修 士 論 文

音声の韻律モデルに基づく基本周波数パターンの再構成

指導教官　木村正行 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

隈田章寛

1997 年 2 月 14 日

要 旨

連続音声認識において、発話文の句境界検出に韻律的な特徴を利用することが注目されている。韻律の分析には重要な特徴量の一つであるピッチパターンを用いることが多い。ピッチパターンはそのままでは、不連続で扱いにくいので、検出の前処理として平滑化が行われる。しかし従来の平滑化手法は数値解析の理論を基にしており、ピッチパターンとしてあり得ないパターンに変換してしまう危険性がある。そこで、本稿ではピッチパターンの生成機構に則したモデルにしたがって再構成することで抽出の高精度化を図る。

目 次

図 目 次

表 目 次

第 1 章

序論

1.1　背景

近年、パーソナルコンピュータや、電子手帳などの急速な普及とともに、一部の専門家だけでなく一般の人も、計算機に接する機会が増えてきた。それとともに、企業から発売される商品には一般の人から敬遠されがちなキーボードに変わり、手書き文字入力などの機能を持ったものが出てきた。それに続いて音声による入力を目指した、パーソナルコンピュータ用のハードウェアまで発売されている。このように、音声認識等のユーザーインターフェースの分野は活発に研究開発されているが、まだ実用的とは言いがたいのが現状である。それは、音声認識の場合、音声の特性上、話者により物理的な特徴量が変動するのみならず、同一話者でも文脈や発話のスピードなどで特徴量が変化し、不変な特徴を抽出するのが困難であることが理由の一つである。音声認識に利用される情報として、次の３つがあげられる。

音韻情報 言語的なものであり、発音記号で表すことができる。

韻律情報 アクセントやイントネーションといった情報であり、音韻情報と密接な関係があると考えらている。

個人情報 各個人で異なる声帯や声道の長さ、大きさ、形などの特徴に起因するものであり、声を聞いただけで、知合いを確認できるのはこの情報のためである。

現在の音声認識のほとんどは音韻情報のみから認識を行なっている。しかし、音韻情報のみでは認識率に限界がみえており、新たに韻律情報の利用が注目されている。韻律情報を

用いることで次のようなことが可能になる。

同音異義語の識別 “橋” と “箸” のような同音異義語の識別が可能になる。

文節に分割 韻律情報で、長文を短い文節に区切ることができる。連続音声認識の際、長い発声に対して直接 HMM(Hidden Markov models) などの処理を行うのは計算量が膨大になり、また誤認識の原因になる。韻律情報により区切られた短い文節に対して、そのような処理を行うことにより、処理速度や認識率の向上が望める。

文節間の修飾関係の推定 アクセントやイントネーションによって文節間の修飾関係の予測も可能であると言われている。

しかし現在のところこれらの用途に使用できる程、正確な特徴量が抽出できない。韻律情報の分析には重要な特徴量の一つであるピッチ周波数パターンを用いる事が多く、それゆえに、より正確なピッチ抽出が望まれている。これまでに様々なピッチの抽出法が提案されているが、いずれの手法も一長一短であり、決定的な手法は確立されていない。

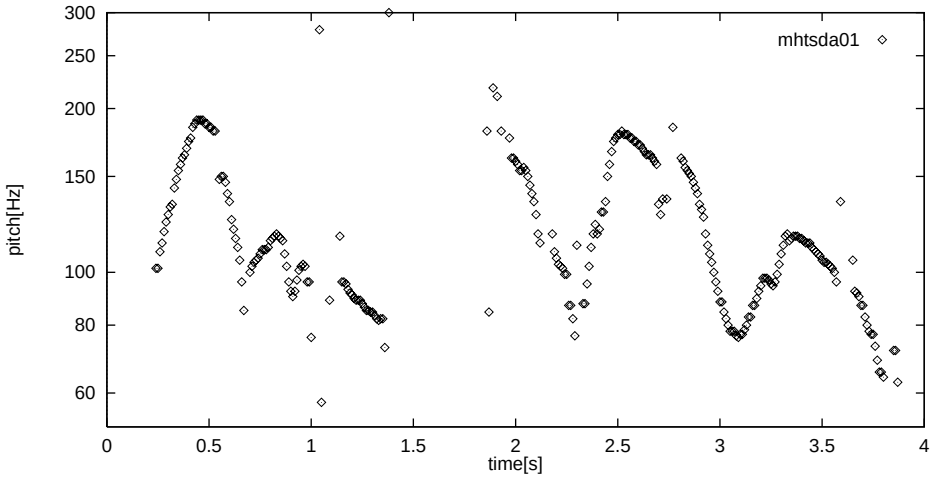

図 1.1: 自動ピッチ抽出法によるピッチパターン

1.2 目的

従来からのピッチ抽出法によって得られたパターンは、ピッチの特性から不連続なものになる（図 1.1)。認識に利用するには、連続で滑らかであるほうが都合が良いため、し

ばしば平滑化の処理が行なわれる。（図 1.2 ）平滑化には線形補間などの手法が用いられ

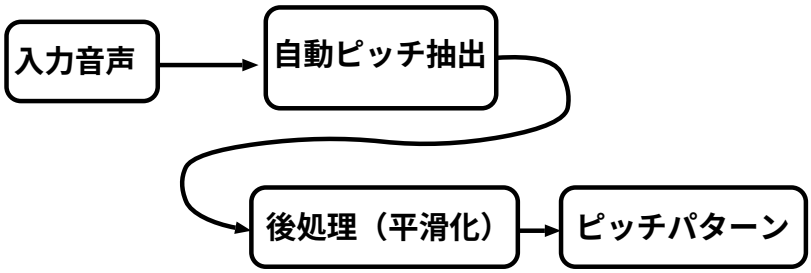

図 1.2: ピッチ抽出の行程

るが、いずれも数値解析的な処理であり、これらの平滑化はピッチパターンとしての物理的な特徴を損なう恐れがある。そこで、本研究では、ピッチパターンを生成機構に則したモデルを用いて再構成することで高精度化を図る。

1.3 本論文の構成

本論文では、第 2章で、ピッチパターンについて、生成過程モデルとその逆フィルタについて述べる。第 3章で再構成の手法と、逆フィルタ以外のフィルタについて述べ、それらの再構成結果の評価を試みる。第 4章は全体の総括及び、今後の改良策の提案を行う。

第 2 章

ピッチパターン

韻律情報のうちの１つにピッチ周波数がある。基本周波数や、F_0周波数とも呼ばれ、声の高さに相当するものである。歌を歌う場合はピッチ周波数を楽譜の音程に合わせている。図 2.1 に示した音声波形では、ピッチ周波数は振幅が大きく比較的長い周期の信号としてあらわれる。また、ピッチパターンとは、ピッチ周波数の時間的な変化を意味する（図 1.1 ）。この章では、ピッチ抽出法について説明した後、ピッチ周波数の物理的な生成

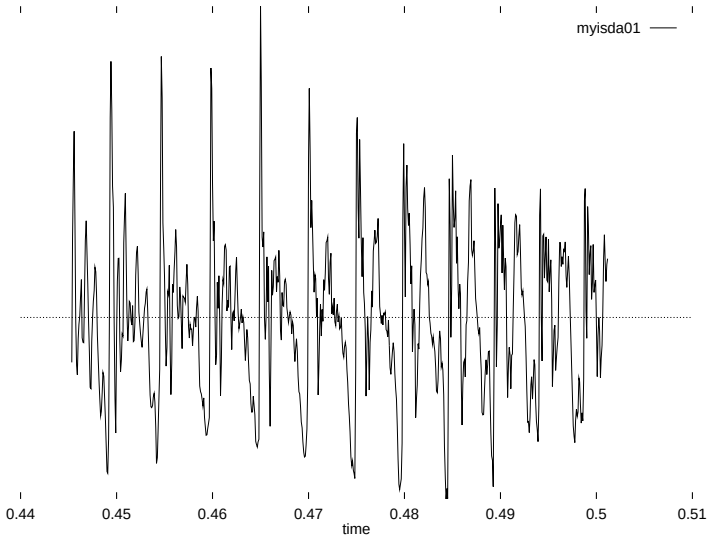

図 2.1: 音声波形

過程を考慮したモデルとその逆フィルタについて説明する。

2.1 ピッチ抽出法

ピッチ抽出は音声波形の繰り返し周期を求める問題である。代表的な自動ピッチ抽出の手法としてケプストラム [9] の最大値から周期を求めるケプストラム法がある。この他にも LPC 分析法、ラグ窓法 [10] など、古くから多数の方法が提案されている [5]。Rabiner らは、異なる発声内容、録音条件、話者等によるさまざまな環境下で収録した音声資料をもとに、それらの F_0抽出精度、有声、無声の弁別能力、計算コストなどの比較評価 [6] を行ったが、各手法それぞれに一長一短があり、すべての評価において他の手法よりも優れている手法を見出すことができなかった。この状況は 90 年代になった現在も相変わらず、それぞれの抽出の長所に応じて使い分けられていて、決定的な手法は確立されていない。

これらの方法は時間軸上の数十 ms 間の音声波形を切り出したものに対して、処理が行なわれる（図 2.2 ）。この処理を取り出すフレームの位置を少しずつ時間軸方向にずらし

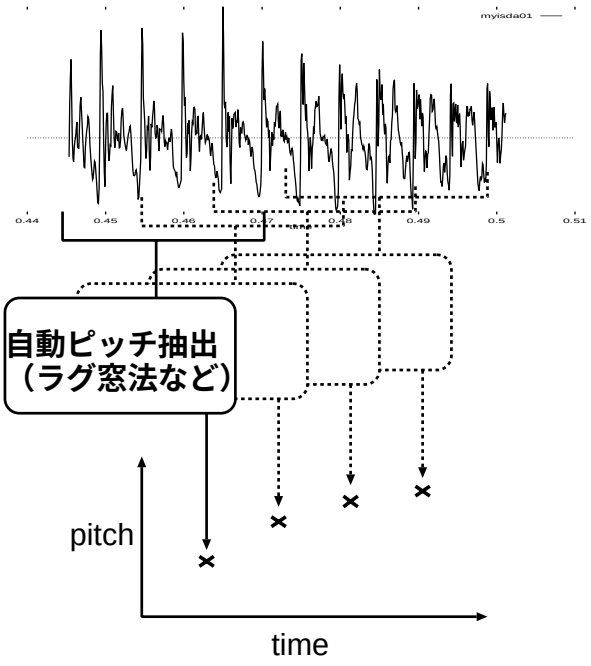

図 2.2: ピッチパターン抽出の動作

ながら処理を行なうことにより、ピッチパターンが得られる。

ここでの問題点はピッチパターン上で、時間的に隣り合ったピッチ周波数どうしが、全く独立して抽出されている点である。ピッチ周波数は声の高さに相当するものであり、自

然な発話では滑らかに変化すると期待できる。しかし、この処理では、数値計算上１フレーム隣のピッチとの関連性はなく、倍ピッチ誤り、半ピッチ誤りなどのピッチ抽出誤りも避けられない。また、ピッチは声帯の振動周波数に起因するものであるので、声帯の振動を伴わない無声音の部分ではピッチは存在しない。そのため、時系列パターンとして連続性が保たれず、韻律情報として利用するには都合が悪い。そこで平滑化の処理が行われる。(図 1.2) 平滑化には従来から以下の手法が用いられる。

- 線形補間
- スプライン補間 [4]
- 平滑化フィルタ [1]
- 直線近似 [8]

これらはいずれも数値解析的な処理であり、ピッチ抽出誤りを正確に訂正することができずに、ピッチパターンとしての物理的な特徴を損なう恐れがある。

ピッチパターンと、その根底にある言語的・パラ言語的情報との関係を定量的に表現する手法として、喉頭制御の物理的機構に根拠をもつピッチパターン生成過程モデルが考案されている。このような生成過程モデルに沿った形にピッチパターンを平滑化することで、先に述べたような従来法での不具合を避けることが可能であると思われる。本研究では藤崎らの提唱した F_0生成過程モデル [3] を用いた平滑化を行うことにする。このモデルは、音声合成において自然なイントネーションを与えるという目的で使用されており、日本語に対して良いモデルであると言われている。次節では F_0生成過程モデルについて説明する。

2.2 F_0生成過程モデル

2.2.1 モデルの概要

藤崎の提唱した F_0生成過程モデルでは、ピッチ周波数パターンが句頭から句末にかけて緩やかに降下するフレーズ成分と、局所的に起伏するアクセント成分との和によって表される。(図 2.3)

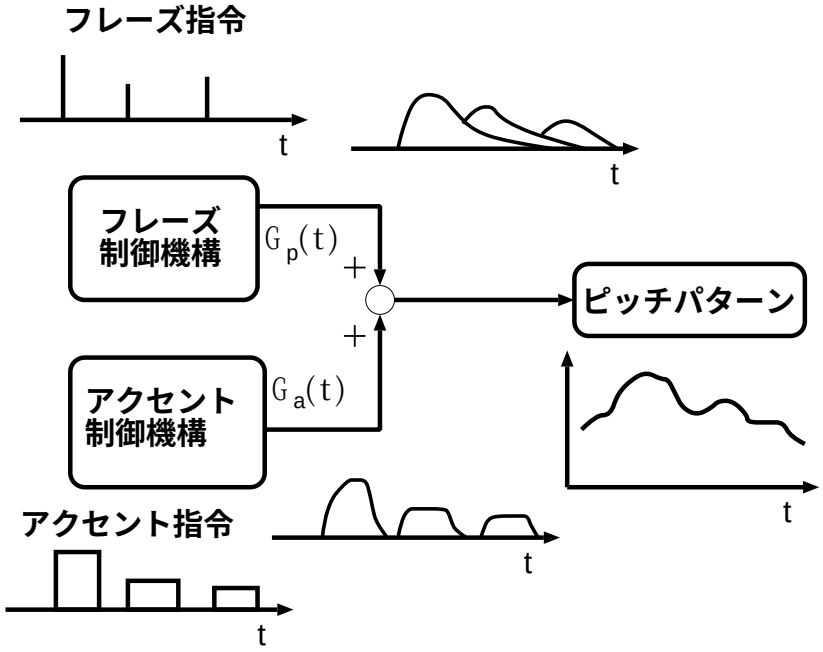

図 2.3: 藤崎モデル

フレーズ指令がI 個、アクセント指令がJ 個ある場合の基本式は以下のように表される。

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

第１項は声帯に可能な最低振動数、A_{pi}, A_{aj}はi 番目のフレーズ指令およびj番目のアクセント指令の大きさ、T_{0i}はi 番目のフレーズ発生時刻、T_{1j}, T_{2j}はj番目のアクセント指令の開始時刻および終了時刻である。また、第２項中の$G_p(t)$ はフレーズ制御機構のインパルス応答関数で、

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & \text{for } t \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

第３項中の$G_a(t)$ はアクセント制御機構のステップ応答であり、

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \theta], & \text{for } t \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

となる。以後、フレーズ、アクセント指令から各制御機構を通して生成されるパターンをそれぞれ、フレーズ指令成分、アクセント指令成分と呼ぶことにする。

また、フレーズ指令とアクセント指令の発生順序には図 2.4 ような制約がある。まず発話の開始は必ずフレーズ指令で始まる。そして矢印に沿った順序で指令が発生し、最終的には必ずアクセント終了指令で終わる。その他にアクセント終了指令の後に負のフレーズ指令が付加されることもある。

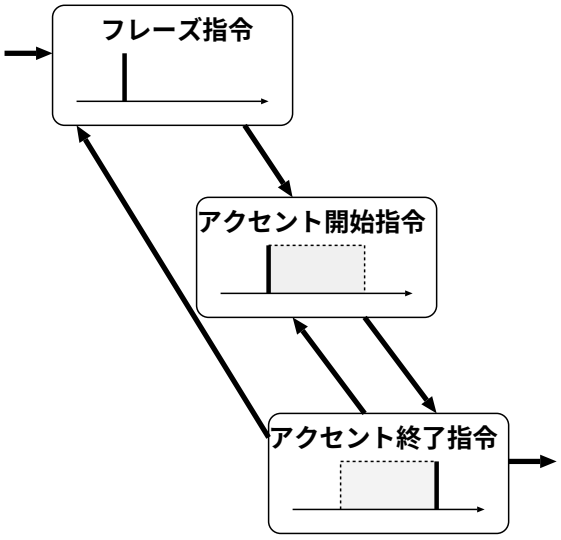

図 2.4: 指令の発生順序の制限

図 2.5 のようなコマンドが与えられたときの生成ピッチパターンは図 2.6 のようになる。

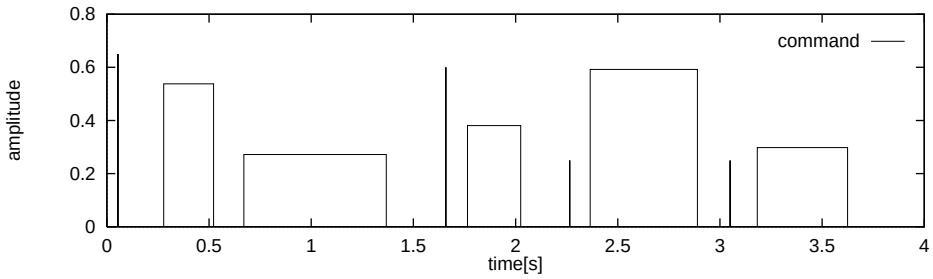

図 2.5: コマンド

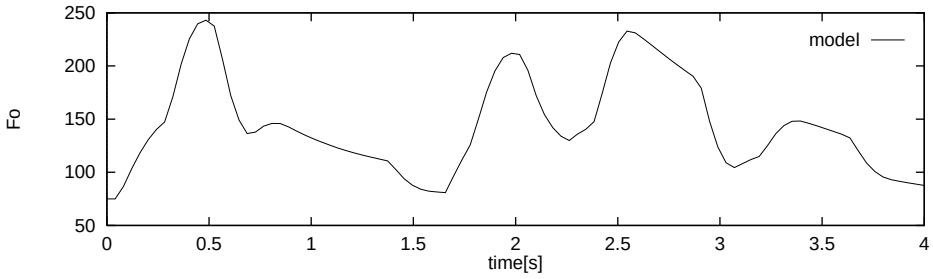

図 2.6: ピッチパターン

本研究では上述の F_0パターン生成過程モデルに基づいた基本周波数パターンの再構成を目的としている。再構成とは、自動抽出されたピッチパターンを一旦、モデルの指令系列として表現し、その指令系列から再びピッチパターンを生成する操作のことを示す。再構成されたピッチパターンは、ピッチ抽出誤りが訂正されており、かつ重要な韻律的特徴が表現されていることが望ましい。従来からの再構成の手法としては A-b-S(Analysis by Synthesis) [7] による方法、Geoffrois が用いた方法 [2] などがある。A-b-S とは入力ピッチパターンに対してパラメータの仮説をたて、そのの評価値が高くなるように反復計算を繰り返し、最適なパラメータへと収束させていく方法である。この方法は入力値として適切なパラメータ仮説を人為的に与えてやる必要がある。Geoffrois が用いた方法は再帰的に指令を決定して行く方法である。この方法は、無声音の部分を取り除くなどの前処理をして、入力とするピッチパターンをある程度選別する必要がある。また、両手法とも時間軸に沿った onepass の処理ではないために入力の終了まで出力を待機する必要がある。

本研究では

- 指令の発生時刻に指令の大きさが推定できるために時間軸を遡った処理を必要とせず onepass の処理が可能であること。
- 初期値を与えずに自動的に指令の推定が可能であること。

の理由から、逆フィルタを用いて指令を検出し再構成する手法を用いることにする。

2.2.2 逆フィルタ

フレーズ、アクセント各制御機構の逆過程を表す伝達関数を求めることで、それぞれの指令成分から各指令の発生タイミング、ならびに大きさを検出する逆フィルタ [17] [12] が

作成できる。

フレーズ制御機構逆フィルタ

基本式のフレーズ制御機構から

$$G_p(t) = \alpha^2 t \exp(-\alpha t)$$

ラプラス変換して

$$H_p(s) = L[G_p(t)] = \frac{\alpha^2}{(s+\alpha)^2}$$

$$H_p^{-1}(s) = \frac{(s+\alpha)^2}{\alpha^2}$$

これに

$$s = \frac{1 - z^{-1}}{T}$$

を代入すると下式のようになる。

$$H_p^{-1}(z) = \frac{z^{-2} - 2(\alpha T + 1)z^{-1} + (\alpha T + 1)^2}{\alpha^2 T^2}$$

この式の zを t に変数変換し、フレーズ制御機構逆フィルタ Y_pは

$$Y_p[t] = b_0 \times \hat{f}_0[t] + b_1 \times \hat{f}_0[t-1] + b_2 \times \hat{f}_0[t-2]$$

となる。ただし、$\hat{f}_0[t] = \ln f_0[t]$ であり、係数 b_kは

$$b_0 = \frac{(\alpha T + 1)^2}{\alpha^2 T^2}$$

$$b_1 = \frac{-2(\alpha T + 1)}{\alpha^2 T^2}$$

$$b_2 = \frac{1.0}{\alpha^2 T^2}$$

である。ここで Tはサンプリング周期である。なお、αは文献 [16] で示された値から、

$$\alpha = 3.0$$

で固定値とする。

アクセント制御機構逆フィルタ

基本式のフレーズ制御機構から

$$G_a(t) = 1 - (1 + \beta t)\exp(-\beta t)$$

ラプラス変換して

$$H_a(s) = L[G_a(t)] = \frac{\beta^2}{s(s+\beta)^2}$$

$$\begin{aligned} H_a^{-1}(z) &= \beta^{-2}T^{-3}\{(\beta T+1)^2 \\ &\quad -(\beta T+1)(\beta T+3)z^{-1} \\ &\quad +(2\beta T+3)z^{-2}-z^{-3}\} \end{aligned}$$

アクセント制御機構逆フィルタ Y_aは

$$\begin{aligned} Y_a[t] &= b_0 \times \hat{f}_0[t] + b_1 \times \hat{f}_0[t-1] \\ &\quad + b_2 \times \hat{f}_0[t-2] + b_3 \times \hat{f}_0[t-3] \end{aligned}$$

となる。式内の係数 b_kは

$$b_0 = \frac{(\beta T+1)^2}{\beta^2 T^3}$$

$$b_1 = -\frac{(\beta T+1)(\beta T+3)}{\beta^2 T^3}$$

$$b_2 = \frac{2\beta T+3}{\beta^2 T^3}$$

$$b_3 = -\frac{1.0}{\beta^2 T^3}$$

である。なお、βもαと同様に文献 [16] で示された値から、

$$\beta = 20.0$$

で固定値とする。

2.2.3　逆フィルタの出力

図 2.5 に示す指令系列から生成されたピッチパターン（以下、理想的ピッチパターン）(図 2.7 ）を逆フィルタに通した時の出力を図 2.8 、 2.9 に示す。0s、1.7s、2.3s、3.1s 付近がフレーズ指令の立上り時刻である。図 2.8 で、丁度その位置に現れている鋭いインパルス状の信号がフレーズ指令検出信号である。またそれ以外の小さな信号はアクセント開始終了の各指令の影響が現れているものである。

図 2.9 では２本重なってみえるうちの早い時刻のインパルス信号が、正確なアクセント指令検出信号である。アクセント指令はステップ状の信号であったが、フィルタ出力では指令の開始時刻に正のインパルス状信号、終了時刻には負のインパルス状信号が検出信号として現れる。

また、2.3s、3.1s 付近に現れている負の鋭いピークはフレーズ指令による影響である。それぞれのフィルタの出力は、互いの成分の影響が現れているものの、所望の発生時刻にそれぞれの指令が推定されていることが分かる。

2.2.4　逆フィルタの特性

フレーズ、アクセントそれぞれの逆フィルタは各指令成分が、1 つのみ含まれている場合には正確に大きさを検出できるが、指令成分どうしに重なりがある場合、つまり、各指令の発生間隔が短い場合は、過去の指令成分の影響を受け、誤った大きさの指令を検出すると考えられる。

図 2.10 〜 2.13 に、隣り合った指令の間隔による誤差の変化を示す。横軸が指令発生間隔、縦軸が正解を１としたときの検出結果を表す。前後の指令の大きさはここでは共に 1 を想定しているが、前の指令の方が大きくなれば誤差は大きく、後ろの指令が大きくなれば、誤差は小さくなると考えられる。

図 2.10で、横軸はフレーズ指令とその後に発生したフレーズ指令の時間間隔を表しており、縦軸は後に発生した指令を検出するときの検出誤差を示す。図 2.11は、アクセント終了指令の後のフレーズ指令を検出する場合の検出誤差、図 2.12は、アクセント指令の後のアクセント指令を検出する場合の検出誤差、図 2.12は、フレーズ指令の後のアクセント指令を検出する場合の検出誤差である。これらの図から次のことがわかる。

- フレーズ成分どうしの重なりは検出にほとんど影響は無い。

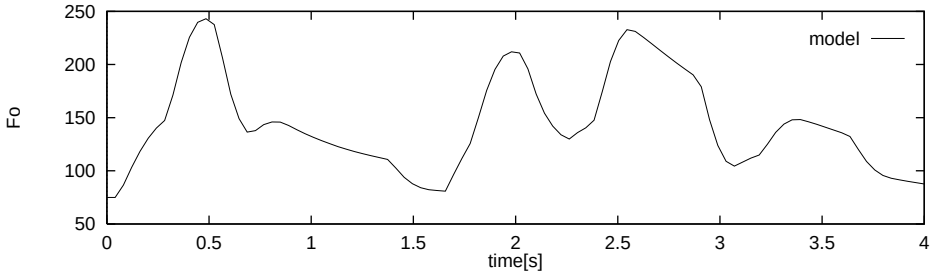

図 2.7: 理想的ピッチパターン

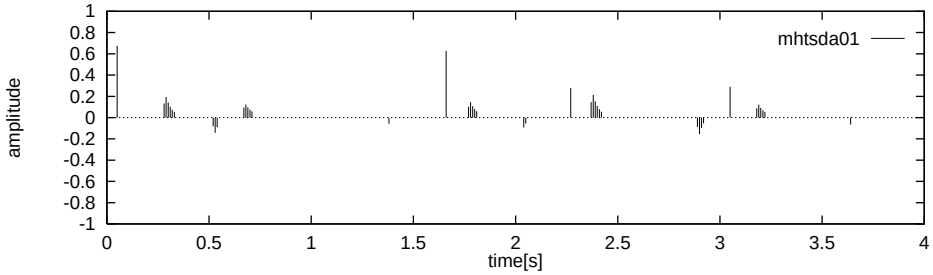

図 2.8: フレーズ逆フィルタの出力

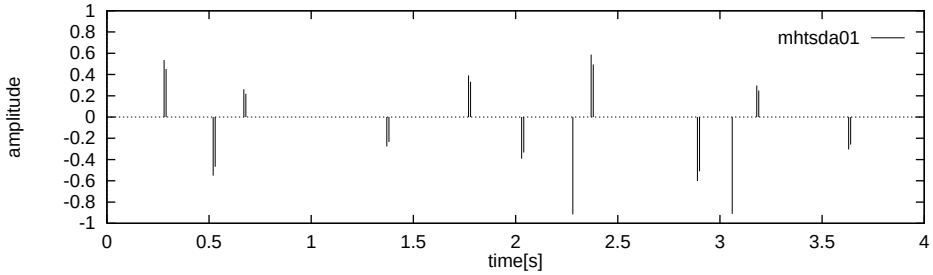

図 2.9: アクセント逆フィルタの出力

- フレーズ指令検出フィルタはアクセント指令成分の影響を大きく受ける。
- いずれのフィルタもアクセント成分の指令発生から 0.2s 以内に他の指令が発生している場合に誤差を受ける。

この結果から、フィルタの出力をそのまま指令の大きさの推定値として用いると、誤差が累積して正しい再構成が行われないということが考えられる。そのため、このような誤差が累積しないような方法が必要である。

次に実際の音声波形から抽出されたピッチパターンを入力として、実験する。ここで、モデルのパラメータ推定に用いるピッチパターンの抽出法としては、

- F_0抽出精度が高いこと。
- 有声、無声を問わず、連続量として F_0の値を抽出できること。(有声、無声の弁別能力は特に必要ではない。)

が条件である [11]。この条件と、ピッチ抽出と同時に 0 から 1 に正規化されたピッチの信頼度が得られる利点から、本研究ではラグ窓法 [10] を用いることにした。今後の実験にも入力となるピッチパターンはすべてラグ窓法により抽出したものを用いる。

図 2.14 に示す実際の音声波形から抽出されたピッチパターンを入力としたときのフィルタ出力を図 2.15 、 2.16 に示す。このフィルタ出力は、極端に大きい指令や小さい指令は取り除いて表示しているが、望ましくない時刻にも多くの指令を検出している。また、0s、1.5s 付近で発生しているはずのフレーズ指令が全く検出されていない。これは一般に発話の先頭のフレーズ指令は発声開始時刻よりも前に発生しているので、原理的に検出できない。そのため先頭のフレーズ指令は別の方法での検出または仮定が必要である。

これらの逆フィルタは、それぞれの時刻において指令が発生していると仮定した上で、推定を行なっており、指令の発生位置を特定することはできない。そのため、これらの指令系列のうち正しい指令のみを選択することが必要である。その方法として、本研究では次章で説明する時間同期型のビーム探索を行なうことにする。

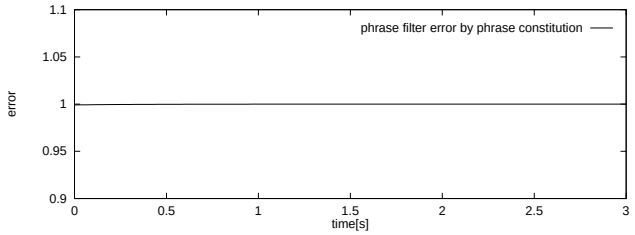

図 2.10: 過去のフレーズ指令成分によるフレーズ指令検出誤差

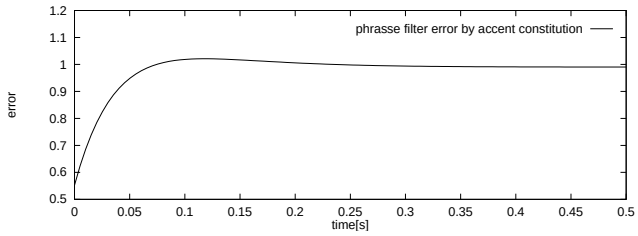

図 2.11: 過去のアクセント指令成分によるフレーズ成分検出誤差

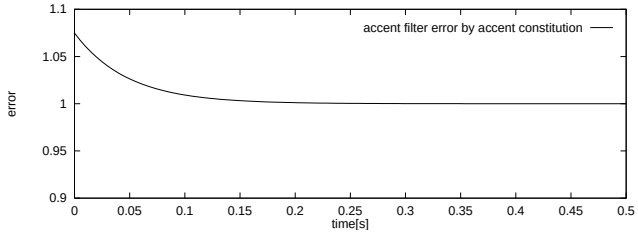

図 2.12: 過去のアクセント指令成分によるアクセント指令検出誤差

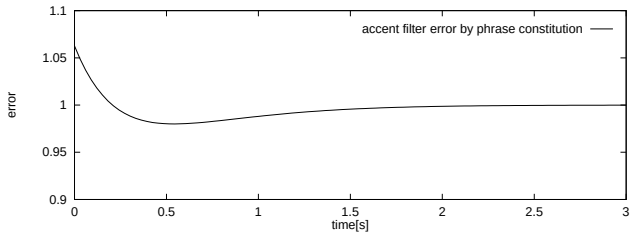

図 2.13: 過去のフレーズ指令成分によるアクセント指令検出誤差

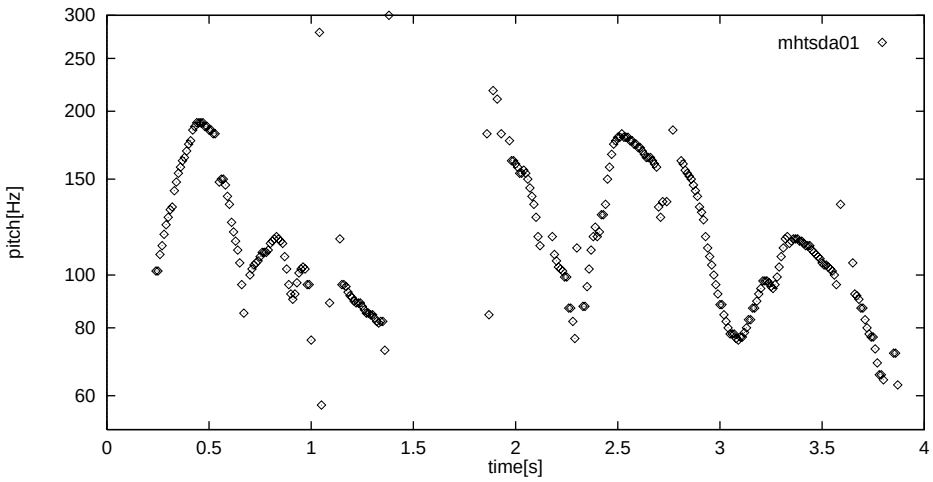

図 2.14: 自動ピッチ抽出法によるピッチパターン

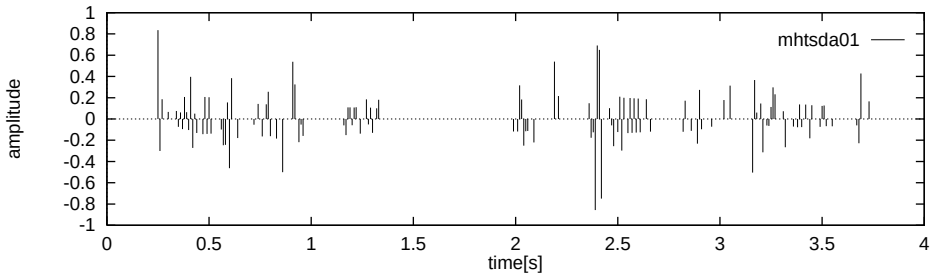

図 2.15: フレーズ逆フィルタの出力（観測ピッチ）

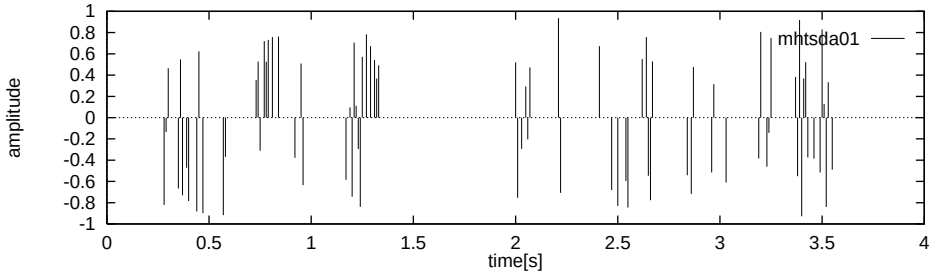

図 2.16: アクセント逆フィルタの出力（観測ピッチ）

第 3 章

ピッチパターンの再構成

3.1 再構成の手法

処理は常に時間軸に沿った one pass のビーム探索法に従い、歪みの小さい指令系列を選択する。探索空間は膨大であるため、各時刻において考慮にいれる指令系列の数を**ビーム幅**として定義し、最大値を Mに制限することで枝切りを行う。

入力となるピッチパターンの時刻 t における観測値を $P(t)$、処理中の時刻 t_p において m 番目の候補の指令系列を再構成したピッチを$\tilde{P}_{m,t_p}(t)$ と定義する。また、歪みは次式により計算する。

$$D_{m,t_p} = \sum_{t=0}^{t_p} W(t)(P(t) - \tilde{P}_{m,t_p}(t))^2$$

ここで、$W(t)$ は、各時刻 t におけるピッチ $P(t)$ の信頼度による重みである。$W(t)$ は時刻 t における信頼度がある閾値より大きい時 1、小さい時 0 の値をとる。信頼度は自動ピッチ抽出の過程で同時に得ることができる。

全体的な処理の流れは次のようになる。（図 3.1 ）

1. 発話の先頭フレーズ指令を近似的に推定する。（ 3.1.1 節）
2. 自動ピッチ抽出法を用いて得られたピッチパターンを F_0生成過程の逆フィルタに通すことにより、処理中の時刻における指令の大きさを推定する。（ 3.1.2 節）
3. 前時刻までの処理において、選択されている指令系列候補と、逆フィルタの推定結果から現時刻において考慮すべき指令系列を選択する。また、現時刻において考慮

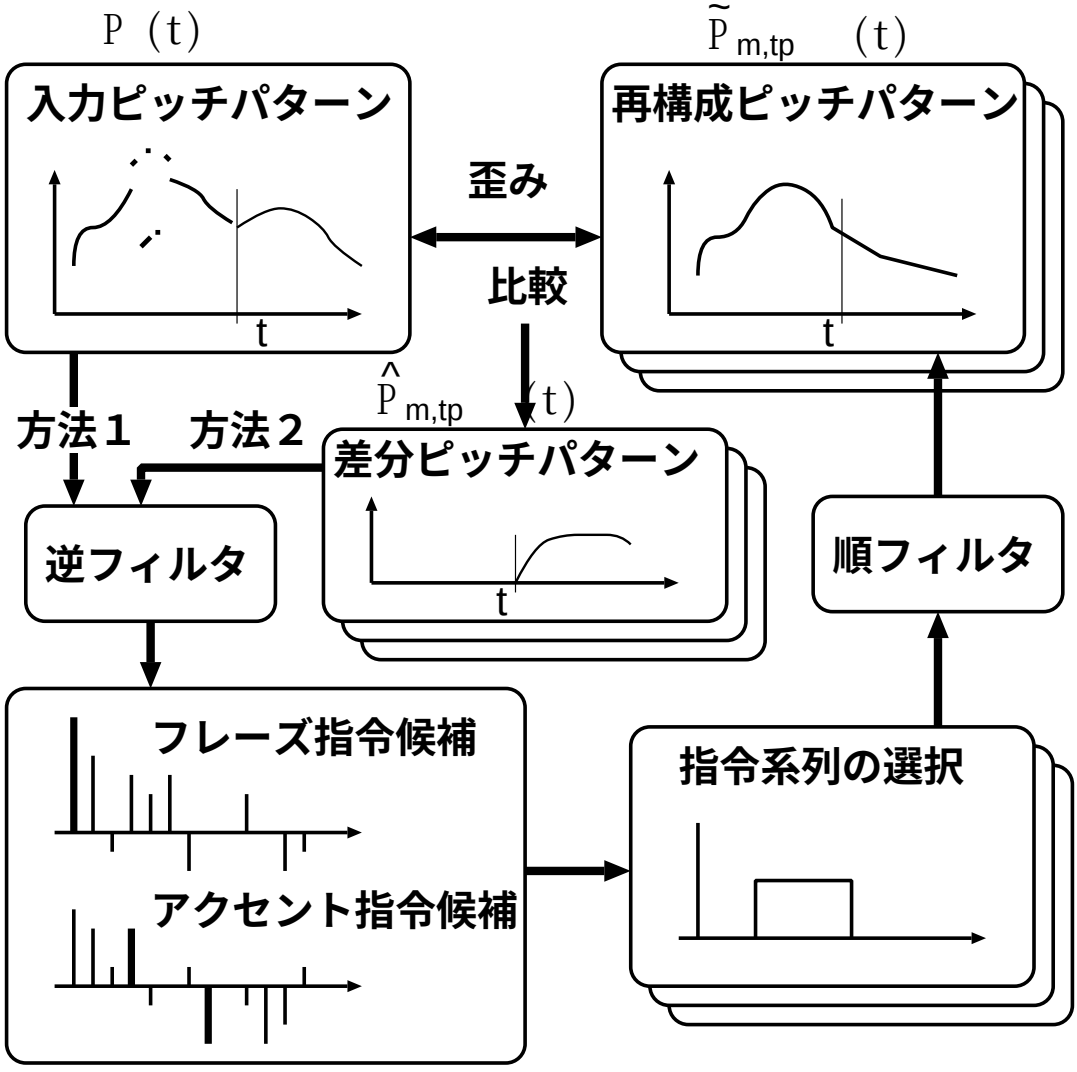
P (t)
入力ピッチパターン
t
歪み
比較
再構成ピッチパターン
t
方法１
方法２
逆フィルタ
差分ピッチパターン
t
順フィルタ
フレーズ指令候補
アクセント指令候補
指令系列の選択

図 3.1: 再構成法の概要

されている指令系列をもとに逐次ピッチパターンを再構成し、入力ピッチパターンとの歪みが小さいものうち M個を選択する。(3.1.3 節)

4. 2. と 3. の処理を 1 フレーム毎に発話終了時刻まで繰り返す。

5. 最終的に最も歪みの小さい系列を正しい指令系列とみなし、再構成を行う。

3.1.1 先頭フレーズ指令の推定

2.2.4で見たように、発話において先頭のフレーズ指令は発話開始時刻よりも前に発生するため、逆フィルタによる指令の推定は不可能である。そこで、発話開始時刻より前の各時刻において、ピッチパターンの極小値から先頭フレーズ指令の大きさを推定しておく。処理中の時刻 t_pにおけるフレーズ指令の大きさの推定値 A_pは、

$$A_p = \min_t P(t_p + t)/G_p(t)$$

3.1.2 指令候補の推定

入力ピッチパターン $P(t)$ から、フレーズ指令、アクセント指令をそれぞれの制御機構の逆フィルタを用いて検出する。(図 3.1)

方法 1

入力ピッチに手を加えずそのままアクセント及びフレーズの逆フィルタに通し、任意の時刻における指令候補をそれぞれ 1 つずつ出力する。理想的ピッチパターンでは指令の推定精度が良いが、観測パターンではピッチの抽出誤り、揺らぎ、量子化誤差があるため推定精度が悪いと思われる。また、 2.2.3 で見たように、指令発生間隔が近接している箇所では先行する指令の成分が大きく影響して、当該時刻における正しい指令の大きさを推定できないという欠点がある。

方法 2

時刻 t_pまでに推定した複数の指令系列候補中の m 位の系列によって生成される成分 $\tilde{P}_{m,t_p}(t)$ を観測ピッチパターン $P(t)$ から除き、その差分パターン$\hat{P}_{m,t_p}(t) = P(t) - \tilde{P}_{m,t_p}(t)$

を求める。この差分パターンを逆フィルタの入力とする。これにより、各指令の成分どうしの重なりによる推定誤りを避けることが可能になる。(図 3.2)

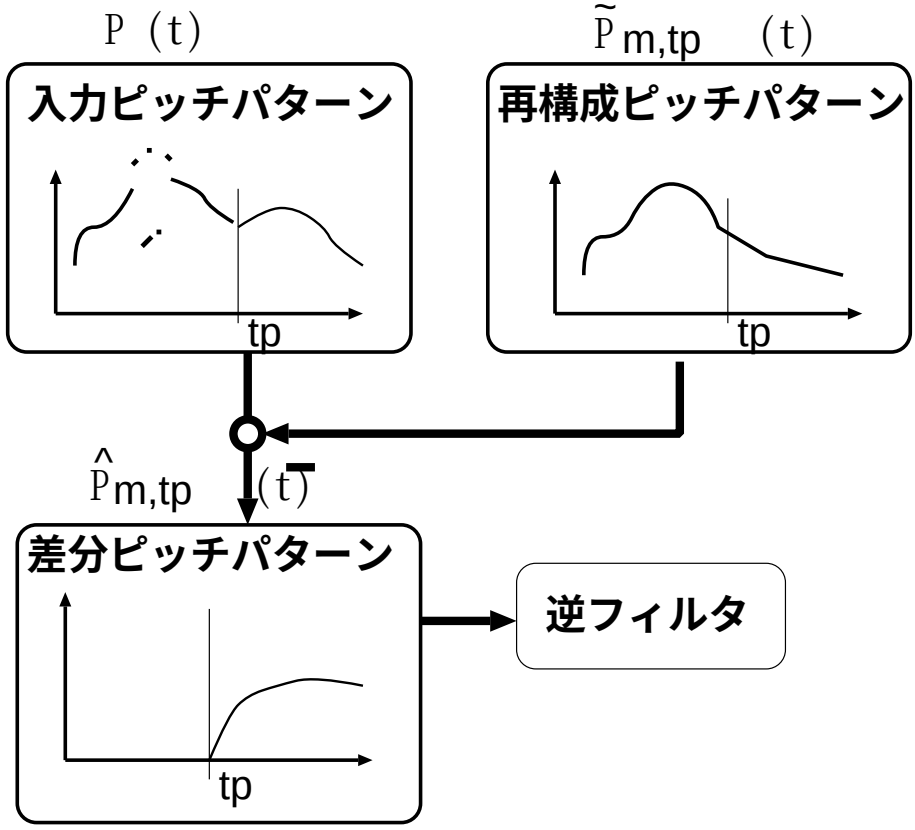

図 3.2: 過去の指令成分の除去

3.1.3 時間同期型の指令探索

探索は時間同期に行なう。時刻 t_pにおける処理の概略を図 3.3 に示す。処理の手順は以下のようになる。

1. M個の指令系列について時刻 t_pにおける歪みを計算する。
2. 逆フィルタの推定結果により、選択すべき指令系列を決める。

時刻 t_pにおいて、フレーズ指令 A_pが推定されたとき 時刻 t_p-1 における M個の指令系列候補の中から指令発生順序の制約（図 2.4 ）に基づいて、次に指令 A_pが

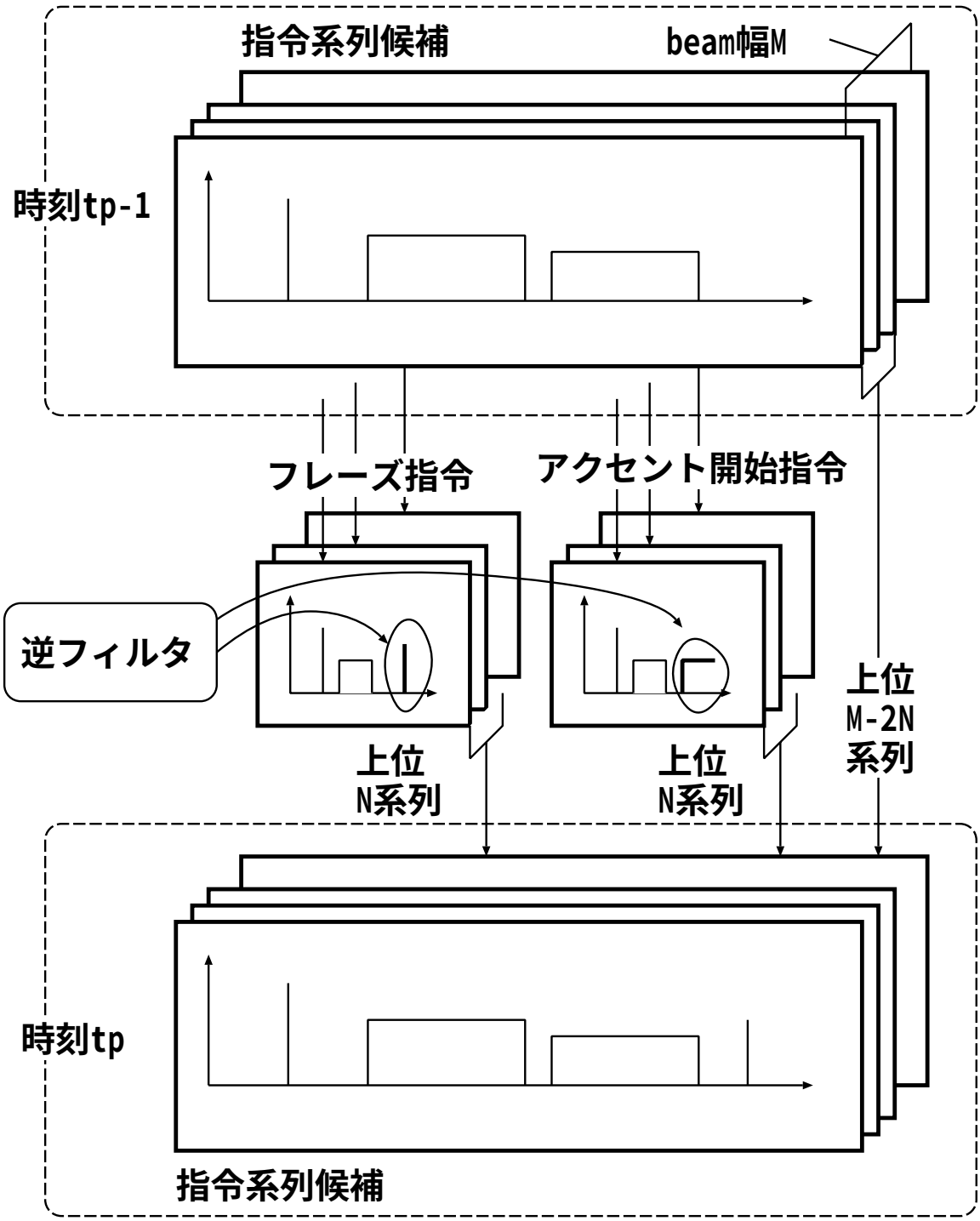

図 3.3: ビーム探索の各時刻に行なう処理

発生し得る指令系列を取り出し、歪み D_{m,t_p} が小さい順に最大 N 個（**n-best**）を選択する。それぞれの系列の最後に指令 A_p を加えた後、時刻 t_p における指令系列とする。

時刻 t_p において、アクセント指令 A_a が推定されたとき 同様に時刻 $t-1$ における M 個の指令系列候補の中から指令発生順序の制約に基づいて、次に指令 A_a が発生し得る指令系列のみを取り出し、歪み D_{m,t_p} が小さい順に最大 N 個（**n-best**）を選択する。また同様にそれぞれの系列の最後に指令 A_a を加えた後、時刻 t_p における指令系列とする。

3. また時刻 t_p において指令が発生しなかったと仮定して、時刻 t_p-1 の系列候補の中から歪みの小さい $M-2N$ 個の系列を残し、常に最大 M 個の指令系列を保つ。

3.1.4 理想的ピッチパターンの再構成

与えられた指令系列から生成された理想的ピッチパターンを入力として再構成を行ったのが、図 3.4 と図 3.5 である。 どちらも正しい時刻に指令が推定されており、ほとんど

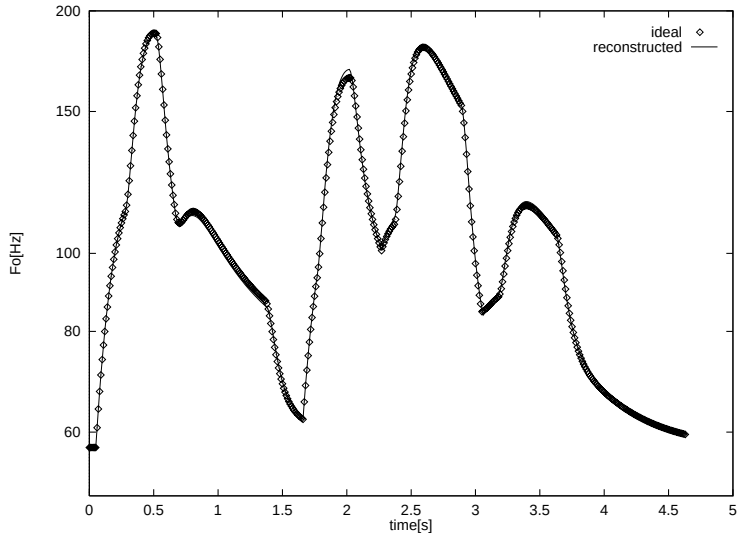

図 3.4: 方法 1 で再構成した結果

歪みのない近似ができている。ただし、方法 1では時刻 $t=2s$ 付近のピークに多少のずれが見られるが、方法 2では正しく近似されている。これは、2.2.4節で見たような過去の指

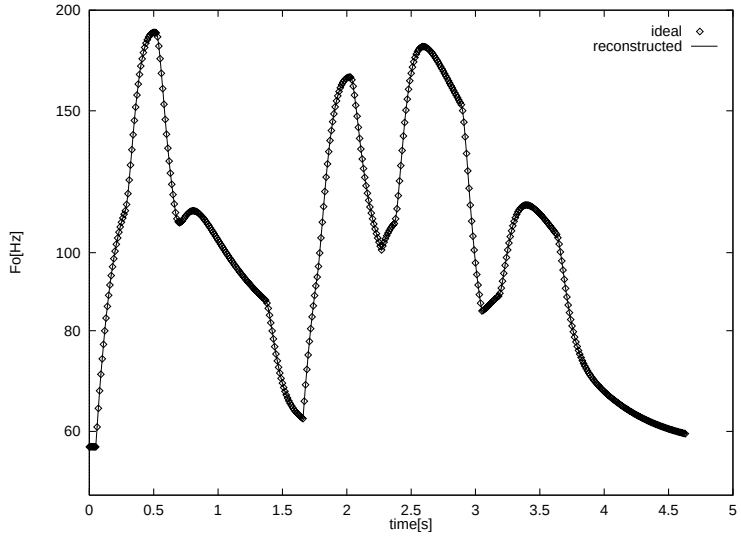

図 3.5: 方法 2 で再構成した結果

令成分による逆フィルタの出力の誤差を打ち消すことができたためであると考えられる。それぞれの歪みの大きさを表 3.1に示す。この歪みの値は入力と再構成ピッチの間の 1 フレーム当りの平均誤差を示す。方法 2の方が歪みが小さく、より正しく近似されていることが確認できた。

表 3.1: 歪み ($\log Hz/flame$)

	逆フィルタ（方法 1）	逆フィルタ（方法 2）
歪み	0.00945	3.76×10^{-6}

3.1.5 実験

ATR 連続音声資料データベース（503 文）のうち、男性話者 MYI の 50 文章について処理を行った。ただし、 2.2.1節で述べた指令の発生順序の制限を使って予備実験を行った結果、良好な結果が得られなかったので、図 3.6のように制限を緩めた。変更箇所は次の点である。

- アクセント開始指令の後にも再びアクセント開始指令を付け加えることができる。
- アクセント開始指令と終了指令の大きさは同じでなくても良い。
- 発話の一番最後のみ負のフレーズ指令を加えることができる。

アクセント開始指令が発生している時刻には、子音が発声されていることが多い。そのためアクセント開始指令が検出されず、1 つ分のアクセントの山が検出されないことになる。これを補うために 1 つ目の変更を行った。2 番目は先頭のフレーズ指令が正確に予測できるとは限らず、重要な情報であるアクセント指令間の谷間を埋めてしまう可能性があるため、付け加えた。3 番目の変更は、日本語は語尾のピッチが極端に下がることがあるため、それをうまく近似するために付け加えた。この変更は一般的にも加えられることが多い。図 3.7 、図 3.9 は「あらゆる現実を全て自分の方へねじ曲げたのだ」という発声に

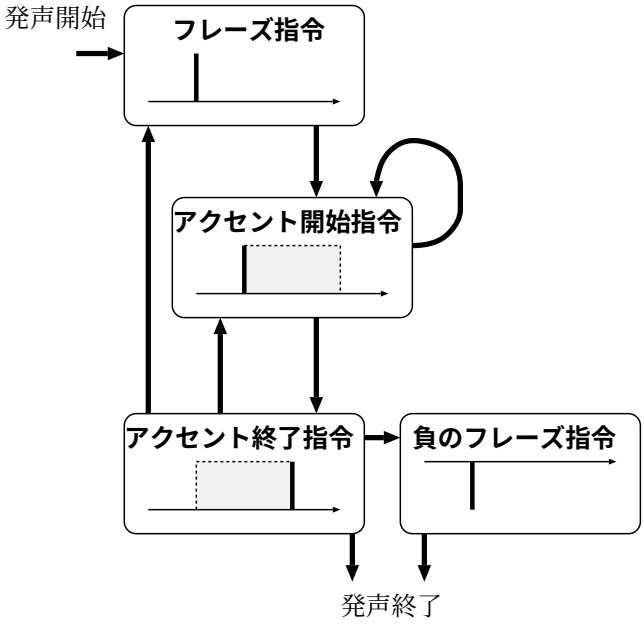

図 3.6: 指令の発生順序の制限

対してビーム幅 8000、n-best 数 140 の条件のもとで再構成を行った結果である。各点が自動抽出ピッチ、実線が再構成ピッチパターンである。また、図 3.8 、図 3.10 はそれぞれ、図 3.7 、図 3.9 のコマンド系列を表したものである。インパルス状の信号がフレーズ指令、ステップ状の立上りがアクセント開始指令、下降が終了指令である。 また、ビー

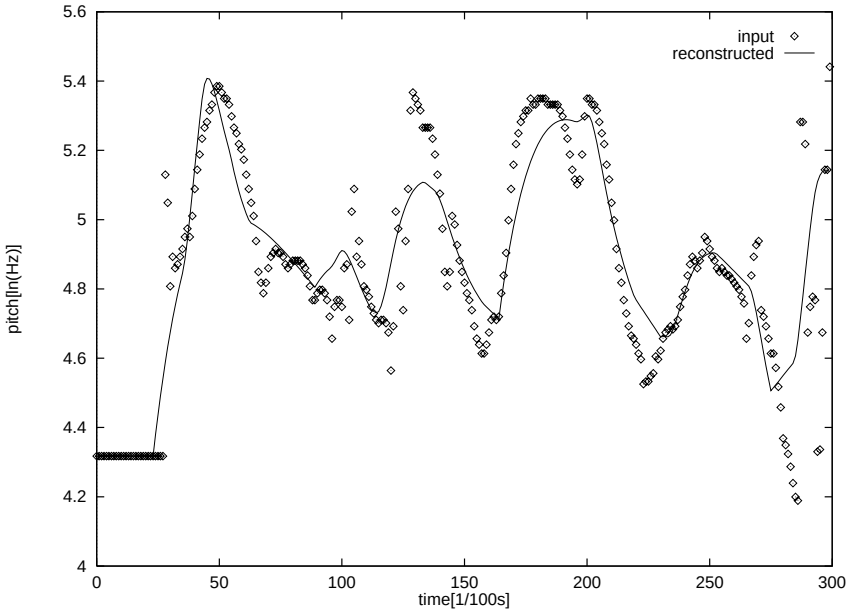

図 3.7: 方法 1 による再構成結果

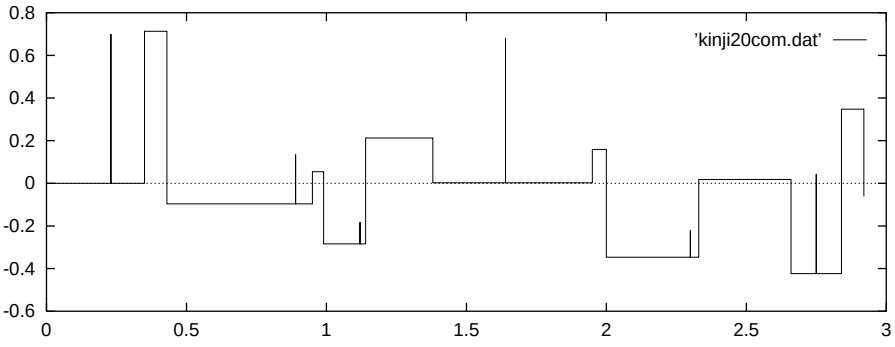

図 3.8: 指令系列

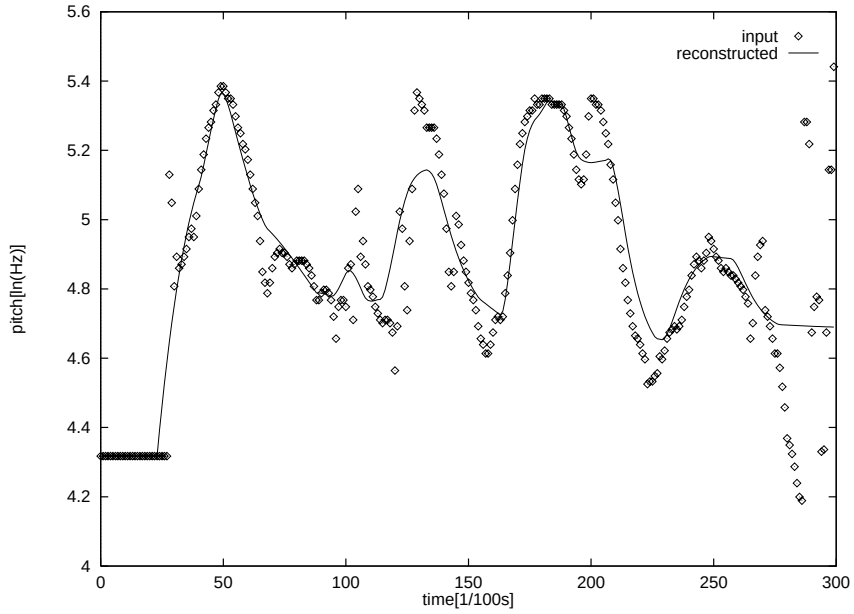

図 3.9: 方法 2 による再構成結果

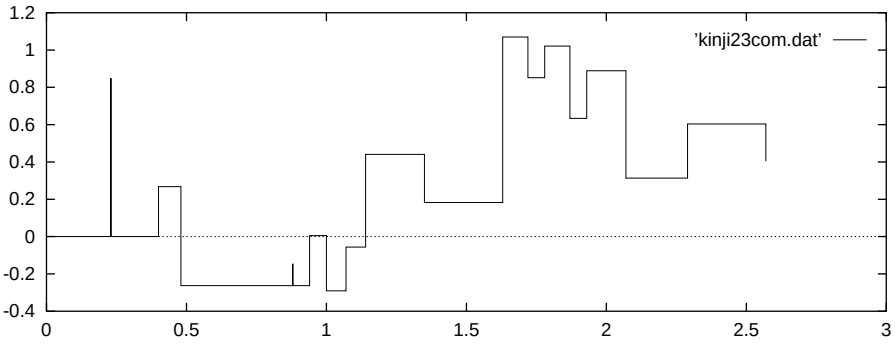

図 3.10: 指令系列

ム幅 Mと n-best 数 Nをそれぞれ変化させた場合の歪みの変化を図 3.11 に示す。横軸はで n-best 数、縦軸は 1 フレームあたりの二乗誤差歪みに平方根をとったものを 50 文章について平均した値である。

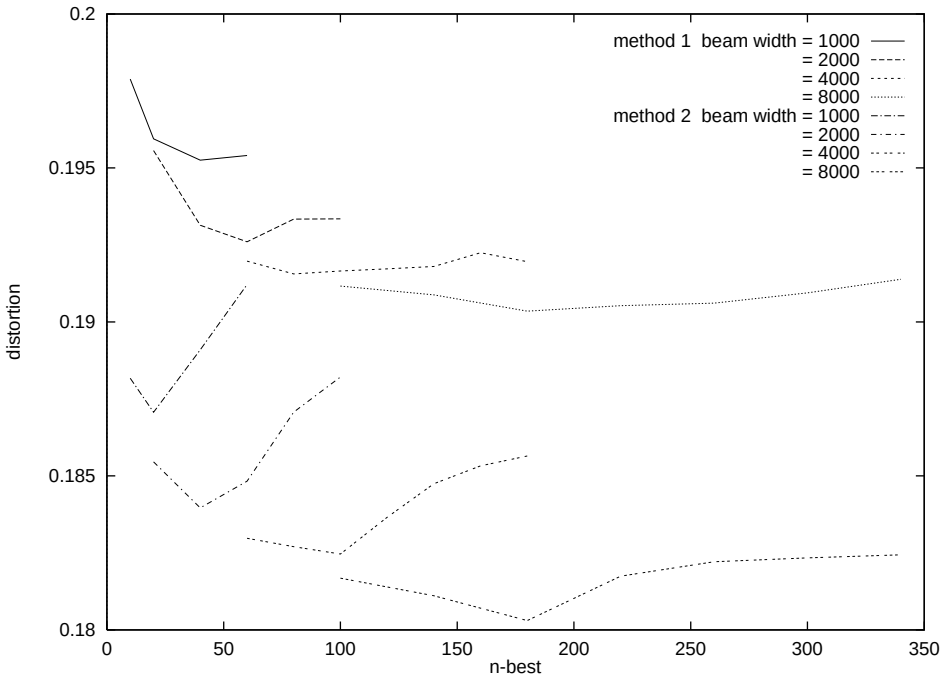

図 3.11: ビーム幅、n-best に対する歪みの変化

3.1.6 考察

図 3.7 、図 3.9 より、部分的に観測ピッチパターンの物理的な特徴が表現されていない箇所があるが、全体的に近似した形が得られていることがわかる。問題点としては、方法 1、方法 2のいずれの手法においても、最小二乗誤差近似を重視しているため、韻律の物理的な構造上、アクセントやフレーズの指令の発生が不適当だと思われる時刻にも、指令を付加してしまっていたり、フレーズ指令の成分をアクセント指令の成分で補っている点である。また、指令系列の図を見るとアクセント開始指令よりもアクセント終了指令が大きいために、底辺がマイナスの領域に落ち込んでいる。これは、先頭のフレーズ指令推定において実際より大きく推定されてしまったためであると考えられる。このような指令系列は本来の藤崎モデルにはない形であるが、先頭のフレーズ指令が正確に推定できない以上、

避けられない。また、アクセント開始指令の発生位置が無声音で指令が検出できないために１つ分のアクセント句が抜け落ちてしまう問題もあった。

図 3.11 からは、ビーム幅を大きくするにつれて歪みが小さくなることが確認できた。また、一定のビーム幅に対しては歪みを最小にする最適な n-best が存在することがわかる。つまり、n-best を大きくすれば歪みは小さくなるが、ビーム幅に対してある一定値を越えるとかえって歪みが大きくなってしまう。これは、n-best に対してビーム幅が小さくなると近視的な探索となり、誤差が積み重なり修正できなくなったものと考えられる。また、過去の指令による成分を除去した方法 2の方が、歪みが小さく良好な再構成が可能になった。これは、 2.2.3 で考察した過去の指令成分による逆フィルタの検出誤差の影響を減少させることができたと考えることができる。

3.2 基本指令成分を利用したフィルタ

逆フィルタの結果では、指令発生位置のピッチ抽出が正確でなければ、良好な結果が得られないため他のフィルタについて考える。

3.2.1 フィルタの概要

概要を、図 3.12 に示す。ここで仮に観測ピッチパターンが、フレーズ指令のみからなるとするとモデルの基本式より、

$$P(t) = A_p G_p(t - tp) + \ln F_b$$

が成り立つ。よって、以下の式が成り立つ。

$$A_p(t_p) = \frac{1}{d} \sum_{t=t_p}^{t_p+d} (P(t) - \ln F_b)/G_p(t - t_p)$$

ここで、d は平均をとるフレーム数である。この式はピッチパターンから時刻 t_pにおける指令の大きさを推定するフィルタと見ることができる。以下、このフィルタを基本指令成分フィルタと呼ぶことにする。

アクセント指令についても同様に、

$$A_a(t_p) = \frac{1}{d} \sum_{t=t_p}^{t_p+d} (P(t) - \ln F_b)/G_a(t - t_p)$$

が成り立つ。ここでは、逆フィルタの代わりにこの基本指令成分フィルタを用いて、再構成を行う。フィルタの入力としては逆フィルタの方法 2と同様、差分ピッチパターンを用いる。

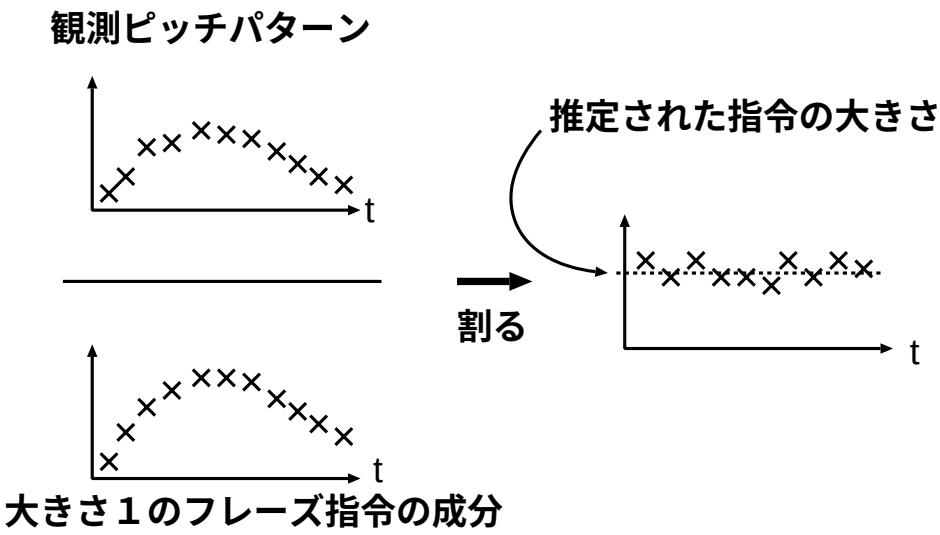

図 3.12: 基本指令成分を利用したフィルタ

3.2.2 理想的ピッチパターンの再構成

与えられた指令系列から生成された理想的ピッチパターンを入力として再構成を行ったのが、図 3.13 である。逆フィルタと同様にほとんど正確な近似ができている。歪みは表 3.2のようになった。ここで、逆フィルタの方法 2 よりも歪みが大きくなっている。入力が理想的なパターンである場合、ピッチが存在しない無声音の部分や、ピッチ誤りが含まれておらず、本フィルタの改善点が生かせない。また基本指令成分フィルタでは評価の対象が広いため、後に発生する指令の成分も評価の対象としてしまい、かえって歪みが大きくなったと考えられる。

3.2.3 実験

ATR 連続音声資料データベース（503 文）のうち、男性話者 MYI の 50 文章について処理を行った。ただし、逆フィルタのときと同様に、図 3.6のように指令発生順序の制限

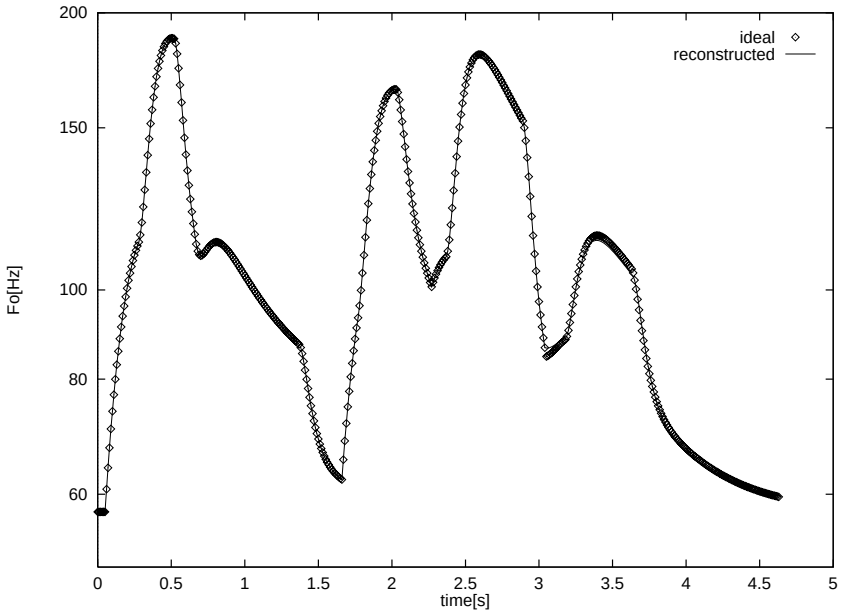

図 3.13: 基本指令成分フィルタを使用して再構成した結果

表 3.2: 歪み $(\log Hz/flame)$

	逆フィルタ（方法 1）	逆フィルタ（方法 2）	基本指令成分フィルタ
歪み	0.00945	3.76×10^{-6}	0.00465

を緩めた。図 3.14 は「あらゆる現実を全て自分の方へねじ曲げたのだ」という発声に対してビーム幅 8000、n-best 数 140 の条件のもとで再構成を行った結果である。 各点が自

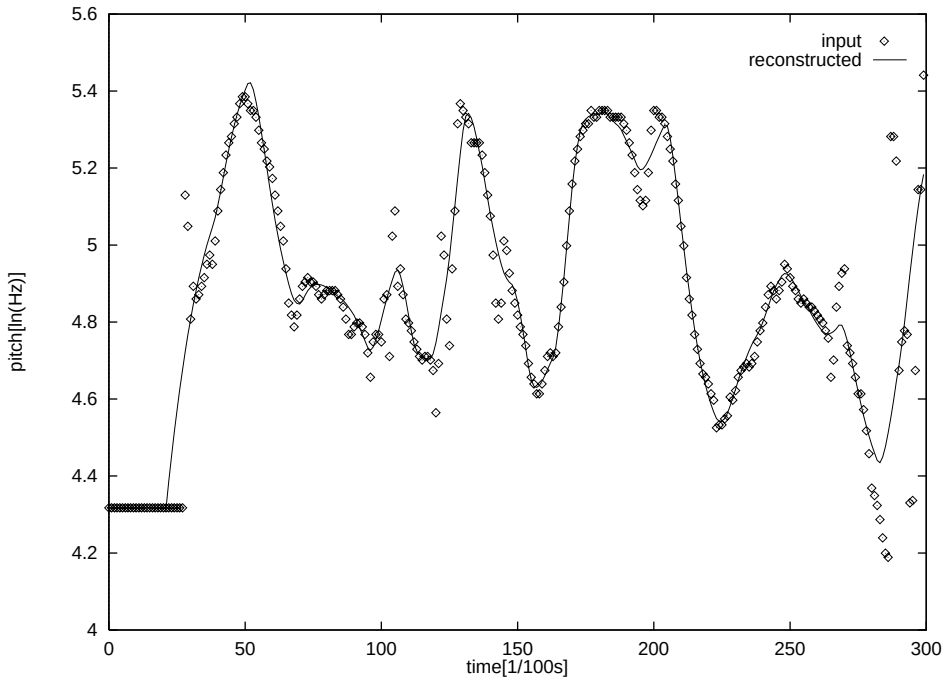

図 3.14: 基本成分フィルタによる再構成結果

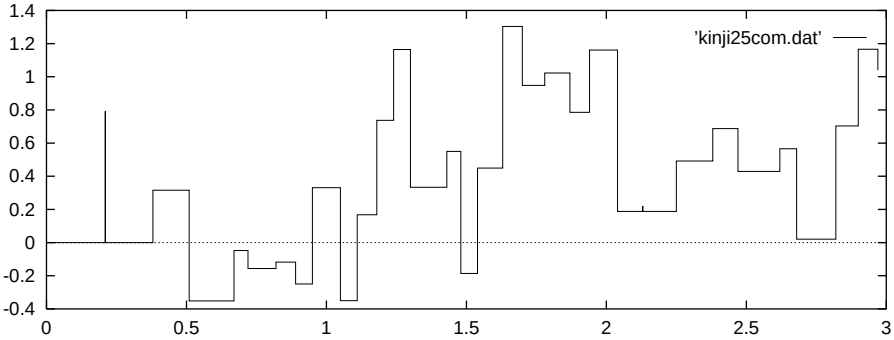

図 3.15: 指令系列

動抽出ピッチ、実線が再構成ピッチパターンである。また、図 3.15 は、図 3.14 のコマンド系列を表したものである。インパルス状の信号がフレーズ指令、ステップ状の立上りがアクセント開始指令、下降が終了指令である。

また、ビーム幅 M と n-best 数 N をそれぞれ変化させた場合の歪みの変化を 3.16 に示す。比較のため、逆フィルタの方法 2による歪みの変化も重ねている。横軸はで n-best 数、縦軸は 1 フレームあたりの二乗誤差歪みに平方根をとったものを 50 文章について平均した値である。

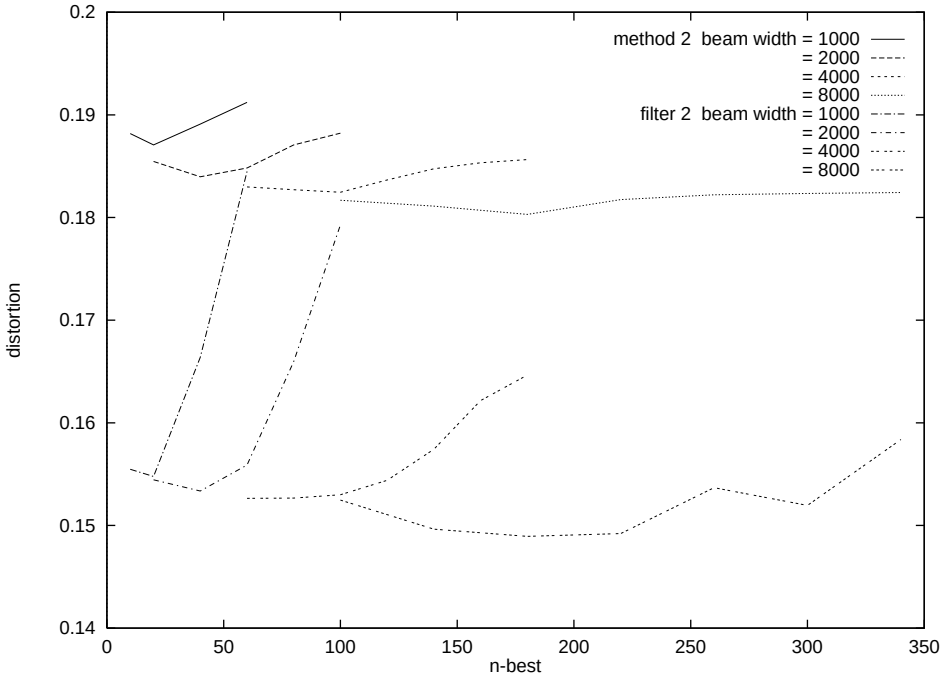

図 3.16: 基本指令成分フィルタによる歪みの変化

3.2.4 考察

図 3.14 より、逆フィルタと比べてより、目的に近い形が得られた。ここでも逆フィルタと同じ問題が現れている。しかし、一つ分のアクセント成分が抜け落ちてしまうことはなくなった。これは、本フィルタが、逆フィルタに比べ、広域的な評価も行っているためであると考えられる。図 3.16 からも、逆フィルタと同様のことが言える。全体的に逆フィルタを用いたときよりも歪みが小さくなった。これも同様に、本フィルタが、逆フィルタに比べ、広域的な評価も行っているためであると考えられる。

3.3　各再構成ピッチパターンの歪みによる評価

各手法によって再構成されたピッチパターンと A-b-S によって得られた理想的ピッチパターンとの歪みを計算することで評価を行う。

3.3.1　実験

ここでは A-b-S によって再構成されたピッチパターンを理想的なものとして扱い、二乗誤差歪みを計算することで評価を行う。ATR 連続音声資料データベース（503 文）のうち、男性話者 MHT の 25 文章について処理を行った。また従来法として以下の方法も同様の条件で処理を行った。

自動抽出ピッチ ラグ窓法により抽出したピッチ。

直線近似 ピッチパターンを複数の直線（折れ線）で近似する方法。

線形補間 ラグ窓法により抽出されたピッチパターンのピッチ信頼度が閾値より低い部分を直線で補間する手法。

移動平均 局所的な時間内でピッチの平均をとる方法である。

計算結果を表 3.3 に示す。これらの値は、無音部分以外の全時刻にわたる二乗誤差の和をフレームの長さで割り、平方根をとったものである。歪み 1 はピッチ信頼度の閾値を適切に与え、信頼度の低いピッチを削除したパターンを入力として与えた場合、歪み 2 はピッチ信頼度の閾値が与えられない場合を示す。線形補間は歪み 1 は適切な閾値 0.27 を与えた。この方法では閾値に 0 を与えた場合は自動抽出ピッチと同じ値になる。移動平均の平均化の幅は 0.16s とした。本手法には、歪み 1 には線形補間と同じ 0.27 の信頼度閾値を用い、歪み 2 は信頼度によるピッチの選別を行わなかった。図 3.17 に線形補間との比較の図を示す。各点が入力としたピッチパターン、実線が理想的ピッチパターンである。全体的に滑らかになっているのが基本指令成分フィルタによる再構成結果、直線で補間されている破線が線形補間によるものである。

表 3.3: 理想的ピッチパターンとの歪みの比較 ($\log Hz/flame$)

	歪み 1	歪み 2
自動抽出ピッチ	-	0.332
直線近似	-	0.259
線形補間	0.083	-
移動平均	-	0.216
逆フィルタ（方法 1）	0.157	0.206
逆フィルタ（方法 2）	0.148	0.192
基本指令成分フィルタ	0.139	0.171

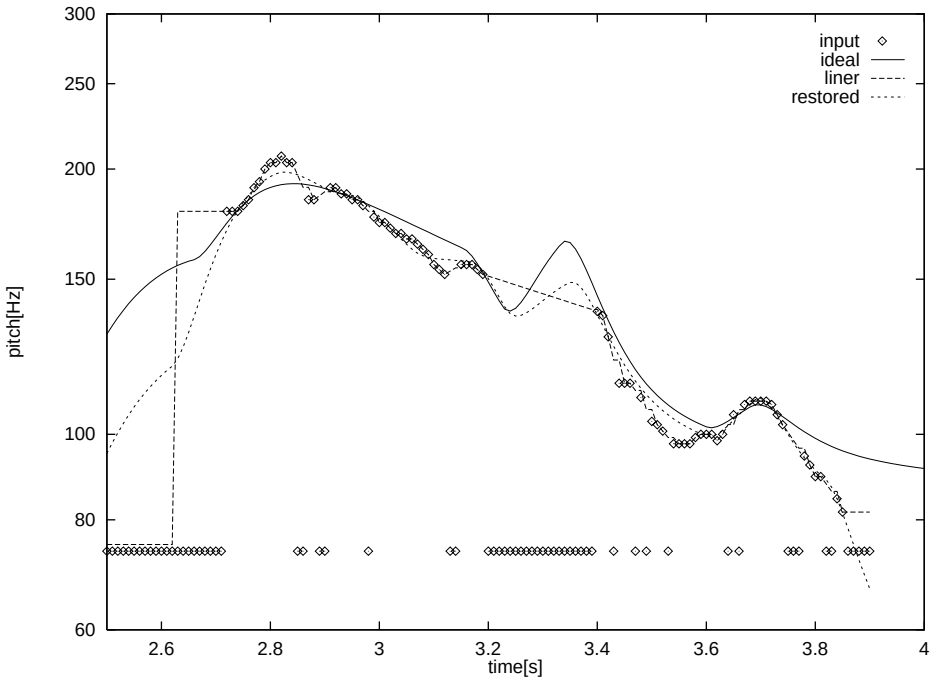

図 3.17: 線形補間との比較

3.3.2 考察

歪み 1 では、線形補間による歪みが、本手法の歪みを下回っているが、歪み 2 では本手法のいづれも従来法より歪みが小さく、理想的ピッチパターンに近い平滑化ができていることが確認できる。このことから、移動平均等の従来法と比べて、ピッチ抽出誤りによる平滑化への影響が少ないことがわかった。これはモデルによる拘束条件により、ピッチ信頼度を用いなくても、ある程度のピッチ誤りを除去できたことによると考えられる。

図 3.17 の時刻 $t = 3.2$ から $3.4s$ 付近を見ると線形補間ではアクセント指令によるピッチの山が完全になくなっている。この手法では信頼度の閾値として全体の平均歪みが最小になる 0.27 を与えたにも関わらず、信頼度が完全に正規化されていないために重要な情報が損なわれることになった。しかし、本手法では同じ入力でも前後のピッチパターンの形状により、アクセント指令の山がある程度再現されていることがわかる。このことからも本手法が入力ピッチパターンの状態に左右されにくいということがわかる。

3.4 各再構成ピッチパターンによる句境界検出

各手法によって再構成されたピッチパターンを対象に F_0連続整合法 [13] による句境界検出を試みる。F_0連続整合法ではピッチパターンの一部分を表したテンプレートを学習によって数種類用意している。入力ピッチパターンとの誤差が最小になるテンプレート系列パターンを求め、テンプレートの継目を句境界として検出する方法である。

3.4.1 実験

ATR 連続音声資料データベース（503 文）のうち、男性話者 MHT の 25 文章について処理を行う。視察句境界はデータベース付属の韻律情報を使用し、これを正解句境界とする。10 位までの句境界検出結果を以下の 3 点によって評価する。

$$\text{累積句境界検出率}\,(R_c^{10}) = \frac{\text{正解検出数}}{\text{視察句境界の総数}}$$

$$\text{平均句境界検出率}\,(\bar{R}_c) = \frac{\text{1 位から 10 位までの延べ正解検出数}}{\text{視察句境界の総数} \times 10(\text{位})}$$

$$\text{平均句境界挿入誤り率}\,(\bar{R}_i) = \frac{\text{1 位から 10 位までの延べ不正解検出数}}{\text{視察句境界の総数} \times 10(\text{位})}$$

ここでは無音区間と発声区間との境界は視察句境界数に含まない。なお、正解検出基準は、視察句境界の$\pm 60ms$とする。その結果、表 3.4 のようになった。ここで、理想的ピッチパターンとは、前節と同様に A-b-S によって得られたピッチパターンを入力とした時の検出精度である。今回の手法の他に従来法として、平滑化しないピッチパターン、線形補間、移動平均、直線近似の処理をしたものも同じ条件で行った。F_0連続整合法ではこの検出率が限界値であると考えられる。

表 3.4: 句境界検出精度 (%)

	R_c^{10}	$\bar{R}_c$	$\bar{R}_i$
理想的ピッチパターン	90	38.4	45.8
自動抽出ピッチ	75	39.7	82.1
移動平均	70	36.0	73.7
直線近似	81	40.9	74.0
線形補間	90	51.3	60.5
逆フィルタ（方法 1）	82	39.1	72.5
逆フィルタ（方法 2）	84	40.7	70.2
基本指令成分フィルタ	87	50.7	54.7

次に句境界検出結果の例を示す。図 3.18 が、本手法の基本指令成分フィルタによる再構成結果と線形補間によるピッチパターンである。各点が入力ピッチ、実線が本手法によるピッチパターン、破線が線形補間によるピッチパターンである。それぞれの結果を用いて句境界検出を行った結果を図 3.19 、図 3.20 に示す。1 つの鍵型で示されているのが、1 つの句で、上から 1 位から 10 位までの検出結果である。この場合の正しい句境界は 2.56s とされている。

3.4.2 考察

本手法では方法 1、方法 2、基本指令成分フィルタの順に累積句境界検出率 R_c^{10}が良くなった。これは、前節の歪みによる評価の結果からも推測できる。線形補間以外の従来法と比較すると検出率、挿入誤り率ともに良い結果が出ている。線形補間の検出率が良好

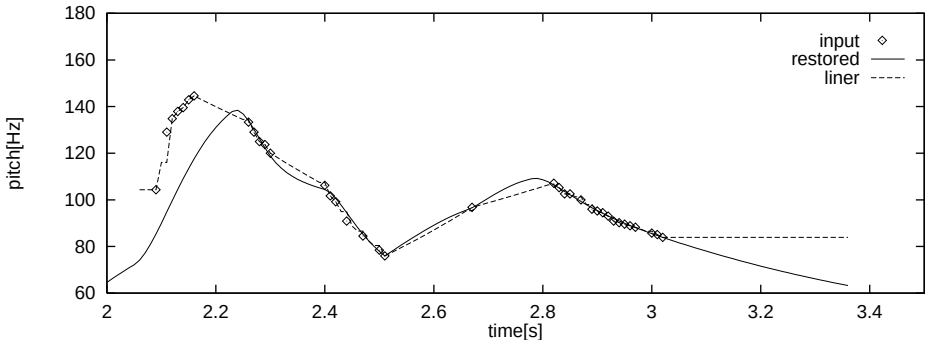

図 3.18: 本手法と線形補間による再構成結果

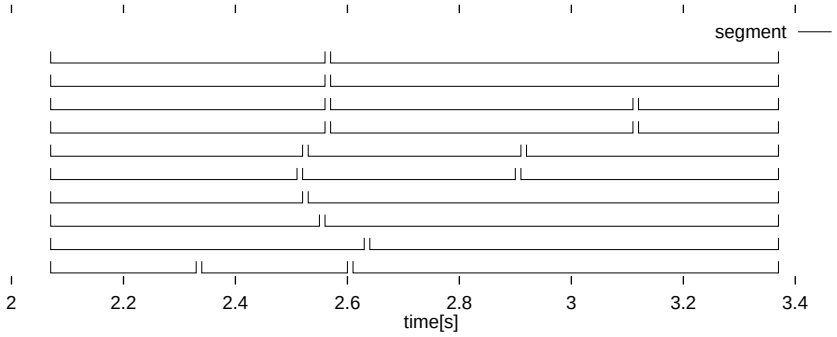

図 3.19: 基本指令成分フィルタによる句境界検出

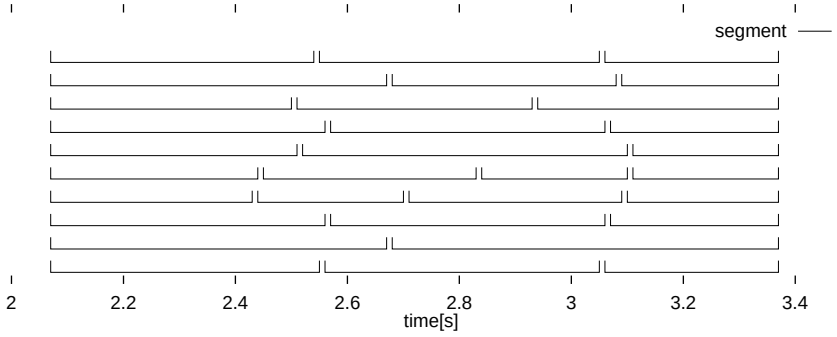

図 3.20: 線形補間による句境界検出

であるが、これはラグ窓法によってピッチ抽出誤りが少なく、またピッチ信頼度の閾値によって無声音区間が適切に除去されている良質なピッチパターンが抽出されている、また抽出誤りが句境界検出に重要でない部分で起こっていた等の理由により、良い結果が出たものと思われる。線形補間と基本指令成分フィルタを比べると、検出率は下がったが、挿入誤り率が改善されている。これは、基本指令成分フィルタでは、ピッチパターンの谷間も滑らかにしてしまうことがあるため検出率が下がり、また、ピッチ抽出誤りによる影響が少なくなっているため、余分な起伏が少なくなり、挿入誤り率が減少しているものと考えられる。

また、直線近似と移動平均を比較した場合、前節の歪みによる評価では移動平均の方が良好であったにも関わらず、本節の句境界検出率では逆転している。二乗誤差近似による歪みの評価が、後段の処理においても必ずしも有効な評価法でないことがわかる。

図 3.18 〜図 3.20 を見ると、線形補間による結果はかなり検出句境界が散らばっているのに比べ、本手法による検出結果は検出誤りが少なく、検出時刻も限定されている事がわかる。これはモデルに基づいたパターンであるため、テンプレートとして記録されているパターンと良く一致し、境界時刻の散らばりが減少したと考えられる。

3.5 各再構成ピッチパターンの推定指令正解率

各種法によって再構成されたピッチパターンを対象に推定された指令が A-b-S(Analysis by Synthesis) によって推定されたものを正解として、どの程度正しいかを調べる。本実験ではモデルによる拘束条件を緩和しており、また指令系列どうしの類似度の定義が困難なため、正確な比較はできないが、参考資料として掲載する。

3.5.1 実験

ATR 連続音声資料データベース（503 文）のうち、男性話者 MHT の 25 文章について処理を行った。AbS によって推定されたものを、各ピッチパターンの正しい指令系列とする。また、ここでは “指令が正しい” とは指令発生時刻が±100ms 以内であることとし、指令の大きさは考慮しなかった。推定指令正解率は次の式で評価する。

$$\text{推定指令正解率}\ (C_c) = \frac{\text{正解指令数}}{AbS\text{による指令の総数}}$$

$$\text{推定指令挿入誤り率}\,(C_i) = \frac{\text{推定された指令の不正解数}}{AbS\text{による指令の総数}}$$

評価結果を表 3.5に示す。右下の添字はそれぞれフレーズ指令 p、アクセント開始指令 ao、アクセント終了指令 ae を表す。

表 3.5: 各再構成ピッチパターンの推定指令正解率 (%)

	$C_{c,p}(C_{i,p})$	$C_{c,ao}(C_{i,ao})$	$C_{c,ae}(C_{i,ae})$
逆フィルタ（方法 1）	27(67)	36(128)	46(81)
逆フィルタ（方法 2）	21(60)	27(92)	40(48)
基本指令成分フィルタ	23(48)	83(180)	73(119)

3.5.2 考察

各フィルタの違いによる良否は、推定された指令の数が違うので、単純に比べることはできないが、歪みの大小関係と同様になった。どの手法も正解率が低く、生成モデルの自動指令推定法に用いるにはまだ至らない。特にフレーズ指令の正解率が悪く、先頭フレーズ指令の正確な推定は重要な課題である。また、どの手法にも共通してアクセント終了指令の検出率が高くなっている。これはピッチパターンのアクセントの立上りより、下降部分の方が良好な抽出が行われていることからもわかる。この特徴は今後指令推定にも考慮にいれるべき点である。

第 4 章

結論

4.1 研究の成果

本研究では、ピッチパターンから F_0生成過程モデルの指令を検出する各種フィルタを用いて、ピッチパターンを生成モデルに基づいて再構成する手法について検討した。各フィルタから検出された指令をもとに時間軸方向のビーム探索を行うことで、ピッチパターンの再構成が可能になった。今回用いた手法は、歪みの観点から比較すると、従来法に比べピッチ抽出誤りや無声音の部分の不正確なピッチによる影響が少ないことがわかった。また、ピッチ信頼度の閾値が不適切で本来あるべきピッチパターンの一部分が削除されていても、その前後のピッチから欠落部分を復元できた。自動で平滑化を行う場合、入力ピッチパターンの状態に左右されにくく、安定した平滑化が可能であると思われる。それぞれの手法によって再構成されたパターンを句境界検出に利用することで、句境界検出率、挿入誤り率が改善された。これらはモデルによる拘束条件により、ピッチ信頼度を用いなくても、ある程度のピッチ抽出誤りを除去できたためと考えられる。評価として、歪みと句境界の検出精度を見たが、それぞれの評価が必ずしも一致していなかった。ピッチパターンの評価法自体もモデルに基づいたものにする必要があると思われる。F_0生成過程モデルの自動指令系列推定法としては良い結果が得られなかったが、アクセントの終了指令の検出率は比較的良いので、今後その特徴を利用した方法が考えられる。

4.2 課題

本研究で用いたモデルは制限を緩めた上で用いているので、モデルを用いる効果が薄くなっていることが考えられる。今後、正式なモデルにしたがった方法を検討することが必要であると思われる。

また、自動抽出によって抽出されるピッチパターンはアクセントの立上りよりもアクセントの下降部分の方がピッチ誤りも少なく、フィルタによる指令の推定誤りも少ないと見られる。そのため、本研究では処理の向きは時間軸方向のみとしたが、アクセントの終了指令を先に推定し、時間を遡って開始指令の位置を決めるなどの方法も検討する必要がある。また、フレーズ指令を正確に検出できるフィルタの作成が必須である。

その他に、指令推定の立場から韻律構造的に正しい指令の選択の可能性についても検討する必要がある。AbS などの方法により求められた指令系列を用いて、指令間隔や隣どうしの指令の大きさの関係などの統計的な拘束条件をしらべ、指令系列の制限に加える等の方法が考えられる。また、二乗誤差以外の歪みの尺度についても検討すべきである。

謝辞

本研究を始めるにあたり、全般的な御指導と御助言を頂いた木村正行教授に心から感謝致します。

また、本研究を行うに当って必要不可欠である音声認識に関する知識とその傾向について御指導、御示唆を頂けた下平博助教授に深く感謝します。

中井満助手には、研究の進行や問題点に対する御助言、御協力を頂きました。深く感謝致します。

木村・下平研究室の高倉健次氏には、研究への御協力を頂きました。深く感謝致します。

さらに木村・下平研究室の皆様には、日頃から御討論、御協力を頂き、意義深い研究生活を送ることができました。深く感謝致します。

最後に、不詳の息子に全面的な援助協力をしてくれた父母へ深く感謝の意を表しつつ、本論文の結びと致します。

参考文献

[1] W. S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots" JASA,Vol. 74,No. 368,pp. 829-836, Dec. 1979.

[2] E. Geoffrois, "Estimation of Prosodic Events from Japanese F_0 Contours" Technical Report of IEICE, SP93-24, Jun. 1993.

[3] K. Hirose and H. Fujisaki, "Analysis and Synthesis of Voice Fundamental Frequency Contours of Spoken Sentences" ICASSP-82,Vol. 2, pp. 950-953, 1982.

[4] R. W. Hamming, "Numerical Method for Scientists and Engineers" Dover Pubns ,2nd ed. ,pp. 349, Apr. 1987.

[5] W. Hess, "Pitch Detection of Speech Signals" Springer-Verlag, 1983.

[6] L. R. Rabiner,M. J. Cheng,A. E. Rosenberg and C. A. McGonnegal, "A Comparative Performance Study of Several Pitch Detection Algorithm" IEEE Trans. Acoust. Speech,Signal Processing,Vol. ASSP-24,5,pp. 399-418, 1976.

[7] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals" Prentice-Hall, 1978.

[8] 小松昭男, 大平栄二, 市川熹, "韻律情報を利用した構文推定及びワードスポットによる会話音声理解方式" 電子情報通信学会論文誌 D,Vol. J71-D,No. 7,pp. 1218-1228,Jul, 1988.

[9] 斉藤, 田中, "音声情報処理の基礎" オーム社, 1981.

[10] 嵯俄山, 古井, "ラグ窓を用いたピッチの抽出の一方法" 信学総全大, 1235, Mar. 1978.

[11] 杉藤, 東川, 板倉, 高橋, “ささやき声におけるアクセントの知覚的、音響的、生理的特徴” 信学技報,SP91-1, 1991.

[12] 寺西秀治, “音声の生成モデルを用いた韻律情報推定” 修士論文,JAIST, May. 1996.

[13] 中井満, “韻律構造を利用した連続音声認識に関する研究” PhD thesis, 東北大学, Mar. 1996.

[14] 萩原昭夫, 米田正次郎, “時間的な連続性を考慮したピッチ候補の選択法” 信学論,Vol. J74-A,No. 7,pp. 948-956, 1991.

[15] 藤崎他, “日本語単語アクセントの基本周波数パターンとその生成機構のモデル” 音響学会誌, Vol. 27, pp. 445–453, 1971.

[16] 藤崎博也, 広瀬啓吉, 高橋登, 杉藤美代子, “共通言語のイントネーションの音響音声学的特徴と方言の影響” 音声研資,S83-36, pp. 277-284, 1983.

[17] 藤崎博也, 大野澄雄, 和田豊, “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推定の一方法” 日本音響学会講演論文集, 2-4-6, 1995.

[18] 三浦種敏, “聴覚と音声” 電子情報通信学会編, 1980.

付録

付録として、ATR 連続音声資料データベース（503 文）のうち、男性話者 MHT の 25 文章について本研究の 3 つの手法により再構成を行った結果を示す。すべて、横軸は時間、縦軸はピッチ周波数に自然対数 ln をかけた値。各点が入力ピッチ、実線が再構成後のピッチパターンである。ピッチ信頼度の閾値を逆フィルタによる手法は 0.20、基本指令成分フィルタは 0.27 に設定した。

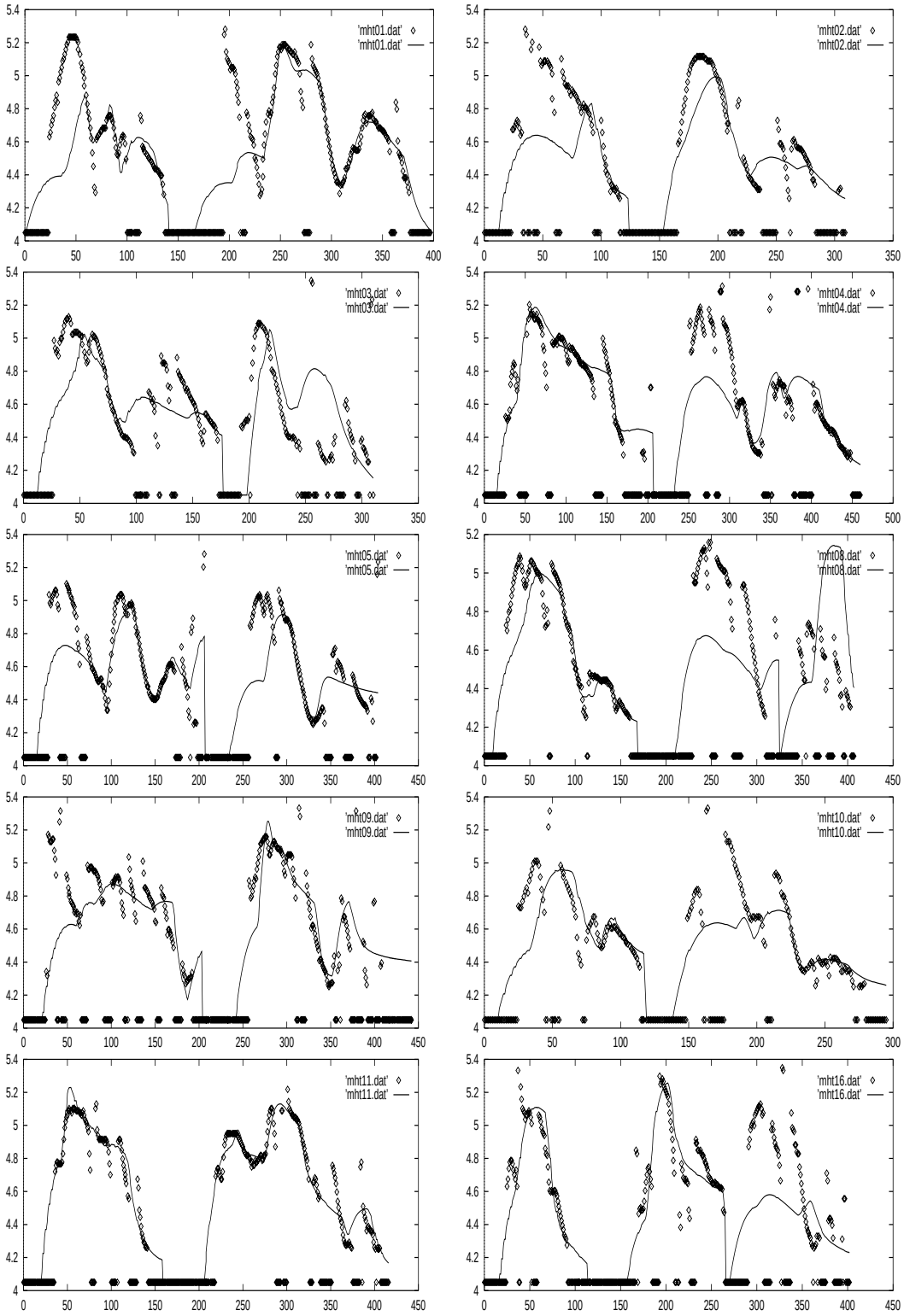

図 4.1: 逆フィルタの方法 1 による再構成結果 (1)

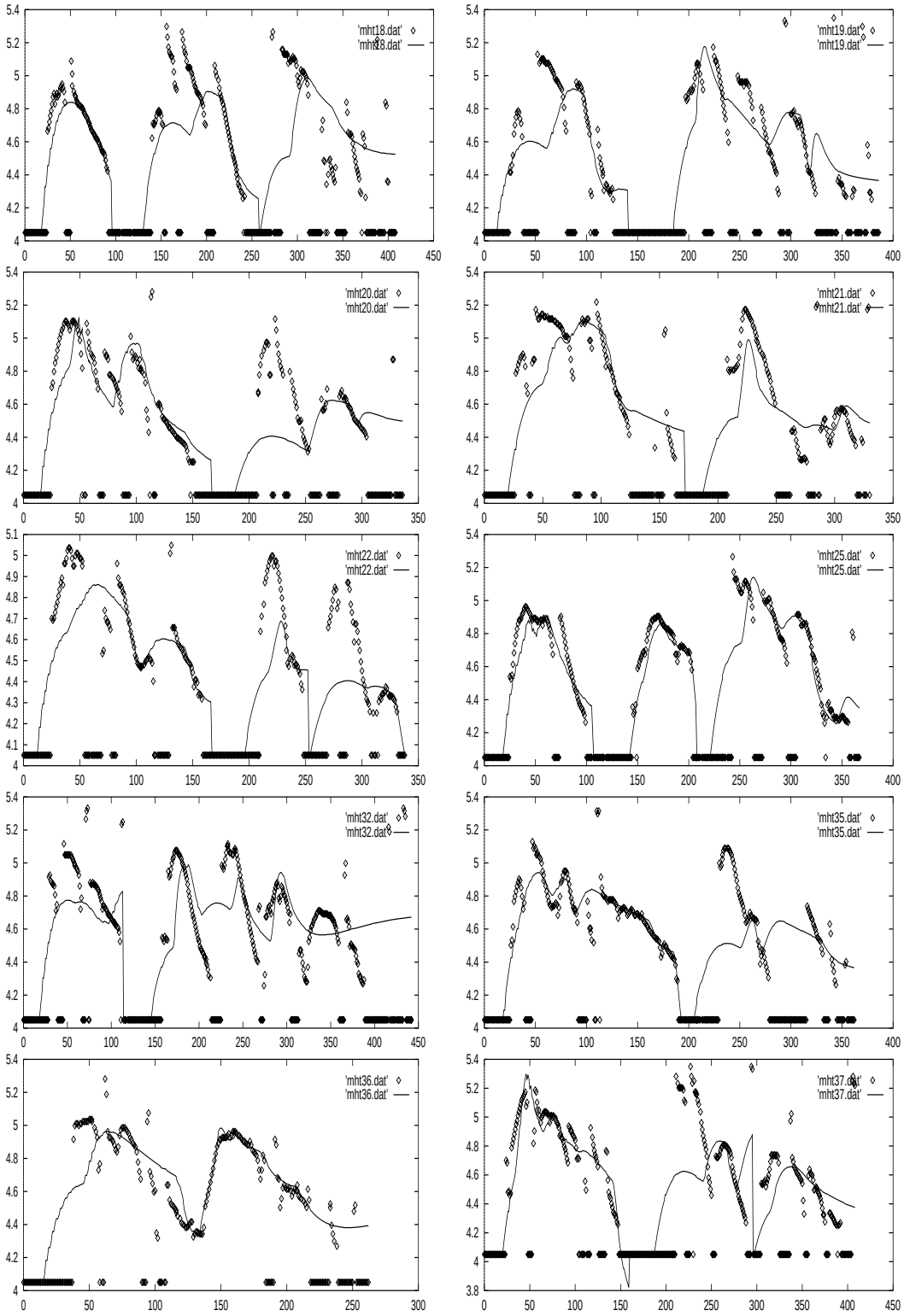

図 4.2: 逆フィルタの方法 1 による再構成結果 (2)

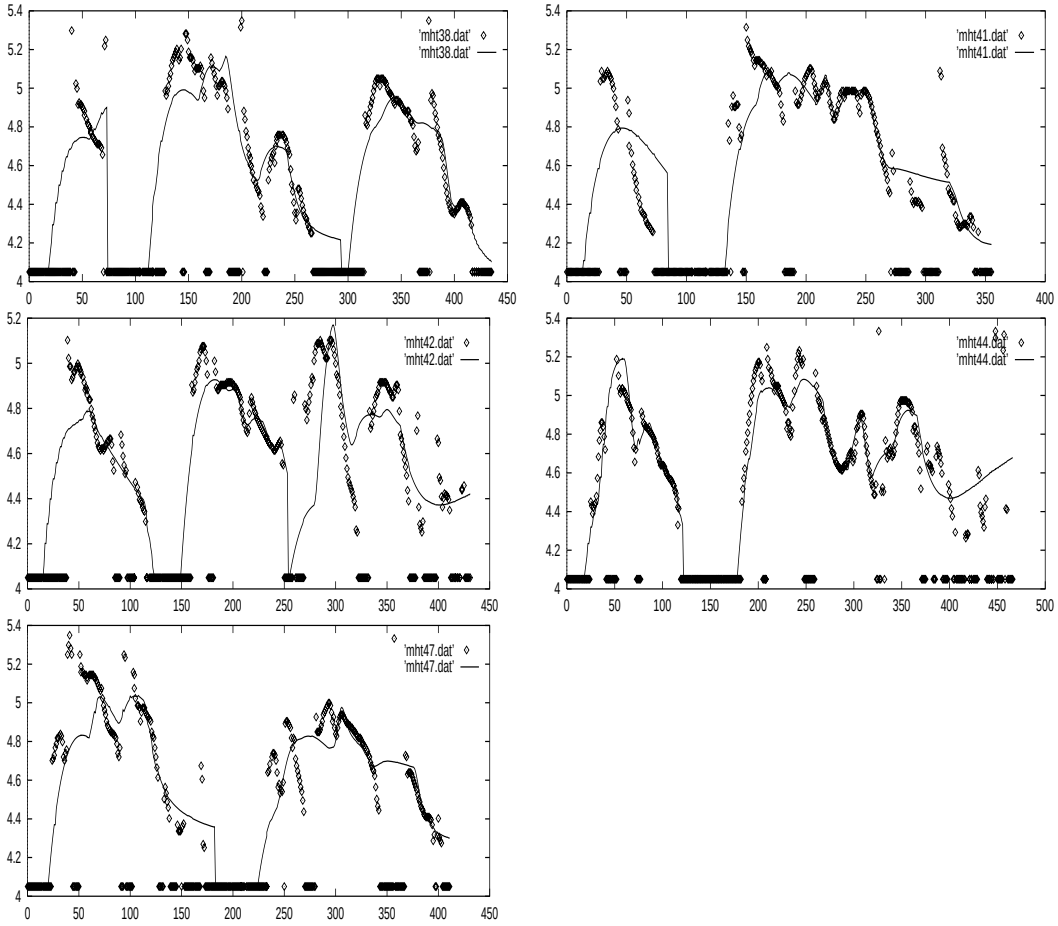

図 4.3: 逆フィルタの方法 1 による再構成結果 (3)

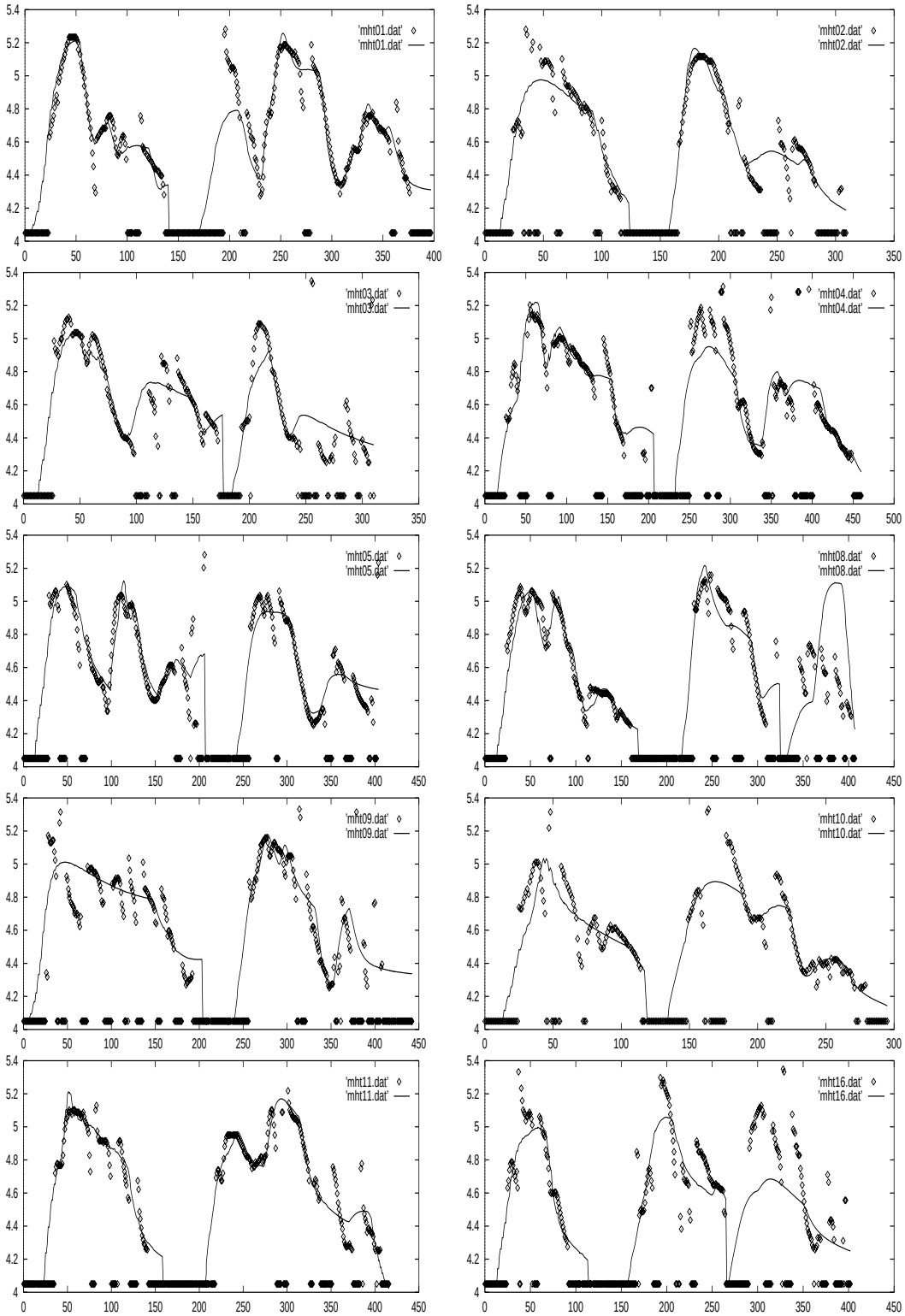

図 4.4: 逆フィルタの方法 2 による再構成結果 (1)

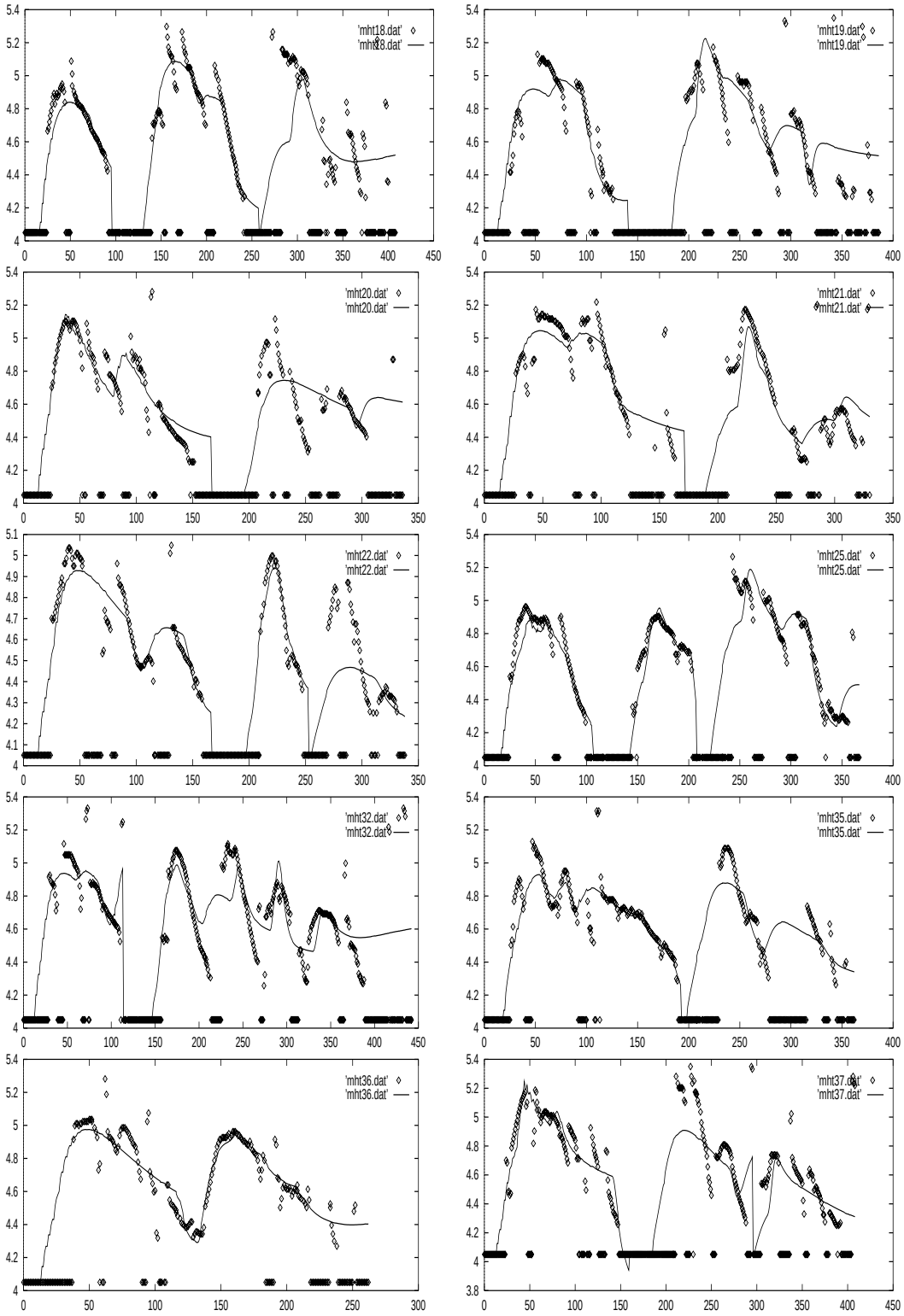

図 4.5: 逆フィルタの方法 2 による再構成結果 (2)

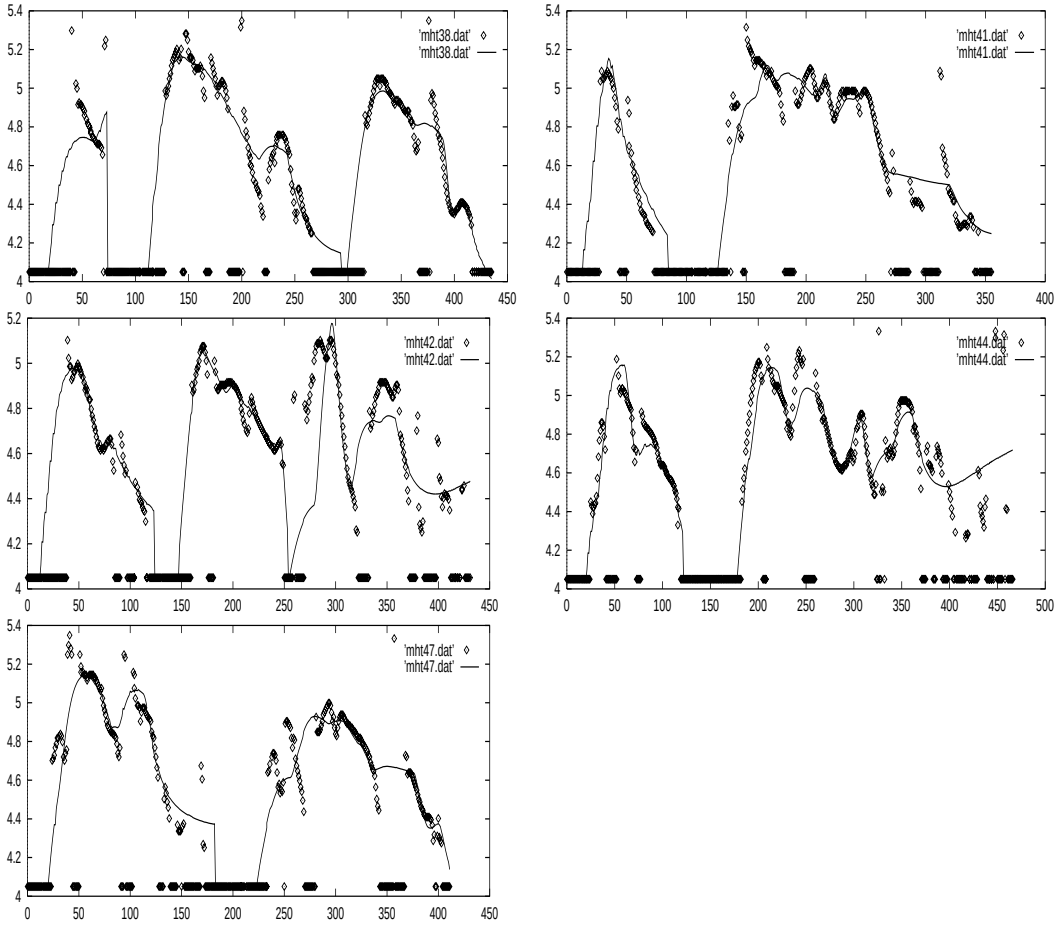

図 4.6: 逆フィルタの方法 2 による再構成結果 (3)

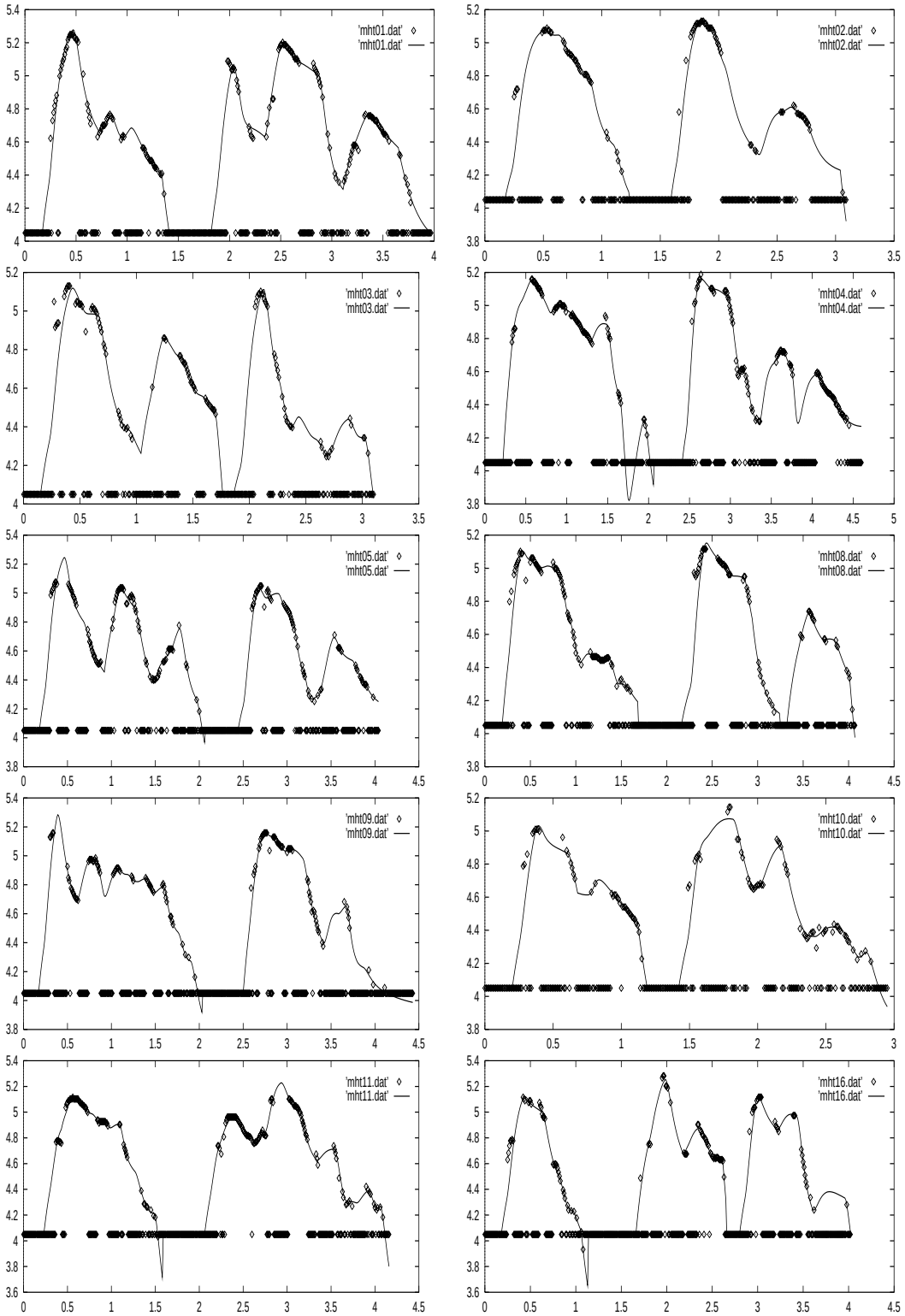

図 4.7: 基本指令成分フィルタによる再構成結果 (1)

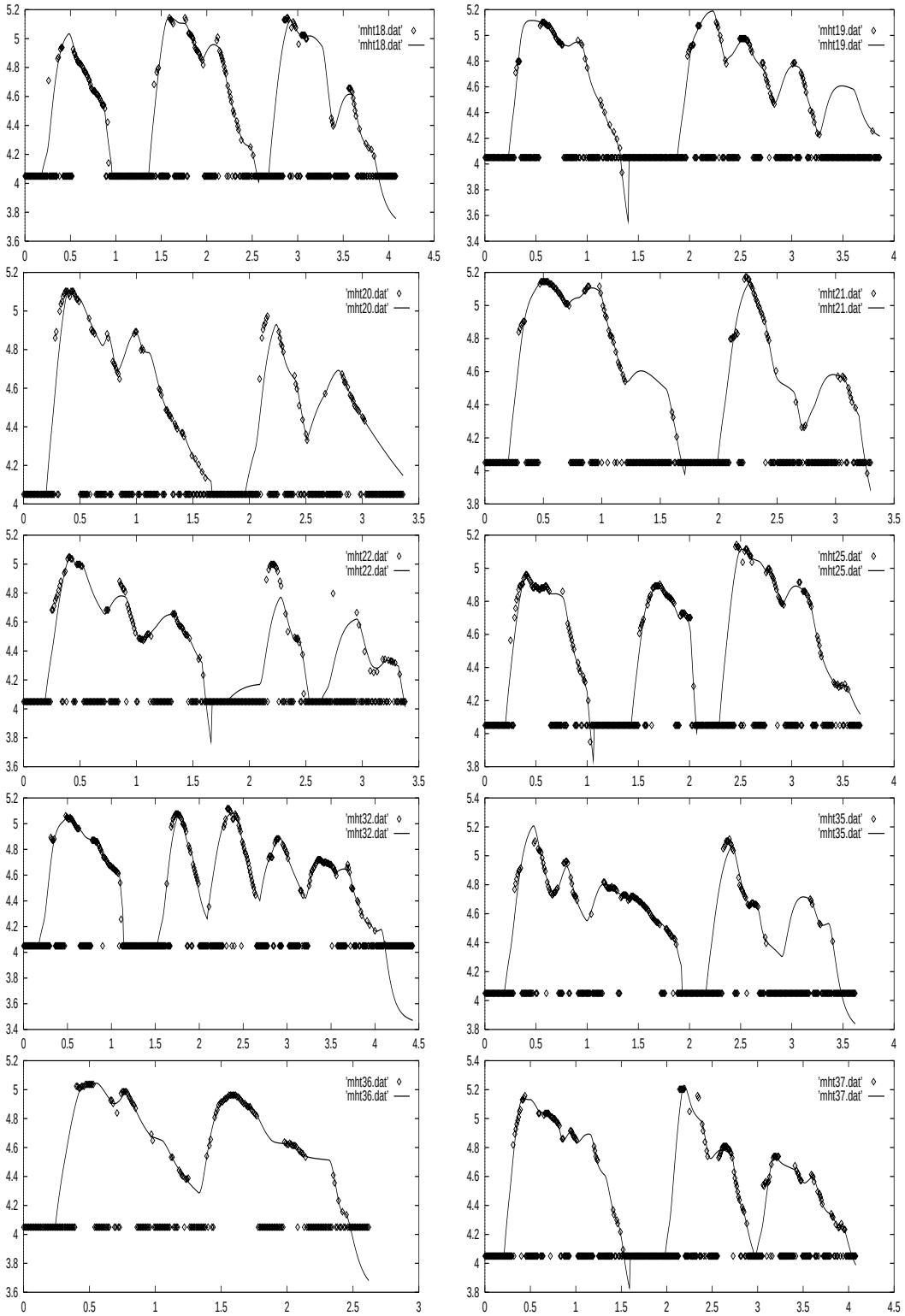

図 4.8: 基本指令成分フィルタによる再構成結果 (2)

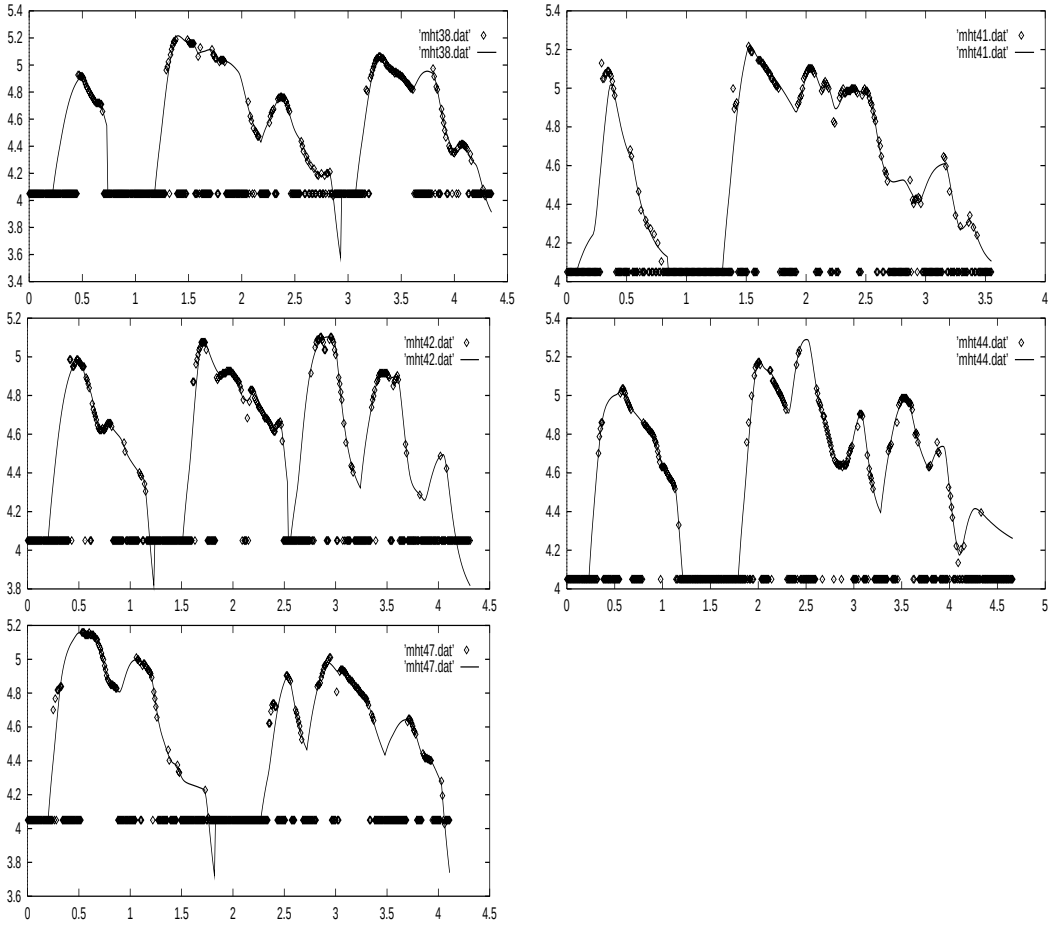

図 4.9: 基本指令成分フィルタによる再構成結果 (3)