修士論文

HMM音声合成における生成過程モデルを用いた基本周波数パターン生成

2011 年 2 月 9 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 48-096422

松田 徹也

内容梗概

近年, 柔軟な音声合成を可能にする手法として HMM 音声合成が注目されている. HMM 音声合成では, 音声スペクトル · 基本周波数 · 継続長がコンテキスト依存 HMM によって同時にモデル化される. 音声の合成に用いる特徴パラメータ列は, この HMM から尤度最大化基準に基づいて生成される. このとき, 動的特徴量や発話内変動についてまで考慮することで, なめらかで自然な合成音声の生成が実現される. HMM 音声合成には, 分節的特徴と韻律的特徴の対応付けが自動的に行われるという特徴があり, これが柔軟な音声合成を可能にする理由の一つになっている. 一方で, 二つの特徴の対応付けはフレーム単位で行われるため, 分節よりも広い時間区間にまたがる韻律的特徴のモデル化は困難な問題となる. 特に学習データの量が十分でないとき, この問題は顕著となり, 誤った基本周波数パターン生成の原因となる.

基本周波数パターンを表現する効果的な方法として生成過程モデルが挙げられる. 生成過程モデルは, 生理学的知見に基づいた数理モデルであり, 基本周波数パターンを, 対数周波数軸上において, フレーズ成分とアクセント成分の重ね合わせとして表現する. フレーズ成分は, 輪状甲状筋の斜部の収縮によって引き起こされる大域的な韻律的特徴を近似する成分であり, アクセント成分は, 輪状甲状筋の直部の収縮によって引き起こされる局所的な韻律的特徴を近似する成分である. このように記述することで, 広い時間区間にまたがる韻律的特徴であっても十分なモデル化が可能となる.

上述の観点に基づき, 本研究では, 生成過程モデルを HMM 音声合成における制約条件として利用することを提案する. HMM 音声合成によって生成される基本周波数パターンを生成過程モデルで表現可能な範囲に絞ることで, 広い時間範囲に渡る韻律的特徴を十分にモデル化できなかったために生じる局所的な変動が平滑化され, 合成音声の自然性は改善される. 提案手法の具体的な有効性については, 主観評価実験の結果を通して考察する.

さらに, 基本周波数パターン中に元来含まれている微小な変動は生成過程モデルによる制約の効果を低減させ得るという観点から, これら微小な変動の検出についても検討する. 微小変動から受ける影響を考慮しつつ生成過程モデルによる制約を適用することで, HMM 音声合成によって生成される基本周波数パターンの自然性はさらに向上する.

目次

図目次

表目次

第 1 章

序論

1.1　研究の背景

過去十数年を通して, Text-to-Speech による合成音声の品質は劇的に向上している. 計算機の飛躍的な能力向上に伴い, 大量の音声コーパスを利用したコーパスベース音声合成が可能となったことがその理由の一つである [1]. 規則に基づく従来の音声合成手法と異なり, 専門的な知識や経験が必要不可欠なものでは無くなったため, コーパスベース音声合成に関する研究は盛んに行われ, 数多くの提案がなされてきた.

コーパスベース音声合成の確定的定義は明らかではないが [2], 本論文では「あらかじめ話者ごとに蓄積された, 大量の音声波形と，これに対応するテキスト, 音素, アクセント位置, イントネーション等の情報を利用して合成音声を得る方式」として扱う. コーパスベース音声合成は, 音声波形を, 音素等に基づく適当な単位 (単位音声) にごとに集約してから利用する. この集約処理の結果として得られるものを音響モデルと呼ぶ. 音響モデルは, 音声合成システムを特徴付ける重要な機能であり, サンプルベース方式と統計量ベース方式の 2 つに分類される.

サンプルベース方式は, 今現在において主流の方式であり, 音声波形を直接つなぎ合わせて合成音声を生成する. 音響モデルも, サンプル (部分波形) を単位音声ごとに直接集める形で構築される. この方式の最大の利点は, 音声コーパス中の自然音声をほとんど加工することなく出力音声に用いるため, 話者の特徴や肉声感を保持した合成音声が得られるところにあり, これがサンプルベース方式が広く用いられている理由でもある. しかし, 音色や音の高さといった音響的な特徴量は, テキストに依存して連続的に変化するものであるため, 任意のテキストから合成音声を生成するのに必要なだけのサンプルをすべて網羅することは不可能である. この問題はサンプルの接続部で顕著となる. 適切な信号処理をサンプルに対して行うことで対処は可能であるが, 十分な音質を得るためには膨大な量の音声コーパスが必要となる [3, 4].

一方, 統計量ベース方式音声合成では統計的な音響モデルの構築が行われる. このとき, 音声波形自体の統計量を求めても音響的な意味合いとの直感的な対応付けができないため, 音声波形から音響的な特徴量を抽出する必要が生じる. 音響特徴量の抽出によく用いられるのが Source-Filter モデルである. Source-Filter モデルは, 音声信号を, 音声信号の周期性を表現する音源信号 (Source) と, 音声信号の音色を表現する線形フィルタ (Filter) との畳み込み演算の結果として定義する. 音源信号は, さらに, 周期性の有無, 基本周波数 (F_0), 継続長の 3 つの要素に分けられる. F_0 とは, 周期性を持つ音源信号の最小周期区間の逆数であり, これはおよそ音の高さに相当する [5]. 一方で, 音色を表現する線形フィル

タは, メルケプストラムと呼ばれるフィルタ係数ベクトルによって表現される [6]. つまり, 周期性の有無, F_0, 継続長, メルケプストラムの 4 つについて統計的なモデリングを行えば, 統計量ベース方式の音声合成が可能となる. 統計量ベース方式を用いることで, 音響特徴量の連続的なモデル化による不連続感の少ない合成音声の生成や, 個々のサンプルが持つ情報の統計的な共有による小規模コーパスからの高品質な音声の合成が実現できる. だが, 統計量ベース方式では, 音響特徴量の平均や分散を用いて音声合成を行うため, 個々のサンプルの持つ詳細な特徴は平均化されてしまい, 合成音声の肉声感は損なわれてしまう. この理由により, 統計量ベース方式はサンプルベース方式ほど利用されることはなかった.

ところが近年, 隠れマルコフモデル (HMM: hidden Markov model) を利用した統計量ベース方式の音声合成システムである HMM 音声合成 [7] に注目が集まっている. 音声信号中には, 先述した音響的特徴の他に, 言語情報を表す分節的特徴が含まれるが, この分節的特徴のコンテキストに基づいて構成されたコンテキスト依存 HMM を用いて, 周期性の有無, F_0, 継続長, メルケプストラムのモデル化を同時に行う点が, HMM 音声合成の特徴である. こうしたモデル化により分節的特徴と韻律的特徴の対応付けが自動的に保たれるため, HMM 音声合成では合成音声の発話スタイルを分節的特徴に基づいて柔軟に制御することが可能となる. 発話スタイルの制御は合成音声の自然性と同様に重視される要素であり, HMM 音声合成が注目されている理由もここにある. また, 従来から統計量ベース方式の課題となっていた合成音声の自然性の低さについても, 音響的特徴の動的特徴量や発話内変動 (GV) を踏まえた合成を行うことで大幅に改善されている [8].

しかし, HMM を用いてモデル化を行うためには, 音響特徴量をフレーム単位で取り扱わなければならない. このような取り扱いは, 特に分節よりも広い時間範囲に跨る韻律的特徴のモデル化を困難にし, F_0 の時間に対する変化パターン (F_0 パターン) 中に局所的な起伏を生むような形でしばしば合成音声の自然性を低下させる [9].

1.2　研究の目的

本研究の目的は, HMM 音声合成における韻律的特徴の取り扱いを工夫し, 合成される音声の自然性を向上させることである. 具体的には, F_0 パターン生成過程モデル (F_0 モデル) [10] を HMM 音声合成に導入することで, フレーム単位でのモデル化に起因する誤った F_0 パターンの生成を抑制する. F_0 モデルは, 生理学的知見に基づいた数理モデルであり, 分節よりも広い時間範囲に跨る韻律的特徴であっても効果的に記述できる [11, 12]. この F_0 モデルで記述可能な範囲に限りながら HMM から F_0 パターンを合成することで,

合成される F_0 パターンに対し生理学的な裏付けを持った平滑化が行われ, フレーム単位でのモデル化が原因で生じた局所的な起伏は低減する. ただし, F_0 パターン中には F_0 モデルの制約としての利用を困難にする微小な変動が元来含まれているため, この微小変動への対策についても検討を行う.

1.3　論文の構成

本論文は全六章から構成される. まず第二章において, 従来手法である HMM 音声合成の要素技術について紹介する. 次の第三節では, 提案手法の要となる技術である F_0 モデルについて説明を行う. その後第四節では, 提案手法の詳細な実装について述べ, 主観評価実験を通してその有効性を検証する. さらに第五節では, 提案手法をより効果的なものとすることを目的とした, F_0 パターン中からの微小変動検出法について検証結果とあわせて説明する. そして最後の第六節で, 本論文をまとめると伴に今後の展望を述べる.

第 2 章

HMM音声合成

2.1　はじめに

現在にいたるまで, 高品質な音声の合成を可能とする数多くの Text-to-Speech システムが提案されてきた. しかし, 個人性や感情に由来するような様々な発話スタイルの変化まで, 細やかに制御可能なシステムは長らく提案されてこなかった. 現在主流となっているサンプルベース方式の音声合成システムでこうした制御を行うためには, 細かな変化に対応した膨大な量の部分波形を用意せねばならず, 現実問題として実現困難であったことがその理由である. HMM 音声合成はこの問題を解決する目的で提案された音声合成システムである. HMM を利用したサンプルベース方式の音声合成システムは, HMM 音声合成が提案される以前からいくつか存在していたが, HMM 音声合成は統計量ベース方式の音声合成システムであり, それらの音声合成システムとは異なったアルゴリズムで音声の合成を行う [13]. HMM は統計モデルの一種であり, HMM 音声合成では, 分節的特徴のコンテキストに基づいて構成される. こうして用意された HMM をコンテキスト依存 HMM と呼ぶ. コンテキスト依存 HMM を用いて音響モデルを構築すると, 分節的特徴と韻律的特徴の対応付けが自動的に保たれるため, 分節的特徴に基づいた発話スタイルの柔軟な制御が可能となる.

コンテキスト依存 HMM の観測シンボルは, メルケプストラムと F_0, さらにこれらの時間変化分である動的特徴量を合わせたものとなる. HMM 音声合成では, 入力された文章に合わせて HMM を繋ぎあわせ, そこから音響特徴量を直接出力する. 動的特徴量についてまでモデル化を行うことは, 接続歪みを和らげる効果をもたらす. 音響特徴量は連続値としてモデル化を行う必要があるが, 音声波形中には F_0 が定義できない区間も存在するため, 通常の HMM では F_0 をモデル化することができない. HMM 音声合成では, 周期性の有無を表現するための状態を持った Multi-Space probability Distribution HMM (MSD-HMM) を用いることで, F_0 のモデル化を可能にしている [14]. HMM からの音響特徴量の出力は尤度最大化基準を用いて行われるが, この方法だけでは適切な継続長を得ることができない. そこで継続長を表現するためのコンテキスト依存 HMM も別に用意 · 学習し, 尤度最大化基準に従ってパラメータ生成を行うことで適切な継続長を取得する [15]. また, 統計量ベース方式の音声合成では, 統計処理によって音響特徴量が平均化されることによる, 合成された音声波形の過剰な平滑化が問題となっていた. HMM 音声合成では, 音響モデルを構築する際, 音声波形の Global Variance (GV) についてもモデル化を行うことで, この問題に対処している. コンテキスト依存 HMM の尤度を最大化する際に, GV の確率モデルの尤度を加味することで, 自然音声中に元々存在していた変動が

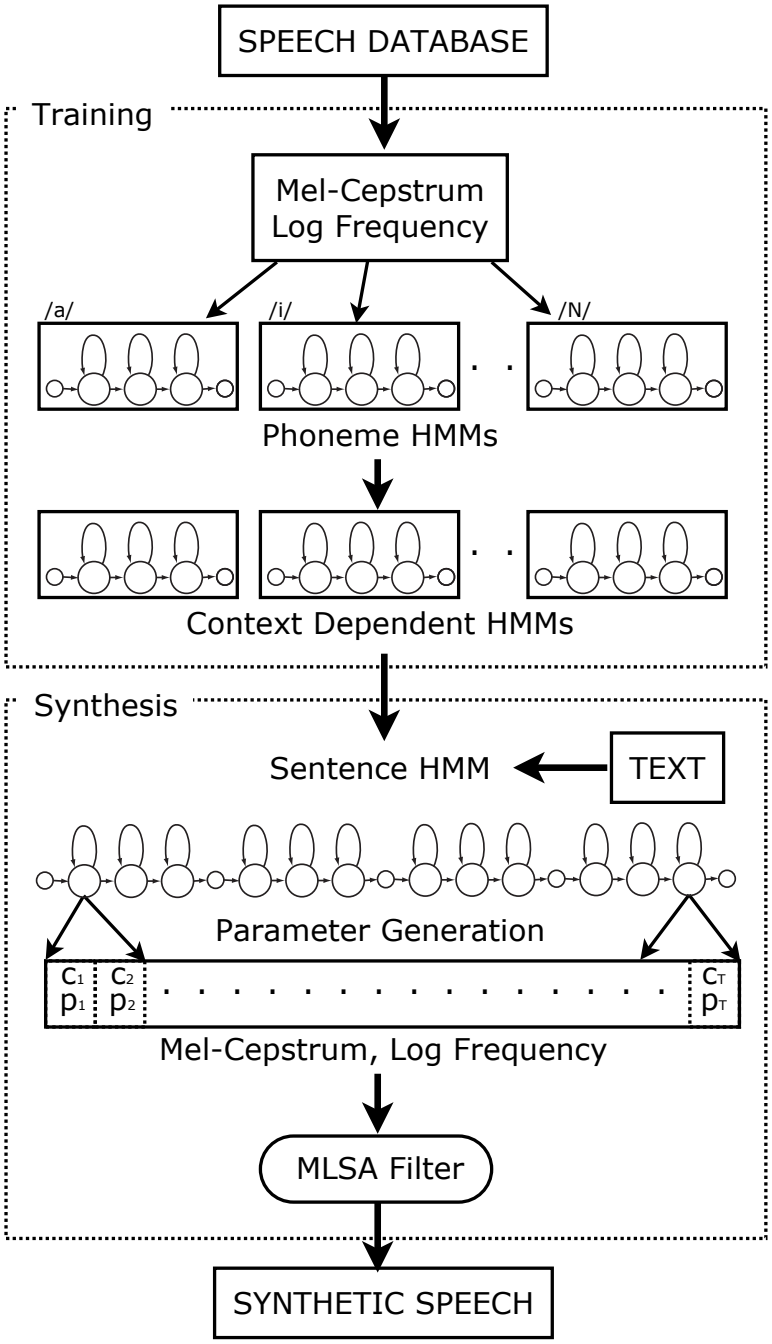

図 2.1　HMM 音声合成の流れ

再現される.

図 2.1 に, HMM 音声合成の流れを示す. 以降, 本章ではこの処理の流れを順に追っていく.

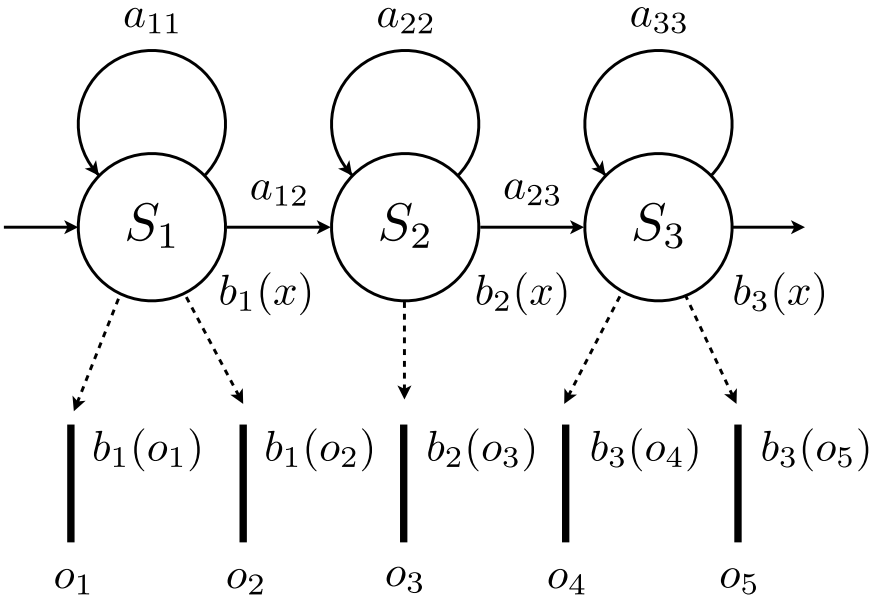

図 2.2　隠れマルコフモデル（HMM）

2.2　隠れマルコフモデル（HMM）

HMM は状態遷移確率と観測シンボルの生起確率分布からなる確率モデルである．図 2.2 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 で音響モデルを構築するとき, 観測シンボルはメルケプストラムや F_0 といった音響的特徴の特徴量ベクトル o_i となる．特徴量ベクトルは連続的なモデル化を行う必要があるため, 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 に基づいた音響モデルは音声認識の分野では既に主流となっており, 音声信号のモデル化における HMM の有効性は広く知られている．

音響モデルの構築とは, 学習用に用意された音声信号コーパスに対し尤度が最大となるよう HMM に学習させることを意味する．学習の対象となるパラメータを $\theta = a_{ij}, b_i(x)$ とおいたとき, HMM の学習は式 (2.1) の尤度最大化問題に帰着される．

$$\hat{\theta} = \arg\max_{\theta} p(\boldsymbol{O}|P, \theta) \tag{2.1}$$

このとき, $\boldsymbol{O}$ は学習コーパスに含まれる音響特徴量ベクトルの時系列であり, P はこれに対応する分節的特徴のコンテキスト情報である．HMM は, 音素などに基づく適切な単位

ごとに複数用意されており, 学習を行う際にはコンテキスト情報 P に対応するよう繋ぎあわせて用いる.

式 (2.1) は, HMM の各状態 S_i と特徴量時系列 $\boldsymbol{O}$ の対応関係が隠れ変数となるため, 解析的に解くことはできない. そこで, Expectation-Maximization アルゴリズムの一種である Baum-Welch アルゴリズムを用いた局所最適解の推定が行われる. Baum-Welch アルゴリズムは, 期待値を計算する Expectation ステップと, 期待値を最大化させるパラメータを求める Maximization ステップから構成される.

Expectation ステップでは, 式 (2.2) と式 (2.3) のように定義される, 前向き確率 $\alpha_t(i)$ と後向き確率 $\beta_t(i)$ をまず求める.

$$\alpha_t(i) = p(o_1, o_2, \ldots, o_t, q_t = S_i | P, \theta) \tag{2.2}$$

$$\beta_t(i) = p(o_{t+1}, o_{t+2}, \ldots, o_T | q_t = S_i, P, \theta) \tag{2.3}$$

ただし, q_t は時刻 t における HMM の状態を表す. 期待値計算の対象となるのは時刻 t における HMM の状態が i である確率 $\gamma_t(i)$ であり, その期待値は式 (2.5) により計算される.

$$\gamma_t(i) = p(q_t = S_i | \boldsymbol{O}, P, \theta) \tag{2.4}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{2.5}$$

Maximization ステップでは, Expectation ステップで求めた値を利用しパラメータの更新を行う. 遷移確率 a_{ij} は式 (2.6) により更新される.

$$\hat{a_{ij}} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{\tau=1}^{T-1} \alpha_\tau(i)\beta_\tau(i)} \tag{2.6}$$

また, 生起確率分布 $b_i(x)$ が単一ガウス分布 $\mathcal{N}(x; \mu_i, \sigma_i^2)$ に従うとすると, 平均 μ_i と分散 σ_i^2 はそれぞれ式 (2.7) と式 (2.8) で更新できる.

$$\hat{\mu_i} = \frac{\sum_{t=1}^{T} \gamma_t(i) o_t}{\sum_{\tau=1}^{T} \gamma_t(i)} \tag{2.7}$$

$$\hat{\sigma_i^2} = \frac{\sum_{t=1}^{T} \gamma_t(i)(o_t - \mu_i)^2}{\sum_{\tau=1}^{T} \gamma_t(i)} \tag{2.8}$$

パラメータ θ と更新されたパラメータ $\hat{\theta}$ の間には式 (2.9) の関係が成り立つ.

$$p(\boldsymbol{O}|P, \theta) \leq p(\boldsymbol{O}|P, \hat{\theta}) \tag{2.9}$$

この事から, Expectation ステップと Maximization ステップを交互に繰り返すことで, パラメータは局所最適解に収束することがわかる.

2.3　音響特徴量ベクトル

統計的な音響モデルの構築を行う際, 音声波形自体の統計量を求めても音響的な意味合いとの直感的な対応がとれないため, 音声波形から音響的な特徴量を抽出する必要が生じる. さらに, 分節的特徴と対応付けたモデル化を行うためには, 音声波形中の局所的な特徴を求めなければならない. そこで行うのが, Source-Filter モデルに基づいた短時間音響特徴量の抽出である.

短時間音響特徴量を抽出するために, まず窓関数による音声信号の切り出しを行う. 窓関数とはある有限区間以外で 0 となる関数である. 式 (2.10) のように定義される矩形窓を使用すると, 単純な音声波形の切り出しとなる.

$$w_R(t) = \begin{cases} 1 & \text{if } 0 \leq \mathrm{t} \leq \mathrm{T} \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

しかし窓関数は, 音声波形を周波数解析する際, 周波数分解能やダイナミックレンジに大きな影響を与えるため, 目的にあった適切なものを使用する必要がある. HMM 音声合成では式 (2.11) のように定義されるハミング窓を使用する.

$$w_H(t) = \begin{cases} 0.54 - 0.46\cos\frac{2\pi t}{T} & \text{if } 0 \leq \mathrm{t} \leq \mathrm{T} \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

Source-Filter モデルでは, 韻律的特徴を表現する音源信号 $x_n(t,p)$ と音韻的特徴を表現する線形フィルタ $h(t,\boldsymbol{c})$ の畳み込み演算の結果として, 音声信号 $s(t)$ を式 (2.12) のように定義する.

$$s(t) = h(t,\boldsymbol{c}) * x_n(t,p) \tag{2.12}$$

$\boldsymbol{c}$ は線形フィルタの係数ベクトルであるメルケプストラムを, p は音源信号の F_0 を, n は音源信号の周期性の有無をそれぞれ表す. 音源信号が周期性を持つのは声帯の振動を音源にして発声を行う場合であり, この時 $x_0(t,p)$ は p を F_0 とする周期パルス列で近似できる. それ以外の場合では, 音源信号が周期性を持つことはなく, $x_1(t)$ は白色雑音で近似される.

まず, 切り出した音声波形中からメルケプストラムを抽出する. 式 (2.16) のように音声波形の対数パワースペクトル $\log|S(\omega)|$ を計算することで, 音源信号の成分と線形フィル

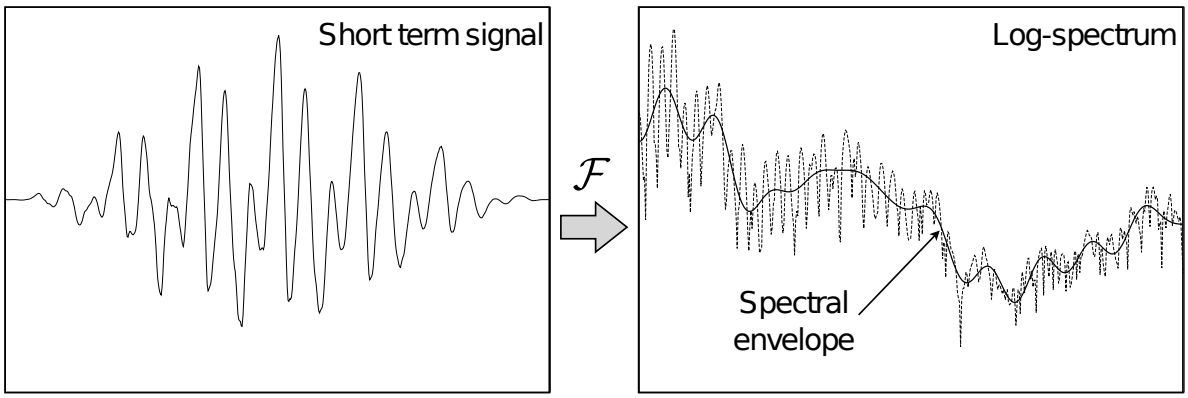

図 2.3　音声波形の対数パワースペクトル

タの成分を分解できる.

$$\log|S(\omega)| = \log|\mathcal{F}[s(t)]| \tag{2.13}$$
$$= \log|\mathcal{F}[h(t,\boldsymbol{c}) * x_n(t,p)]| \tag{2.14}$$
$$= \log|H(\omega)X(\omega)| \tag{2.15}$$
$$= \log|H(\omega)| + \log|X(\omega)| \tag{2.16}$$

図 2.3 は対数パワースペクトルの例である. 音源信号 $x_n(t,p)$ が周期的であるとき, その対数パワースペクトル $\log|X(\omega)|$ も周期信号となる. そのため, 対数パワースペクトル $\log|S(\omega)|$ を低域通過フィルタに通した時に得られるスペクトル包絡が $\log|H(\omega)|$ におよそ対応する. しかし, 音声波形の切り出しに使用した窓関数の窓幅に応じて, スペクトル包絡 $\log|H(\omega)|$ の精度は帯域ごとにばらついてしまう. そこで, 周波数軸を均一な帯域幅に区切り, それぞれの帯域ごとに適切な窓幅の窓関数を使用して対数パワースペクトルを計算し, 得られた対数パワースペクトルから式 (2.17) の三角窓を用いて目的の帯域を切り出し, それらを一つの対数パワースペクトルに足し合わせるというフィルタバンク分析を行う.

$$w_B(\omega) = \begin{cases} 1 - 2|\omega - 0.5| & \text{if } 0 \le \omega \le \Omega \\ 0 & \text{otherwise} \end{cases} \tag{2.17}$$

この時, 三角窓の窓幅は隣接する切り出し区間と半分だけ重複するよう設定する. ただし, 人間の聴覚特性は低周波数域になるほど対数的に分解能が高くなるため, 等間隔にフィルタバンク分析を行うのは効果的とは言えない. そこで, フィルタバンク分析は式 (2.18) のように定義されるメル周波数軸上で行われる.

$$\tilde{f} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.18}$$

こうして得られたスペクトル包絡 $\log|H(2\pi\tilde{f})|$ に対し離散コサイン変換を行った結果がメルケプストラム $\boldsymbol{c}$ となる．なお, 標準的な HMM 音声合成では, 線形フィルタ $h(t, \boldsymbol{c})$ に MLSA フィルタを利用している [16].

次に, 切り出した音声波形中から F_0 を抽出する．F_0 の抽出法は数多く存在するが, ロバスト性の高さから, 自己相関関数法が好んで使用される [17]. 標準的な HMM 音声合成で利用している RAPT アルゴリズム [18] も自己相関関数法の一種である．RAPT アルゴリズムでは, $t = m$ から始まる i 番目の切り出し区間において, 式 (2.19) のように定義される正規化相互相関関数 (NCCF) を計算する．

$$\Phi_i(\tau) = \frac{\sum_{t=m}^{m+n-1} s(t)s(t+\tau)}{\sqrt{e_m e_{m+\tau}}} \tag{2.19}$$

ここで, n は切り出し区間の幅であり, e_j は式 (2.20) である．

$$e_j = \sum_{t=j}^{j+n-1} s^2(t) \tag{2.20}$$

$\Phi_i(\tau)$ が極大となる遅れ時間 τ の逆数が F_0 の候補値となる．切り出し区間の幅は, メルケプストラムを抽出するときと同様に, F_0 抽出の精度に大きな影響を与える．そこで RAPT では, 候補の中で $\Phi_i(\tau)$ が最大となるものを仮の F_0 の推定値とし, その推定値に合わせた切り出し幅でもう一度 NCCF を計算する．二度目の NCCF から得られた F_0 の候補値から動的計画法によって推定値を決定する．各候補値の選択コスト d_{ij} は式 (2.21) のように定義される．

$$d_{ik} = 1 - \Phi_i(\tau_{ik})(1 - 0.3\frac{\tau_{ik}}{\tau_{\max}}) \tag{2.21}$$

τ_{ik} は i 番目の切り出し区間における k 番目の候補値の遅れ時間であり, $\tau_{\max}$ は NCCF 中からの極大探索における τ の上限である．また, 切り出し波形に周期性が無い場合もあるため, その選択コストとして式 (2.22) を加える．

$$d_{i0} = \max_k \Phi_i(\tau_{ik}) \tag{2.22}$$

候補 l から候補 k への遷移コストには式 (2.23) を利用する．

$$\delta_{ikl} = 0.02 \cdot \min[\xi_{kl}, 0.35 + |\xi_{kl} - \log 2|] \tag{2.23}$$

$$\xi_{kl} = \left|\log \frac{\tau_{ik}}{\tau_{(i-1)l}}\right| \tag{2.24}$$

周期性を持たない状態を含む場合の遷移コストは式 (2.25) から式 (2.27) である.

$$\delta_{i0l} = 0.005 + 0.5 \cdot \frac{0.2}{d_S(i) - 0.8} + 0.5 \cdot r_{\mathrm{RMS}}(i) \tag{2.25}$$

$$\delta_{i00} = 0 \tag{2.26}$$

$$\delta_{ik0} = 0.005 + 0.5 \cdot \frac{0.2}{d_S(i) - 0.8} + \frac{0.5}{r_{\mathrm{RMS}}(i)} \tag{2.27}$$

$d_S(i)$ は式 (2.28) のように定義される最尤スペクトル距離 [19] である.

$$d_S(i) = \log\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|S_i(\omega)|}{|S_{i-1}(\omega)|}d\omega\right) \tag{2.28}$$

$|S_i(\omega)|$ は切り出された音声波形のパワースペクトルであり, 自己相関関数の係数を用いて Yule-Walker 方程式を解くことで得られる LPC 係数 a_k から, 式 (2.29) のように求まる.

$$|S_i(\omega)| = \frac{1}{|1 + \sum_k a_k e^{-j\omega k}|^2} \tag{2.29}$$

また, $r_{\mathrm{RMS}}(i)$ は式 (2.30) のように定義される.

$$r_{\mathrm{RMS}}(i) = \sqrt{\frac{\sum_{t=0}^{n-1}\left(w(t)s(t + m_i + 0.5n)\right)^2}{\sum_{t=0}^{n-1}\left(w(t)s(t + m_{i-1} - 0.5n)\right)^2}} \tag{2.30}$$

$w(t)$ はハン窓と呼ばれる式 (2.31) のような窓関数である.

$$w(t) = \begin{cases} 0.5 - 0.5\cos\frac{2\pi t}{T} & \text{if } 0 \leq \mathrm{t} \leq \mathrm{T} \\ 0 & \text{otherwise} \end{cases} \tag{2.31}$$

以上のコスト関数を用いて, 式 (2.32) を動的計画法により解くことで F_0 の推定値の選択は行われる.

$$D_{ik} = \begin{cases} 0 & \text{if i} = 1 \\ d_{ij} + \min\limits_{k}[D_{(i-1)k} + \delta_{ijk}] & \text{otherwise} \end{cases} \tag{2.32}$$

そして最後に, 以上のようにして求めたメルケプストラムと F_0 の動的特徴量を求める. メルケプストラム $\boldsymbol{c}_i$ における m 次の係数 c_{im} の動的特徴量は, 式 (2.33) と式 (2.34) より求められる.

$$\Delta c_{im} = \frac{\partial^2 c_{im}}{\partial i} \tag{2.33}$$

$$\Delta^2 c_{im} = \frac{\partial^2 c_{im}}{\partial i^2} \tag{2.34}$$

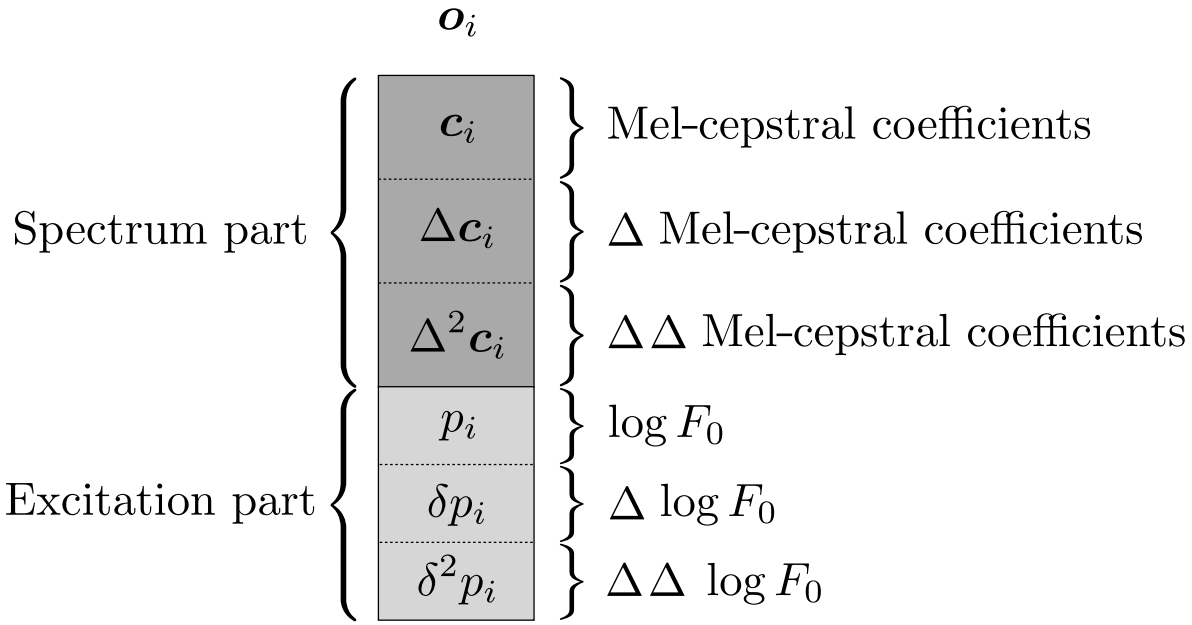

図 2.4　音響特徴量ベクトルの構成

F_0 についても動的特徴量の計算方法は同じであるが, HMM によるモデル化の対象となるのは対数 F_0 であるため, 動的特徴量の算出は対数 F_0 に対して行われる.

こうして求めた音響特徴量から, 音響特徴量ベクトルは図 2.4 のように構成される. この特徴量ベクトルが, HMM によるモデル化の対象であり, 生起確率分布 $b_n(o_i)$ に基づいて出力される観測シンボルである. ところで, 式 (2.33) と式 (2.34) はそれぞれ式 (2.35) と式 (2.36) により近似できるため,

$$\Delta c_{im} \approx 0.5c_{(i+1)m} - 0.5c_{(i-1)m} \tag{2.35}$$

$$\Delta^2 c_{im} \approx c_{(i+1)m} - 2c_{im} + c_{(i-1)m} \tag{2.36}$$

動的特徴量の計算は式 (2.37) のように行列形式で記述できる.

$$\boldsymbol{O}_C = \boldsymbol{W}\boldsymbol{C} \tag{2.37}$$

ただし, $\boldsymbol{C}$ はケプストラム係数の時系列, $\boldsymbol{O}_c$ は音響特徴量ベクトルのケプストラム部の時系列であり, それぞれ式 (2.38), 式 (2.39) となる.

$$\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_i, \ldots, \boldsymbol{c}_I]^\top \tag{2.38}$$

$$\boldsymbol{O}_c = \left[\boldsymbol{o}_{c1}^\top, \boldsymbol{o}_{c2}^\top, \ldots, \boldsymbol{o}_{ci}^\top, \ldots, \boldsymbol{o}_{cI}^\top\right]^\top \tag{2.39}$$

$$\boldsymbol{o}_{ci} = [\boldsymbol{c}_1, \Delta\boldsymbol{c}_1, \Delta^2\boldsymbol{c}_1]^\top \tag{2.40}$$

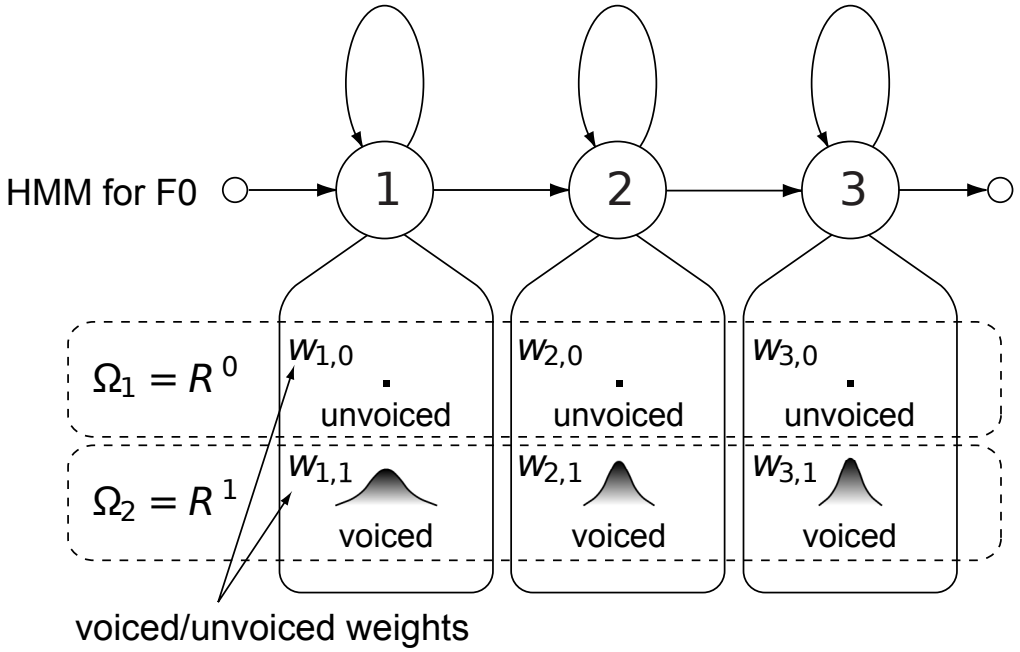

図 2.5　多空間確率分布 HMM

また, 行列 $\boldsymbol{W}$ は式 (2.41) のように定義される.

$$\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_i, \ldots, \boldsymbol{w}_I] \tag{2.41}$$

$$\boldsymbol{w}_i = \left[\boldsymbol{w}_i^{(0)}, \boldsymbol{w}_i^{(1)}, \boldsymbol{w}_i^{(2)}\right] \tag{2.42}$$

$$\begin{aligned}\boldsymbol{w}_i^{(n)} = \Bigg[& \underbrace{\boldsymbol{0}_{M\times M}}_{1\text{st}}, \ldots, \boldsymbol{0}_{M\times M}, \underbrace{w^{(n)}(-L^{-(n)})\boldsymbol{E}_{M\times M}}_{(i-L^{-(n)})-\text{th}}, \\ & \ldots, \underbrace{w^{(n)}(0)\boldsymbol{E}_{M\times M}}_{i-\text{th}}, \ldots, \\ & \underbrace{w^{(n)}(L^{+(n)})\boldsymbol{E}_{M\times M}}_{(i+L^{+(n)})-\text{th}}, \underbrace{\boldsymbol{0}_{M\times M}, \ldots, \boldsymbol{0}_{M\times M}}_{\text{I}-\text{th}} \Bigg]\end{aligned} \tag{2.43}$$

$w^{(n)}(\tau)$ は式 (2.35) や式 (2.36) における n 次微分計算のための係数である. 実際の動的特徴量の計算は, 上述した行列計算によって行われる.

2.4　多空間確率分布HMM

音声波形中には F_0 を定義できない区間が存在するため, 通常の HMM では F_0 のモデル化を行うことができない. そこで HMM 音声合成では, コンテキスト依存 HMM として, 図 2.5 に示すような多空間確率分布 HMM (MSD-HMM) を利用する [21]. MSD-HMM

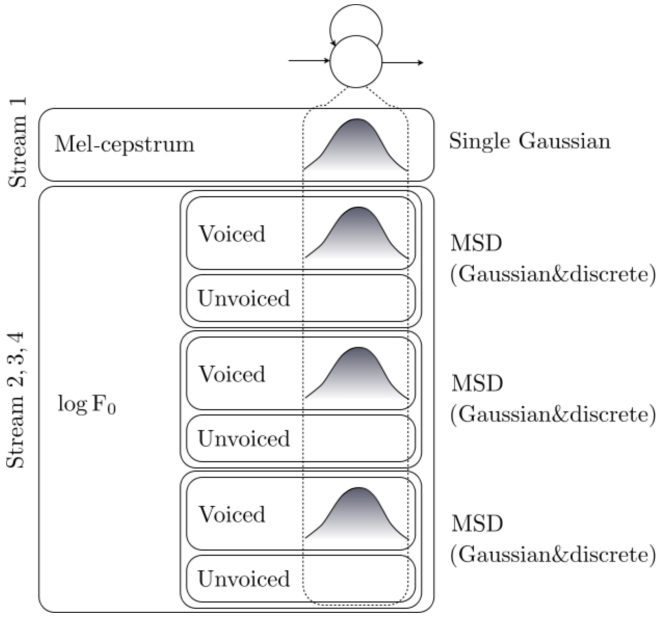

図 2.6　音響特徴量ベクトルのマルチストリーム化

は, 離散的なシンボル列をモデル化する離散分布 HMM と, 連続値を持つベクトル系列をモデル化する連続分布の二つを同時に内包する HMM であり, F_0 パターンを, 周期を持つ区間に対応し F_0 値を出力する空間 Ω_2 と, 周期を持たない区間に対応する空間 Ω_1 の二つの空間から観測される事象としてモデル化する. 観測シンボルを $\boldsymbol{o}_p = [n, [p, \delta p, \delta^2 p]^\top]$ としたとき, 状態 i における $\boldsymbol{o}_p$ の生起確率 $b_i(\boldsymbol{o}_p)$ は式 (2.44) となる.

$$b_i(\boldsymbol{o}_p) = \sum_{n \in \{1,2\}} w_{in} \mathcal{N}_{in}(V(\boldsymbol{o}_p)) \tag{2.44}$$

このとき, $V(\boldsymbol{o}_p) = [p, \delta p, \delta^2 p]^\top$ であり, w_{in} は各空間に対する重みである.

音響特徴量ベクトルの生起確率分布の学習は, 図 2.6 に示すように, 特徴量ベクトルを四つのストリームに分割して行う. 音韻的特徴と韻律的特徴の時間的対応が取れるようにメルケプストラムと $\log F_0$ を単一のベクトルにまとめているが, これらは独立性の高い事象であるため, 生起確率分布の学習は別々に行われる. また, HMM 音声合成では, $\log F_0$ の空間選択の重み w_{in} の大小関係だけで周期性の有無を決定しているため, ロバスト性の低い動的特徴量とは独立させて $\log F_0$ のモデリングを行う. 以上より, 特徴量ベクトル $\boldsymbol{o}$ の正規確率分布 $b_i(\boldsymbol{o})$ は式 (2.45) のように定義できる.

$$b_i(\boldsymbol{o}) = \prod_s \left(b_{is}(\boldsymbol{o}_s)\right)^{w_s} \tag{2.45}$$

ここで, s はストリームの添字, w_s はストリームの重みである.

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

図 2.7　コンテキストラベルの構成

2.5　コンテキスト依存モデル

音響特徴量ベクトルのモデル化に用いる HMM は, 基本的には音素ごとに分類されるが, 前後環境や言語的な構造に依存した精度の高いモデル化を行うため, コンテキストラベルと呼ばれる図 2.7 のような分節的特徴のコンテキストを加味した音素表記に基づいてより詳細に分類される [20]. 表 2.1 はコンテキストラベルの表記に含まれるコンテキスト情報の一覧である. 文中の位置についてはモーラ (拍) を単位として記述される. コンテキストラベルは音声波形ごとにあらかじめ学習コーパスに用意されており, そのままコンテキストラベルごとに HMM を用意して音響モデルを構築することも可能であるが, 詳細すぎる分類が過学習を引き起こすことがあるため, 一旦単純な音素 HMM でモデル化を行った結果をコンテキスト依存 HMM に変換するといった手順を取る.

コンテキストの組み合わせは膨大であり, そのすべてを網羅することは現実的ではない. 学習コーパスから構築可能なコンテキスト依存 HMM だけで任意の文章から音声合成を行う方法として, 決定木に基づくコンテキストクラスタリングが挙げられる [22]. ある 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, その分割を適応するか否かを決定する基準となるのは式 (2.46) に示す最小記述長 (MDL) 基準である.

$$\Delta_q = \frac{1}{2}\left\{\Gamma(S_{q+})\log|\Sigma_{S_{q+}}| + \Gamma(S_{q-})\log|\Sigma_{S_{q-}}| - \Gamma(S)\log|\Sigma_S|\right\} + K\log|\Gamma_{S_0}| \quad (2.46)$$

$\Gamma(S)$ 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. このように構築された決定木は, 単に未知のコンテキストラベルに対応する HMM の決定にだけでなく, パラメータ共有によるロバスト性の高い HMM の学習にも利用される.

表 2.1　コンテキストの分類

p_L	先行音素
p_C	当該音素
p_R	後続音素
a_{C1}	アクセント句内モーラ位置 (単位:モーラ)
a_{C2}	アクセント型とモーラ位置との差 (単位:モーラ)
b_{L1}	先行品詞 ID
b_{L2}	先行品詞の活用形 ID
b_{L3}	先行品詞の活用型 ID
b_{C1}	当該品詞 ID
b_{C2}	当該品詞の活用形 ID
b_{C3}	当該品詞の活用型 ID
b_{R1}	後続品詞 ID
b_{R2}	後続品詞の活用形 ID
b_{R3}	後続品詞の活用型 ID
c_{L1}	先行アクセント句の長さ (単位:モーラ)
c_{L2}	先行アクセント句のアクセント型
c_{L3}	先行アクセント句と当該アクセント句の接続強度
c_{L4}	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
c_{C1}	当該アクセント句の長さ (単位:モーラ)
c_{C2}	当該アクセント句のアクセント型
c_{C3}	先行アクセント句と後続アクセント句の接続強度
c_{C4}	当該呼気段落でのアクセント句の位置
c_{C5}	疑問文かそうでないか
c_{R1}	後続アクセント句の長さ (単位:モーラ)
c_{R2}	後続アクセント句のアクセント型
c_{R3}	後続アクセント句と当該アクセント句の接続強度
c_{R4}	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
d_{L1}	先行呼気段落の長さ (単位:モーラ)
d_{C1}	当該呼気段落の長さ (単位:モーラ)
d_{C2}	文中での当該呼気段落の位置
d_{R1}	後続呼気段落の長さ (単位:モーラ)
e	文の長さ (単位:モーラ)

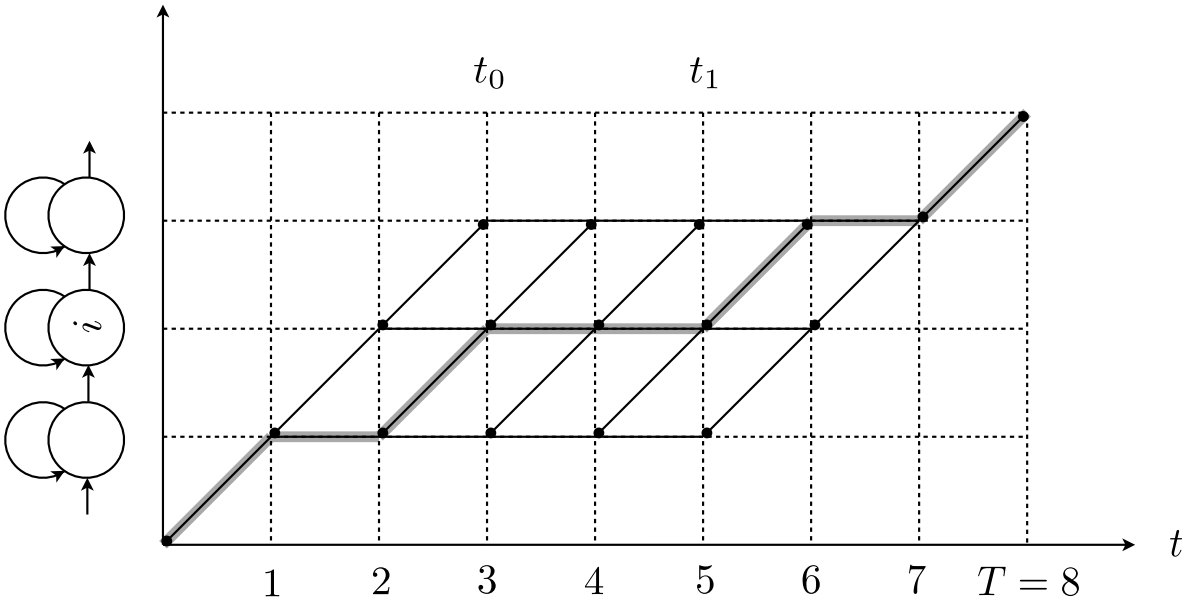

図 2.8　状態継続長の推定

2.6　状態継続長モデル

音響特徴量ベクトルのモデル化と並行して状態継続長の学習も行われる. ガウス分布で状態継続長をモデル化する場合, 図 2.8 において, 状態 i における状態継続長の平均値 $\xi(i)$ と分散 $\sigma^2(i)$ は, それぞれ式 (2.47) と式 (2.48) となる [23].

$$\xi(i) = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0 t_1}(i)(t_1 - t_0 + 1)}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0 t_1}(i)} \tag{2.47}$$

$$\sigma^2(i) = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0 t_1}(i)(t_1 - t_0 + 1)^2}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0 t_1}(i)} - \xi^2(i) \tag{2.48}$$

$$\chi_{t_0 t_1}(i) = (1 - \gamma_{t_0-1}(i)) \cdot \prod_{t=t_0}^{t_1} \gamma_t(i) \cdot (1 - \gamma_{t_1+1}(i)) \tag{2.49}$$

この時, $\chi_{t_0 t_1}(i)$ は, 時点 t_0 から t_1 の間, 状態 i にとどまり続ける確率であり, $\gamma_t(i)$ は, 式 (2.5) で定義した, 時点 t において状態が i である確率である. 連続分布による状態継続長のモデル化は合成音声の話速制御を容易にする. また, 特徴量ベクトルのモデル化の過程においてモデル化を行えてしまうため, 状態継続長の学習用に別途ラベルデータを用意する必要がないことも, この手法の長所である.

2.7　系列内変動（ＧＶ）

統計量ベース方式では, 音響特徴量が連続的にモデル化されるため, 不連続感の少ない合成音声が生成可能となる．さらに HMM 音声合成では, 先述したコンテキストクラスタリングにより, 似た属性や音響的特徴をもつ音響特徴量ベクトルを平均化して用いるため, 個々の特徴量ベクトルが持つ情報が共有され, 小規模の学習コーパスからでも高品質な音声が合成できる．また, 特徴量ベクトルの平均化は信号処理に対するロバスト性も向上させる [24]．しかしながら, 平均化処理は, 音声波形が持つ詳細な特徴を除いてしまい, 合成音声の肉声感は損なわれることになる．この点が, サンプルベース方式が統計量ベース方式より好んで用いられる理由となっている．

HMM に基づいて構築された音響モデルは, 時間方向に離散的な状態系列からなり, それぞれの状態ごとに平均化処理が行われる．音響的特徴の動的特徴量のモデル化は, こうした平均化処理の一つの課題を解決し, 時間方向に滑らかな音響特徴量の生成を実現した．しかし, 特徴量系列全体に渡って詳細な特徴を再現するには至っておらず, 肉声感は過剰に平滑化されたままである．コンテキスト依存モデルをより複雑にすることで平均化処理の影響を低減させることもできるが, HMM の状態数の増加は過学習を引き起こす要因ともなる．過学習は, 特に動的特徴量のモデル化で問題となり, 時間方向に滑らかな音響特徴量の生成を阻害する．

そこで, 特徴量系列全体に渡る変動をモデル化することを考える．音響特徴量ベクトル系列 $\boldsymbol{c}$ の系列内変動 $\boldsymbol{v}(\boldsymbol{c})$ は発話ごとに式 (2.50) によって計算される．

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

$$v(d) = \frac{1}{I}\sum_{i=1}^{I}\left(c_{id} - \bar{c}_d\right)^2 \tag{2.51}$$

$$\bar{c}_d = \frac{1}{I}\sum_{i=1}^{I} c_{id} \tag{2.52}$$

系列内変動を考慮した音声合成とは, 尤度最大化基準に基づいて音響特徴量系列を生成する際, 系列内変動 $\boldsymbol{v}(\boldsymbol{c})$ の尤度についても考慮することを意味する．詳細な特徴を持つ自然音声は, 過剰に平滑化された合成音声よりも変動が大きいことが予想されるが, $\boldsymbol{v}(\boldsymbol{c})$ の尤度についても考慮することは, 合成音声の特徴量系列の変動を大きくすることに相当し, ひいては合成音声の肉声感の向上へとつながる．

2.8　音響特徴量生成アルゴリズム

先述のように構築した音響モデル θ に対し入力文章に対応したコンテキスト情報 P が与えられたとき, 音響特徴量系列 $\boldsymbol{O}$ の尤度は式 (2.53) となる.

$$p(\boldsymbol{O}|P,\theta) = \sum_{\text{all}\boldsymbol{Q}} p(\boldsymbol{O},\boldsymbol{Q}|P,\theta) \tag{2.53}$$

ここで $\boldsymbol{Q}$ は, 式 (2.54) のように HMM の状態を表す系列である.

$$\boldsymbol{Q} = \{(i_1,n_1),(i_2,n_2),\ldots,(i_t,n_t),\ldots,(i_T,n_T)\} \tag{2.54}$$

(i_t,n_t) は, 時点 t における HMM の状態の添字 i_t と空間の添字 n_t を表す. i_t の時系列 $\boldsymbol{i} = [i_1,i_2,\ldots,i_t,\ldots,i_T]$ は, 状態継続長モデル θ_d から, 尤度 $p(\boldsymbol{i}|P,\theta_d)$ を最大化させることで求まる. 一方, n_t の時系列 $\boldsymbol{n} = [n_1,n_2,\ldots,n_t,\ldots,n_T]$ は重みの大きい空間を選択することで求まる. こうした手順を踏むことなく, 尤度 $p(\boldsymbol{O}|P,\theta)$ の最大化によって, 直接的に音響特徴量系列 $\boldsymbol{O}$ を出力することも可能ではあるが, 計算量を削減するため, こうした準最適な状態系列 $\boldsymbol{q}$ の決定を行っている.

ここで, 系列内変動 $\boldsymbol{v}(\boldsymbol{C})$ について考慮しながら, ケプストラム系列 $\hat{\boldsymbol{C}}$ を合成すること考えると, 式 (2.55) のような尤度最大化を行えばよいことがわかる.

$$\hat{\boldsymbol{C}} = \arg\max_{\boldsymbol{c}} \ \log\{p(\boldsymbol{O}_C(\boldsymbol{C})|P,\theta,\boldsymbol{Q})^{\omega}\cdot p(v(\boldsymbol{C})|\theta_v)\} \tag{2.55}$$

$$= \arg\max_{\boldsymbol{c}} \ \log\{p(\boldsymbol{W}\boldsymbol{C}|P,\theta,\boldsymbol{Q})^{\omega}\cdot p(v(\boldsymbol{C})|\theta_v)\} \tag{2.56}$$

ただし, 定数 ω は尤度のバランスの操作に用いられる重み付け係数であり, その値は $\boldsymbol{O}$ と $v(\boldsymbol{p})$ の次元数の比率によって決定される. 式 (2.55) の計算には, 最急降下法やニュートン法のような勾配法が使用される. $p(v(\boldsymbol{C})|\theta_v)$ は系列内変動の減少に対する制約として機能する. ここではケプストラム系列を例に音響特徴量生成アルゴリズムの説明を行ったが, F_0 パターンの生成を行う場合も同様の手順を用いる. HMM 音声合成では, 合成音声の自然性を向上させるための制約として, 動的特徴量と系列内変動の統計的なモデル化を行っているが, ケプストラム系列を生成する際にはどちらの制約も大幅な改善をもたらしているのに対し, F_0 パターンを生成する際は主に動的特徴量に関する制約が自然性の向上に貢献している.

2.9　まとめ

本章では, 従来の HMM 音声合成の基礎技術について紹介を行った. HMM 音声合成では, 音響的特徴と分節的特徴が自動的に対応付けられるため, 分節的特徴に基づいた柔軟な発話スタイルの制御が可能となることを説明した. また, 動的特徴量や系列内変動に関する制約を導入したことによって, HMM 音声合成のような統計量ベース方式の音声合成システムで予てから問題となっていた合成音声の自然性の低下が抑制されることを示した.

第 3 章

基本周波数パターン生成過程モデル

3.1　はじめに

音声の生成過程は, 言語的情報を伝えるためのメッセージを計画する段階, メッセージに付与する意図や態度を計画する段階, 計画に沿った音声波を発声すべく音声器官の運動指令を生成する段階, そして制御指令に従い音声波を生成する段階の四つの段階にわけて考えることができる. メッセージの計画は分節的特徴を持つ文法規則に従い, 意図や態度の計画は超分節的特徴を持つ韻律規則に従うが, 運動指令を生成する段階において二つの規則の役割は分離したものではなくなり, 音声波を生成する段階において生成される音声波形には分節的特徴と超分節的特徴が同時に含まれる.

音声波形に含まれる各種音響的特徴と, それらが担う情報との対応関係を定量的に把握することは, 音声言語情報処理を行う上で極めて重要となる. しかし, 処理を行う際に観測できるのは分節的特徴と超分節的特徴が混在した音声信号であり, そこから遡って各種の情報を推定することは至難である. 特に韻律的特徴に関して言えば, 音声波形からの運動指令の推定と, 推定された運動指令からの音声波形に付与された意図や態度などの推定, という二つの段階を踏む必要がある.

F_0 パターン生成過程モデル (F_0 モデル) は, 生理学的 · 物理学的な知見に基づき, 韻律的特徴の一つである F_0 パターンの生成過程を数理的にモデル化したものであり, フレーズ指令とアクセント指令と呼ばれる二つの指令によって制御される. これらの指令は, 音声器官の運動指令に相当し, 音声波形に含まれる分節的特徴や超分節的特徴と密接な関係を持つ [25]. 観測された F_0 パターンからのフレーズ指令やアクセント指令のパラメータ推定は, 解析的には解けない逆問題となってしまうため, Analysis-by-Synthesis による逐次近似により行われる. そのため, 最終的に得られるパラメータの精度は初期値に大きく依存してしまうが, F_0 パターンに対する区分的三次曲線近似の結果から解析的に初期値を求めることで, 高い精度でのパラメータの推定が可能となる [26].

一方, F_0 モデルの各指令と分節的特徴のコンテキスト情報との対応関係から, パラメータ推定を行う手法も提案されている [27]. Tone and Break Indices (ToBI) システム [28] は, 韻律構造とイントネーションパターンを記述することを目的としたラベルであり, F_0 モデルの初期値を求める際に必要となる情報を含んでいる.

本章では, 音声の生成過程における F_0 モデルの位置付け, F_0 モデルの生理学的 · 物理学的な背景, F_0 モデルのパラメータ抽出法, F_0 モデルの各指令とコンテキスト情報との対応関係について順に説明を行う.

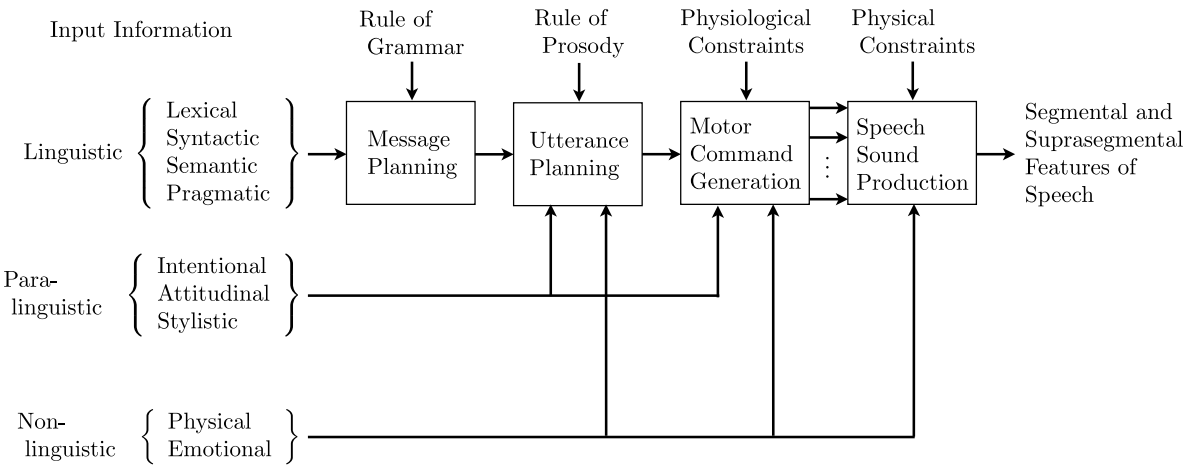

図 3.1　音声による情報の表出課程

3.2　音声による情報の表出課程

音声波形には, 言語的情報, パラ言語的情報, 非言語的情報の三種類の情報が含まれており, それぞれ図 3.1 に示す過程に従い音声波形に付与される. 言語的情報とは, 単語の読みや統語構造といった, 言語によって規定される分節的な情報のことを指す. 一方, パラ言語情報は意図や感情を表現する情報であり, 同じ範疇に分類される情報であっても量的な差を持ちうる点が分節的な情報との違いである. そして, 非言語的情報は, 上記二つとは異なり, 一般的には意識的な選択や制御が不可能な, 話者の個人的な特徴を表現する.

文法規則に基づき統合された言語的情報に対しパラ言語的情報を付加する段階において, 韻律を踏まえた発話の計画が立てられる. 韻律とは, 複数の言語単位にまたがって発話に一貫性を持たせるものとして定義される. 言語的情報には韻律的特徴としてアクセントが含まれているが, これらは発話計画の段階において韻律規則に基づき結合される. また, 言語的情報の句切りの単位となる韻律的特徴であるフレーズについても, アクセント同様に, 韻律規則に基づき決定される. 言語的情報に対するパラ言語的情報や非言語的情報の付加は, アクセントやフレーズに対しての情報付加という形で実現される.

F_0 モデルは, 上述のように決定されたアクセントやフレーズの情報と, 物理的な音響特徴量である F_0 パターンとを対応付ける生理学的過程を記述した数理モデルであると言える. F_0 モデル上において, 言語的情報やパラ言語的情報がアクセントやフレーズといった形でのみ扱われるのに対し, 非言語的特徴については直接的な物理量としても出現する.

3.3　咽頭制御機構の物理的特性

F_0 パターン生成の生理学的過程を考える上で, まず咽頭の物理的特性について考える. 声帯筋を含む骨格筋の長さ l と張力 T の関係は, 式 (3.1) で近似できることが知られている.

$$\frac{dT}{dl} = a + bT \tag{3.1}$$

これを解くと式 (3.2) が求まる.

$$T = \left(T_0 + \frac{a}{b}\right) e^{b(l-l_0)} - \frac{a}{b} \tag{3.2}$$

ただし, l_0 は骨格筋の長さ l の初期値である. ここで $T \ll \frac{a}{b}$ と仮定すると, 式 (3.2) は式 (3.3) で近似できる.

$$T \simeq T_0 e^{bx} \tag{3.3}$$

ただし, x は声帯の伸び $(l - l_0)$ である.

一方, 任意の弾性膜の基本周波数は式 (3.4) であるため,

$$F_0 = c_0 \sqrt{\frac{T}{\sigma}} \tag{3.4}$$

声帯を音源とする音声波形の F_0 は式 (3.5) となる.

$$\log F_0 = \log \left(c_0 \sqrt{\frac{T_0}{\sigma}} \right) + \frac{b}{2} x \tag{3.5}$$

ただし, σ は膜の密度, c_0 は膜の大きさによって決まる定数である. 対数 F_0 が声帯の伸び x に比例して変化する成分を持つことは立体内視鏡を用いた観察によって確かめられており, x が時間的に変化する場合も式 (3.6) に示すように成り立つ.

$$\log F_0(t) = \log F_b + \frac{b}{2} x(t) \tag{3.6}$$

ここで, $F_b = c_0 \sqrt{T_0/\sigma}$ は, 話者の骨格筋の物理的性質によって決まる固定値であり, F_0 パターンの基底値となる.

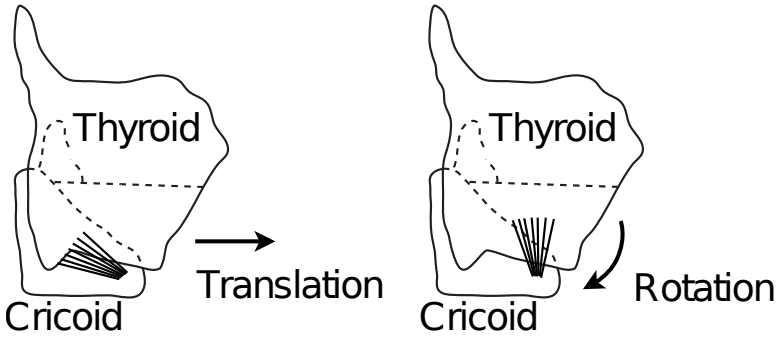

図 3.2　輪状甲状筋による声帯の長さの制御

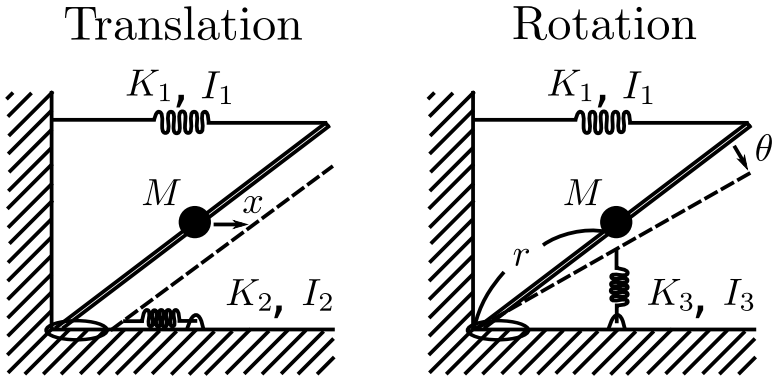

図 3.3　二次線形系による輪状甲状筋の近似

3.4　咽頭制御機構のモデル化

音声波形の F_0 の制御において最も大きな役割を果たすのは声帯の長さの変化であり, 声帯の長さの変化は, 輪状甲状筋 (*cricothyroid muscle*) の働きによる甲状軟骨 (*thyroid cartilage*) と輪状軟骨 (*cricoid cartilage*) の相対位置の変化によって生じる. 図 3.2 は, 甲状軟骨の輪状軟骨を基準とした平行移動と回転の様子を表している. 甲状軟骨の平行移動は輪状甲状筋の斜部 (*pars obliqua*) の収縮によって, 回転は直部 (*pars recta*) の収縮によってそれぞれ生じる. これらの運動は図 3.3 のように 2 次線形系で近似することが可能である. このとき, 輪状甲状筋斜部の収縮速度が系の応答速度と比較して十分に大きく, かつその持続時間が十分に短ければ, 声帯の長さの微小変化分 $x_1(t)$ は系のインパルス応答で近似できる. 一方, 輪状甲状筋直部の収縮速度が系の応答速度と比較して十分に大きく, またその収縮が持続的なものであるとすれば, 甲状軟骨の回転によって生じる声道長の微小変化分 $x_2(t)$ は系のステップ応答で近似できる. そして, $x_1(t)$ と $x_2(t)$ が互いに独立と

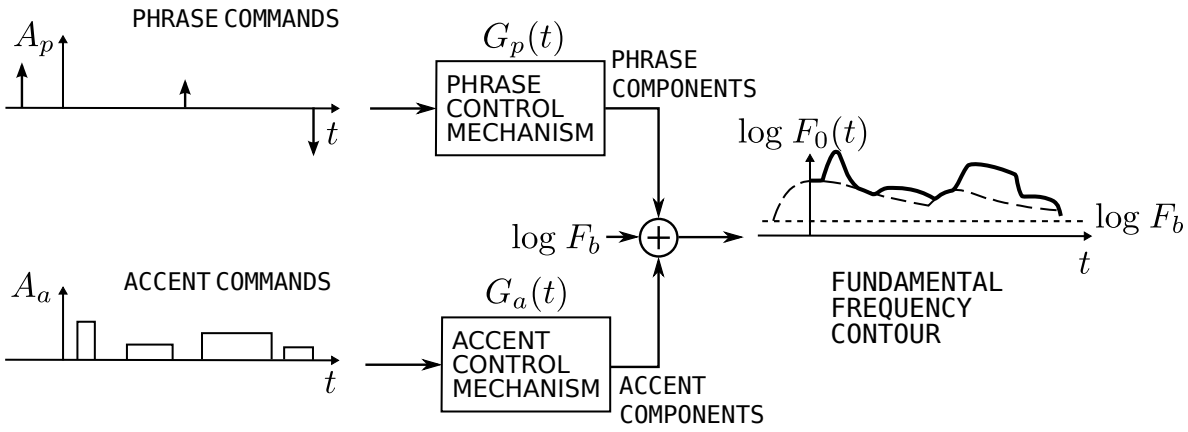

図 3.4　基本周波数パターン生成過程モデル

みなせる範囲において, 声帯を音源とする音声波形の F_0 は式 (3.7) となる.

$$\log F_0(t) = \log F_b + \frac{b}{2}\left(x_1(t) + x_2(t)\right) \tag{3.7}$$

なお, 甲状軟骨の平行移動の時定数は, 回転の時定数よりもはるかに大きいため, $x_1(t)$ はフレーズ単位の比較的緩やかな音調の表現に, $x_2(t)$ はアクセントのように急激で局所的な音節単位の音調の表現に用いられることが多い.

図 3.4 は, 上述した F_0 パターン生成の生理的 · 物理的過程を簡略化したモデルである. ここでは, 輪状甲状筋斜部の瞬間的な活動をインパルス関数で近似して表したものをフレーズ指令と定義し, 輪状甲状筋直部の持続的な活動をステップ関数で近似して表したものをアクセント指令と定義する. これらはそれぞれ, 甲状軟骨の平行移動を模した二次線形系であるフレーズ制御機構と, 甲状軟骨の回転を模した二次線形系であるアクセント制御機構に入力される. そして, それぞれの出力であるフレーズ成分とアクセント成分を足しあわせたものに, さらに F_0 パターンの基底周波数である $\log F_b$ を加えたものが, 式 (3.8) で表される最終的な F_0 パターン $\log F_0(t)$ となる.

$$\begin{aligned}\log F_0(t) = \log F_b &+ \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) \\ &+ \sum_{j=1}^{J} A_{aj}\left\{G_a(t - T_{1j}) - G_a(t - T_{2j})\right\}\end{aligned} \tag{3.8}$$

ここで, $G_p(t)$ はフレーズ制御機構のインパルス応答, $G_a(t)$ はアクセント制御機構のス

テップ応答であり, それぞれ式 (3.9) と式 (3.10) のように定義される.

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{3.9}$$

$$G_a(t) = \begin{cases} \min[1-(1+\beta t)e^{-\beta t}, \gamma] & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{3.10}$$

また, I は発話中のフレーズ指令の数, J は発話中のアクセント指令の数, A_{pi} は第 i 番目のフレーズ指令の大きさ, A_{aj} は第 i 番目のアクセント指令の振幅, T_{0i} は第 i 番目のフレーズ指令の生起時点, T_{1j} は第 j 番目のアクセント指令の始点, T_{2j} は第 j 番目のアクセント指令の終点, α はフレーズ制御機構の固有角周波数, β はアクセント制御機構の固有角周波数, γ はアクセント成分の相対飽和レベルをそれぞれ表す. 式 (3.9) と式 (3.10) は, 図 3.3 の二次線形系モデルを臨界制動系と仮定して求めたものである. 図 3.3 が臨界制動系となる確証は無いが, 実験的にこの近似が成り立つことが確かめられている. また, α や β は, 話者の声帯の物理的性質によって決まる定数であり, 厳密には話者ごとに異なる値ではあるが, 個人差や言語の違いによる差が小さいことが実験的に確かめられており, $\alpha = 3.0[\mathrm{s}^{-1}]$, $\beta = 20[\mathrm{s}^{-1}]$ として扱うことができる. 同様に, $G_a(t)$ が飽和する値である γ も話者や言語によらず 0.9 として扱うことができる.

3.5　モデルパラメータの推定

観測された F_0 パターンからのフレーズ指令やアクセント指令のパラメータ推定は, 解析的には解けない逆問題ではあるものの, 適切な初期値から Analysis-by-Synthesis による逐次近似を行うことで実現可能となる. このとき, 正確な初期値抽出を行うため, F_0 パターンに対し, F_0 抽出ミスの修正, マイクロプロソディの除去, 周期性を持たない無声区間の補間, F_0 パターンの平滑化を前処理として順に行っていく.

式 (3.11) の条件を満たすとき, t_i において $\log F_0$ の抽出ミスが起きたと判定し, 線形補間により $\log F_0(t_i)$ の値を修正する.

$$\left| \frac{\log F_0(t_i)}{Mdn[\log F_0(t_{i-m}), \log F_0(t_{i-m+1}), \ldots, \log F_0(t_{i+m}))]} - 1 \right| > S \tag{3.11}$$

F_0 抽出のステップ幅が 2 [ms] であれば, $m = 2$, $S = 0.01$ が実験的に好ましい値となる. なお, 時点 t_i の前後 m 個のフレームにおいて抽出された $\log F_0$ の数が m 個以下であった場合, $\log F_0(t_i)$ は誤検出であったとして修正される.

声帯を音源とする区間である有声区間と, それ以外の区間である無声区間の境界には, しばしばマイクロプロソディと呼ばれる F_0 パターンの微小な変動が観測される. $\log F_0(t)$

の勾配を $G_0(t)$, 有声区間直後のフレームの添字を i としたとき, 式 (3.12) を満たす最小の区間 $[t_{i-n_1}, t_{i-1}]$ はマイクロプロソディであると判定され, F_0 パターン中から除去される.

$$|G_0(t_{i-1})| > 2|G_0(t_{i-n_1})| \tag{3.12}$$

ただし, $G_0(t_{i-1})$ と $G_0(t_{i-n_1})$ の極性は一致しており, n_1 は最大でも 10 未満であるとする. マイクロプロソディは有声区間の開始時にも生じている可能性があり, そちらについても上述したものと同様の基準で除去を行う.

ある無声区間 $[t_{i+1}, t_{j-1}]$ は式 (3.13) のような三次曲線近似によって補間される.

$$\log F_0(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \tag{3.13}$$

ここで, 係数 $[a_0, a_1, a_2, a_3]$ は式 (3.14) を満たすものである.

$$\begin{cases} \log F_0(t_i) &= a_0 + a_1 t_i + a_2 t_i^2 + a_3 t_i^3 \\ G_0(t_i) &= a_1 + 2a_2 t_i + 3a_3 t_i^2 \\ \log F_0(t_j) &= a_0 + a_1 t_j + a_2 t_j^2 + a_3 t_j^3 \\ G_0(t_j) &= a_1 + 2a_2 t_j + 3a_3 t_j^2 \end{cases} \tag{3.14}$$

また, 区間幅が 333 [ms] を超えるような場合は, 無声区間ではなく無音区間として取り扱われるため, 補間は行わない.

前処理の仕上げとして, F_0 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. 三次曲線による近似には無声区間の補間と同様に式 (3.13) を用いるが, 係数の計算方法は異なる. まず, 有声区間の先頭区分においては式 (3.15) より係数を求める.

$$\begin{cases} \sum_{i=1}^{N_1} \log F_0(t_i) &= N_1 a_0 + \sum_{i=1}^{N_1} a_1 t_i + \sum_{i=1}^{N_1} a_2 t_i^2 + \sum_{i=1}^{N_1} a_3 t_i^3 \\ \sum_{i=1}^{N_1} \log F_0(t_i) t_i &= \sum_{i=1}^{N_1} a_0 t_i + \sum_{i=1}^{N_1} a_1 t_i^2 + \sum_{i=1}^{N_1} a_2 t_i^3 + \sum_{i=1}^{N_1} a_3 t_i^4 \\ \sum_{i=1}^{N_1} \log F_0(t_i) t_i^2 &= \sum_{i=1}^{N_1} a_0 t_i^2 + \sum_{i=1}^{N_1} a_1 t_i^3 + \sum_{i=1}^{N_1} a_2 t_i^4 + \sum_{i=1}^{N_1} a_3 t_i^5 \\ \sum_{i=1}^{N_1} \log F_0(t_i) t_i^3 &= \sum_{i=1}^{N_1} a_0 t_i^3 + \sum_{i=1}^{N_1} a_1 t_i^4 + \sum_{i=1}^{N_1} a_2 t_i^5 + \sum_{i=1}^{N_1} a_3 t_i^6 \end{cases} \tag{3.15}$$

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, 式 (3.15) より求まっ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, 式 (3.16) から求めた係数を用いて, 150 [ms] ごとに平滑化を行う.

$$\begin{cases} \log F_0(t_j) = a_0 + a_1 t_j + a_2 t_j^2 + a_3 t_j^3 \\ G_0(t_j) = a_1 + 2a_2 t_j + 3a_3 t_j^2 \\ \sum_{i=j}^{j+N_2} \log F_0(t_i) = N_2 a_0 + \sum_{i=j}^{j+N_2} a_1 t_i + \sum_{i=j}^{j+N_2} a_2 t_i^2 + \sum_{i=j}^{j+N_2} a_3 t_i^3 \\ \sum_{i=j}^{j+N_2} \log F_0(t_i) t_i = \sum_{i=j}^{j+N_2} a_0 t_i + \sum_{i=j}^{j+N_2} a_1 t_i^2 + \sum_{i=j}^{j+N_2} a_2 t_i^3 + \sum_{i=j}^{j+N_2} a_3 t_i^4 \end{cases} \tag{3.16}$$

表 3.1　トーン層における J_ToBI ラベル

H*+L	アクセント核の位置
<	アクセント核の後方で F_0 が遅下がりする場合の位置
H-	アクセント句初頭における F_0 上昇の飽和位置
%L	アクセント句の始端
%wL	アクセント句の始端であり, F_0 が始めから高い
L%	アクセント句の終端
wL%	アクセント句の終端であり, F_0 が下がりきらない
H%	イントネーション句の終端
*?	アクセント核の位置でありながらアクセントが存在しない

ここで, 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. 以上の前処理を行った F_0 パターンは, 無音区間を除き, 任意の場所で微分可能である.

前処理を行った F_0 パターンを $F_0'(t)$, $\log F_0'(t)$ の微分を $G_0'(t)$ としたとき, 式 (3.10) より, $G_0'(t)$ が極大や極小となる時点の $\frac{1}{\beta}$ [s] 手前が, それぞれアクセント指令の立ち上がり時点 T_{1j} や立ち下がり時点 T_{2j} となる. さらに, そのアクセント指令の大きさは式 (3.17) となる.

$$A_{aj} = \frac{e}{2\beta}\left(G_0'(T_{1j}+\frac{1}{\beta}) - G_0'(T_{2j}+\frac{1}{\beta})\right) \tag{3.17}$$

フレーズ指令についても同様で, $F_0'(t)$ からアクセント成分を除去したものを $R_0(t)$ としたとき, $R_0(t)$ の微分値が極大となる時点の $\frac{1}{\alpha}$ [s] 手前がフレーズ指令の立ち上がり時点 T_{0i} であり, 指令の大きさは式 (3.18) より求まる.

$$A_{pi} = \frac{e}{\alpha}\left(\log R_0(T_{0i}+\frac{1}{\alpha}) - \log F_b\right) \tag{3.18}$$

このとき, 基底周波数 F_b は $F_0'(t)$ 中の最小値である.

3.6　分節的特徴との対応

アクセント指令やフレーズ指令の位置の推定に分節的情報を利用することもできる. J_ToBI は, 韻律情報のラベリングシステムとして代表的なものの一つであり, 韻律情報をトーン層と Break Index (BI) 層に分けて記述する. トーン層では, 表 3.1 に記したラベル

を用いて, F_0 パターンの特徴を記述する. 一方, BI 層では, 0 から 4 までの 5 段階の数字を用いて, 単語の結合度を記述する. このとき, 値が大きいほど単語同士の結合度は低くなる.

F_0 モデルの各指令との対応を考えたとき, アクセント指令の位置を推定する上で有用なのが, トーン層におけるラベルである, “H-” と “H*+L” である. “H-” はアクセント句初頭における F_0 上昇の飽和位置であるため, アクセント指令の立ち上がり位置と対応付けることができ, ラベル位置 $t_{\mathrm{H}-}$ から 0.2 [s] 前の時点が指令の立ち上がり位置となることが知られている. また, “H*+L” はアクセント核の位置であり, F_0 低下の開始時点であるため, ラベル位置 $t_{\mathrm{H*+L}}$ をそのままアクセント指令の立ち下がり位置として利用出来る. なお, アクセント位置が単語の比較的前方にある場合, ラベル “H-” は省略されが, “H*+L” を F_0 上昇の飽和位置とみなせるため, $t_{\mathrm{H*+L}} - 0.2$ [s] をアクセント指令の立ち上がり位置として用いる.

一方, フレーズ指令の位置は BI 層におけるラベル情報から推定できる. フレーズ指令は韻律句とよばれる分節的単位の開始時点に発生すると考えられているが, BI 層において韻律句境界と対応関係を持つのが “BI 3” と記述されるラベルである. フレーズ指令の生起位置は韻律句境界直後に現れる最初の母音の開始時点から 0.21 [s] 手前であるため [29], ラベル “BI 3” の直後に現れる最初の母音の開始時点から 0.21 [s] 手前がフレーズ指令の生起位置であるといえる.

3.7　まとめ

本章では, F_0 モデルが, 生理的 · 物理的な背景に裏打ちされた数理モデルであり, アクセントやフレーズといった言語的 · パラ言語的情報と, 物理的な音響特徴量である F_0 パターンとを効果的に対応付けることを示した. また, 観測された F_0 パターンからモデルパラメータを抽出する方法や, その際に分節的特徴を効果的に利用出来ることを説明した.

第 4 章

生理学的制約を加味した音声合成

4.1　はじめに

HMM 音声合成では, 分節的特徴のコンテキストと対応付けられた HMM を用いて統計的に構築された音響モデルから, 音響特徴量の時系列を生成することで音声の合成が行われる. こうした統計量ベース方式の音声合成システムでは, 従来, 分節境界における不連続な音響特徴量の生成や, 統計処理による音響特徴量の過剰な平滑化が生じてしまい, 十分な自然性を持つ合成音声の生成は困難な課題とされていた. しかし, 動的特徴量や GV の導入によりこれらの問題が低減されたため, HMM 音声合成は, 分節的特徴に基づいた柔軟な発話スタイルの制御が可能な音声合成手法として注目されるに至った.

だが, 音響的特徴の一つである韻律的特徴は, 分節を跨って表れるものが多いため, 分節的特徴と対応付けてモデル化を行うと好ましくない結果となる場合がある. 特に音響特徴量の時間微分である動的特徴量はロバスト性が低いため, この問題は F_0 パターン中においては急峻な変動として表れる. F_0 パターンの変動は種々の情報を伝える上で重要な役割をになっているため, このような誤ったモデル化に起因した変動は合成音声の自然性を大きく損なわせてしまう. しかし, 本来の F_0 パターン中にもアクセントやフレーズといった形で変動は含まれているため, 単純な平滑化では誤って生じた変動だけを除去することは出来ない.

そこで提案するのが, HMM 音声合成における F_0 モデルの生理的 · 物理的制約としての利用である. F_0 モデルは, 音声による情報表出の生理的過程に基づく数理モデルであり, F_0 パターンの変動を言語情報やパラ言語情報と対応付けてモデル化するため, 言語情報やパラ言語情報と関連を持たない F_0 パターンの変動を生成することがない. 提案手法では, HMM 音声合成によって生成された F_0 パターンを F_0 モデルを用いて記述しなおすが, この処理によって, 韻律的特徴の誤ったモデル化に起因して生じた F_0 パターンの変動の平滑化が期待できる.

HMM 音声合成によって生成された F_0 パターンからの F_0 モデルのパラメータ抽出は, Analysis-by-Synthesis によって行われるため, 提案手法は, 適切な初期パタメータの決定と, 逐次近似によるパラメータの最適化の二段階で構成される. 本章では, これら二つの処理につてそれぞれ説明を行った後, 主観評価実験とその結果についての報告を通して, 提案手法の有効性について考察を行う.

4.2　初期パタメータの決定

モデルパラメータの逐次近似に用いる初期値は F_0 パターン単体からでも求めることはできるが, 提案手法が対象としている F_0 パターン中には誤ったモデル化に起因した F_0 の変動が含まれている可能性があるため, 副次的な情報を与えなければ十分なロバスト性を確保することができない. そこで, HMM 音声合成に対し入力として与えられた文章から得られるコンテキスト情報も利用して初期パラメータの決定を行う.

コンテキスト情報を利用した初期パラメータの決定法としては J_ToBI を用いた方法が挙げられる. J_ToBI はアクセント句の情報や単語間の接続強度などを記述するラベリングシステムであり, 初期パラメータの決定に際しては, アクセント句初頭における F_0 上昇の飽和位置を示す “H-”, アクセント核の位置を示す “H*+L”, 韻律句境界を示す “BI 3” の三つのラベルが重要となる. つまり, これらのラベルと HMM 音声合成に対する入力文章から得られるコンテキスト情報との対応関係を求めれば, J_ToBI を用いた方法と同様にして初期パラメータを決定できる. “H-” は, コンテキト以外の情報にも依存したラベルであるが, 東京方言の場合, アクセント句の 2 モーラ目に該当することが多いと知られている [28]. 図 2.7 に示したコンテキストラベルにおいて, アクセント句内のモーラ位置に対応するのは a_{C1} である. つまり, $a_{C1} = 2$ となるコンテキストラベルを “H-” に対応するラベルとしてみなすことができる. また, a_{C2} はアクセント核との距離を示しているため, $a_{C2} = 1$ となるコンテキストラベルが “H*+L” に対応するといえる. 最後に “BI 3” についてであるが, HMM 音声合成に対する入力文章の情報からだけでは韻律句境界を求めることができないため, 厳密にはコンテキストラベルを “BI 3” と対応付けることはできない. ただし, “BI 3” から求まる F_0 モデルのフレーズ指令の生起位置は呼気段落の開始直前であることが多いため, 当該音素 p_C が, 音声波形初頭における無音区間を示す記号であった場合や, 音声波形中のポーズ区間を示す記号であった場合, そのラベルを “BI 3” に代わるものとしてみなすことができる.

以上をまとめると, アクセント指令の数はアクセント句の数と等しく, 指令の立ち下がり時間はアクセント核の直後, 立ち上がり時間はアクセント句の第二モーラから 0.2 [s] 前方となる. ただし, アクセント核がアクセント句の先頭である場合, 立ち上がり時間はアクセント核から 0.2 [s] 前方となる. 一方, フレーズ指令の数は音声波形中のポーズの数 +1 個となり, 指令の位置は, 無音区間やポーズの直後に出現する母音から 210 [ms] 前方となる. なお, コンテキストラベルの位置は HMM に基づく継続長モデルから決定される. また, 各指令の大きさは式 (3.17) と式 (3.18) から求まる.

4.3　パラメータの最適化

逐次近似によるモデルパラメータの最適化には, 目的の F_0 パターンに対する近似度を表す評価関数が必要となる. 従来手法におけるモデルパラメータの最適化では, F_0 モデルによって記述される F_0 パターンと解析対象である F_0 パターンとの間の平均二乗誤差を評価関数として利用していた [30]. しかし, 提案手法では, HMM によってモデル化された F_0 の統計量を利用することができるため, 評価関数には尤度関数に基づいたものを用いる.

HMM 音声合成では, ガウス分布に基づいた音響特徴量のモデル化を行っているため, 観測ベクトル $\boldsymbol{x}$ の尤度は式 (4.2) から求められる.

$$\log p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \left\{ \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp \left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) \right\} \tag{4.1}$$

$$= -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{d}{2} \log(2\pi) \tag{4.2}$$

$\boldsymbol{\mu}$ は分布の平均ベクトルであり, $\boldsymbol{\Sigma}$ は分散共分散行列である. 提案手法では, この尤度計算式に基づき, 式 (4.3) を評価関数として利用する.

$$\boldsymbol{p}' = \arg\min_{\boldsymbol{p}} (\boldsymbol{p} - \tilde{\boldsymbol{p}})^\top \boldsymbol{U}^{-1} (\boldsymbol{p} - \tilde{\boldsymbol{p}}) \tag{4.3}$$

ただし, $\boldsymbol{p}$ は F_0 モデルによって記述される F_0 パターン, $\tilde{\boldsymbol{p}}$ は HMM 音声合成によって生成される F_0 パターン, $\boldsymbol{p}'$ は最終的に出力される F_0 パターンである. $\tilde{\boldsymbol{p}}$ を生成する際には動的特徴量や GV のモデルも利用している. $\boldsymbol{U}$ は F_0 の分散の対角行列であり, F_0 パターンの各時点に対する重みとして機能する. なお, 式 (4.3) はマハラノビス距離としても知られている.

式 (4.3) を満たす F_0 パターン $\boldsymbol{p}$ の探索は, F_0 モデルのパラメータを微小変化させることで行われる. パラメータの初期値にはそれぞれ前節にて求めた値を用いる. フレーズ指令の生起時間, アクセント指定の立ち上がり時間, アクセント指令の立ち下がり時間の探索範囲はいずれも初期値 ± 0.2 [s] である. フレーズ指令の大きさとアクセント指令の大きさの探索は 0.1 から 0.9 までの範囲で行う. 指令の数については最適化の対象に入れない. また, モデルパラメータの探索には最急降下法を利用する.

表 4.1　音響モデルの構築における音響分析条件

サンプリング	16 [bit] / 16 [kHz]
フレーム窓	窓幅 25 [ms] のハミング窓
シフト長	5 [ms]
学習データ	450 文
特徴量ベクトル	パワーを含む 25 次元のメルケプストラム係数と, 対数 F_0, さらにそれぞれの一次微小変化分と二次微小変化分の計 78 次元
HMM の種類	5 状態 3 分布の left-to-right 型コンテキスト依存 HMM
生起確率分布	単一ガウス分布

4.4　実験的検証

提案手法の有効性を実験的に検証するため, 表 4.1 の条件のもと, F_0 モデルを利用する場合 (提案手法) と利用しない場合 (従来の HMM 音声合成) で, それぞれ HMM に基づく F_0 パターンの生成を行い, 合成された音声を自然性の観点から評価した.

学習や合成に用いる文章には, ATR 日本語音声データベースに含まれている音素バランス文を使用した. ATR 日本語音声データベースには, 50 文の音素バランス文から構成される文章セットが 9 個, 53 文の音素バランス文から構成される文章セットが 1 個含まれており, それぞれの文章セットには A から J までのアルファベットが割り当てられている. つまり, 学習に用いたのは 50 文の文章セットである A セットから I セットまでに含まれる 450 文であり, 音声合成を行う際の入力文章として用いたのは 53 文の文章セットである J セットに含まれる 53 文である. 学習に用いる 450 文にはそれぞれに対応する読み上げ音声が用意されており, 音響モデルの構築にはこれらの音声波形を利用する. なお, いずれの文章に対しても表 2.7 に示したコンテキスト情報はあらかじめ与えられている.

合成音声は韻律的特徴の自然性の観点から日本人 12 名よって評価された. 評価方法には RAB 法が使用され, 評価者は, 合成のターゲットとなる自然音声 (R) を聞いた後, 上述の二種類の音声 (A と B) を聞き, 提案手法の方が良い (2), 提案手法の方がやや良い (1), 差はない (0), 提案手法の方がやや悪い (-1), 提案手法の方が悪い (-2) の 5 段階で評価を行った.

なお, 以上の評価実験は, HMM の学習に男性話者による読み上げ音声を利用した場合と女性話者による読み上げ音声を利用した場合についてそれぞれ行った.

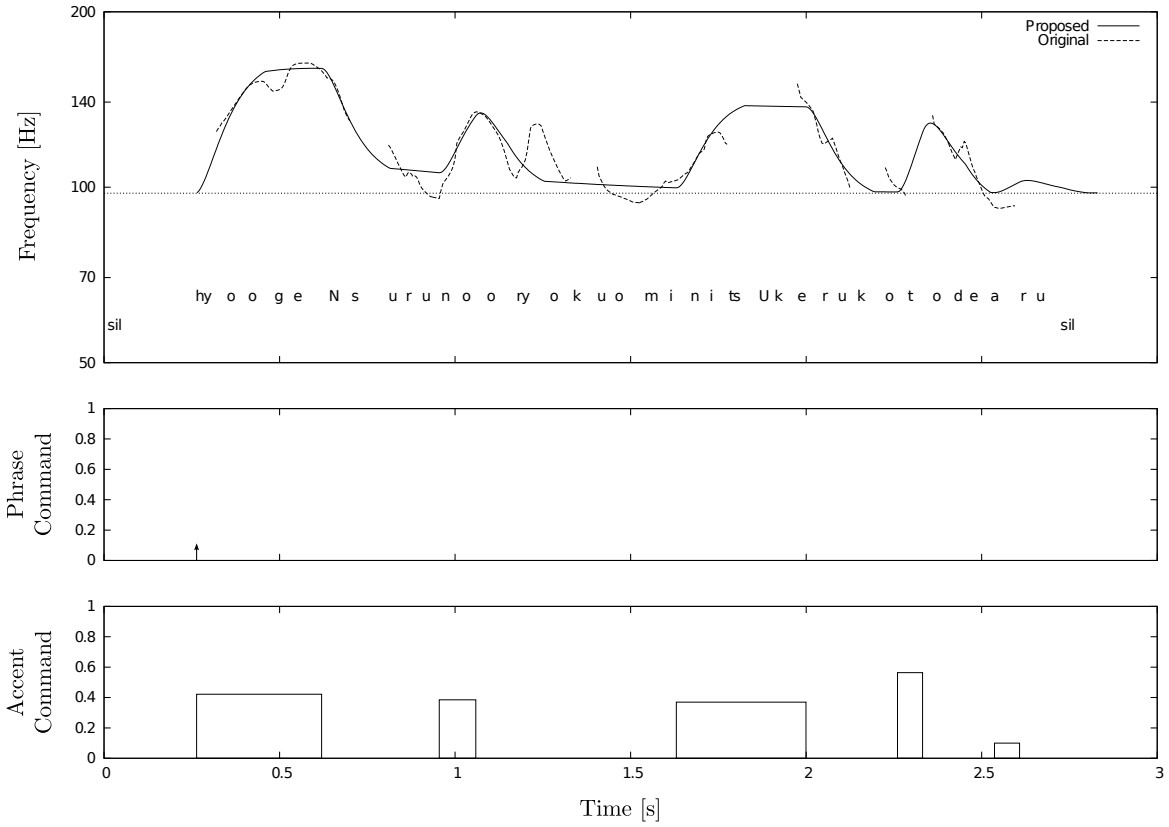

図 4.1　提案手法により生成された F0 パターン（男性話者）

図 4.1 は, 男性話者を用いた評価実験において, 提案手法を用いたことにより有意な自然性の向上が確認された合成音声の F_0 パターンである. 読み上げられているのは, J セットに含まれる 6 番目の文章, “表現する能力を身につけることである。” である. この文章は, 一つの呼気段落と五つのアクセント句から構成されているため, フレーズ指令の総数は 1 個, アクセント指令の総数は 5 個である.

提案手法により生成された F_0 パターンと従来手法により生成された F_0 パターンとを比較すると, 従来手法により生成された方には 1.2 [s] から 1.3 [s] にかけて, F_0 パターンの隆起が生じているのがわかる. この区間を含むアクセント句は “能力を” であるが, このアクセント句のアクセント核位置は先頭モーラであり, 先頭モーラが読み上げられているのは 1.1 [s] 付近であるため, この F_0 パターンの隆起は明らかな誤りである. 一方, 提案手法により生成された F_0 パターンでは, F_0 モデルはアクセントやフレーズと対応を持たない F_0 パターンの変動を記述しないため, このように明らかに誤った F_0 パターンの隆起は平滑化されている.

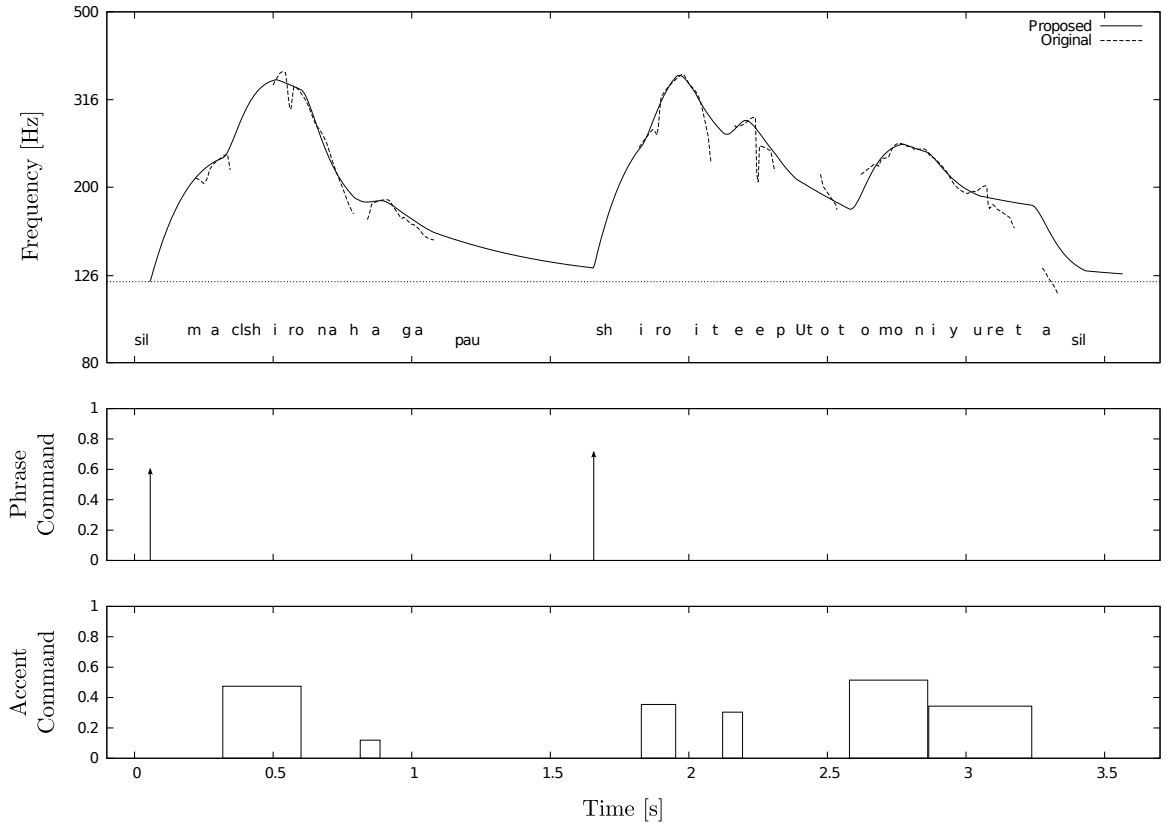

図 4.2　提案手法により生成された F0 パターン（女性話者）

図 4.2 は, 女性話者を用いた評価実験において, 提案手法を用いたことにより有意な自然性の向上が確認された合成音声の F_0 パターンである. 読み上げられているのは, J セットに含まれる 20 番目の文章, “真っ白な歯が、白いテープと共にゆれた。” である. この文章は, 二つの呼気段落と六つのアクセント句から構成されているため, フレーズ指令の総数は 2 個, アクセント指令の総数は 6 個である.

提案手法により生成された F_0 パターンと従来手法により生成された F_0 パターンとを比較すると, 従来手法により生成されたものの末尾において, F_0 パターンの急激な降下が生じていることがわかる. 平静に文章を読み上げた場合, このような F_0 パターンの降下が生じることは生理学的な制約により考えにくいため, この F_0 パターンの降下は誤りであるといえる. 一方, 提案手法により生成された F_0 パターンでは, F_0 モデルは生理的 · 物理的な背景に基づいた時定数に従って F_0 パターンの降下を記述するため, このように誤った F_0 パターンの降下は抑制されている.

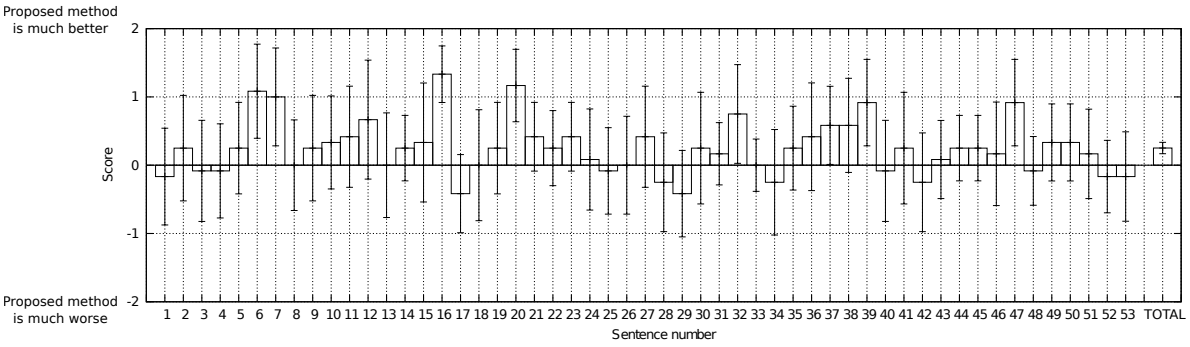

図 4.3　RAB テストにおける文書ごとの平均スコアと 95% 信頼区間（男性話者）

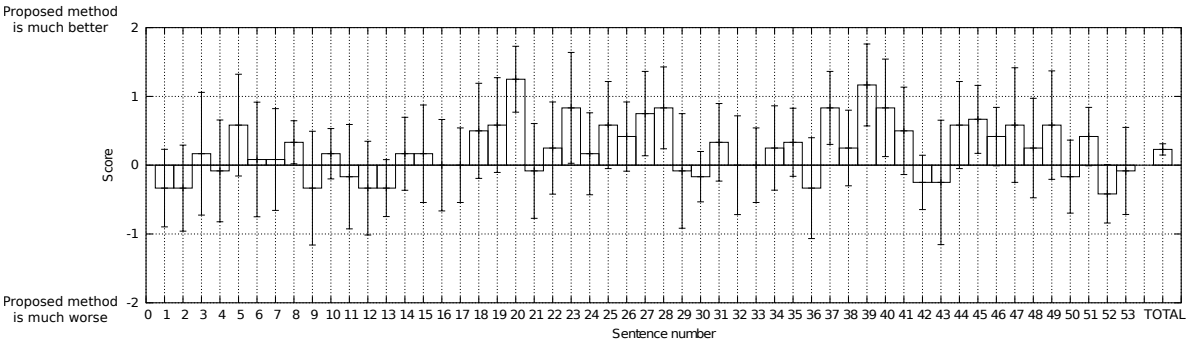

図 4.4　RAB テストにおける文書ごとの平均スコアと 95% 信頼区間（女性話者）

評価実験によって得られたスコアの文書ごとの平均と 95[%] 信頼区間を計算したところ, 男性話者については図 4.3, 男性話者については図 4.4 となった. 男性話者の場合は文書 6, 7, 16, 20, 32, 37, 39, 47 の 8 個において, 女性話者の場合は文書 8, 20, 23, 27, 28, 37, 39, 40, 45 の 9 個において, 提案手法を用いたことによる顕著な改善が得られた. 残りの読み上げ文については, 男女とも目立った変化はなかった. また, 全読み上げ文合計の平均スコアと 95[%] 信頼区間を求めたところ, 男性話者の場合はそれぞれ 0.252 と [0.165, 0.335], 女性話者の場合はそれぞれ 0.230 と [0.148, 0.311] となり, いずれの場合も有意な改善が得られていることがわかった.

上述の結果は, 提案手法を用いることによって, F_0 パターンの自然性が有意に改善されることを示している. 特に, 図 4.1 や図 4.2 に示した例のように, 音声波形中に含まれる言語情報や咽頭制御機構の物理的性質を考慮すれば明らかに誤りであることがわかる F_0 パターンの変動が, F_0 モデルの持つ生理学的な制約によって抑制されたことから, 提案手法において F_0 モデルは期待した通りの働きをしているといえる. また, 提案手法を用いたことによりかえって自然性が低下してしまう合成音声がみられなかったことも, 提案手法を実用する上で好ましい結果であった.

4.5　まとめ

HMM 音声合成のための F_0 パターン生成手法を提案した．従来手法における分節単位に基づいた F_0 の取り扱いは, 分節をまたがるような広い時間の範囲に渡る韻律的特徴には不向きであった．提案手法では, この問題に対し, F_0 モデルを制約として用いることで解決を図った．F_0 モデルは, 生理学的な知見に基づいた数理モデルであり, 時間の広い範囲に渡る韻律的特徴を効果的に記述することができる．この F_0 モデルを HMM 音声合成に導入することで, 分節単位での F_0 の取り扱いに起因する不自然な F_0 パターンの生成の抑制が期待できる．男女各一名について音声合成を行い, 主観評価実験を行ったところ, 男性話者については 53 文中 8 文において, また女性話者については 53 文中 9 文において, 提案手法を用いたことによって大幅な改善が得られた．この結果から, F_0 パターンの自然性の有意な改善が提案手法の使用により期待できるといえる．

第 5 章

微細変動の検出による高精度化

5.1　はじめに

HMM 音声合成によって生成される F_0 パターンに対する F_0 モデルを用いた平滑化処理は, 誤った韻律的特徴のモデル化に起因して生じた変動を F_0 パターン中から除去し, F_0 パターンの自然性を改善する. ところで, F_0 パターン中にはマイクロプロソディを始めとする微細変動が数多く含まれている. これらの微細変動は, 聴覚的にはほとんど影響力を持つことはないが, F_0 モデルの生理学的背景に含まれるものではないため, F_0 モデルの記述力で対応できるものではない. F_0 モデルを用いた F_0 パターンの平滑化は尤度最大化基準に基づく Analysis-by-Synthesis によって行われるが, 微細変動の存在は尤度を不当に低下させてしまう. そのため, 平滑化処理の有効性を高めるためには微細変動に配慮することが重要となるが, F_0 パターン中には様々な規模の変動が入り交じって存在しているため, そこから微細変動だけを検出するには局所性の高い検出法を利用する必要がある.

そこで, Difference-of-Gaussian (DoG) [31] を用いた F_0 パターンからの微細変動検出手法を提案する. DoG は, 画像からの特徴点抽出アルゴリズムである SIFT でも用いられている関数であり [32], ガウス関数に基づいた局所性やロバスト性の高い特徴点検出を可能にする [33–35]. SIFT では, DoG の極値の検出, 検出結果の局所化 [36], オリエンテーションの正規化, 特徴量の記述の四つの段階を踏むことで特徴点抽出を実現している [37]. DoG を F_0 パターンの微細変動検出に利用することを考えたとき, これらの処理の中で重要となるのが DoG の極値の検出と検出結果の局所化である. DoG の極値の検出は, カスケードフィルタリングによる実装が可能なため, 高速な微細変動の検出を実現する. また, 検出結果の局所化は, サブピクセル位置の推定やコントラストによる絞り込みによって, 微細変動の検出をノイズに対して頑健なものにする. 検出された微細変動に対しては, Analysis-by-Synthesis のための評価関数を計算する際, 影響が少なくなるように重みが付加される.

本章ではまず, F_0 パターンにおける DoG の計算方法と極値の検出について説明する. 次に, サブピクセル位置の推定やコントラストによる絞り込みに相当する処理によって検出結果のロバスト性が向上することを説明した後, F_0 モデルを生理学的制約として利用した HMM 音声合成において, 検出された微細変動の情報をどのように利用するかを述べる. そして最後に, 主観評価実験とその結果についての報告を通して, 微細変動について配慮することで得られる効果や, 従来の HMM 音声合成や F_0 モデルだけを利用した HMM 音声合成と比較した際の合成音声の自然性の変化について考察を行う.

5.2　微細変動の検出

微小変化分の計算による信号中からの極値検出は, 微細変動の特定を含め, 信号の特徴を量的に捉える際の有効な手段となる. ここで問題となるのが, どの程度の範囲における変化分を計算するかである. 微小変化分の計算は中心となる標本値に隣接するある程度の範囲に含まれる標本値を用いて行われるが, 計算に用いる範囲を決定するための根拠は多くの場合存在しない.

そこで用いるのが, スケール空間フィルタリングである [38]. スケール空間フィルタリングでは, 信号はガウシアンフィルタとの畳み込みによってスケール軸へ連続的に拡張される. ガウシアンフィルタは, 左右対称であり, 平均値に対して狭義の単調減少となるため, 信号に付加される重みは距離に従って滑らかに減少していく. また, スケールのパラメータである分散の値を極限まで大きくしたり小さくしたりした場合でも, ガウシアンフィルタによる畳込みであれば極端な値となることがない. その上, ガウシアンフィルタは, 容易に微分や積分ができるため, スケール軸への拡張に用いるフィルタとして非常に都合が良い.

代表的なスケール空間フィルタリングを利用した特徴点検出アルゴリズムである SIFT では, 入力画像を $I(x, y)$ としたとき, 入力画像の微小変化分 $\nabla^2_{\mathrm{norm}} L(x, y; \sigma^2)$ を式 (5.1) のように計算する.

$$\nabla^2_{\mathrm{norm}} L(x, y; \sigma^2) = \sigma^2 \nabla^2 L(x, y; \sigma^2) \tag{5.1}$$

ただし, $L(x, y; \sigma^2)$ は式 (5.2) に示す畳み込み演算により得られた $I(x, y)$ の平滑化画像であり,

$$L(x, y; \sigma^2) = G(x, y, \sigma^2) * I(x, y) \tag{5.2}$$

$G(x, y, \sigma^2)$ は式 (5.3) のようなガウス関数である.

$$G(x, y, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp - \frac{x^2 + y^2}{2\sigma^2} \tag{5.3}$$

$\nabla^2_{\mathrm{norm}} L(x, y; \sigma^2)$ は Scale-normalized Laplacian-of-Gaussian (LoG) と呼ばれる.

LoG は, 局所性の高い極値検出を実現する一方で, 計算コストが高いという問題も抱えている. そこで, 式 (5.4) の拡散方程式に基づき,

$$\nabla^2_{\mathrm{norm}} L(x, y; \sigma^2) = \frac{\partial L(x, y; \sigma^2)}{\partial \sigma^2} \tag{5.4}$$

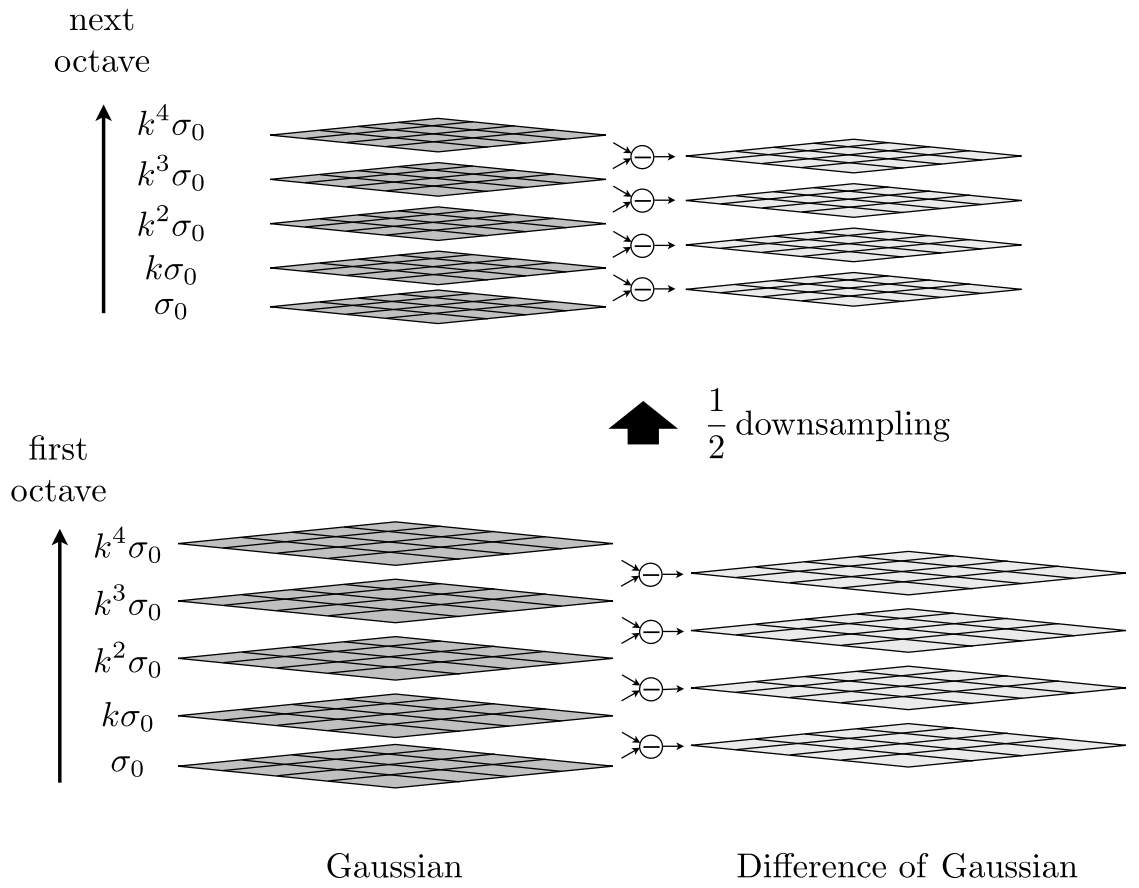

図 5.1　Difference-of-Gaussian 処理の流れ

式 (5.5) のように近似して計算する方法を用いる.

$$\nabla^2_{\mathrm{norm}} L(x, y; \sigma^2) \approx \frac{L(x, y; k\sigma^2) - L(x, y; \sigma^2)}{(k-1)\sigma^2} \tag{5.5}$$

この手法は, 式 (5.6) のようなガウス関数の差分との畳み込み演算に相当することから, Difference-of-Gaussian (DoG) による方法とも呼ばれる.

$$\nabla^2_{\mathrm{norm}} L(x, y; \sigma^2) \approx \frac{\big(G(x, y, k\sigma^2) - G(x, y, \sigma^2)\big) * I(x, y)}{(k-1)\sigma^2} \tag{5.6}$$

極値の検出に用いる DoG は図 5.1 の流れに沿って計算される. 基本的には, DoG の計算ごとにスケールパラメータである σ を k 倍することで, 入力画像をスケール軸方向へ拡張していく. しかし, σ が大きくなると, ガウシアンフィルタの窓幅は広くなり, それに伴い計算コストも増加してしまう. ところで, 入力画像 $I(x, y)$ の平滑化画像 $L(x, y; \sigma^2)$ と,

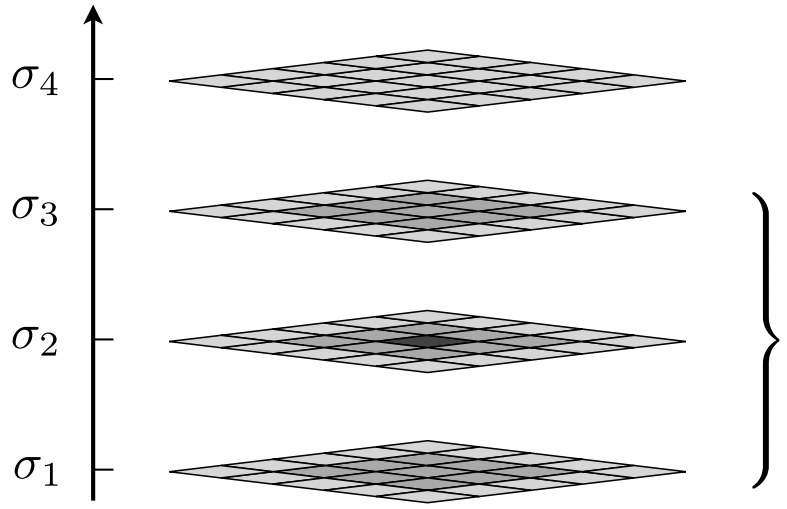

図 5.2　極値検出の流れ

$I(x, y)$ を $\frac{1}{2}$ のサイズにダウンサンプリングした画像 $I_{\frac{1}{2}}(x, y)$ の平滑化画像 $L_{\frac{1}{2}}(x, y; \sigma^2)$ との間には, 式 (5.7) の関係が成り立つ.

$$L(2\sigma_0) \approx L_{\frac{1}{2}}(\sigma_0) \tag{5.7}$$

つまり, スケールパラメータ $k^s\sigma_0$ が $2\sigma_0$ となるタイミングでダウンサンプリングを行えば, スケールパラメータの上限は $2\sigma_0$ となり, 計算量の増加は抑制される. ダウンサンプリングは平滑化画像を 6 回計算するごとに行われるのが標準的であるが, 6 回中 3 回は極値の検出を連続的に行うための重複分であるため, 増加率 k には $\sqrt[3]{2}$ を用いる.

極値の検出は図 5.2 のように DoG を 3 枚用いて行う. DoG 中の極値とは, x 軸, y 軸, σ 軸とそれらを組み合わせた 26 近傍に対する極値を意味する. σ 軸についての極値の検出は σ の値が小さい方から順に行っていき, 一旦極値が検出されると, その位置においては極値の検出を終了する.

以上の手順を踏むことで, 入力信号中の特徴について, その位置だけでなく規模に関する情報も同時に抽出することができる. しかし, スケール空間フィルタリングは, DoG の微小変化分に基づいて行われているため, ロバスト性の面で問題を抱えている. そこで, 極値の値を基準に特徴点の絞り込みを行うが, DoG は離散化されているため, 極値の正確な値を求めるためには補間が必要となる. $\boldsymbol{x} = [x, y, \sigma]^\top$ とし, DoG を $D(\boldsymbol{x})$ とおくと, 補間

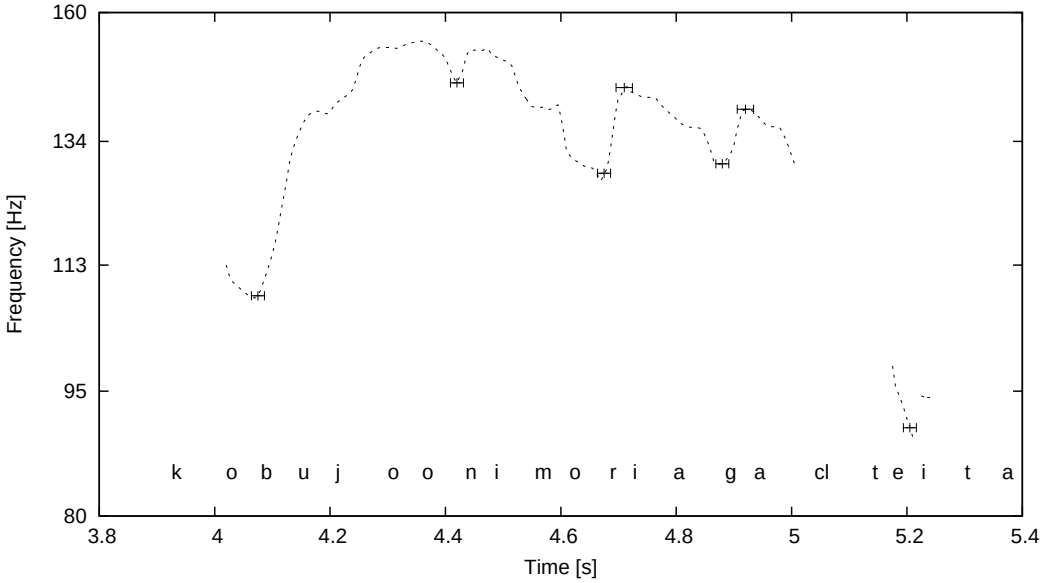

図 5.3　Difference-of-Gaussian による微細変動の検出（男性話者）

された正確な DoG の極値 $\hat{D}$ は式 (5.8) となる.

$$\hat{D} = D - \frac{1}{2}\frac{\partial D}{\partial \boldsymbol{x}}^{\top}\frac{\partial^2 D}{\partial \boldsymbol{x}^2}^{-1}\frac{\partial D}{\partial \boldsymbol{x}} \tag{5.8}$$

このようにして求まった $\hat{D}$ が 0.03 以下であった場合, その極値はノイズとして無視する.

以上は SIFT における入力画像からの特徴点抽出の流れであるが, F_0 パターンに対しても, 単に二変数が一変数となるだけで, 同様に適用できる. 特に, スケールパラメータ σ の上限を平均音素継続長よりも十分に短い 17.85 [ms] に設定したとき, 検出される特徴点は F_0 パターンの微細変動に相当する. 図 5.3 は, 男性話者による読み上げ音声 “両手の指は変形し、関節のあたりが、瘤状に盛り上がっていた。” から検出した微細変動の位置と範囲の例である.

5.3　微細変動に配慮した生理学的制約の適用

式 (4.3) における対角分散行列 $\boldsymbol{U}$ に対し, 式 (5.9) のように重み付けを行うことで微細変動から受ける影響を低減させる.

$$\boldsymbol{U} = \begin{bmatrix} w_1 u_1^2 & 0 & \cdots & 0 \\ 0 & w_2 u_2^2 & \cdots & 0 \\ \vdots & \vdots & w_j u_j^2 & \vdots \\ 0 & 0 & \cdots & w_N u_N^2 \end{bmatrix} \tag{5.9}$$

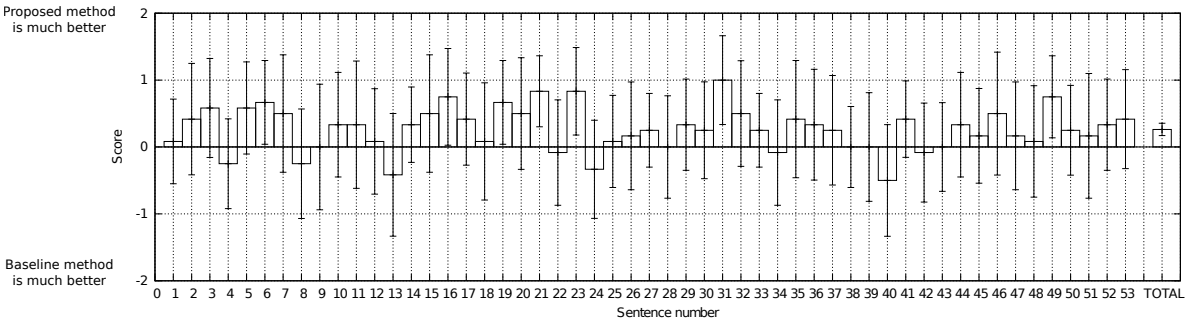

図 5.4　微細変動に配慮した HMM 音声合成の RAB テストの結果

u_j^2 は HMM によってモデル化された F_0 値の分散である．w_j は, 微小変動の影響を低減させるための重みパラメータであり, 検出された微小変動位置の $\pm\sigma_j$ の区間では 4.0, それ以外の区間では 1.0 の値をとる．

5.4　実験的検証

微細変動への配慮の有効性を実験的に検証するため, 表 4.1 の条件のもと, 微細変動に配慮しながら F_0 モデルを利用する場合 (提案手法) と利用しない場合 (従来の HMM 音声合成) で, それぞれ HMM に基づく F_0 パターンの生成を行い, 合成された音声を自然性の観点から評価した．学習や合成に用いる文章には, ATR 日本語音声データベースに含まれている音素バランス文を使用した．合成音声は韻律的特徴の自然性の観点から日本人 12 名よって RAB 法に基づき評価された．なお, HMM の学習には男性話者による読み上げ音声を利用した．

評価実験によって得られたスコアの文書ごとの平均と 95[%] 信頼区間を計算したところ図 5.4 となった．特に, 文書 6, 16, 19, 21, 23, 31, 49 の 7 個において顕著な改善が得られた．残りの読み上げ文については, 男女とも目立った変化はなかった．また, 全読み上げ文合計の平均スコアと 95[%] 信頼区間を求めたところ, それぞれ 0.263 と [0.170, 0.355] であった．微細変動に配慮する点以外は同条件である 4.1 の結果と比較すると, 有意差こそないものの全読み上げ文合計の平均スコアは改善されている．しかし, 顕著な改善が得られた文の総数は減少してしまった．

図 5.5 と図 5.6 は, 微細変動に対して重み付けを行ったことにより F_0 パターンの自然性が向上した典型的な例である．どちらも, “いつもの休日のパターンを過ごして、日が暮れる。” という J セットに含まれる 17 番目の文章を合成した結果であるが, 重み付けを行っていない図 5.5 の F_0 パターンが従来手法によって生成されたものに近い形となっている

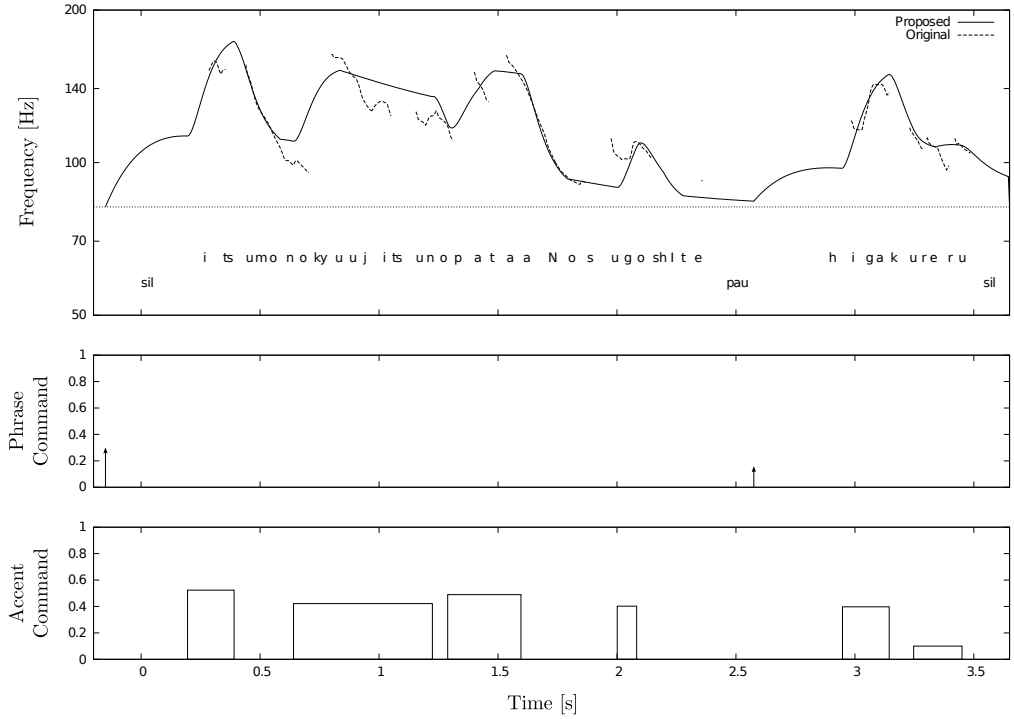

図 5.5　F0 モデルによる制約のもと生成された F0 パターン（重み付け無し）

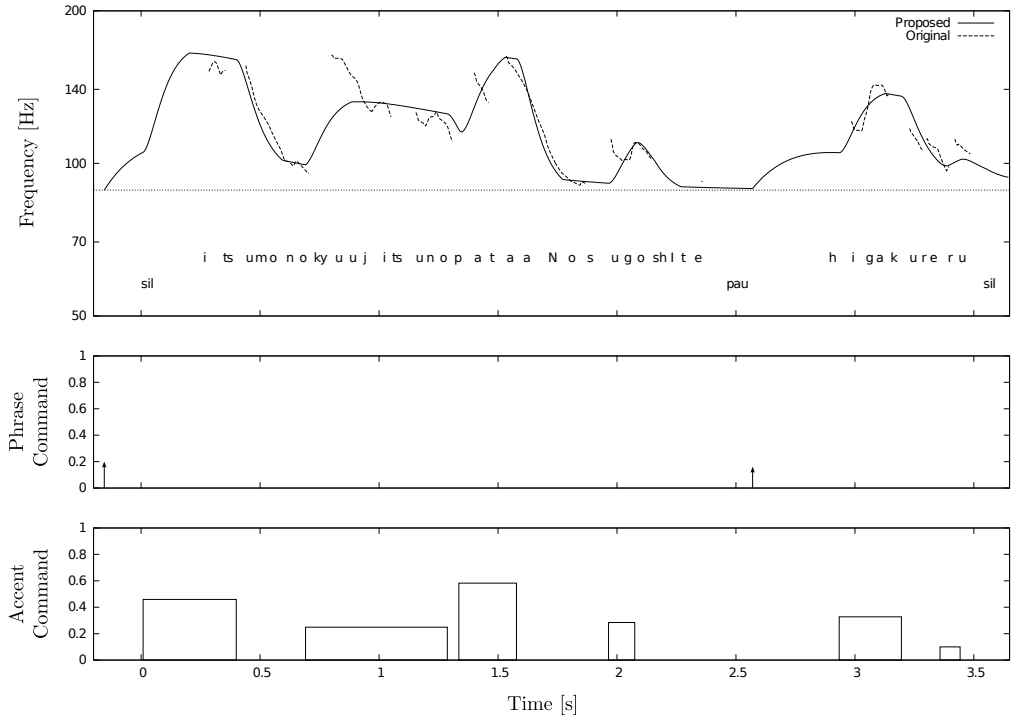

図 5.6　F0 モデルによる制約のもと生成された F0 パターン（重み付け有り）

のに対し, 図 5.6 の F_0 パターンは言語情報に沿うように変化している. これは微細変動に重み付けを行ったことで F_0 モデルの言語的な制約が強まった結果であると考察できる.

5.5　まとめ

F_0 パターンの微細な変動に配慮することで F_0 モデルによる制約効果を向上させようと試みた. F_0 パターン中には F_0 モデルの記述力で対応できるものではない微細変動が含まれており, これらの微細変動は F_0 モデルによる制約の要となる尤度計算の精度を低下させてしまう. そこで, 局所性の高い特徴点検出が可能な DoG を用いて微細変動を検出し, 尤度計算を行う際に, これらの微細変動の影響が低下するよう重み付けを行った. 男性話者一名について合成音声の主観評価実験を行ったところ, 有意な差ではなかったものの, 微細変動に対し重み付けを行った場合のほうが自然性の向上の度合いは大きかった. また, 重み付けを行うことで, F_0 モデルによる言語的な制約が強まる傾向も確認された.

第 6 章

結論

6.1　まとめ

本論文では, F_0 モデルを生理学的制約として利用することで, HMM 音声合成による合成音声の自然性を有意に改善させる手法についての提案を行った. また, F_0 パターン中に元来含まれている微細な変動について考慮することで, F_0 モデルによる制約の効果を向上させる方法についても検討を行った.

HMM 音声合成は, 分節的特徴と対応付けて統計的にモデル化された音響特徴量に基づいて音声の合成を行うことから, 柔軟な発話スタイルの制御が可能な音声合成手法として注目されている. また, こうした統計量ベース方式の音声合成システムにおいて常々問題となっていた, 分節境界における不連続な音響特徴量の生成や, 統計処理による音響特徴量の過剰な平滑化に対し, 動的特徴量や GV の導入による対処を行うことで, 十分な自然性を持った合成音声の生成も実現している. だが, 分節的特徴と対応付けたモデル化は, 分節を跨って表れる音響的特徴である韻律的特徴には不向きであり, 誤った F_0 パターンの変動を生み出す原因ともなる. F_0 パターンの変動は, 言語情報を伝える上で重要な役割を果たしているため, その生成が正しく行われなければ合成音声の自然性を大きく低下させてしまう.

HMM 音声合成によって生成された F_0 パターンに対する F_0 モデルを用いた平滑化処理は, 上述の問題を低減させる方法として有効に作用する. F_0 モデルは, 生理学的な知見に基づいた数理モデルであり, 分節を跨るような広い時間範囲に渡る韻律的特徴であっても効果的に記述する. また, F_0 モデルは, アクセントやフレーズと対応付けながら韻律的特徴の記述を行うため, 言語情報やパラ言語情報と関係を持たない F_0 パターンの変動は抑制される. 平滑化処理の有効性を検証すべく, 男女各一名の音声を用いて評価実験を行ったところ, 男性話者については 53 文中 8 文において, また女性話者については 53 文中 9 文において, F_0 モデルを用いたことによって大幅な改善が得られた.

ただし, F_0 パターン中にはマイクロプロソディを始めとする微細変動が数多く含まれており, これらの微細変動は F_0 モデルの生理学的制約としての効果を低下させてしまう. そこで本論文では, F_0 パターン中から微細変動を検出し, それらの微細変動に配慮して F_0 モデルに基づく平滑化を行う手法についても検討した. 男性話者一名を用いた評価実験では, 微細変動に配慮したことで, F_0 モデルの言語的な制約が強化される傾向が見られた.

以上のことから, F_0 モデルの生理学的制約としての利用は, HMM 音声合成よって生成される音声の自然性を向上させる効果的な手段であるといえる.

6.2　今後の展望

本論文では, F_0 パターン中に含まれる微細変動についても考慮した F_0 モデルの利用法についても検討を行ったが, 実験結果からも十分な効果が得られているとは言えないものであった. 改善案として, HMM 音声合成における GV の取り扱いのように, 微細変動をその都度検出するのではなく, 統計的なモデル化によって取り扱う方法を検討している.

また, F_0 モデルを制約として利用する際, 尤度に基づく評価関数を使用したが, 実際の音声合成システムにおいては制約の強さを制御できることが望ましいため, 制約の適用の仕方については見直しの余地が残っている.

そして, HMM 音声合成は日本語に限らず様々な言語での音声の合成が可能であり, F_0 モデルもまた数多くの言語に適応するモデルであることが確かめられているため, 本論文にて提案した手法は日本語音声合成に限定して用いられるものではない. 現在までのところで, 中国語での HMM 音声合成における F_0 モデルの利用については検討を進めているが, 今後, 英語を始めとする別の言語における提案手法の有効性についても調査していく必要がある.

謝辞

まず, 研究を進めるに際し, 多大なご指導を賜りました指導教員である広瀬啓吉教授に深く感謝致します. 様々な形で身に余るほどの機会を与えていただきました. また, 数多くのご鞭撻を頂きました峯松信明准教授にもお礼を申し上げます. もし本論文に何かしら価値ある知見が含まれているならば, それはお二方のご助力に依るものです.

加えて, 研究環境の充実に尽力してくださった高橋登技官, 事務補佐員の磯部史子氏, 池上恵氏らにも感謝致します. 何不自由無く研究活動に取り組むことができました.

そして, 広瀬 · 峯松研究室の皆様には大変お世話になりました. 特に, 博士課程の齋藤大輔氏には, 実験環境の構築から論文の執筆まで, 幾度と無くご支援をいただきました. また, 英字論文を執筆した際, 丁寧に原稿に目を通して頂いた同じく博士課程の Greg Short 氏にも感謝致します. さらに, 度重なる長時間の聴取実験に参加して頂いた方々にも改めてお礼申し上げます. 研究室の皆様からの多岐にわたるご協力なくして, 本論文を書き上げることは出来ませんでした.

最後に, これまでの研究生活を支えてくれた両親に感謝します.

2011 年 2 月 9 日

松田 徹也

参考文献

[1] K. Tokuda, T. Kobayashi, and S. Imai, “Speech parameter generation from HMM using dynamic features,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, I, pp.660–663 (1995–5)

[2] 河井 恒,“音声合成用大規模音声コーパスの構築”, 電子情報通信学会音声研究会, vol.105, no.97, pp.19–24 (2005–9)

[3] Y. Stylianou,“Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. Speech and Audio Processing*, vol.9, no.1, pp.21–29 (2001–3)

[4] 戸田 智基, 河井 恒, 津崎 実, “素片接続型テキスト音声合成における韻律変形の有効性”, 日本音響学会研究発表会講演論文集, 1-8-10, pp.201–202 (2003–9)

[5] 古井 貞熙, “発声の仕組と音声の特徴”, ディジタル音声処理, 東海大学出版会, pp.5–34 (1985)

[6] 今井 聖, 住田 一男, 古市 千枝子, “音声合成のためのメル対数スペクトル近似フィルタ”, 電子情報通信学会, J66-A, no.2, pp.122–129 (1983–2)

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1315–1318 (2000–6)

[8] T. Toda and K. Tokuda, “Speech parameter genaration algorithm considering global variance for HMM-based speech synthesis,” *Proc. Interspeech*, pp.2801–2804 (2005–9)

[9] K. Hirose, “Speech prosody in spoken language technologies,” *Journal of Signal Processing*, vol.12, no.1, pp.7–16 (2008–3)

[10] H. Fujisaki and 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)

[11] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” *Journal of the Acoustical Society of Japan*,

vol.5, no.4, pp.233–242 (1984-10)

[12] H. Fujisaki, "In Search of Models in Speech Communication Research," *Proc. Interspeech*, pp.1–10 (2008–9)

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proc. Eurospeech*, pp.2347–2350 (1999–9)

[14] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Information & System*, vol.E85-D, no.3, pp.455–464 (2002–3)

[15] 吉村 貴克, 徳田 恵一, 益子 貴史, 小林 隆夫, 北村 正, "HMM に基づく音声合成のための状態継続長モデルの構築", 電子情報通信学会音声研究会, vol.98, no.264, pp.45–50 (1998–9)

[16] 今井 聖, 住田 一男, 古市 千枝子, "音声合成のためのメル対数スペクトル近似(MLSA)フィルタ", 電子情報通信学会論文誌 A, vol.J66–A, no.2, pp.122–129 (1983–2)

[17] T. Shimamura and H.Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Trans. Speech and Audio Processing*, vol.9, no.7, pp.727–730 (2001–10)

[18] D. Talkin, "A robust algorithm for pitch tracking," *Speech Coding and Synthesis*, pp.495–518 (1995)

[19] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.ASSP–23, pp.67–72 (1975–2)

[20] 酒向 慎司, "VoiceMaker-1.1 – HMM 音声合成用音響モデルの構築 –", `http://iij.dl.sourceforge.jp/galateatalk/26798/VoiceMaker-doc-1.1.pdf` (2007–8)

[21] 全 炳河, 徳田 恵一, 益子 貴史, 小林 隆夫, 北村 正, "動的特徴量を考慮したピッチの高精度モデル化手法", 日本音響学会研究発表会講演論文集, vol.1–2–7, pp.219–220 (2001–9)

[22] H. Zen, K. Tokuda, and T. Kitamura, "Decision tree distribution tying based on a dimensional split technique," *Proc. International Conference on Spoken Language Processing*, pp.1257–1260 (2002–9)

[23] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling in HMM-based speech synthesis system," *Proc. International*

Conference on Spoken Language Processing, vol.2, pp.29–32 (1998–9)

[24] 籠嶋 岳彦, 赤嶺 政巳, “閉ループ学習に基づく最適な素片選択の解析的生成”, 電子情報通信学会, vol.J83-D-II, no.6, pp.1405–1411 (2000–2)

[25] H. Fujisaki, “Information, prosody, and modeling – with emphasis on tonal features of speech,” *Proc. Speech Prosody*, pp.1–10 (2004–3)

[26] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujiaski, “A method for automatic extraction of model parameters from fundamental frequency contours of speech,” *Proc. IEEE International Conference on Acoustics, Speech, & Signal Processing*, vol.I, pp.509–512 (2002-5)

[27] 平井 俊男, 樋口 宣男, “韻律ラベリング・システム J_ToBI を用いた基本周波数制御規則の自動抽出”, 電子情報通信学会音声研究会, vol.97, no.64, pp.27–32 (1997–5)

[28] N. Campbell, “Tones and Break Indices (ToBI) システムと日本語への適用”, 日本音響学会誌, vol.53, no.3, pp.223–229 (1997–3)

[29] 広瀬 啓吉, 藤崎 博也, 河合 恒, 山口 幹雄, “基本周波数パターン生成過程モデルに基づく文章音声の合成”, 電子情報通信学会音声研究会, vol.J72-A, no.1, pp.32–40 (1989–1)

[30] 成澤 修一, 峯松 信明, 広瀬 啓吉, 藤崎 博也, “音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, pp.2155–2168 (2002–7)

[31] D. G. Lowe, “Object recognition from local scale-invariant features,” *Proc. IEEE International Conference on Computer Vision*, vol.2, pp.1150–1157 (1999–9)

[32] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. Journal of Computer Vision*, vol.60, no.2, pp.91–110 (2004)

[33] J. J. Koenderink, “The structure of images,” *Biological Cybernetics*, vol.50, no.5, pp.363–370 (1984)

[34] T. Lindeberg, “Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention,” *Int. Journal of Computer Vision*, vol.11, no.3, pp.283–318 (1993)

[35] T. Lindeberg, “Scale-space theory: A basic tool for analysing structures at different scales,” *Journal of Applied Statistics*, vol.21, no.2, pp.224–270 (1994)

[36] M. Brown, D. Lowe, “Invariant features from interest point groups,” *In British Machine Vision Conference*, pp.656–665 (2002–9)

[37] 藤吉 弘亘, “Gradient ベースの特徴抽出 : SIFT と HOG”, 電子情報通信学会 パター

ン認識・メディア理解研究会, vol.107, no.206, pp.211–224 (2007–8)

[38] A. P. Witkin, “Scale-space filtering”. *Proc. International Joint Conference on Artificial Intelligence*, vol.2, pp.1019–1022 (1983)

発表文献

学術論文

[1] T. Matsuda, K. Hirose, and N. Minematsu, "HMM-based synthesis of fundamental frequency contours using the generation process model," *Journal of Signal Processing*, vol.14, no.4, pp.281–284 (2010–7)

[2] N. Sunada, T. Matsuda, K. Hirose, and N. Minematsu, "Use of paired white noises with inverse phases in ensemble empirical mode decomposition," *Journal of Signal Processing*, vol.14, no.4, pp.277–280 (2010–7)

国際会議論文

[3] T. Matsuda, K. Hirose, and N. Minematsu, "Quality improvements by SIFT in HMM-based F0 contour synthesis using generation process model," *Proc. Int. Workshop on Nonlinear Circuits, Communication and Signal Processing*, (2011–3, accepted)

[4] T. Matsuda, K. Hirose, and N. Minematsu, "Control of fundamental frequency contours using the generation process model in HMM-based speech synthesis," *Proc. IEEE Int. Conf. on Signal Processing*, pp.617–620 (2010–10)

[5] K. Hirose, K. Ochi, M. Wang, T. Matsuda, M. Wen, and N. Minematsu, "Using F0 contour generation process model for improved and flexible control of prosodic features in HMM-based speech synthesis," *Proc. Int. Conf. on Electronic Speech Signal Processing*, pp.84–93 (2010–9)

[6] M. Watanabe and T. Matsuda, "Development of a CALL system using speech recognition technology and its use for prosody learning", *Proc. Int. Conf. on Japanese Language*, pp.1614-1–1614-4 (2010–7)

[7] T. Matsuda, K. Hirose, and N. Minematsu, "HMM-based synthesis of fundamental frequency contours using the generation process model," *Proc. Int. Workshop on Nonlinear Circuits, Communication and Signal Processing*, pp.464–467 (2010-3)

[8] N. Sunada, T. Matsuda, K. Hirose, and N. Minematsu, “Use of paired white noises with inverse polarity in ensenble empirical mode decomposition,” *Proc. Int. Workshop on Nonlinear Circuits, Communication and Signal Processing*, pp.512–515 (2010–3)

国内研究会論文

[9] 松田徹也, 広瀬啓吉, 峯松信明, “生成過程モデルを用いた HMM に基づく基本周波数パターン生成”, 電子情報通信学会音声研究会, SP2010–34，pp.73–78 (2010–6)

[10] 松田徹也, 広瀬啓吉, 峯松信明, “経験的モード分解による主構造抽出を介した雑音環境下における音声信号の基本周波数推定”, 電子情報通信学会音声研究会, SP2009–15, pp.49–54 (2009–5)

国内全国大会論文

[11] 松田徹也, 広瀬啓吉, 峯松信明, “HMM に基づく生成過程モデルを用いた F0 パターン生成における品質改善”, 日本音響学会春季講演論文集, (2011–3, 発表予定)

[12] 松田徹也, 広瀬啓吉, 峯松信明, “HMM 音声合成における生成過程モデルを用いた F0 パターン生成法”, 日本音響学会秋季講演論文集, 2-1-5, pp.233–236 (2010–9)

[13] 砂田宜宏, 松田徹也, 広瀬啓吉, 峯松信明, “アンサンブル経験的モード分解における効果的なノイズ付与の検討”, 日本音響学会春季講演論文集, 2-Q-33, pp.519–522 (2010–3)

学位論文

[14] 松田徹也, “経験的モード分解による主構造抽出を介した雑音環境下における音声信号の基本周波数推定”, 東京大学工学部電子情報工学科卒業論文 (2008)