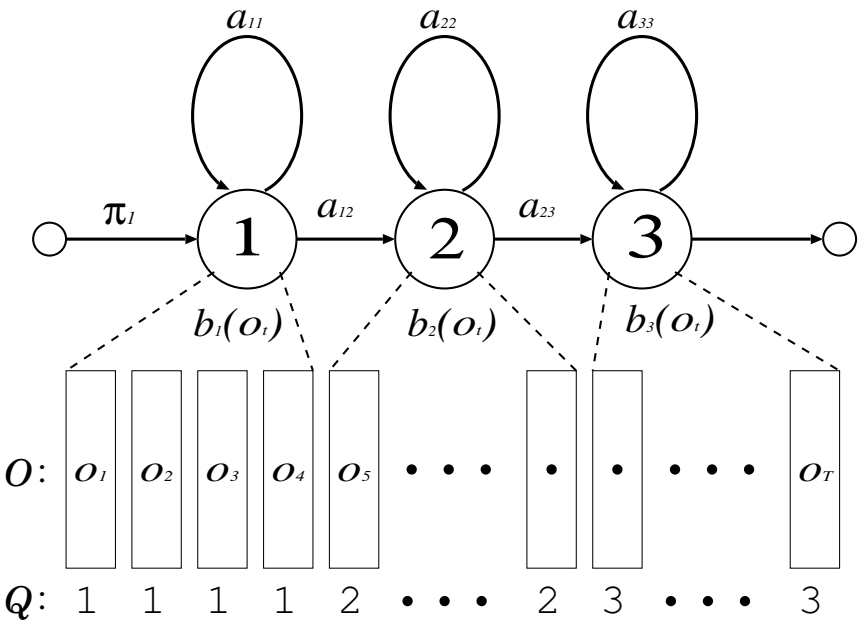


図 2.2: HMM の例

HMM は、図 2.2 に示すように、出力ベクトル o_t を出力する確率が $b_i(o_t)$ であるような

[1]cepstrum: スペクトラム (spectrum) をもじった造語。いったん時間窓を掛けて短時間フーリエ変換を行い、その絶対値 (振幅スペクトル) の対数をとる。これに逆フーリエ変換を行なったものをケプストラムといい、その次元は時間と同じになる。単位はケフレンシー (quefrency) で、これも frequency をもじったもの。スペクトルの微細構造 (基本周波数成分: 声帯に依存) はケフレンシの高いところにピークが現れ、スペクトル包絡 (声道、すなわち舌、顎や唇の位置や形などに依存) はケフレンシの低いところに集中する。

信号源が状態遷移確率 $a_{ij} = P(q_t = j|q_{t-1} = i)$ を持って接続されたものとして定義される。ただし、i, j は状態番号である。音声関連で利用される場合では、出力ベクトル o_t は、MFCC(Mel Frequency Cepstral Coefficients)、LPC ケプストラムなどの音声の短時間的なスペクトルを表現するパラメータとなる。HMM は時間方向とスペクトル方向の変動を統計的にモデル化しており、様々な要因で変動する音声パラメータ系列の表現として適しているといえる。出力確率分布としては、多次元ガウス分布の重み付き和で表される多次元ガウス混合分布が用いられることが多いが、ここでは、簡単のため、単一の多次元ガウス分布を仮定することにする。すると、出力確率分布 $b_i(o)$ は、

$$\begin{aligned} b_i(o) &= \mathcal{N}(o|\mu_i, U_i) \\ &= \frac{1}{\sqrt{(2\pi)^N |U_i|}} \exp\left\{\frac{1}{2}(o-\mu_i)' U_i^{-1}(o-\mu_i)\right\} \end{aligned} \tag{2.1}$$

のように表される。この場合、ガウス分布の平均ベクトル μ_i と共分散行列 U_i が、出力確率分布 $b_i(o)$ を特徴づけるパラメータとなる。

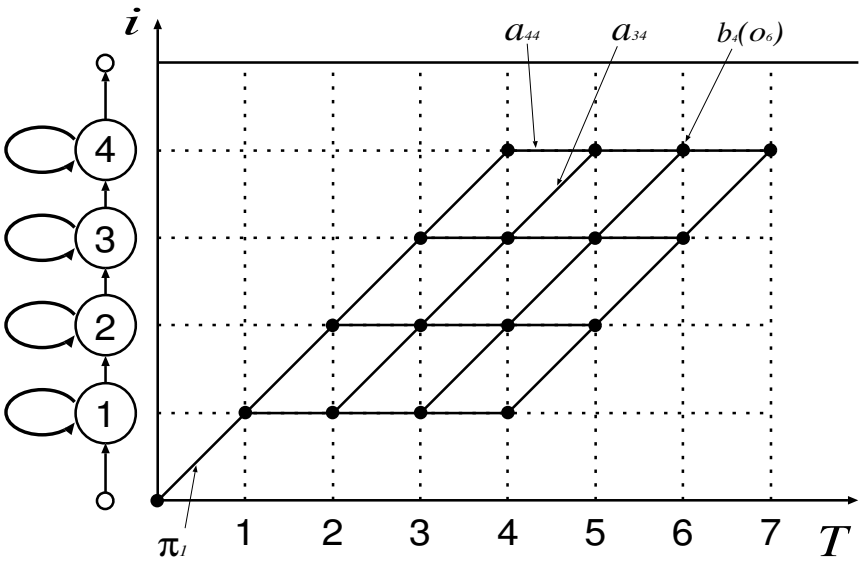


図 2.3: HMM の状態遷移

HMM の状態数を N とした時、HMM のパラメータ λ は、初期状態確率 $\pi = \{\pi_i\}_{i=1}^N$、状態遷移確率 $A = \{a_{ij}\}_{i,j=1}^N$、各状態 i での出力確率 $B = \{b_i(o)\}_{i=1}^N$ により、$\lambda = (A, B, \pi)$ と与えられる。この時、状態が $Q = \{q_1, q_2, \cdots, q_T\}$ と遷移して、出力ベクトル系列 $O = \{o_1, o_2, \cdots, o_T\}$ が出力される確率は、遷移確率と各状態での出力確率を掛け合わせることにより、

$$P(O, Q|\lambda) = \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \tag{2.2}$$

で与えられる。ただし、$a_{q_0 i} = \pi_i$ とおいている。

したがって、出力ベクトル系列 $O = \{o_1, o_2, \cdots, o_T\}$ が λ から出力される確率は、すべての可能な状態遷移の組合せについて和をとることになり、

$$\begin{aligned} P(O|\lambda) &= \sum_{\mathrm{all}_Q} P(O, Q|\lambda) \\ &= \sum_{\mathrm{all}_Q} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \end{aligned} \tag{2.3}$$

と書くことができる。式 (2.2) は、図 2.3 において、左下端のノードから、右下端のノードへ至る 1 本のパス上の確率をすべて掛け合わせたものである。一方、式 (2.3) は、可能なすべてのパスに対応する確率を加え合わせたものとなる。

2.3.2　決定木・回帰木

韻律制御に用いられる統計的手法には、重回帰分析・ニューラルネットワークなど様々なものがあるが、今回の実験では、決定木を使用した。決定木を用いた韻律生成は、他の統計的手法と比べても、同程度の結果を得ており [17]、また、構築されたモデルに関する解析が容易であるという利点を持つ。

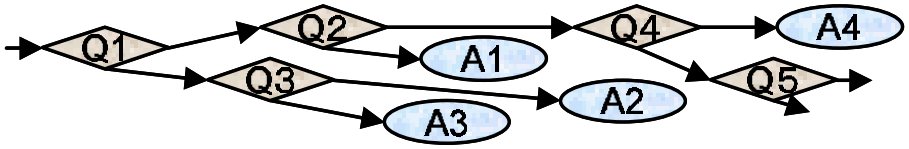


図 2.4: 決定木の概略図

決定木は、図 2.4 に概略を示すような統計モデルであり、各ノードに用意されている質問に答えて階層を進んでいくと、最終的に必ずひとつの答えにたどり着き、所望の値が得られるというものである。また最終的得られる答えが、離散的な値あるいはクラスといったようなものの場合のモデルを決定木、連続値を答えに持つような木を回帰木と呼ぶ。本稿ではこれら 2 つをまとめて決定木と呼ぶことにする。図 2.4 に示したのは、2 分木形態の場合の決定木である。

決定木は要因の組み合わせによってデータを逐次的に分割してゆくことによって作成される。学習は各ノード作成時点で最も効果的な要因による分割が選ばれるという最良探索法によるため、要因の組み合わせすべてによって作成可能な木からの全数探索はなされていない。結果として得られる木の全体的な最適性は必ずしも保証されないが、データ中に見られる要因による分布の偏りをもとに統計的に有為な範囲でモデルが得られるためデータ量に応じたモデル化が可能である。また、各要因による効果を線形回帰モデルのように一定値で表すことをしないため、要因間にまたがった効果を表現する上で自由度が高い。

決定木構築手法としては、CART(Classification And Regression Trees) [18] によるモデル化が挙げられる。今回用いる決定木もこの CART の手法を用いた The Edinburgh Speech Tools Library [19] の wagon を採用している。

決定木構築の際のリーフ数は、細かくすればするほど、学習データを良く再現するが、未学習データについての推定精度が悪くなってしまう。本稿では、最小リーフ数を 40 に設定して実験を行っている。

2.4　基本周波数パターン生成過程モデル

基本周波数は、F_0 とも表現され、声の高さに当たる情報である。ピッチアクセント型言語である日本語においては、韻律的特徴量のうちでも、基本周波数が特に重要となる。基本周波数は一般的に対数軸で取り扱われるが、声帯の特性として、ある線形の微少伸縮に対し、周波数が対数的に変化することが知られているからである。

本研究では、抽出した基本周波数パターンを分析するにあたって、下記の基本周波数パターン生成過程モデルを用いている [16]。本モデルは、少ないパラメータで F_0 パターンを良く近似するため、TTS システムに用いる上で利点が多い。

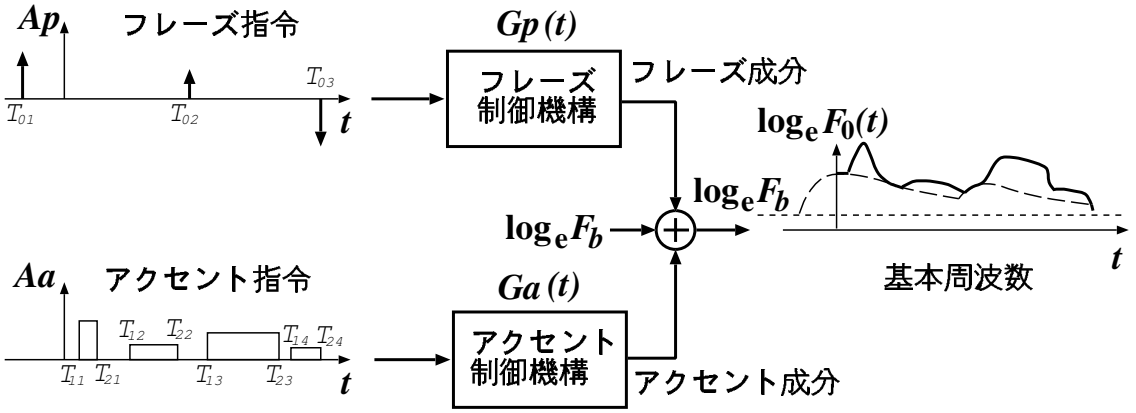


図 2.5: F_0 パターン生成過程モデル

図 2.5 に示すように、このモデルでは、比較的ゆっくりとした土台の起伏部分（フレーズ成分）と、比較的急速に上下する成分（アクセント成分）とに分けて考えている。式で表わすと、対数基本周波数の時間パターン $\ln F_0(\mathrm{t})$ は、

$$\log_e F_0(t) = \log_e F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\} \qquad (2.4)$$

と表される。右辺第 2 項がフレーズ成分、第 3 項がアクセント成分にあたり、A_{pi} と T_{0i} はそれぞれ i 番目のフレーズ指令（インパルス）の大きさと生起位置、A_{aj} と T_{1j} と T_{2j} はそれぞれ j 番目のアクセント指令（ステップ）の振幅と立上り位置、立下り位置である。また、G_{pi}、G_{aj} は

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t), & t \geq 0, \\ 0, & t < 0, \end{cases} \tag{2.5}$$

$$G_{\alpha j}(t) = \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma], & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{2.6}$$

と近似される。ここで α、β はそれぞれの制御機構の固有角周波数、γ はアクセント成分が有限時間内に一定値に達することを保証する相対飽和値である。α、β および γ の話者ごと、発話ごとの変動は比較的小さいため、初期値としては、それぞれ $\alpha = 3.0$rad/s、$\beta = 20.0$rad/s、$\gamma = 0.9$ を用いることができ、本研究ではこの値に固定してモデル化している。

2.5　言語情報処理に用いるツール

韻律コーパスの作成に先立ち、以下に利用したツールの概要を示す。

2.5.1　F_0 モデルパラメータ自動抽出ツール AXPF

F_0 モデルパラメータ自動抽出ツール AXPF[20] では、F_0 パターンファイルと音声ファイルから、F_0 モデルパラメータを自動抽出する。F_0 パターンファイルには、各フレームごとの F_0 の値の他、有声/無声音の表記もなされている。音声ファイルは、10kHz(または 16kHz)、16bit、little endian の RAW ファイルと、非圧縮、48kHz の WAV ファイルをサポートしている。表 2.1 に得られるファイル形式の例を示す。

14 行目の F_0MSE は、式 (5.1) に示す平均二乗誤差である。また、21 行目以降に生成過程モデルのパラメータが記されているが、(21+フレーズ指令総数 (8 行目の数値)−1) 行目までがフレーズ指令のパラメータで、(21+フレーズ指令総数) 行目から (21+フレーズ指令総数+アクセント指令総数 (9 行目の数値)−1) 行目までがアクセント指令のパラメータである。

2.5.2　日本語形態素解析ソフト茶筌・JUMAN と日本語構文解析システム KNP

日本語形態素解析システム茶筌 [21] は、漢字仮名混じりテキストが入力として与えられると辞書を参照してそれを形態素単位に分解し、その情報を出力する。茶筌の解析結果からは、それぞれの形態素に含まれる言語情報を得ることができ、出力形式は様々にカスタマイズすることができるが、本研究では、漢字仮名混じり表記、カタカナ表記、読み、品詞、活用型、活用形、アクセント型、アクセント結合様式・アクセント価に関して出力するように設定した。

表 2.1: PAC ファイルの形式

行番号	内容	意味
1	v04001	ファイル名
2	100	フレーム長
3	03.12.24	作成日
7	241	フレーム総数
8	2	フレーズ指令総数
9	4	アクセント指令総数
10	60.9090	F_b
11	0.010000	シフト長
14	0.001384 => 0.001123	AbS 前の F_0MSE => AbS 後の F_0MSE
21	-0.2458 2.3939 0.5055 3.0000	T_0 T_0' A_p α
22	0.5277 2.3939 0.5056 3.0000	
23	0.0625 0.2222 0.6622 20.0000	T_1 T_2 A_1 β
24	0.7712 1.4757 0.3689 20.0000	
25	1.6553 1.9512 0.1576 20.0000	
26	2.0765 2.1951 0.2130 20.0000	

一方、テキストを日本語形態素解析ソフト JUMAN[22] で分析すると、形態素とよばれる言葉の一部として意味をなす最小単位に分割する。日本語構文解析システム KNP[23] は、その結果を受けて、文節境界と文節間の係り受けの様子を推定し、その様子を表示する。

2.5.3　大語彙連続音声認識システム Julius

大語彙連続音声認識システム Julius[24] は、単語 N-gram に基づく高性能音声認識ソフトウェアであるが、本研究においては Julius の音素セグメンテーション機能を使って、アラインメント[2] を取るのに使用した。サンプリング周波数 16kHz の音声ファイル (ここでは WAV ファイルを用いた) と、発話内容テキストを、1 行につき 1 つの形態素のカタカナ表記の読みが記述された形にしたファイルを Julius に与えると、1 つ 1 つの音素がどの時刻からどの時刻までの間に発話されているのか、という情報を得ることができる。

[2]alignment: 音声と発話内容テキストとの対応付けを行い、最も適切な音素の発声時間の組を得ること。

2.6　韻律コーパス

2.6.1　韻律コーパスの自動生成

適切なパラメータを生成する統計モデルを構築するためには、十分な量の学習用コーパスを用意する必要がある。それらを人手によって行うのは非常に手間がかかるため、本研究では学習用コーパスを自動で作成し、韻律コーパスとした。

F_0 モデルパラメータの学習・推定の際には、韻律語[3] という単位で推定を行っているため、韻律コーパスでは、漢字仮名混じり文を韻律語単位に分ける必要がある。

韻律コーパスの自動作成処理は、図 2.6 のような流れで行う。

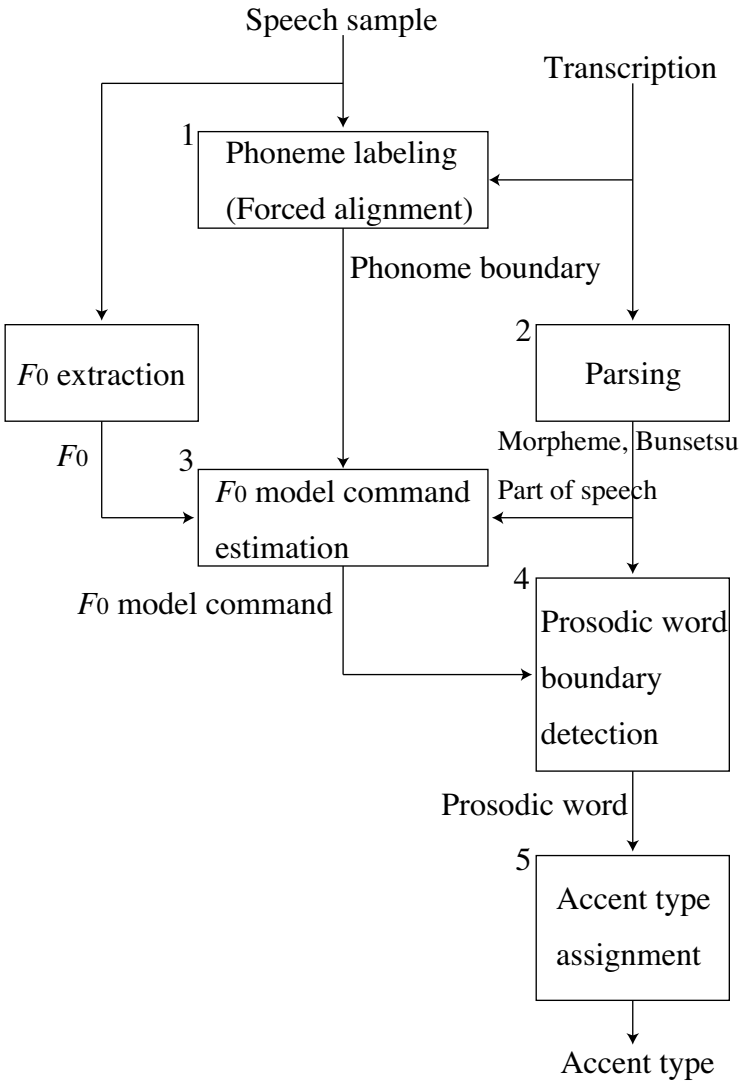


図 2.6: 韻律コーパスの自動作成

[3]ここでは、F_0 モデルパラメータにおけるアクセント指令を 1 つだけ持つ語と定義する。

以下に、具体的な処理の流れを示す。

1. 音声ファイルから基本周波数値を抽出し、F_0 モデルパラメータ自動分析システムにより F_0 モデルパラメータファイルを得る。
2. 音声ファイルと、漢字仮名交じり文から得た発音ファイルから Julius を利用して、音素アライメントファイルを得る。なお、漢字仮名交じり文から発音ファイルを得るには、茶筌の機能を利用した。
3. 漢字仮名交じり文から、茶筌を用いて形態素、品詞情報を得るとともに、JUMAN+KNP を用いて文節、統語情報を得て、言語情報ファイルとする。
4. 音素アライメントファイルと言語情報ファイルを参照して、実際の音声の時間情報と個々の形態素のマッチングを取る。
5. 言語情報ファイルと F_0 モデルパラメータファイルを参照し、韻律語境界の抽出規則に基づき、形態素系列を韻律語系列へと変換する。
6. アクセント結合規則 [25] に基づき、各韻律語のアクセント型を決定する。アクセント核をなす各モーラ母音開始時点とアクセント指令の立ち下がり位置との差分を T_{2off} として定義する。また、アクセント指令の生起タイミング T_{1off} を韻律語の先頭モーラの母音開始時点からの差分とする。
7. フレーズ指令はアクセント指令の間に 0 または 1 個存在すると仮定して検索を行い、存在した場合、時間的に後続する韻律語の先頭モーラの母音開始時点からの差分をフレーズ指令のタイミング T_{0off} とする。
8. 以上の作業を、用意したファイルすべてについて行い、韻律コーパスを作成する。

2.6.2　基底周波数の設定

F_0 パターンを作成するためには、フレーズ指令、アクセント指令の他に基底周波数 F_b を定めることが必要である。

F_b もアクセント指令、フレーズ指令と同様、統計モデルによる推定の対象とすることも考えられる。しかし、F_b そのものを F_0 パターンから自動的に推定することは難しい。また、F_b は感情ごとに変動はあるが、同一話者、同一感情内では変動が少ないと考えられるため、各感情で一定であると仮定した。

本研究の枠組みでは韻律コーパスを自動的に作成することを目指しているため、抽出された F_0 パターンから自動的に F_b を算出できる枠組みが必要となってくる。F_b は次のような求め方が考えられる。

- すべての抽出された F_0 の最低値
- 「各文の F_0 の最低値」の平均値
- すべての抽出された F_0 の平均値から 3σ を引いたもの

今回は F_0 の抽出結果に手を加えておらず、扱う音声が感情音声であることも手伝って、F_0 の抽出エラー、主に整数倍ピッチエラーなどが朗読の場合に比べ、多く起きていることが考えられるため、その F_0 の中での最低値をそのまま使用するのは危険であると考えられ

る。したがって各感情において、抽出された F_0 すべての平均値から 3σ（ただし σ は標準偏差）を引いたものを F_b として採用した。

2.6.3　言語的制約を加えた F_0 モデルパラメータの自動抽出

学習・推定において使用するコーパスの作成は自動で行っており、推定の対象となる F_0 モデルパラメータは、音声から自動抽出したものを用いている。

音声から F_0 モデルパラメータを得ることは逆問題であり、得られるパラメータは一義的な解ではない。合成の際に、テキストから F_0 モデルパラメータを推定することを考えると、言語情報との相関の高い解を、学習の対象とする方が、より推定精度が向上するはずである。

本研究では、フレーズ指令に関して言語的な制約をかけた F_0 モデルパラメータ自動抽出ツールを用いる。具体的な条件としては、フレーズ指令は必ず文節境界の前に存在するとし、

1. 文節境界の直前にポーズがある場合は、文節境界から-300〜-100msec. の区間
2. 文節境界の直前にポーズがない場合は、文節境界から-100〜0msec. の区間

のように、挿入位置に対して制約が設けられている。

2.7　まとめ

本章では、TTS システムの概観について述べ、TTS システムの各処理の中で使用する統計モデルについて述べた。また、F_0 パターン作成時に、生成するパラメータの自由度を減らすために用いる F_0 モデルについても、その原理を示した。加えて、統計モデル学習の際に必要となる韻律コーパスの自動生成と、その際用いるツールについて説明した。

第3章

F_0パターンの推定

3.1　はじめに

本研究では、F_0 パターンの推定に、F_0 モデルを用いる。既に、先行研究 [14] [15] において、F_0 モデルパラメータ推定のための統計モデルが提案されており、本研究ではそれを利用して F_0 パターン作成の枠組みを構築する。

本章では、まず実際にテキストから F_0 パターンを作る枠組みとして構築される F_0 モデルパラメータ推定の統計モデルについて説明する。次に F_0 モデルパラメータの推定として、フレーズ指令推定・韻律語境界推定・アクセント指令推定について順に説明する。

3.2　推定するパラメータ

F_0 パターンを作成するために、推定する必要がある F_0 モデルパラメータは表 3.1 の通りである。

表 3.1: F_0 モデルパラメータ推定の出力項目

出力項目	カテゴリ数
先頭のフレーズ指令有無 PF	2 値
フレーズ指令の大きさ A_p	連続値
フレーズ指令のタイミング T_{0off}	連続値
アクセント指令の大きさ A_a	連続値
アクセント指令の生起タイミング T_{1off}	連続値
アクセント指令の終了タイミング T_{2off}	連続値

表 3.1 の各項目の詳細は以下の通りである．PF は当該韻律語の先頭にフレーズ指令が存在するかどうかのフラグであり，PF=1 ならば，韻律語の先頭モーラの母音開始時刻から距離 T_{0off} の位置に大きさ Ap のフレーズ指令が立つとする．PF=0 ならば，フレーズ指令は存在しないとする．アクセント指令は，上述のように，韻律語内には必ず 1 つ存在するとするので，存在の有無を示すフラグは必要なく，韻律語の先頭モーラの母音の開始時刻から T_{1off} の位置に生起し，アクセント核モーラの終点の位置を基準に T_{2off} の位置で終了するとし，その大きさを A_a とする．以上をまとめたものが図 3.1 である．

なお、アクセント核位置、韻律語境界については、F_0 モデルパラメータの推定の前に決定しておく必要がある。それらについては、3.4、3.5 節で述べる。

3.3　F_0 パターン推定の枠組み

F_0 パターンの推定および作成手順について図 3.2 にまとめる。図 3.2 では、推定したフレーズ指令を韻律語境界の推定に利用することを考え、フレーズ指令の推定を一番始めに行っている。

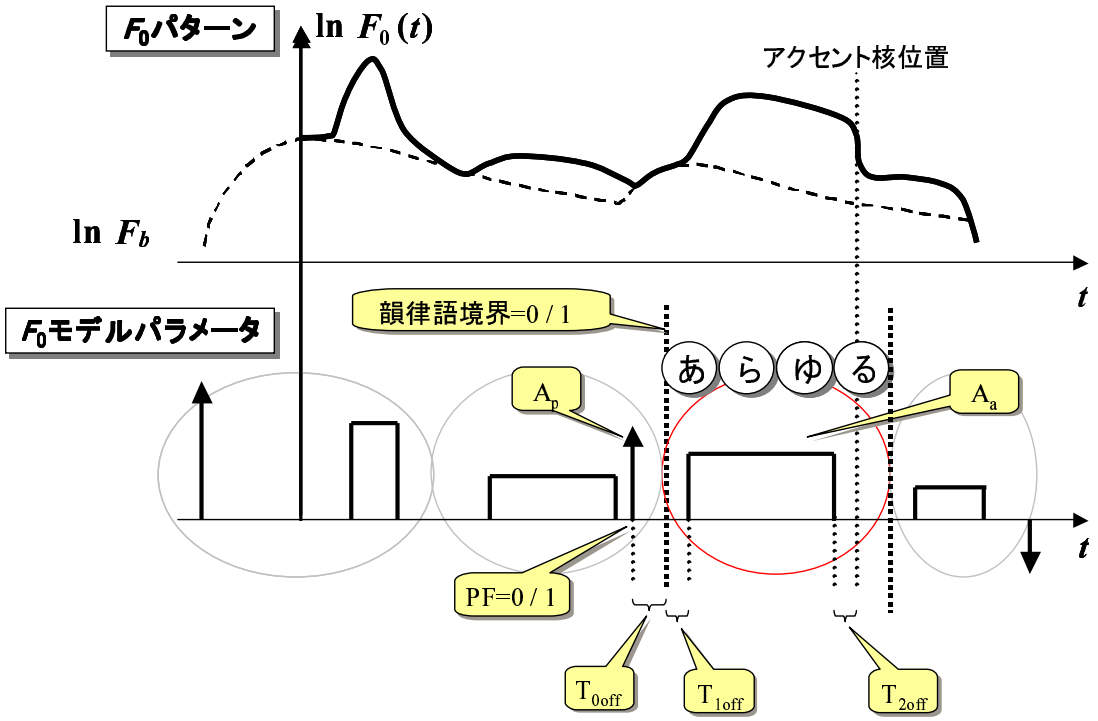


図 3.1: 推定の対象となるパラメータの概略

まず、入力された漢字仮名混じりテキストに対し形態素解析を行い、同時に、構文解析などのテキスト解析により、他の言語情報も得る。次に、本手法では推定したフレーズ指令を韻律語境界の推定に利用するため、フレーズ指令の推定を行なう。その後、フレーズ指令の大きさや有無などの情報も入力項目に加え、決定木を用いた学習によって韻律語境界を推定する。さらに韻律語境界が決定されたら、得られた韻律語系列にアクセント型辞書と結合規則を用いてアクセント型を付与する。最後に、決定木を用いた学習によってアクセント指令を推定し、得られた生成過程モデルパラメータから F_0 パターンを生成する。

漢字仮名混じりテキストを形態素列に変換することは、一般的な形態素解析ツールによって行う。形態素系列から韻律語系列への変換、および韻律語系列への生成過程モデルパラメータの付与の際の「規則」は、人間の韻律制御方法を模したものであるが、本手法では、これらの規則を統計を利用したモデル (以下、統計モデル) によって実現する。つまり、韻律コーパスを用いた学習によりあらかじめ統計モデルを構築し、音声の合成時にはそれらの統計モデルを規則として用いる。

また、それらの統計モデルの入力には、品詞の種類やモーラ数など、漢字仮名混じりテキストから自動的に得られる情報を用いる。

2.6.3 節で述べた自動推定の制約から、このコーパスではフレーズ指令は文節境界の前にしか存在しない。そのため、フレーズ指令は韻律語単位で推定するのではなく、文節単位で推定することを提案する。この手法の利点としては

- フレーズ指令挿入位置は文節境界の直前に限られるので、フレーズ指令の時間的推定精度が上がる。
- 韻律語境界推定エラーの影響を受けない。
- 韻律語境界の推定に、推定したフレーズ指令の情報を利用できる。

などが考えられる。

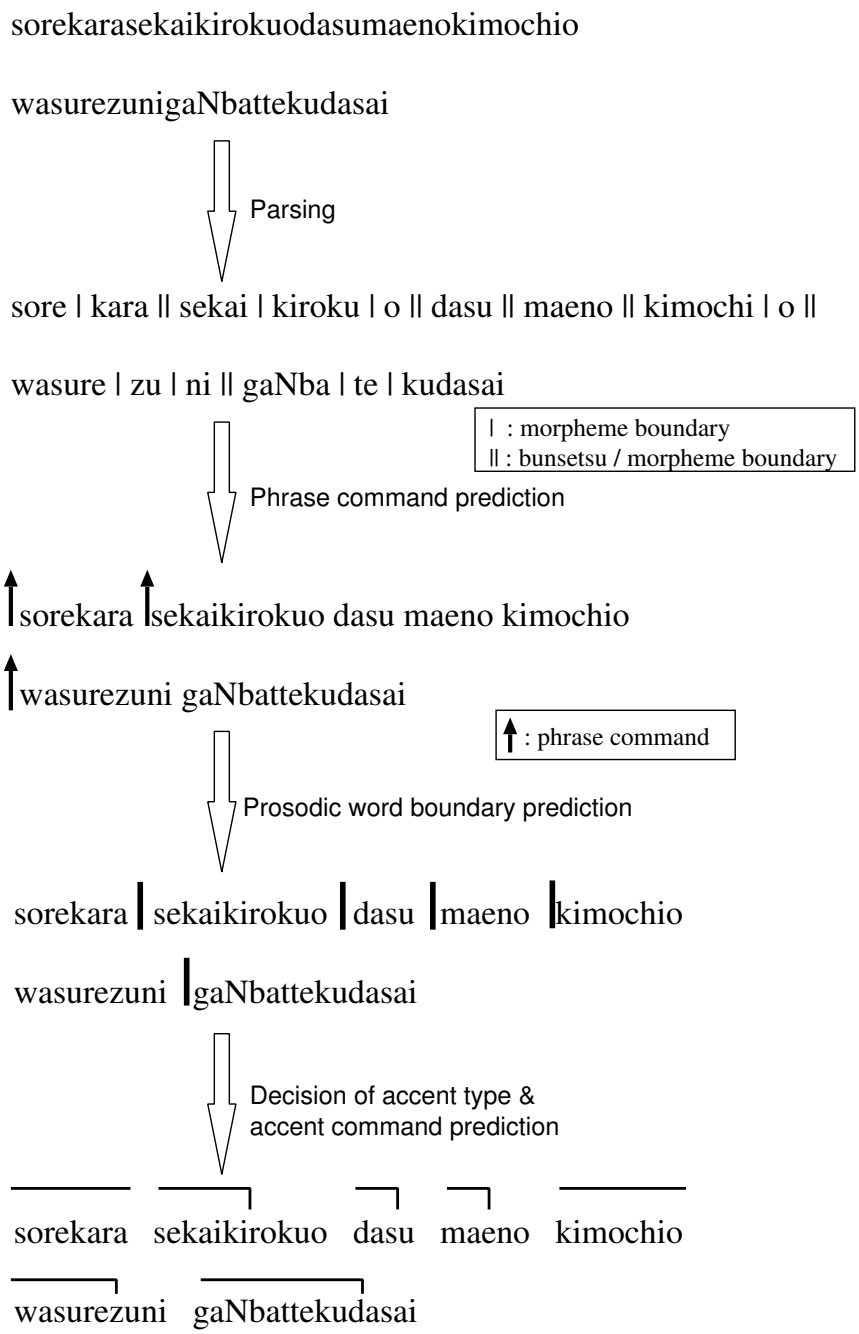


図 3.2: コーパスベース韻律生成の枠組み

3.4　韻律語境界の推定

F_0 モデルパラメータの学習・推定の単位は韻律語である。韻律語とは、アクセント成分を一つ含む日本語の発声単位である。韻律語は文字言語上ではおおむね、文節として表わされる。しかし、韻律語は、音声言語特有の韻律現象であるため、話者ごと、あるいは同じ話者でも発話ごと、感情ごとに異なる場合がある。

TTS システムを考える際には、言語情報から韻律語を推定しなくてはならない。本実験では韻律語境界を統計的手法によって推定するという方法をとった。統計モデルは計算によって自身の説明に有力な入力要因を取り込んで自動的に学習されるが、そのためには、あらかじめ必要と予想される要因を準備し、モデル化されやすいように適切にコード化しておくことが重要である。また、用いる情報は全て漢字仮名交じりテキストから自動的に得られるものでなければならない。

先行研究である朗読音声の場合を参考にして、表 3.2 に示す項目を韻律語推定モデルの入力項目とした。構築に用いる言語情報の単位は形態素である。また、カテゴリ数というのは、用いた韻律コーパス中に出現した該当要因の取り得た範囲内の値の種類数である。当該形態素に比べ、同種の項目において先行形態素でカテゴリ数が 1 つずつ多くなっているのは、当該形態素が文の先頭である場合、先行形態素というものが存在しないため、そのことを表すダミーのカテゴリを追加した分である。

表 3.2: 韻律語推定の入力項目

形態素の情報	カテゴリ数
当該 (先行) 形態素の品詞	15(16)
当該 (先行) 形態素の活用形	24(25)
当該 (先行) 形態素の活用型	35(36)
当該 (先行) 形態素のモーラ数	9(10)
当該形態素の文内位置	63
当該形態素の属する文節の境界コード	22
当該形態素の先頭の文節境界の有無	2 値

なお、境界コードは、統語構造の情報を表す入力項目である。用いる統語構造情報は、漢字仮名交じりテキストから自動的に得られるものである必要があるが、これには日本語構文解析システム KNP[23] を用いた。KNP は、漢字仮名交じりテキストを文節に区切るとともに、それぞれの文節がどの文節に係っているかという情報を出力する。そこでまず、各文節について係り先文節への距離 (単位：文節) を求めて、後続文節との文節境界の深さとする。そして、実際に統計モデルの学習で用いる韻律語の先頭の境界コードとして、そこに存在する文節境界の深さを割り当てる。例を図 3.3 に示す。

図 3.3 中の境界コードのうち、F は文節境界が定義されない、文章の先頭であることを示す。また、1 は境界が左枝分かれ境界であることを示し、2 以上の数字は右枝分かれ境界とその深さに相当するものを表していることになる。この例ではたまたま文節区切りと韻

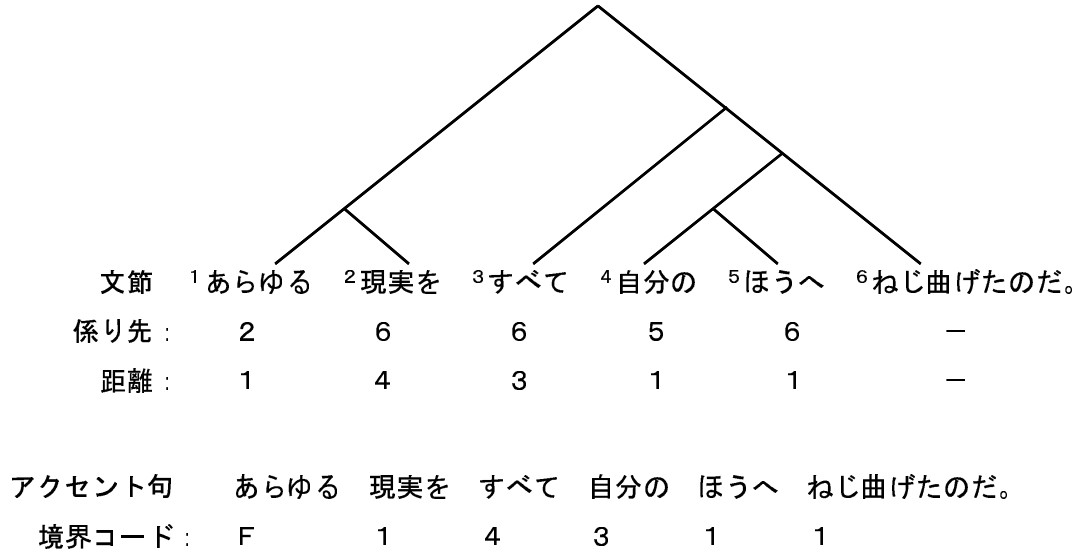


図 3.3: 境界コード

律語区切りが一致しているが、韻律語境界に文節境界が存在しない場合は先頭の形態素が属する文節の境界コードを割り当てる。

3.5　アクセント核の推定

生成過程モデルパラメータの推定において、アクセント指令の立ち下がり位置 T_{2off} は、アクセント核モーラの終点の位置とアクセント指令の立ち下がりの位置の差分として定義している。韻律語の終了位置からの差分とすることもできるが、それでは、時間的な分散が非常に大きくなり、T_{2off} の推定精度が著しく悪化してしまうからである。

日本語では、個々の語句に対しアクセント型が定義されており、そのアクセント型に従って発話される。アクセント核位置では、F_0 パターンの降下が見られるため、その周辺にアクセント指令の立ち下がりが存在するはずである。これを利用して、アクセント指令の立ち下がり位置 T_{2off} をアクセント核位置からの差分とした。

本研究での韻律語の定義から、1 つの韻律語の中には、アクセント指令は 1 つしか存在しない。そのため、韻律語に対して、アクセント型が正しく指定できれば、アクセント指令の時間的な推定精度を大きく上げることができる。しかし、韻律語は、複数の語句から構成される場合がほとんどであり、それに対して、1 つのアクセント型を定める規則が必要である。

先行研究 [25] では、複数の語句が結合した場合に起こるアクセント型の変化を記述している。本研究では、このアクセント結合規則を利用し、推定された韻律語に対して、アクセント核モーラ位置を決定した。

3.6　F_0 モデルパラメータの推定

3.6.1　共通項目

表 3.3 中の無印の項目群は、漢字仮名交じり文から得られる言語的情報のみから構成される入力項目である。おおまかにこれらの入力パラメータが導入された背景を説明する。

まず言語情報としてはじめに考えられるのが、扱う対象が韻律語であるので、当該韻律語に関する情報であった。しかしそれだけでは各パラメータを推定するには精度が悪いことが指摘されていた。

そこで、韻律現象はひとつの韻律語より広い範囲に関連して起こることが想像されるので、当該韻律語の情報のほか、ひとつ前の韻律語の情報を追加する。

次に統語情報を考慮することが考えられる。テキストの統語構造は F_0 パターンの制御に大きく関与している。例えば、先行句が隣接する後続句に係る境界 (左枝分かれ境界) では F_0 パターンの上昇は見られず、隣接する後続句に係らない境界 (右枝分かれ境界) では上昇が見られる。後者に関して、F_0 モデルにおいてはフレーズ成分の生起によってこれが実現されているといえる。統計モデルの入力として上記のようなテキストの統語構造を考慮に入れ、フレーズ指令に関連するパラメータの推定精度を向上させることを目的としている。

表 3.3: F_0 モデルパラメータ推定の入力項目

出力項目	カテゴリ数
当該句の文内位置	27
当該 (先行句) の有するモーラ数	17(18)
当該 (先行句) のアクセント型	16(17)
当該 (先行句) の有する単語数	9(10)
当該 (先行句) の最初の単語の品詞	14(15)
当該 (先行句) の最初の単語の活用形	23(24)
当該 (先行句) の最後の単語の品詞	14(15)
当該 (先行句) の最後の単語の活用形	23(24)
先頭の境界コード	22
当該句の PF	2 値 *
当該句の A_p	連続値 *
当該句の T_{0off}	連続値 *

*:二段階推定（A_a,T_{1off},T_{2off} のみ）

無印の項目群は、統計モデルの入力項目として言語情報のみをとり、それぞれの F_0 モデルパラメータを個別に一括して推定するものである。これに対し、推定を 2 段階に分けて、第 1 のステップで推定した項目を今度は入力項目として用いて第 2 のステップの推定を行う方法が考えられる。(* の項目) もし最終的な出力項目の間に相関があるならば、当然それらのうちのいずれかが他の項目の有力な説明変数となるということであり、第 2 ステッ

プの統計モデルの入力項目として採用することでその性能を向上させ、全体として推定精度を上げられる可能性がある。

自然発声においてはフレーズ成分が大きい時点ではアクセント成分はそれほど大きくはできないなどといった人間の発声機構の制約に由来する相関が考えられる。特に、怒りや喜びなどの感情音声においては基底周波数、フレーズ指令、アクセント指令などが朗読音声に比べて大きくなるため、より強い発声機構の制約を受けるなどといったようなことが考えられる。このことは F_0 モデルにおいてはフレーズ指令とアクセント指令のパラメータ間の相関として現れることが考えられる。以上を踏まえ、まずフレーズ指令についてのパラメータを推定したあとで、改めてそのパラメータをコード化したものを入力項目に加え、アクセント指令関連パラメータを推定する。

3.6.2　フレーズ指令推定のための入力項目

フレーズ指令推定のための入力項目・カテゴリ数を表 3.4 に示す。2.5 で紹介したツールによって、漢字仮名混じり文のみから自動的に抽出される情報に加え、先行フレーズ指令の考慮や文節境界におけるポーズの有無の考慮などを行なっている。

表 3.4: フレーズ指令推定の入力項目

出力項目	カテゴリ数
当該句の文内位置	28
当該 (先行句) の有するモーラ数	21(22)
当該 (先行句) のアクセント型	18(19)
当該 (先行句) の有する単語数	10(11)
当該 (先行句) の最初の単語の品詞	14(15)
当該 (先行句) の最初の単語の活用形	19(20)
当該 (先行句) の最後の単語の品詞	14(15)
当該 (先行句) の最後の単語の活用形	16(17)
先頭の境界コード	20
先行文節のフレーズ指令の有無	2 値
当該文節開始点から先行フレーズ指令までのモーラ数	25
先行フレーズ指令の大きさ	連続値
当該文節境界直前のポーズの有無	2 値

3.6.3　韻律語境界推定のための入力項目

韻律語境界の推定のための入力項目として、漢字仮名混じり文のみから自動的に抽出される情報で構成した入力項目に加え、フレーズ指令に関して考慮した情報を追加した。

表 3.5: 韻律語境界推定の入力項目

形態素の情報	カテゴリ数
当該 (先行) 形態素の品詞	15(16)
当該 (先行) 形態素の活用形	24(25)
当該 (先行) 形態素の活用型	35(36)
当該 (先行) 形態素のモーラ数	9(10)
当該形態素の文内位置	63
当該形態素の属する文節の境界コード	22
当該形態素の先頭の文節境界の有無	2 値
当該形態素のフレーズ指令の有無	2 値
当該形態素開始点から当該フレーズ指令までのモーラ数	31
当該フレーズ指令の大きさ	連続値

3.6.4　アクセント指令推定のための入力項目

アクセント指令の推定のための入力項目として、漢字仮名混じり文のみから自動的に抽出される情報で構成した入力項目に加え、フレーズ指令と先行アクセント指令の情報を考慮した。

表 3.6: アクセント指令推定の入力項目

出力項目	カテゴリ数
当該句の文内位置	27
当該 (先行句) の有するモーラ数	17(18)
当該 (先行句) のアクセント型	16(17)
当該 (先行句) の有する単語数	9(10)
当該 (先行句) の最初の単語の品詞	14(15)
当該 (先行句) の最初の単語の活用形	23(24)
当該 (先行句) の最後の単語の品詞	14(15)
当該 (先行句) の最後の単語の活用形	23(24)
先頭の境界コード	22
当該韻律語開始点から当該フレーズ指令までのモーラ数	23
当該フレーズ指令の大きさ	連続値
先行アクセント指令の大きさ	連続値
先行アクセント指令の持続モーラ数	13
当該韻律語開始点から先行アクセント指令の立ち下がりまでのモーラ数	17

3.7　まとめ

本章では、テキストから F_0 パターンを推定するために構築した枠組について説明した。推定はフレーズ指令推定、韻律語境界推定、アクセント核推定、アクセント指令推定の順に行い、F_0 モデルパラメータの推定は当該句の文内位置やモーラ数、当該 (先行)F_0 モデルパラメータの有無や大きさなどを入力に用いる。

第4章

感情の種類がF_0モデルパラメータに与える影響

4.1　はじめに

感情音声では、朗読音声に比べ、韻律の果たす役割が非常に大きく、その制御は大きな課題である。また、一口に感情と言ってもその表現は実に多様であり、話者や状況によっても表現方法や観測される音声情報も非常に多岐にわたるため、その分析や合成には困難が伴う。

本章では、先行研究の手法に一部変更を加えて利用し、「平静・怒り・喜び・悲しみ」の 4 感情の F_0 モデルパラメータの推定を行い、その際に作成された学習木を用いた 2 種類の実験を行う。この実験により、感情の種類と韻律的特徴量の関係を調べる。

4.2　F_0 パターンの作成

本章では、前章で説明した先行研究 [15] によって示されている手法を用い、4.3.2 節で説明する音声試料に対し、F_0 モデルパラメータの推定と F_0 パターンの作成実験を行なった。

各パラメータの推定結果を、他パラメータ推定の入力項目として扱うため、作成は

1. フレーズ指令の推定
2. 韻律語境界の推定
3. アクセント指令の推定

の順に行う。本枠組みの精度を測るため、音素の時系列は正解のデータ (音声より自動抽出したもの) を与えた。従って、入力項目は漢字仮名混じり文と音素時系列、出力項目は F_0 パターンである。

4.3　全感情音声を入力に用いた学習木の作成

4.3.1　実験の目的

先行研究の F_0 パターン推定実験では、「平静・怒り・喜び・悲しみ」の 4 感情について各感情ごと別々に推定を行っていたが、本実験ではこの枠組みを変更し、4 感情分の音声試料を同時に使用して推定木を作成する。この場合、入力に「感情の種類 (4 値)」という項目を新たに追加する。この実験で作成された推定木を分析する事によって、感情音声の F_0 モデルパラメータ推定に感情がどのような影響を与えているのかが調べられる。

従来の手法と今回行った実験の模式図を図 4.1 と図 4.2 にそれぞれ示す。

4.3.2　用意した感情音声試料

本実験で使用される音声試料は次の通りである。

読み上げられた文としては、朗読音声に関しては ATR 音素バランス文セット 503 文を使用し、感情音声に関してはそれぞれの感情を込めやすいように、Web の掲示板から、怒

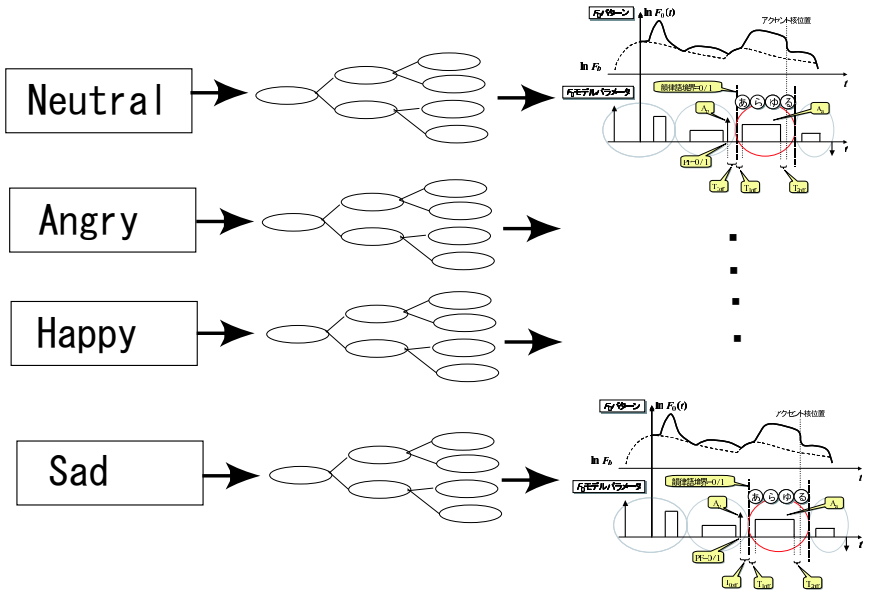


図 4.1: 従来手法における各感情ごとの学習木作成

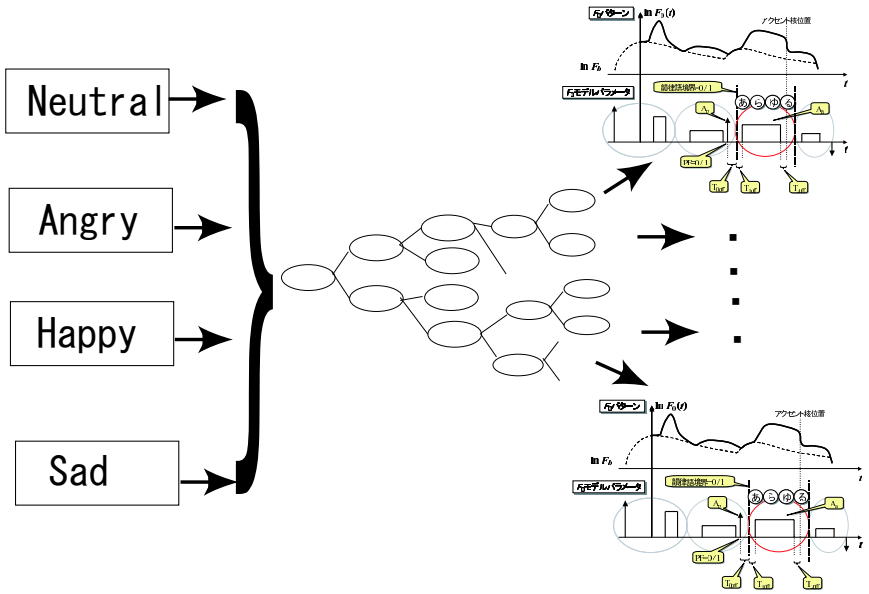


図 4.2: 全感情を入力に用いた学習木の作成

り、喜び、悲しみの体験が書き込まれた文を収集したものを用いた。朗読音声として収録された音声を本研究では平静音声として扱う。以下、平静音声という場合にはこの朗読音声を指すものとする。

用意された文の例を表 4.1 に示す。

発話者はプロの女性ナレーター 1 名で、原則として 1 日に 1 感情が収録されている。収

表 4.1: 用意した各感情音声の文の例

平静	あらゆる現実をすべて自分の方へねじ曲げたのだ。
怒り	ずうずうしいにも程がある。
喜び	大きな勇気と大きな感動をどうもありがとうございました。
悲しみ	私は夫を病気で亡くして一年になります。

録環境は遮音室で行われ、量子化ビットは 16bit、サンプリング周波数は 10kHz である。

用意された文の数を表 4.2 に示す。

表 4.2: 用意した各感情音声の文の数

感情	平静	怒り	喜び	悲しみ
文数	503	688	523	587

4.3.3　結果

作成された推定木のうち、図 4.3 から図 4.5 に、アクセント指令 (A_a、T_1、T_2) の推定木を示す。まず、推定の精度については、従来の手法と比較して本実験による明確な変化はなかった。図 4.3 はアクセント指令の大きさを示すパラメータ A_a の推定木であるが、これを見ると、一番上のノードにある質問項目に「感情が Sad か否か」が現れている。これは「感情の種類」という入力項目が、アクセント指令の大きさを決定づける上で最も重要項目であるということを示している。同様に、アクセント指令の立ち上がりのタイミングを示す T_1 の推定木においては「フレーズ指令からのモーラ数が 1 か否か」が、アクセント指令の立ち下がりのタイミングを示す T_2 の推定においては「アクセント型が 0 型か否か」が、値を決定する最も大きな要素であることを示している。

図 4.3 からわかるように、大きさ成分のパラメータ A_a では感情のタイプについての分岐が一段目に現れているのに対し、図 4.4,4.5 に示されたタイミングパラメータ T_1, T_2 に関してはそうではないという結果になった。

この結果から、感情の違いは大きさ成分により顕著に現れることがわかった。

4.4　同じ文で生成された指令パラメータの比較

4.4.1　実験の目的

本節ではもうひとつの実験として、同じ文で生成された指令パラメータの比較を行う。ここで行う実験は、平静文での F_0 パラメータ推定が良い結果を示した文を、怒り・喜び・悲しみの各感情ごとに作成された推定木を用いて F_0 パラメータを推定すると、その結果

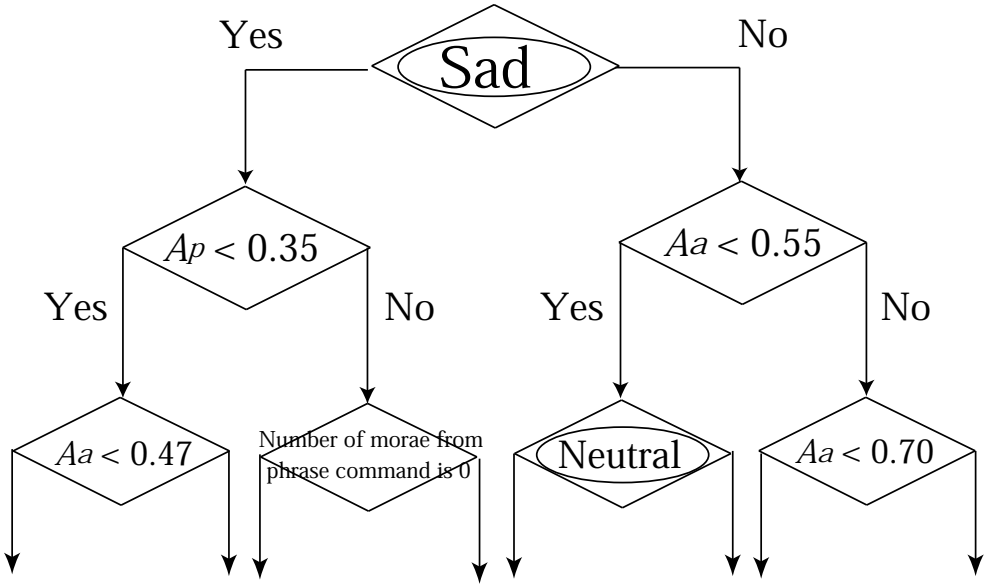


図 4.3: パラメータ A_a の推定木

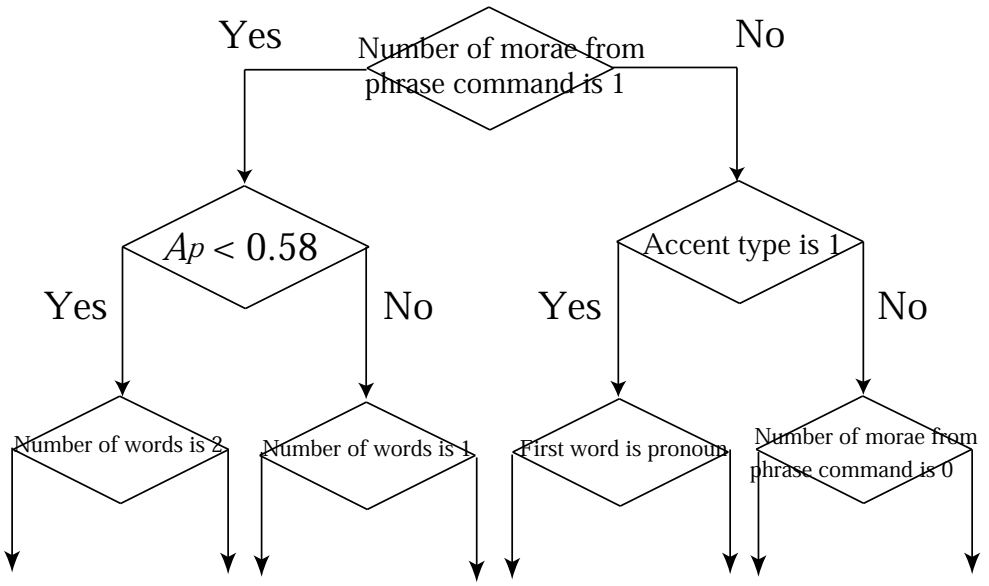


図 4.4: パラメータ T_1 の推定木

は平静文の推定木によって得られた結果とどのような差異があるかを調べるというものである。

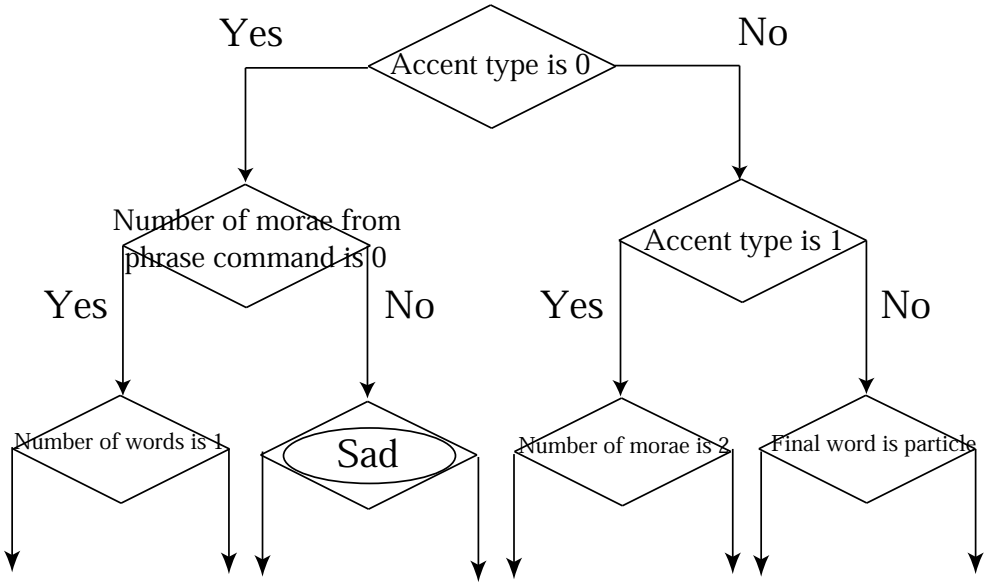


図 4.5: パラメータ T_2 の推定木

4.4.2　実験条件

まず、平静文のうちで推定結果がよかったもの（F_0MSE の値が低かったもの）を 503 文の中から 28 文選び出した。その 28 文について、「怒り」の推定木、「喜び」の推定木、「悲しみ」の推定木を用いて、F_0 モデルパラメータを推定した。各フレーズ指令、各アクセント指令について、平静文の対応するものとパラメータがどの程度違っているか、各パラメータごとに、平均と標準偏差を求めた。ここでの「対応する」とは、文中での同じモーラ位置にある指令ということである。

この際、F_0 モデルパラメータのうち、フレーズ指令に関しては、文の先頭のものと、それ以外のものを区別して平均と標準偏差を求めた。アクセント指令については、フレーズの先頭のものと、それ以外のものを区別し、同じく平均と標準偏差を求めた。

ここで、文あるいはフレーズの先頭の語とそれ以外の語を分けて考えているのは、平静・感情音声では先頭の語とそれ以外の語では強さ/大きさ成分において違いがあることが先行研究 [27] によって示されているからである。

4.4.3　結果

表 4.3 に、「平静・怒り・喜び・悲しみ」の 4 感情の推定木を用いて推定された、28 文の F_0 の平均値を示す。

「怒り・喜び」に関しては基底周波数 (F_b) が高いため、F_0 の平均値も平静音声の結果に比べ高くなっていることが分かる。

表 4.4 に、文の先頭のフレーズ指令 (A_p,T_0) に関しての結果と、フレーズ指令の先頭のアクセント指令 (A_a,T_1,T_2) に関しての結果を示す。各項目の一段目の数値は平静文を「平静・

表 4.3: 各感情音声の推定木による結果の F_0 平均値

平静	怒り	喜び	悲しみ
243.6 Hz	262.7 Hz	293.5 Hz	240.1 Hz

怒り・喜び・悲しみ」の 4 感情のそれぞれの推定木を用いることによって得られた値の平均値を示し、二段目の () でくくられた数値はその標準偏差を示す。三段目の数値は「平静」の推定木で得られた値と「怒り・喜び・悲しみ」の推定木で得られた値の差分を各 F_0 パラメータごとに求め、その値を平均したものであり、四段目はその標準偏差である。

表 4.5 は、文の先頭以外のフレーズ指令 (A_p,T_0) に関しての結果と、フレーズ指令の先頭以外のアクセント指令 (A_a,T_1,T_2) に関しての結果である。一段目から四段目までの各数値の意味は、表 4.4 と同様である。

数値の正/負は、強さ・大きさ指令 (A_p, A_a) については感情音声の推定木による推定結果が平静音声の推定木による推定結果に比べ大きい/小さいことを示し、タイミング指令 (T_0, T_1, T_2) については同じく、数値の正/負は感情音声の推定木による推定結果が平静音声の推定木による推定結果に比べ遅い/早いものとなっていることを示す。

表 4.4: 文の先頭/フレーズ指令の先頭についての平均値と標準偏差

	平静	怒り	喜び	悲しみ
A_p	0.463	0.284	0.272	0.203
	(0.054)	(0.033)	(0.052)	(0.029)
	-	-0.18	-0.191	-0.260
	-	(0.057)	(0.060)	(0.057)
T_0	-0.165	-0.169	-0.180	-0.176
	(0.020)	(0.019)	(0.018)	(0.021)
	-	-0.004	-0.015	-0.011
	-	(0.018)	(0.017)	(0.017)
A_a	0.419	0.393	0.405	0.269
	(0.071)	(0.061)	(0.045)	(0.039)
	-	-0.039	-0.007	-0.151
	-	(0.081)	(0.078)	(0.059)
T_1	-0.021	-0.014	0.006	0.000
	(0.052)	(0.060)	(0.090)	(0.074)
	-	0.010	0.019	0.018
	-	(0.048)	(0.097)	(0.085)
T_2	0.050	-0.006	0.049	0.002
	(0.160)	(0.106)	(0.138)	(0.161)
	-	-0.062	-0.007	-0.052
	-	(0.076)	(0.086)	(0.109)

表 4.5: 文の先頭以外/フレーズ指令の先頭以外についての平均値と標準偏差

	Calm	Anger	Joy	Sadness
A_p	0.184	0.167	0.128	0.145
	(0.081)	(0.096)	(0.049)	(0.085)
	-	-0.020	-0.047	-0.042
	-	(0.096)	(0.090)	(0.088)
T_0	-0.059	-0.085	-0.046	-0.095
	(0.039)	(0.057)	(0.035)	(0.070)
	-	-0.017	-0.002	-0.029
	-	(0.031)	(0.020)	(0.037)
A_a	0.405	0.356	0.423	0.266
	(0.082)	(0.056)	(0.055)	(0.041)
	-	-0.045	0.024	-0.137
	-	(0.061)	(0.114)	(0.082)
T_1	-0.040	-0.042	-0.020	-0.024
	(0.063)	(0.038)	(0.068)	(0.088)
	-	-0.002	0.020	0.019
	-	(0.051)	(0.057)	(0.081)
T_2	0.018	-0.013	0.002	-0.038
	(0.146)	(0.104)	(0.138)	(0.175)
	-	-0.033	-0.016	-0.057
	-	(0.072)	(0.084)	(0.093)

4.4.4　考察

表 4.4、表 4.5 の結果を見ると、タイミング指令に関しては特に目立った傾向が見受けられないものの、強さ・大きさ指令成分に関しては次の 2 点の特徴を見ることができる。まず一つめに、「悲しみ」の推定木による結果が、最も強さ・大きさ指令が小さくなるという事、二つめに、平静音声の推定木に関する結果では特に、文およびフレーズ成分の先頭での指令が、先頭以外での指令に比べ、大きくなっている事、である。2 点目に関しては、感情音声の推定木による結果でも同様の傾向は見られるが、その差は平静に比べ、小さい。この結果は自然発声の分析の結果 [27] とも一致する。

4.5　まとめ

本章では、先行研究の手法に一部変更を加え、「平静・怒り・喜び・悲しみ」の 4 感情の F_0 モデルパラメータの推定を従来手法により行い、その際に作成された学習木を利用して「感情」が F_0 モデルパラメータ推定に与える影響について 2 種類の実験を行った。アクセント指令の大きさ成分を決定づける要素として「感情の種類」が重要な意味を持っている事と、文およびフレーズ指令の先頭での指令が、先頭以外の指令に比べ大きくなっていることなどが示された。

第5章

文節単位での感情の程度を考慮したF_0パターンの推定

5.1　はじめに

本章では、より高精度な韻律制御を行なうことを目的とし、先行研究で提案された F_0 モデルパラメータ推定の枠組みを改良した。F_0 パターンの推定のための入力項目の検討を行い、得られた F_0 パターンについての評価を行った。

具体的には、F_0 パターン推定の入力項目に文節を単位とした発話者の感情の程度の情報を追加することで、感情音声における韻律制御のさらなる高精度化を図った。

5.2　用意した感情音声試料

まず、本研究のために用意された音声試料について述べる。

用意した感情は「朗読音声 (平静)・怒り・喜び・悲しみ」の 4 感情。読み上げる文としては、ATR 音素バランス文セット 503 文を使用した。前述の音声試料とは異なり、平静文・感情文ともに同じ文章を読み上げている。

用意された文の例を表 5.1 に示す。

表 5.1: 用意した感情音声の文の例

あらゆる現実をすべて自分の方へねじ曲げたのだ。
一週間ばかりニューヨークを取材した。
野球のあとのビールぐらいうまいものはない。
昼から夜までおよそ子供の考えつくありとあらゆる遊びをやった。

発話者はプロの女性声優 1 名である。収録環境は遮音室で行われ、量子化ビットは 16bit、サンプリング周波数は 10kHz である。

用意された文の数を表 5.2 に示す。

表 5.2: 用意した各感情音声の文の数

感情	平静	怒り	喜び	悲しみ
文数	503	503	503	503

5.3　F_0MSE による評価

F_0 パターンの作成については前章と同じ手法を取るが、本章では 2 つの F_0 パターンの違いを定量的に表す尺度として式 5.1 に示す F_0MSE を用いる。ただし、t は有声フレーム数、T は有声フレーム総数である。

$$F_0MSE = \frac{\sum_t (\Delta \ln F_0(t))^2}{T} \tag{5.1}$$

自動抽出により得られた韻律語境界・F_0 モデルパラメータ、統計モデルより推定した韻律語境界・F_0 モデルパラメータのそれぞれから F_0 パターンを作成し、それらで比較を行う。

5.4　発声者の感情の程度の情報

5.4.1　発声者の感情の程度

本研究で用いる音声試料は 5.2 節で説明した通りであるが、その音声の収録の際に、音声と同時に発声者の感情の程度の情報に関する記録を取った。

記録は、収録を終えた発声者に、読み上げた文のテキストを見せ、発声者自身が発声内のどの部分に「特に感情を込めたか」を鉛筆で記録してもらうという手法を取った。記録してもらった紙データを元に、電子データにおこした。この電子データは文章を形態素単位に分割し、発声者が特に感情を込めた部分にはラベルを付ける意味で “1” を、それ以外の部分には “0” を記録した。

本研究では F_0 モデルパラメータの推定を文節単位で行なうため、形態素で記述されている感情の程度情報を文節単位に改めてまとめ直した。その際、2 つ以上の形態素で構成されるひとつの文節内で、感情の程度ラベルが “1” であるものと “0” であるものの 2 種類が混在した場合は、その部分は文節ごと特に感情を込めたものとみなし、“1” を記録した。以降の実験では、発声者の感情の程度情報としてこのデータを用いる。

5.4.2　入力項目の検討

従来の手法では、F_0 モデルパラメータ推定のための入力項目は形態素の品詞や活用形などの言語情報、および先行文節・アクセント句の F_0 モデルパラメータ情報などであった。本稿ではこれに、表 5.3 に示すように、前節で説明した発話者が特に感情を込めて読み上げた箇所の情報を追加する。先行文節との感情の程度差が当該文節に影響する可能性も考え、当該文節の程度情報だけでなく先行文節の程度情報も入力に追加する。

表 5.3: 指令推定の追加入力項目

入力項目	カテゴリ数
当該文節の強調感情有無	2 値
先行文節の強調感情有無	2 値

5.5　感情音声データの検証

この音声試料を使用して F_0 パターンの推定を行なう前に、第三者に発声者の感情表現と込められている感情の程度などを判定してもらった。判定をお願いしたのは音声の日本

語話者の発声指導を行なっているスピーチセラピストの櫻庭京子さんである。彼女は感情表現に関する論文 [28] も発表するなど、音声研究の分野において優秀な研究者であると同時に、幾人もの生徒への指導を通して感情音声の発声に関して十分なスキルを身につけているため、収録した音声データの聴取による検証をしてもらうのに適した人物であると考えた。

この検証によって、発声者が記録した感情の程度情報の有効性を確認することも期待できる。

まず、平静文と 3 種類の感情文から 10 文程度ずつ聴取してもらい、込められている感情の程度を聴取者の立場で判定してもらった。また各音声について感じる事はないか、感想をまとめてもらった。以下、その感想からの抜粋である。

- 怒り：文全体で感情を表現。話速が速い。感情の強弱は声の大きさ・強度によって表現しているようだ。
- 喜び：あまり喜びに感じない。推定対象から外すことをお薦めする。
- 悲しみ：語中の助詞の声質の変化、語尾にかけて声を小さくする事によって表現しているようだ。

感想より、この話者の場合「怒り」の感情が最も F_0 モデルによる統計的韻律制御が有効に働くだろう事が予想される。

また、聴取者として、発声者がどこに特に感情を込めていると感じるかのラベリングをしてもらった記録を見ると、ほぼ発声者が自身で記録した箇所と同じ位置に、感情の程度ラベルが入っていた。話者・聴取者ともに同じ箇所に強い感情の程度があると判断したことから、本データは客観的にも有効なものであることが確かめられた。

以上のことから、発声者の感情の程度の情報を F_0 パターン推定に利用するとともに、今後の実験では「怒り」の感情音声のみを用いて合成を行なうことにした。

5.6　推定実験条件

推定する F_0 パラメータのうち、フレーズ指令推定・アクセント指令推定それぞれに対する感情の程度情報の利用について、実験条件として以下に示す 3 条件を設定した。

- 従来手法：フレーズ指令・アクセント指令どちらの推定にも感情の程度情報を用いない
- 条件 1：フレーズ指令推定にのみ感情の程度情報を用いる
- 条件 2：アクセント指令推定にのみ感情の程度情報を用いる
- 条件 3：フレーズ指令・アクセント指令どちらの推定にも感情の程度情報を用いる

この 4 条件をもとに F_0 パターンを作成し、その評価を行った。本実験の評価は 5.3 節で説明した F_0MSE を用いる。

表 5.4: F_0MSE の平均：怒りの感情音声

	従来手法	条件 1	条件 2	条件 3
close	0.0696	0.0713	0.0692	0.0714
open	0.0755	0.0750	0.0745	0.0742

5.7　推定結果

5.7.1　F_0MSE

発話音声から自動抽出により得られた F_0 モデルパラメータと、5.7.2 節で推定した F_0 モデルパラメータのそれぞれから F_0 パターンを作成し、F_0MSE による比較を行った。なお、韻律語境界は発話音声から自動抽出で求めたものを利用している。表 5.4 に推定結果を示す。ここで、close は学習に用いたデータ、open は学習に用いていないデータの推定結果を示している。表 5.4 を見ると、アクセント指令推定のみに感情の程度情報を用いた 条件 2 の手法による推定精度が、close, open データともに従来手法よりも向上している。一方、フレーズ指令の推定に感情の程度情報を用いた 条件 1, 条件 3 の手法による推定精度は、従来手法に比べ悪化するという結果となった。

5.7.2　推定した F_0 パターンの例

図 5.1 に、「私達は静かに歩みより頭を下げた」という文章について、話者の感情の程度情報と F_0 パターンを推定した例を示す。感情は、怒りである。

図には、従来手法で作成した F_0 パターンとアクセント指令、および節で説明した 条件 2 の手法で作成した F_0 パターンを示している。図 5.1 の F_0 パターンの結果において、黒色線で示されているのが原音声の F_0 であり、赤色線で示されているものが音声から自動抽出した値 (正解)、青色線で示されているものが今回推定した結果である。アクセント指令パラメータにおいても同様で、赤色が自動抽出した値 (正解)、青色が推定した値である。

発声者は「頭を下げた」という部分に特に感情をこめて発声しているが、この部分の感情の程度フラグによって、アクセント指令の値がより正解に近いものとなっていることがわかる。感情の程度情報は、特にアクセント指令の大きさ成分について作用していることがこの結果から見て取れる。

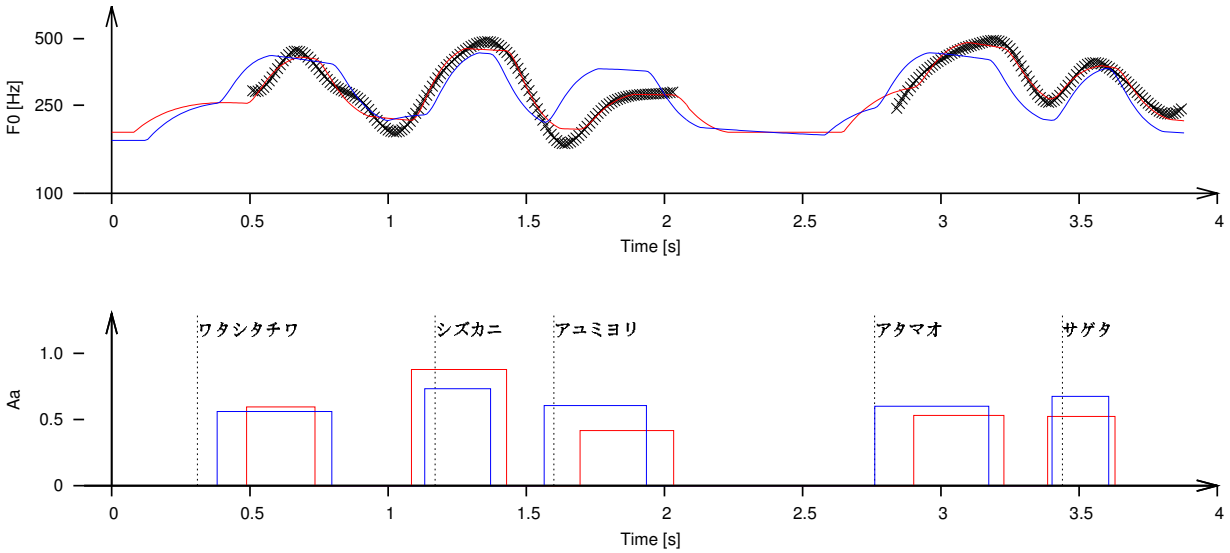
5.8　考察

表 5.4 に示す通り、表 5.3 の項目をアクセント指令推定にのみ追加した結果が最も良好であった。文節単位での感情の程度の情報を利用した制御によって、局所的な F_0 の立ち上がりに影響を与えるアクセント指令の推定精度が改善された結果であると考えられる。作成

(1) 合成音声の文章と感情の程度のフラグ

	私達は	静かに	歩みより	頭を	下げた
Strong Emotional Flag by the Speaker	0	0	0	1	1

(2) 従来手法により推定した F_0 パターンとアクセント指令パラメータ



(3) 条件 2 の手法により推定した F_0 パターンとアクセント指令パラメータ

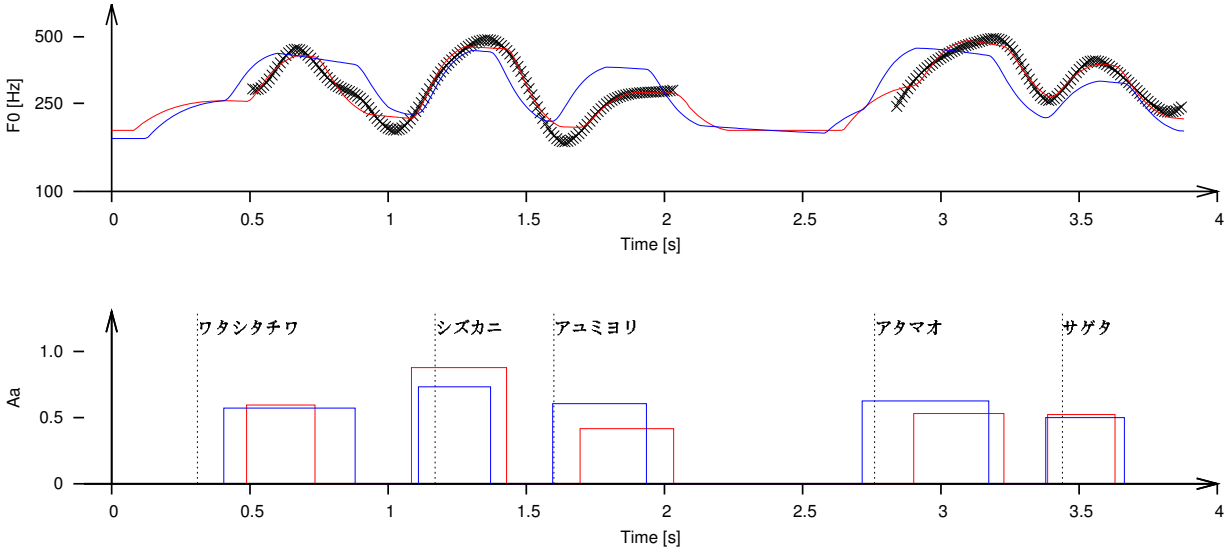


図 5.1: 推定した F_0 パターンの例

された F_0 パターンと F_0 モデルパラメータを見ると、従来手法に比べ、提案手法の推定結果がより正解に近いものとなっていることがわかった。

5.9　まとめ

本章では、先行研究の F_0 モデルパラメータ推定の枠組を改良した。より高精度な韻律制御を行なうことを目的とし、F_0 パターンの推定のための入力項目の検討を行ない、得られた F_0 パターンについての評価を行った。

F_0 モデルパラメータ推定の入力項目に文節を単位とした発話者の感情の程度の情報を追加することで、感情音声における韻律制御のさらなる高精度化が実現できた。感情の程度情報をアクセント指令の推定に利用した結果が最も推定精度が向上した。

また、推定した F_0 パターンの結果より、提案手法では従来手法に比べアクセント指令の大きさ成分について、より一致度が高まることが確かめられた。

第6章

主観的評価実験

6.1　はじめに

これまでの実験においては、F_0MSE という評価尺度を用いてきた。F_0MSE と知覚実験との相関関係は、先行研究により示されているものの、聴覚上重要であるピッチの時間変化を直接考慮している訳ではないこと、基底周波数の違いを吸収できないことなどの問題点がある。

そのため、本章ではテキストからの感情音声合成システムを構築し、知覚実験による F_0 パターンの評価を行った。

6.2　感情音声合成システムの構築

6.2.1　感情音声合成システムの概略

F_0 パターン作成の枠組みを用い、実際にテキストからの感情音声合成システムを構築した。図 6.1 に概略を示す。

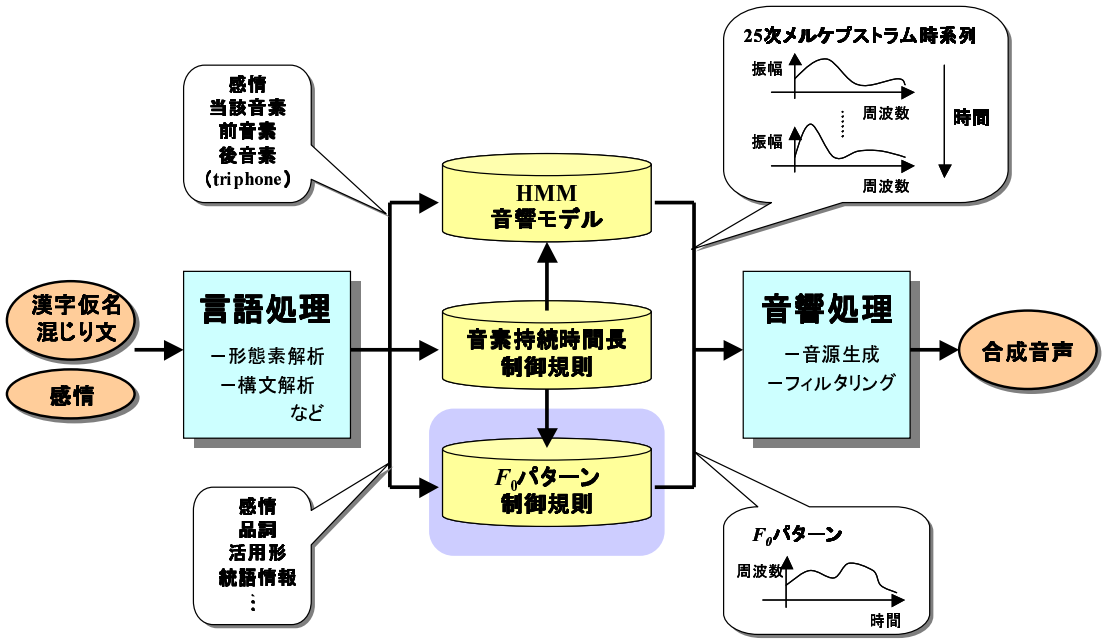


図 6.1: 構築した感情音声合成システムの概略

分節的特徴に関しては HMM を利用してメルケプストラム[1] の時系列を生成し、韻律的特徴に関しては決定木を用いて音素持続時間長・F_0 パターン推定に必要なパラメータを指

[1]音程の感覚を表す尺度であるメルスケール上での対数パワースペクトルの逆フーリエ変換として定義される。人の聴覚は、音の高さに関して、メル (mel) 尺度と呼ばれる対数に近い非線形の特性を示し、低い周波数では細かく、高い周波数では荒い周波数分解能をもつ。このため、メルケプストラムが広く用いられている。

定し、音声合成を行う。入力は漢字仮名混じり文と合成する感情で、指定された感情により制御規則である統計モデルが選択される。

6.2.2　HMM 音響モデルの作成

HMM の状態数を N とした時、HMM のパラメータ λ は、初期状態確率 π、状態遷移確率 A、各状態での出力確率 B により $\lambda = (A, B, \pi)$ と与えられる。

HMM の学習とは、与えられた学習用ベクトル系列 $O = \{o_1, o_2, \ldots, o_T\}$ に対し、観測尤度 $P(O|\lambda)$ を最大にする λ を求めることである。このためのアルゴリズムは、EM アルゴリズムに基づいて導出することができ、Baum-Welch 再推定式 [29] と呼ばれる。

本実験では、HMM の学習に HTK[30] を用いた。その際、HMM の学習用ベクトル系列として、メルケプストラム・Δ ケプストラム[2]・Δ^2 ケプストラム[3] を与えた。ベクトル系列の次元数はそれぞれ 25 次元 (パワー項を含む) の計 75 次元である。使用した HMM は、7 状態の left-to-right モデルとし、音素モデルは triphone で作成した。音素のスペクトルは、同一の音素でも前後の音素の影響を受け、違ったスペクトルとなるため、音素モデルとして、前後の音素を考慮した triphone を選択した。

6.2.3　HMM からの音声パラメータの生成

音声パラメータ生成に基づいた HMM からの音声合成は、音素 HMM を連結することにより得られる文に対応する HMM のパラメータ λ に対して、尤度最大となる長さ T の出力ベクトル系列を求めることである。

本実験では、各音素の持続時間長 (フレーム数) をあらかじめ決定木によって推定するという手法を取った。従って、HMM からの音声パラメータ生成における入力項目は音素 triphone モデルと、音素時間系列であり、出力項目は時系列のメルケプストラムとなる。

6.3　第一の実験：文節単位での感情表現

6.3.1　実験の目的

発話者の感情の程度情報を F_0 パターン推定の入力項目に加えることによって作成された合成音声の感情表現がどのようなものになるかを調べた。

本手法では文節を単位として発話者の感情の程度情報を入力項目に追加しており、合成音声が発話者の意図と同じ様に感情を表現できているかを調べることが必要である。

[2]ケプストラムの時間変化の微係数。ただし、微係数をそのまま用いたのではゆらぎが大きすぎるので、50ms 程度の時間幅におけるケプストラム時系列の線形回帰係数によって平滑化した値を用いる。

[3]Δ ケプストラムの時間変化の微係数。

6.3.2　実験条件

本実験には、5.7.2 節で作成した合成音声のうち、条件 2 の合成音声 (以下、提案手法) と従来手法による合成音声を用い、知覚実験による F_0 パターンの評価を行った。実験に用いる音声試料は、提案手法による合成音声を 20 文・従来手法による合成音声を 10 文の計 30 文で、被験者にはどちらの手法による合成音声かは提示しない。この 30 文をランダムな出現順序で被験者に示し、実験を行った。被験者は日本語話者 12 名である。

実験には、図 6.2 に示すフォームを用意した。中央部に合成音声を再生するボタンを配置し、上部には流される音声の読み上げ文章、左部にはその文章を文節単位で区切ったテキストと、各文節ごとにチェックボックスを一つずつ用意した。右部には「どの部分にも程度差を感じない」チェックボックスを用意した。また、右下部に各合成音声に対して感じたことがあった場合に被験者がコメントを入れられるように、テキストボックスを用意した。被験者にはまず「合成音再生」ボタンをクリックすることで流れる音声を聞いてもらい、その音声中のどの部分に特に感情がこめられているかを判断し、該当文節のチェックボックスにチェックを入れてもらった。この際、複数の文節を選択することを可能にした。どの部分にも特に感情の程度の違いを感じなかった場合は、右部にある「どの部分にも程度差を感じない」チェックボックスにチェックを入れてもらった。この項目にチェックを入れると、文節ごとに選択したチェックマークは全てクリアされる。すべての操作がマウスクリックのみで行なえるよう、インターフェースを工夫した。

6.3.3　評価方法

本実験では、合成音声の精度を示す指標として図 6.3 に示す一致度を用いる。

図 6.3 に示すように、話者が特に感情を込めて読み上げた部分と、聴取者が感情がこもっているように聞き取った部分が一部でも一致した場合にその回答は正解とし、各文章に対し何名が正答したかの正解率を計算した。

6.3.4　結果

提案手法による合成音声と従来手法による合成音声それぞれについて正解率を求めた。その平均値を表 6.1 に示す。

表 6.1: 正解率の平均

	提案手法	従来手法
正解率	92.7%	78.2%

提案手法による合成音声に関しては、正解率は平均 92.7%であった。一方、従来手法による合成音声については、一致度は 78.2%にとどまった。

提案手法による正解率が高かったことは、発声者の感情表現が被験者により正確に伝わったと言う事であり、作成された合成音声の品質が良かったことが示された。

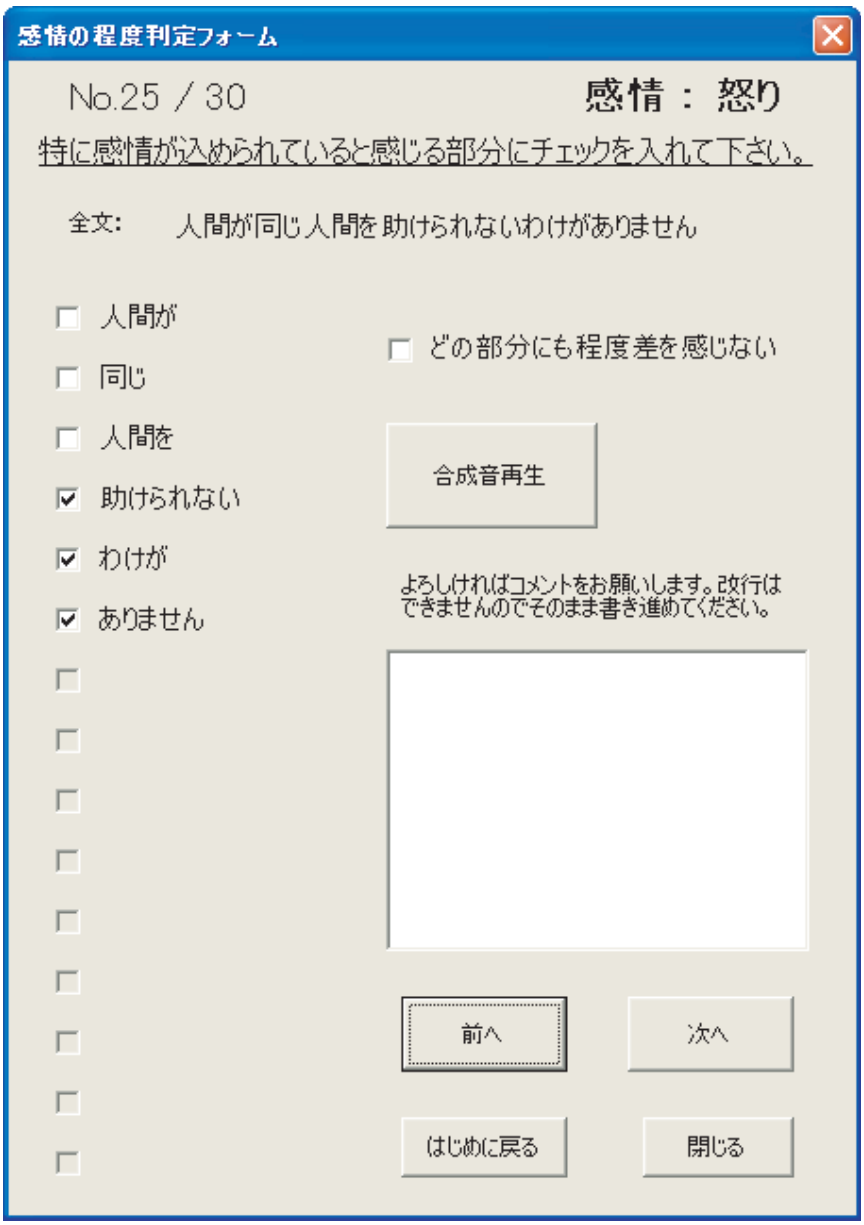


図 6.2: 第一の実験フォーム

6.4　第二の実験：文章全体の感情表現

6.4.1　実験の目的

第一の実験では、文節ごとの感情の程度を聞き取ってもらったが、ここでは第二の実験として、従来手法と提案手法、それぞれの合成音のうち、「より怒りを感じるのはどちらの合成音声か」「より自然な音声と感じるのはどちらか」を判定してもらった。この実験によって、提案手法によって文章全体の印象がどう変化したかを問うた。

図 6.3: 発話者が感情をこめた部分との一致

6.4.2　実験条件

従来手法と比べ、提案手法で作成された合成音声の文章全体の印象がどう変化したかを調べるために、第二の実験を行った。ひとつの文章に対し、提案手法によるものと従来手法によるもの、両方の合成音声を被験者に聞いてもらい、どちらの音声により明確に怒りの感情が込められているように聞こえるか、およびどちらの音声がより自然な音声と感じるかを評価してもらった。被験者は日本語話者 9 名で、用いた音声試料は 30 文章分の提案手法・従来手法による合成音声である。

実験には、図 6.4 に示すフォームを作成した。上部に流される音声の読み上げ文章と 2 種類の合成音声を再生するボタンをそれぞれ配置し、被験者がそのボタンをクリックすることで合成音声を何度も聞けるようにした。合成音 A,B どちらのボタンに提案手法／従来手法どちらの音声を割り与えるかは、ランダムに決定した。まず、ひとつめの質問として、「どちらの合成音に、より怒りがこもっているか」を評価してもらった。この質問に対し、被験者は合成音 A、および合成音 B のどちらか一方のみを選択することができる。同様に、ふたつめの質問として「どちらの合成音が、より自然な音声と感じるか」を評価してもらった。下部には各合成音声に対して感じたことがあった場合に被験者がコメントを入れられるように、テキストボックスを用意した。本実験でも、すべての操作をマウスクリックのみで行なえるよう、インターフェースを工夫した。

6.4.3　結果

ひとつめの項目である「どちらにより怒りの感情が込められているか」という質問に対しては、30 文すべての音声に対し、提案手法による合成音声の方に、より怒りの感情が込められていると判断する被験者が多数を占めるという結果となった。また、各文章ごとにそれぞれの合成音声がどれだけ指示されたかの選択率を求め、その平均値を表 6.2 にまとめた。提案手法による合成音声を選択する被験者が明らかに多いという結果になった。

表 6.2: 選択率の平均：怒りの感情

	提案手法	従来手法
選択率	79.3%	20.7%

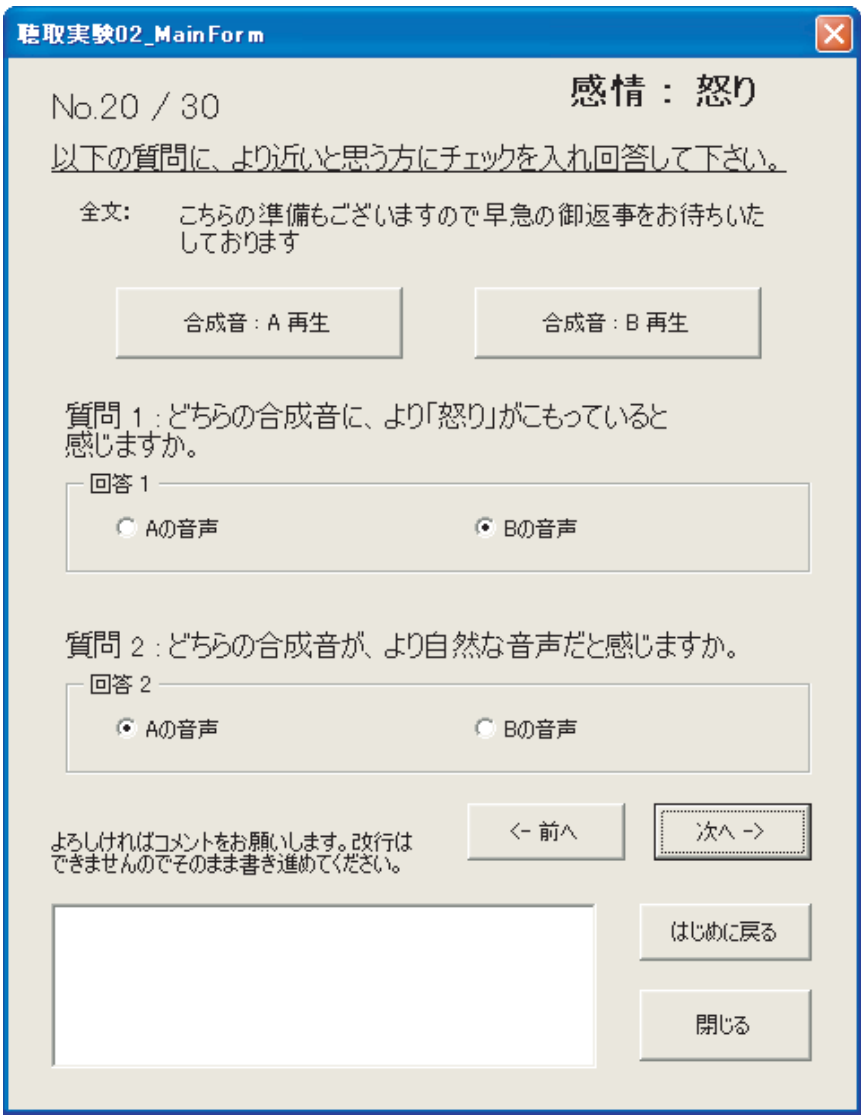


図 6.4: 第二の実験フォーム

一方、ふたつ目の項目である「どちらの合成音声がより自然か」という質問に対しては、30 文章のうちの 22 文章について、従来手法による合成音の方が自然だと判断する被験者が多数を占めるという結果となった。ひとつ目の質問と同様に、表 6.3 に各文章ごとにそれぞれの合成音声がどれだけ指示されたかの選択率の平均を示す。これを見ても、自然性については、従来手法による合成音声を指示する被験者が多かった事がわかり、文節単位での局所的な制御を行なわない方が自然と判断される結果となった。この、感情の表現をより豊かにした合成音声の自然性を高めるという手法を検討する事については、今後の課題である。

表 6.3: 選択率の平均：自然性

	提案手法	従来手法
選択率	39.3%	60.7%

6.5　第三の実験

6.5.1　実験の目的

第一の実験は音声と同時に発話内容のテキストも被験者に提示していたため、被験者がどこに感情が込められているかを判定する際に、合成音声によってではなくテキストから判断したのではないかという疑念が残る。この懸念について検証を行なうために、第三の実験を行った。

6.5.2　実験条件

本実験では、被験者には文章のみを提示し、自身であればどこに怒りの感情を特にこめて読み上げるか、を判断してもらった。実験に使用する文章は第一、第二の実験で用いたのと同様のものである。各被験者が第一の実験でチェックを入れた箇所と本実験でチェックを入れた箇所を比較し、全く同一箇所にチェックが入っているのであれば、その被験者は第一の実験において、合成音声を聴取しての判断でなく、文章の内容から判断したことになる。なお、本実験は第一の実験の実施日から十分に日を空けて行なっているため、第一の実験で聞いた音声の印象が残って判断されるとは考え難い。

実験には、図 6.2 に示すフォームを用意した。上部には全体の文章を提示し、左部にはその文章を文節単位で区切ったテキストと、各文節ごとにチェックボックスを一つずつ用意した。右部には「どの部分にも特に感情をこめないだろう」と考える場合に選択するチェックボックスを置いた。また、右下部に各文章に対して感じることがあった場合に被験者がコメントを入れられるように、テキストボックスを用意した。被験者には、まず文章の全文に目を通し、自身であればその文章中のどの部分に特に感情がこめて読み上げるかを判断し、該当文節のチェックボックスにチェックを入れてもらった。この際、複数の文節を選択することを可能にした。どの部分にも特に感情をこめないだろうと考える場合は、右部にある「どの部分にも特に感情をこめない」チェックボックスにチェックを入れてもらった。この項目にチェックを入れると、文節ごとに選択したチェックマークは全てクリアされる。この実験に置いても、全ての操作をマウスクリックのみで行なえるよう、インターフェースを工夫した。

6.5.3　結果

第一の実験および第三の実験の両方に参加してもらった被験者 10 名、全員について、第一の実験における回答と第三の実験における回答が完全に一致するという事はなかった。

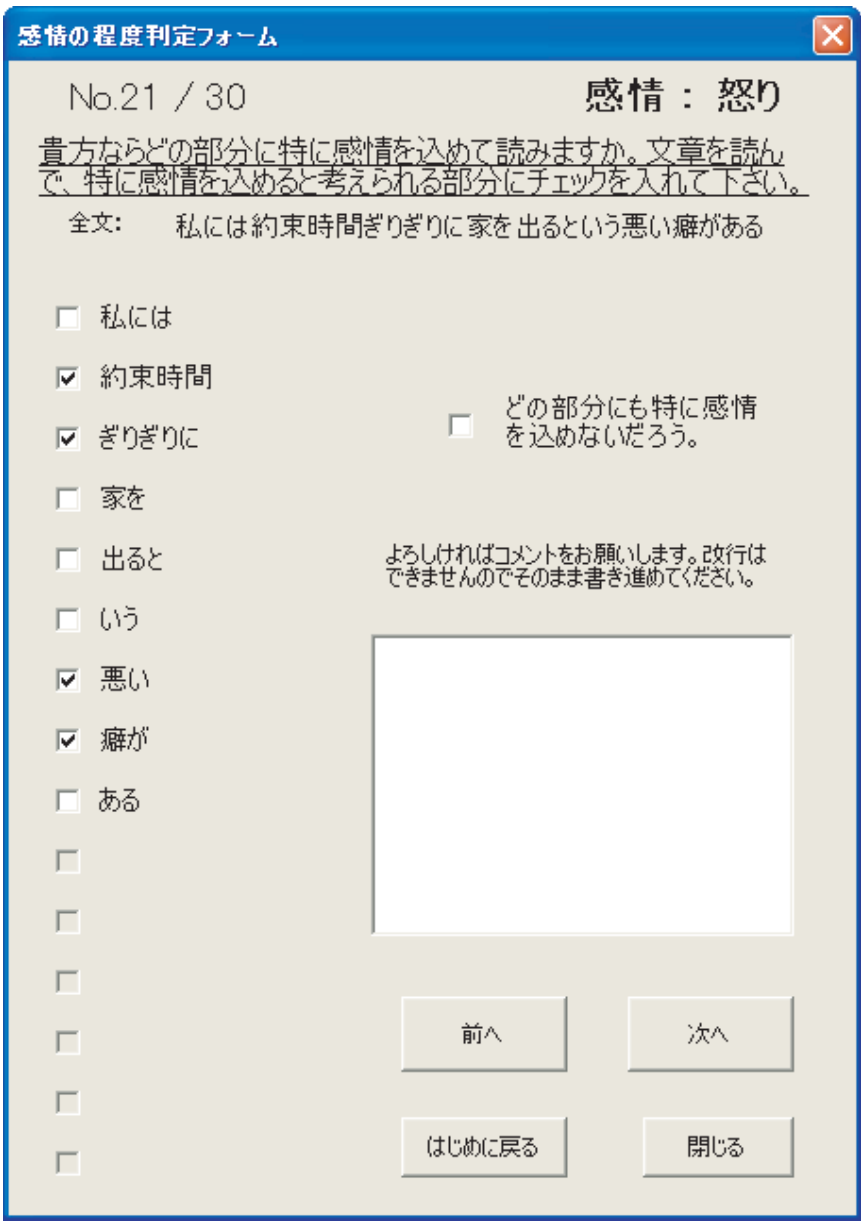


図 6.5: 第三の実験フォーム

この実験結果により、第一の実験結果の有効性が確かめられた。

6.6　考察

第一の実験より、文節単位での感情の程度情報を入力項目に加えることによって、より元の感情表現に近い合成音声を作成できた事が確かめられた。また、第二の実験より、提案手法による合成音声の方がより明確に怒りの感情を表現できたことが分かった。最後に、第三の実験により、被験者が第一の実験においてテキストから回答を判断したわけではな

い、という事が確かめられ、合成音声の韻律制御の精度が確実に向上しているという事が確かめられた。

6.7　まとめ

本章では、考案した F_0 パターンの生成法を元に感情音声合成システムを構築し、主観的評価実験を行った。

感情音声合成システムは、HMM でのスペクトル生成と決定木による音素持続時間長推定の枠組みを利用することで構築した。主観的評価実験では、提案手法によって、従来手法よりも元の音声の感情表現に近く、明確に怒りの感情を表現できているという評価を得た。

第7章

結論

本研究では、韻理的特徴量の中でも F_0 パターンの制御に着目し、それを適切に制御する手法について検討した。朗読音声以外の異なる発話スタイルの中でも、感情に焦点を起き、検討したコーパスベース韻律制御手法を適用することによって、感情音声合成の実現を目指した。

第 1 章では、まず、本研究の背景と、研究の目的、本稿の構成などについて説明し、音声に含まれる情報や感情音声の合成手法、強調が用いられた音声の分析の研究動向などについてまとめた。また、感情音声が主に韻律的特徴量によって表されることを説明した。それと共に、音声合成のための基礎技術、感情音声合成についての現状を述べた。

第 2 章では、感情音声合成に必要な要素技術について述べた。統計モデルを用いたテキスト音声合成システムの概観と、F_0 パターン制御の要素技術である基本周波数パターン生成過程モデルについてその原理を示した。また F_0 モデルパラメータの自動抽出ツールや形態素および構文解析システムなど言語情報処理に用いるツールと、本研究の枠組みで必要となる韻律コーパスについても合わせて説明した。

第 3 章では、F_0 モデルパラメータの推定や韻律語・アクセント核の推定のための統計モデルの概要について概要を示した。F_0 モデルパラメータの推定に関しては、フレーズ指令およびアクセント指令の推定で用いられる学習の入力項目について、順に説明を加えた。

第 4 章では、先行研究において提案されていた F_0 モデルパラメータ推定の枠組みを一部変更し、「感情」が韻律的特徴量に与える影響について、実験とともに分析を行った。推定される F_0 モデルパラメータにおいて、感情の種類はタイミング成分よりは大きさ成分について重要な項目となること、感情音声においては文およびフレーズ成分の先頭での指令が、先頭以外での指令にくらべ、大きくなる事を示した。

第 5 章では、感情音声とともに発声者の感情の程度が文節ごとに記録されているデータを用い、F_0 パターンを作成するシステムを構築した。F_0 モデルパラメータのうち、アクセント指令の推定の入力項目に感情の程度情報を追加する事によって、韻律推定の精度が向上する事を示した。

第 6 章では、第 5 章で示した方法の有用性を検証するため、感情音声合成システムを構築し、3 種類の主観的評価実験を行った。第一の実験の結果、提案手法によってより元の音声に近い感情表現が実現できたことが示された。第二の実験の結果、提案手法の合成感情音声が、より全体として怒りの感情を明確に表現できていることが示された。しかし、韻律の自然性という観点では、従来手法で作成された合成音声に軍配があがり、より自然で聞き取りやすい韻律生成の手法について課題が残された。第三の実験の結果、第一の実験結果の有効性が確認された。

今後の課題としては、まずポーズを制御する枠組みの問題がある。構築したシステムでは、ポーズの位置は既知として扱っていたため、本当の意味でのテキスト音声合成システムは完成していない。単純に、句読点の位置をポーズとすることもできるが、感情音声においては、特に悲しみなどでは、ポーズの果たす役割が大きく、その制御は単純なものではない。また、音素持続時間長の推定に関しても課題が残る。本来、音素の時系列は推定項目であるが、本研究では自動抽出された音素時系列を与えている。この、ポーズを制御する枠組みと音素持続時間長の推定手法とは今後検討が必要な課題である。

また、感情の程度を考慮した韻律制御における韻律の自然性の向上も今後の課題として挙げられる。第 6 章の第二の実験によって、提案手法の合成音声が従来手法の合成音声にくらべ、韻律の自然性という観点では支持される割合が低かった。今回音声試料として使用した声優の発声は、原音声を聞いても抑揚の差がかなり大きく、この音声を元に合成音声を作成したため、その影響が色濃く反映されてしまった事も原因として考えられるが、本提案手法による合成音声は韻律の抑揚が従来手法に比べ激しく、それが自然性が低いと判断される原因となったのではないかと考えられる。

本来の感情音声 TTS システムとは、完全にテキストのみから感情音声を作成するシステムであることが理想である。本研究では文章中に特に感情を込める箇所を話者の感情の程度情報として入力項目に追加したが、本来はこの「どこに感情を強くこめるか」という情報も自動的に付与されるべきである。そこで、現在は品詞の種類や文法的係り受け等にとどまっている言語情報に、単語の意味情報や意味的な係り受け情報、感情音声中に登場する頻度情報などの情報を追加して推定のための学習を行うなど、テキストから全て自動で音声を生成する TTS システムの構築を目指すのも将来的に興味深い研究であると考える。この点も、最後に今後の課題として付け加えておく。

謝辞

この研究を進めるにあたって、いつも温かい御指導、御鞭撻を賜わりました広瀬啓吉教授、峯松信明助教授には心より感謝いたします。特に広瀬教授には私の指導教官という事もあり、研究の方針立てや進め方など親身に相談に乗って下さり、私自身が楽しみながら研究に取り組むことができました。本当にありがとうございました。

技術的・設備的な面をはじめとして私達学生の研究の支援に多大なる御尽力を賜わりました高橋登技官、秘書の武田祥子さん、笠島恵美さんに深く御礼申し上げます。余計なストレスを感じることなく研究に専念できましたのは、皆様のおかげだと思っています。

先行研究となる偉大な研究成果と感情音声合成の美しいプログラムを残して下さいました佐藤賢太郎さんに深く感謝いたします。直接お礼を申し上げられないのが残念ですが、この場を借りて心よりの感謝の意をお伝えいたしたいと思います。

御自身のお仕事・研究でお忙しい中にもかかわらず、私の研究に快く御協力を下さいましたスピーチセラピストの櫻庭京子さんに深く感謝いたします。櫻庭さんのお話はどれも興味深く、お話を通してたくさんの知的な刺激をいただきました。

最後に、聴取実験に御参加下さいました皆様、本当にありがとうございました。本日こうして研究成果としてまとめることができたのも、皆様の御協力があったからこそです。素晴しい諸先輩方、同級生、後輩に恵まれた私は幸せ者です。

いつの日か皆様にこのご恩をお返しする事ができますよう、今後とも精進してまいりたいと思います。何か私で力になれる事がございましたら、いつでも御連絡下さい。それでは、皆様の今後ますますの御健勝を心よりお祈り申し上げます。

2006 年 1 月
浅野泰史

参考文献

[1] 藤崎博也: “韻律研究の諸側面とその課題,” 日本音響学会講演論文集, 2-5-11, pp.287–290 (1994).

[2] M.Kienast, W.F.Sendlmeier: “Acoustical Analysis of Spectral and Temporal Changes in Emotional Speech,” *Proceedings of ISCA Workshop on Speech and Emotion*, pp.92–97 (2000).

[3] 小川孝, 中尾光志, 重永実: “単語音声に含まれる感情情報について,” 日本音響学会平成 7 年度春季研究発表会講演論文集, Vol.1, 2-4-2, pp.267–268 (1995).

[4] 藤崎博也, 大野澄雄, 富田修, 丸山英晃: “話者の感情が音声の韻律的特徴に及ぼす影響,” 日本音響学会平成 7 年度春季研究発表会講演論文集, Vol.1, 2-4-3, pp.269–270 (1995).

[5] 重永実: “感情の判別分析からみた感情音声の特性 (III),” 電子情報通信学会技術研究報告, SP97-66, pp.65–72 (1997).

[6] 川波弘道, 広瀬啓吉: “態度・感情音声における韻律的特徴の考察,” 電子情報通信学会技術研究報告, SP97-67, pp.73–80 (1997).

[7] 武田昌一, 西澤良博, 大山玄: “「怒り」の音声の特徴分析に関する 1 考察,” 信学技報, SP2000-164, pp.33–40 (2001).

[8] A.Iida, F.Higuchi, N.Campbell, M.Yasumura: “Corpus-based speech synthesis system with emotion,” *Proceedings of Speech Communication*, Vol.40/1-2, pp.161–187 (2002).

[9] J.Yamagishi, K.Onishi, T.Masuko, T.Kobayashi: “Modeling of Various Speaking Styles and Emotions for HMM-Based Speech Synthesis,” *Proceedings of EUROSPEECH*, Vol.4, pp.2461–2464 (2003).

[10] 都築亮介, 全炳河, 徳田恵一, 北村正, Murtaza Bulut, Shrikanth S.Narayanan: “HMM に基づく感情音声合成に関する検討,” 日本音響学会 2003 年秋季講演論文集, Vol.1, 1-8-31, pp.241–242 (2003).

[11] 大野澄雄、藤崎博也、高橋浩生: ”日本語音声における強調の韻律的特徴に与える影響について”, 日本音響学会講演論文集, 平成 10 年 9 月, 1-2-14, pp.201-202(2001)

[12] 杉山佳三、藤崎博也、酒巻宏、大野澄雄: ”韻律的特徴によって表現される強調の定量的分析”, 日本音響学会講演論文集, 2001 年 3 月, 1-6-19, pp.265-266(2001)

[13] 立花隆輝、西村雅史: ”読み上げ韻律との差分を使った強調韻律の学習”, 日本音響学会講演論文集, 2005 年 3 月, 1-1-13, pp.179-180(2005)

[14] 桂聡哉, 広瀬啓吉, 峯松信明: “感情音声のための生成過程モデルに基づくコーパスベース韻律生成とその評価,” 電子情報通信学会技術研究報告, SP2002-184, pp.31–36 (2003).

[15] K.Hirose, K.Sato, N.Minematsu: “Emotional Speech Synthesis with Corpus-Based Generation of F_0 Contoures Using Generation Process Model”, Proceedings of Speech Prosody 2004

[16] H.Fujisaki, S.Nagashima: “A model for synthesis of pitch contours of connected speech,” *Annual Report of Engineering Research Institute, University of Tokyo*, Vol.28, pp.53–60 (1969).

[17] 江藤雅哉, 広瀬啓吉, 峯松信明: “テキスト音声合成システムのための統計モデルによる F_0 パターン生成の改良,” 日本音響学会 2002 年春季講演論文集, Vol.1, 1-10-8, pp.245–246 (2002).

[18] L.Brieman, J.H.Friedman, R.A.Olshen, and C.J.Stone: “Classification and Regression Trees,” *Wadsworth, Paciic Grove, California* (1984).

[19] The Edinburgh Speech Tools Library.

http://www.cstr.ed.ac.uk/projects/speech_tools/

[20] 成澤修一, 峯松信明, 広瀬啓吉, 藤崎博也: “声の基本周波数パターン生成過程モデルのパラメータ自動抽出法,” 情報処理学会論文誌, Vol.43, No.7, pp2155–2168 (2002).

[21] 形態素解析システム 茶筌.

http://chasen.aist-nara.ac.jp

[22] 日本語形態素解析システム JUMAN.

http://www-nagao.kuee.kyoto.u.ac.jp/ml-resource/juman.html

[23] 日本語構文解析システム KNP.

http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/

[24] 大語彙連続音声認識デコーダ Julius.

http://winnie.kuis.kyoto-u.ac.jp/pub/julius/

[25] 喜多竜二, 峯松信明, 広瀬啓吉: “日本語テキスト音声合成を目的としたアクセント結合規則の構築と改良,” 電子情報通信学会技術研究報告, SP2002-26, pp.13–18 (2002).

[26] 古山悠介, 成澤修一, 広瀬啓吉, 峯松信明, 藤崎博也: “言語情報を用いた基本周波数生成過程モデルパラメータ自動抽出の高精度化,” 電子情報通信学会技術研究報告, SP2003-70, pp.37–42 (2003).

[27] Hirose,K., Minematsu,N., Kawanami,H.,: ”Analytical and perceptual study on the role of acoustic features in realizing emotional speech.” In: International Conf. on Spoken Language Processing,Beijing, Vol.2, pp.378-381.(2000)

[28] 櫻庭京子，今泉敏，筧一彦，Donna Erickson : ”日米語の音声的制約が感情表現に及ぼす影響の音響的比較— ピカチュウ vs pikachu — ,” 日本音響学会聴覚研究会資料 H-2001-22, Vol.31 (2), pp.157-164, (2001).

[29] 北研二 : “確率言語モデル,” 東京大学出版会 (1999).

[30] Hidden Markov Model Toolkit.

http://htk.eng.cam.ac.uk/

発表文献

[1] Keikichi Hirose, Yuji Yagi, Seiya Takada, Yasufumi Asano and Nobuaki Minematsu: "Generation of Speech Reply in a Dialogue System," Annaul Project Report Grant-in-Aid for Creative Basic Research "Language Understanding and Action Control," pp.209-218, 2005-3.

[2] K.Hirose, K.Sato, Y.Asano, and N.Minematsu: "Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora; application to emotional speech synthesis," Speech communication, vol.46, no.3-4, pp.385-404 (2005-7).

[3] 浅野泰史, 広瀬啓吉, 峯松信明: "文節単位での感情の程度を考慮した統計的韻律制御," 日本音響学会 2006 年春季講演論文集 (2006.3) 発表予定.

[4] Keikichi Hirose, Yuji Yagi, Seiya Takada, Yasufumi Asano and Nobuaki Minematsu: "Speech Synthesis from Concept and its Prosodic Control for Reply Speech Generation in a Spoken Dialogue System," Annual Project Report Grant-in-Aid for Creative Basic Research "Language Understanding and Action Control," 2006-3. (予定)