ATR Technical Report(非公開)
002

TR-IT-0276

聴取実験を考慮した部分信号処理による CHATR の韻律改善の検討
Improving prosody of CHATR output speech based on partial PSOLA and a MOS decision tree

丸本 徹　　丁 文
Toru Marumoto　　Wen Ding
ニック・キャンベル
Nick Campbell

1998.9.11

本研究では第2研究室の多言語音声合成システム CHATR の韻律向上を目指し、部分信号処理を用いて韻律の改善を行う。まず、CHATR 合成音声の韻律を評価するために、聴取実験を行い、MOS 値を予測できる決定木を構成した。それに部分 PSOLA 法を用い、フレーズ単位に最適な韻律を持つ音素を探索し、信号処理を行う。実際に CHATR システムに組み込み、検討を行った。

目次

第 1 章

序章

1　現在の CHATR と問題点

これまでの音声合成システムでは、音声は比較的簡単な物理的なモデル化が可能であり、韻律とは切り離して考え、音素の系列のみから音声波形の合成が可能であるという暗黙の了解に基づいていた。これに対し我々は、音声波形は音響的及び韻律的な環境によって一意に定まるものであるとする立場から、音声合成システム自身が音声波形を生成することをせず、音響的及び韻律的な環境が最も適する音素単位の音声波形を何らの信号処理も行わずに接続し合成音を得るという手法を取っている。このため通常の音声合成システムより多くの音声データを必要とするが、極めて自然性の高い音声の合成が可能であることが確認されている [1]。

そもそもこの自然音声波形接続型任意音声合成システム CHATR の特長は高い自然性のまま、何らの信号処理も行わずに接続し、連続音声として出力することであった。話者性や発話様式等の特徴を保存したまま任意の音声を合成することができ、かつ選択された波形自体には劣化が全く無い。しかし、この方法では有限なるデータベースから、合成に必要なすべての特徴 (連続性など) を予測して選択しているので、合成音の韻律は目標とする韻律から大きく異なることもあり得る。イントネーションの自然性にはまだまだ改善の余地がある。

そこで本研究では音声単位接続後の韻律補正を行うために、選択後の波形に最小限の信号処理部を加えることを目的とした。信号処理は処理後も高い品質が保てる TD-PSOLA(Time Domain Pitch Synchronous Overlap and Add) 法を用いた。ただ、不必要に信号処理することで本来の自然性を損ねるのを防ぐために、ある基準によるコストを設け、それが最小になる位置での処理・合成をした。基準とは信号処理をすることによって劣化するコストと、モーラの母音部分毎に対する基本周波数 (以下 F0) との傾き差の平均距離や最大値、ピッチの標準化

得点 (以下 F0z-score) などを要素とする決定木を用いた予測評価値との組合わせである。

2 本論文の構成

本論文は 5 章より構成されており、各章の概要は以下の様になる。

まず、2 章で聴取実験の実験方法及び結果を示し、それから各パラメータを求める。

次に 3 章では、各パラメータを要素とした決定木による評価値の予測と、対象となるフレーズの位置、要因の種類による実データとの相関を示し、今後用いる決定木を決定する。

そして 4 章では PSOLA 法の概要と、聴取実験のデータをもとに予測される評価値、信号処理の音質劣化を考慮した韻律補正のアルゴリズムを提案する。そして、実際 CHATR に実装し、処理前後の結果を示す。

最後に 5 章において以上で得られた結果を基に本論文の結論、また今後の課題を述べる。

なお付録に、全データとサンプル文の紹介、本研究で用いた CHATR の所在を示している。

第 2 章

MOS 評価実験からパラメータの抽出

1　実験方法

まず、サンプル文 (.wav ファイル) から図 2.1にあたるフレーズを数サンプル区切りだす。各文節に対して CHATR による合成を施し、その部分に対する韻律 (prosody) と不連続性 (discontinuity) を MOS 評価してもらう。合成に関して従来 CHATR が音素選択の候補のうち採用コストの高いもの 10 個から組み合わせたものを使う。

テスト文は付録に示す 20 文から、56 部分を選択し、被験者 11 人である。

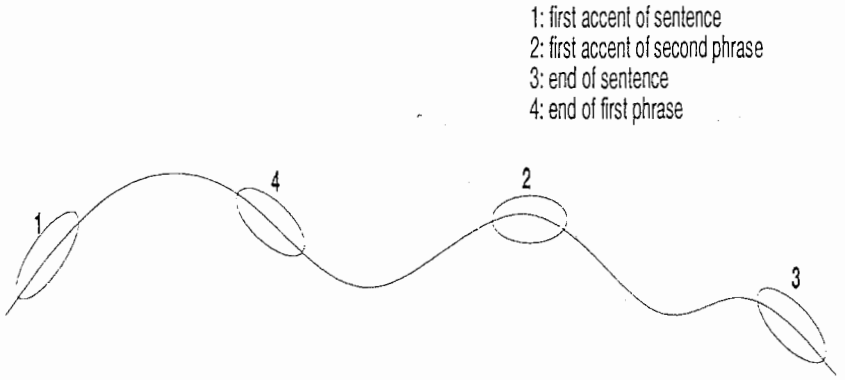


図 2.1: 文中におけるサンプルの位置の分類

2　データ処理、パラメータ抽出

集計したデータと、CHATR の合成時に使われるデータを用いて以下のパラメータを算出する。詳細は付録に示すプログラムを参照されたい。

mean of MOS (以下 Score) 集計したデータから各サンプルにおける prosody と discontinuity の有効回答者による平均を求める。以後、prosody における Score を使用する。なお、回答は

2: 非常に良い
1: 良い
0: 普通
-1: 悪い
-2: 非常に悪い
5: 未回答

である。

distance of slope (以下 Slope) あるモーラの基本周波数 (以下 f0) を次のモーラの f0 から引いた傾きをだし、目標とする (以下 target) 傾きから CHATR によって生成された (以下 unit) 傾きを引いた 2 乗和平均を (モーラ数 −1) で割る。

モーラ数を N、i 番目の f0 を $f0^i$ とすると、式 (2.3) のように表せる。

$$slope^i_{target} = f0^{i+1}_{target} - f0^i_{target} \tag{2.1}$$

$$slope^i_{unit} = f0^{i+1}_{unit} - f0^i_{unit} \tag{2.2}$$

$$Slope_{distance} = \frac{\sqrt{\sum_{i=1}^{N-1}(slope^i_{target} - slope^i_{unit})^2}}{N-1} \tag{2.3}$$

distance of logslope (以下 Logslope) Slope とほぼ同様だが、傾きを求める際、聴覚的影響を考慮して対数を取り入れた。

式 (??) に示す。

$$logslope^i_{target} = \log f0^{i+1}_{target} - \log f0^i_{target} \tag{2.4}$$

$$logslope^i_{unit} = \log f0^{i+1}_{unit} - \log f0^i_{unit} \tag{2.5}$$

$$Logslope_{distance} = \frac{\sqrt{\sum_{i=1}^{N-1}(logslope^i_{target} - logslope^i_{unit})^2}}{N-1} ?? \tag{2.6}$$

maximum of diffrence between target_f0 and unit_f0 (以下 delta)　モーラ毎における target と unit の傾きの最大差。

$$Delta = \max_{i=1,M} |logslope^i_{target} - logslope^i_{unit}| \tag{2.7}$$

この値が大きいとそのサンプル中に傾き差が最大、つまりイントネーションに大きなズレが生じていることがわかる。

f0z_score (以下 f0zs)　各モーラにおける unit_f0 の標準化得点を CHATR から抽出した。各母音部分で母音毎の基本周波数の平均を $\bar{f}_0$、標準偏差を σ_{f_0} としたとき、$f_{0_{zs}} = \frac{f_0 - \bar{f}_0}{\sigma_{f_0}}$ であらわされる。

mean of target_cost (以下 meancost)　各モーラ毎の target_cost の平均値。

target_cost$C^t(t_i, u_i)$ とは音声単位の特徴ベクトルと音声データベース中 (unit) から選ばれた音声単位 u_i の特徴ベクトルとの各要素における差分の重み付き合計である。合成音声として実現したい (target) の音声単位を t_i、特徴ベクトルの次元数を p、各 targetsub_cost$C^t_j(t_i, u_i)$ の重み w^t_j が与えられたとき、

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w^t_j C^t_j(t_i, u_i) \tag{2.8}$$

で表される。[?]

大きく異なると韻律が目標にしている合成音全体的にと似つかわないことになる。

maxmum of target_cost (以下 maxcost)　各モーラ毎の target_cost の最大値。大きく異なる部分が目標合成音と大きくずれているところである。

target_cost の値は CHATR のコマンドから抽出している。

3　結果及び分析

結果の一部を以下に示す。左からサンプル名、アクセントの種類 (図 2.1)、target_f0 の最大値、unit_f0 の最大値、f0zs、Slope、Logslope、(下段へ)delta、maxcost、meancost、prosody の MOS 値 (11 人分)、有効回答の平均 (Score) である。

```
FTK_SD_A01_1_0.list 1 348 334 1.550 13.715360  0.043447
```

```
0.109828  1.235305  1.157844   2  1  2  2  2  1  2  2  5  2  2   1.800
 FTK_SD_A01_1_1.list 1 348 340 1.620 13.482499  0.046544
0.090031  1.281861  1.249067   0  0  1  1  0  0  0  0  1  5  5   0.333
 FTK_SD_A01_1_2.list 1 348 310 1.130 20.278615  0.061393
0.154373  1.339392  1.308161   0 -1  1  1 -1 -1  0  0  5  5  5  -0.125
 FTK_SD_A01_1_3.list 1 348 342 1.650 29.845901  0.097002
0.221840  1.378598  1.350712  -1 -2  1 -1 -1 -1 -2  0 -1  5  5  -0.889
 FTK_SD_A01_1_4.list 1 348 333 1.510  4.484541  0.016898
0.034670  1.403364  1.364434  -1  0  0  2 -1  1 -1  1  2  5  5   0.333
 FTK_SD_A01_1_5.list 1 348 291 1.050 17.688666  0.055333
0.147896  1.487097  1.404785  -1 -1  1  0  0 -1 -2  1  5  1  2   0.000
 FTK_SD_A01_1_6.list 1 348 305 1.050  7.241854  0.017196
0.050717  1.490171  1.432588   0  0  1  2  0  0 -1  0  0  5  5   0.222
 FTK_SD_A01_1_7.list 1 348 310 1.130 10.472185  0.038361
0.102851  1.490567  1.447366  -1 -1  1  1  1  0  1 -1  5  5  5   0.125
 FTK_SD_A01_1_8.list 1 348 274 0.550 22.373099  0.080863
0.231489  1.547309  1.480424  -1 -1  1 -1  0  0 -1  0  0  5  5  -0.333
 FTK_SD_A01_1_9.list 1 348 274 0.550 24.884176  0.081988
0.166134  1.555603  1.495014  -1 -1  0  0  0  0 -2 -1  1  5  5  -0.444
 FTK_SD_A01_2_0.list 2 325 331 1.770 26.580068  0.108103
0.210933  1.102074  1.015048  -2  1  2  2  2  2  5  5  5  2  2   1.375
 FTK_SD_A01_2_1.list 2 325 287 0.960 20.359273  0.120593
0.235068  1.148132  1.062725   0  2  2  2  2  2 -1  5  0  5  5   1.125
 FTK_SD_A01_2_2.list 2 325 328 1.490  2.500000  0.013391
0.024829  1.150146  1.093600   0 -1  1  2 -1  0  0  0  5  5  5   0.125
 FTK_SD_A01_2_3.list 2 325 322 1.620 23.194827  0.074315
0.148282  1.199218  1.118579  -1 -1  1 -1  0  0  0  0 -1  5  5  -0.333
 FTK_SD_A01_2_4.list 2 325 272 0.680 13.200379  0.037281
0.069037  1.238635  1.142421   0  0  1  5  1  1 -1  0  2  5  5   0.500
 FTK_SD_A01_2_5.list 2 325 323 1.340 20.651876  0.074293
0.143672  1.293036  1.174249   0  0  1  0  1 -1  5  5  5 -1  2   0.250
 FTK_SD_A01_2_6.list 2 325 273 0.530 35.531676  0.133578
0.267150  1.304486  1.202406  -1 -1  0  0  0 -1 -2  5 -1  5  5  -0.750
 FTK_SD_A01_2_7.list 2 325 315 1.210  6.264982  0.018734
0.030389  1.336497  1.218839  -1 -1  0  1 -1  0 -2  0  5  5  5  -0.500
 FTK_SD_A01_2_8.list 2 325 327 1.410 39.246019  0.136030
0.234145  1.373068  1.239000   0 -1  0 -1  0  1 -1  0 -1  5  5  -0.333
 FTK_SD_A01_2_9.list 2 325 292 1.050 25.544080  0.091494
0.132526  1.428171  1.263113   0  0  1  1 -1  1  0  0  1  5  5   0.333
 FTK_SD_A01_3_0.list 3 257 265 0.540 25.733679  0.778980
4.646290  3.084955  1.596045  -2 -1  0  0  1 -2  5  5  5 -1 -2  -0.875
```

これらから Slope(Logslope) と Score の相関を求めてみる。図 2.2、図 2.3に示す。図中の数字は負の相関係数である。

target と unit との傾きの差が小さいほど MOS 値は良くなると予想されたが、相関度にはさほど表れていない。これは聴取被験者の個人差と少ない有効回答数からの平均 MOS 値より、多少、値が揺らいだものと思われる。

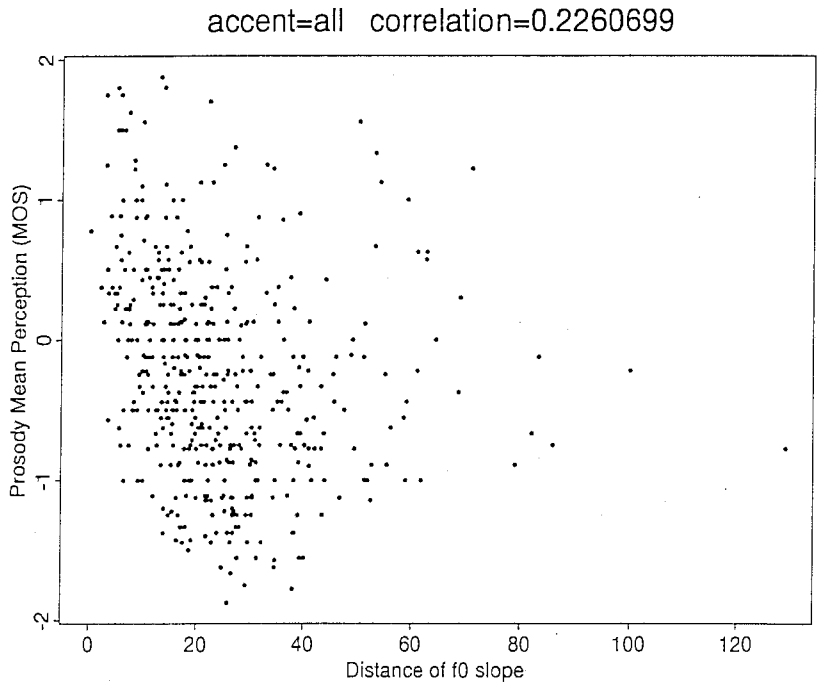


図 2.2: 傾きの差における MOS 値への影響 (Slope)

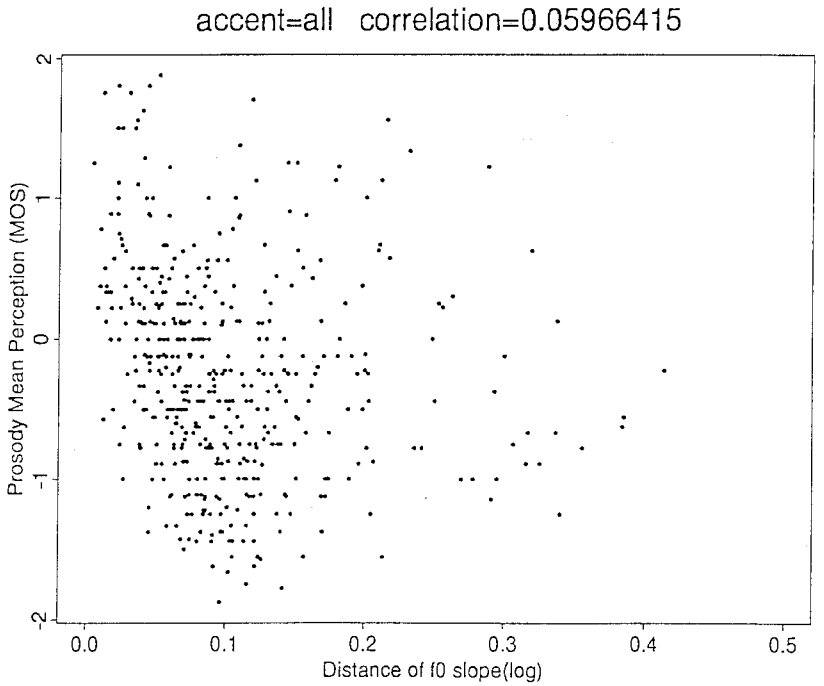


図 2.3: 傾きの差における MOS 値への影響 (Logslope)

第 3 章

決定木による予測

1　決定木の構築

2章で求めた各パラメータを要素とする決定木を作成する。各要素の組合わせを以下のようにする。

1. Logslope,f0zs
2. Logslope,f0zs,delta
3. maxcost,meancost
4. Logslope,f0zs,delta,maxcost,meancost
5. Slope,f0zs

各要素ごとかつ、各アクセントの種類 (all,1,2,3,4) 別に、決定木を構成した。一例を示す (図 3.1～図 3.5)。全データは付録参照。

残差平均 (図中 Residual) が小さいほど信頼が高い。

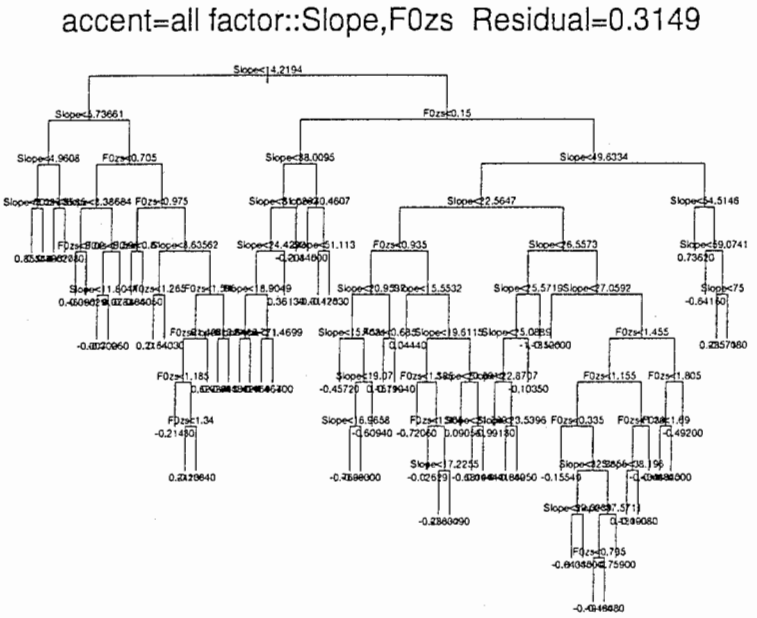


図 3.1: 聴取実験による決定木

2 決定木による予測

2.1 予測値と MOS 値の相関性

前節で作成した決定木を用いて、逆に各要素の値が入力されたときの MOS 値を予測する。聴取データをそのまま用いているので閉区間での予測となる。

その予測値と実際のデータの値との相関を示す。図 3.6、図 3.7はその一例である。図中 cor は相関係数を示す。全データは付録参照。

2.2 結果

以上のデータの数値のみを参照するため、全データに対する決定木の残差平均、それから予測した値と実データとの相関計数をまとめ、アクセント、要素ごとに分類した表を図 3.8～図 3.12に示す。

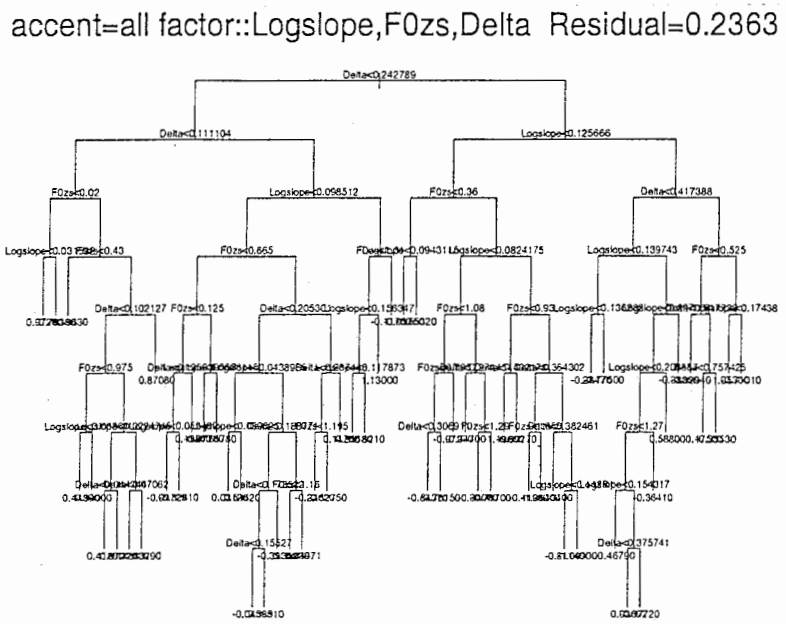


図 3.2: 聴取実験による決定木

相関計数についてまとめたものを点グラフで表示する。(図 3.13、図 3.14) 横軸は相関係数、縦軸は要素とサンプル位置をそれぞれ比較項目別に並び替えたものである。

3 考察

まず、聴取実験の各パラメータから決定木を構成したが、残差平均の値が (図 3.8) 最小の要素 (Logslope,f0zs,delta,maxcost,meancost)、すなわち図 3.3がもっとも決定木としての信頼性が高いとわかる。これは、各要素が全て有効で MOS 値との依存性が強いということになる。

この決定木から予測された値は実際の MOS 値との相関ももっとも大きい。ただ、この木を利用するには target_cost の値が不可欠になるが、これはデータベース中に存在する値なので、以後使用する PSOLA 組み込みのプログラムに用いることは難しい。

したがって、次に信頼性の高い要素 (Logslope,f0zs,delta) の決定木 (図 3.2) を当初、使用していたが、 f0zs の算出も困難であることが発覚したので (p26.2参照) 、要素 (Logslope,delta) を特別に構成し、以後使用することにした。

なお、聴取データには当然個人差、回答数の少なさのため、ばらついた値を取ることが多い。そういう値の処理、決定木の正規化は行ってないのでどうしても理論に矛盾した結果が起

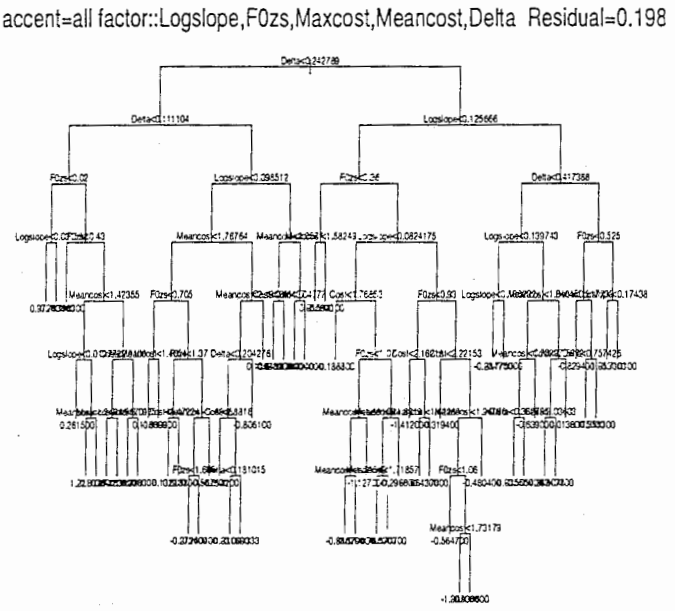


図 3.3: 聴取実験による決定木

きることが多々ある。

今後使用する決定木も要素数2つより、あまり細分化せず、少々信頼性は落ちても行き過ぎた予測をしないようにpruning(葉を刈り取る、すなわち複雑な形にしない)を施した。処理した決定木は図3.15、相関性は図3.16に示す。terminal nodes数(葉の数)は9個とアクセントが全部の場合にしては、簡単な構造になっている。

相関係数を比較すると(図3.13, 図3.14)、やはりtarget_costがより正確な決定木の構成に貢献していることがわかる。また、目標音声との傾きを要素に選ぶ場合にも、聴覚的影響を考慮して対数をとりこんだほうが、信頼度の高い結果がでている。

acc_type=1、すなわち文頭フレーズだけに着目すると、target_costの影響最もが大きく出ている。つまり、文の発生時の韻律の良し悪しは特に反映されやすく、目標に近い必要性があると考えられる。

逆に韻律が低い文末(acc_type=3)では、韻律から離れていることが評価につながることは少なく、Logslope,f0zs,delta特にdelta(targetとの傾きの最大差)に注意しなければならない。つまり、イントネーション違うようにならないことが重要である。

第1、2アクセント付近(acc_type=4,2)では、総合的に相関係数が高いので、target_cost、

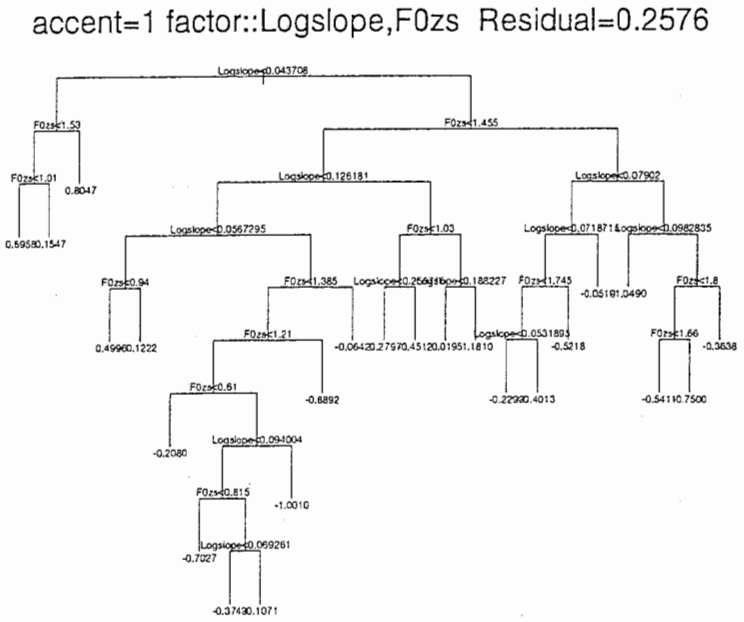


図 3.4: 聴取実験による決定木

イントネーション共に高い精度が要求される。(例: 図 3.7)

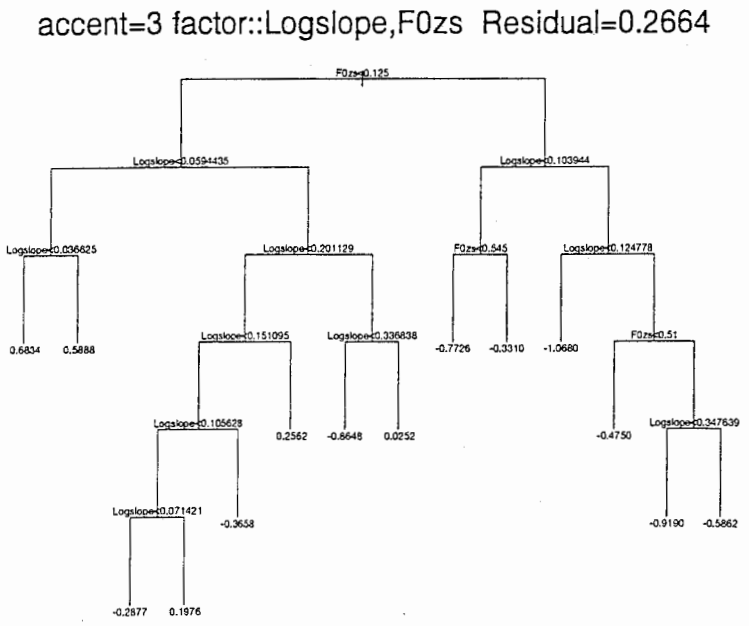


図 3.5: 聴取実験による決定木

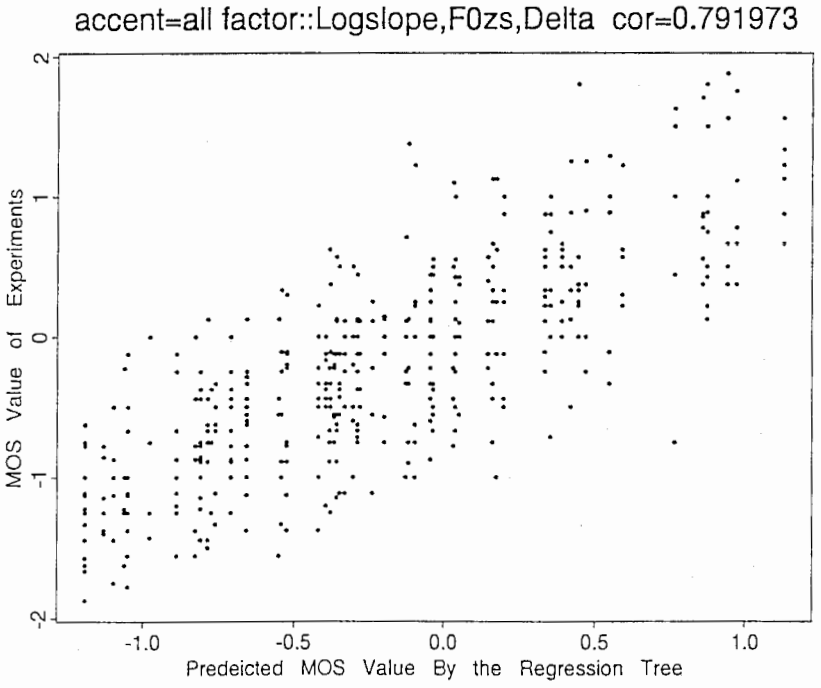


図 3.6: 決定木による予測値の信頼性

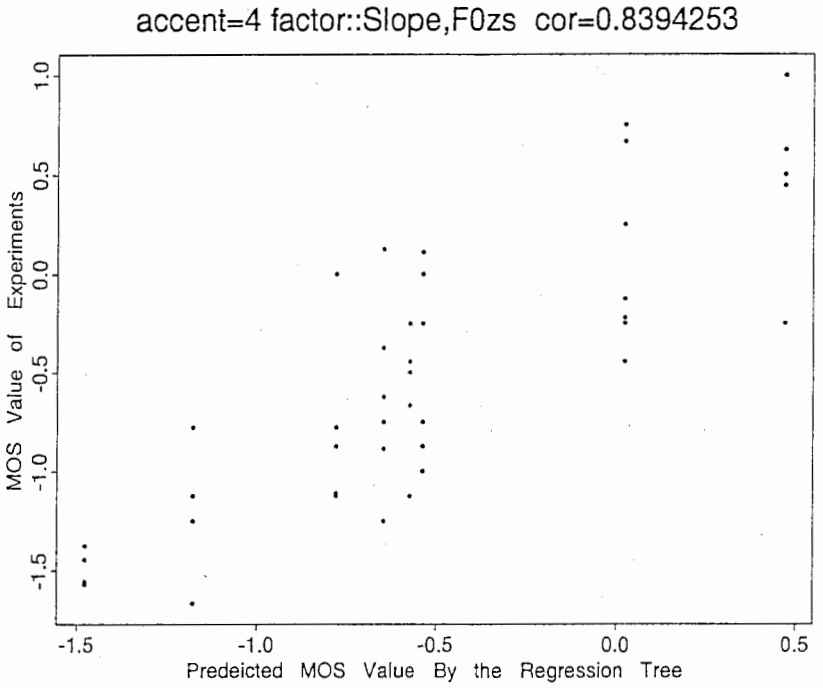


図 3.7: 決定木による予測値の信頼性

Prediction by Regression Tree *accent−type: all*

Factor	Residual	Correlation
logslope,f0zs	0.3029	0.7299141
logslope,f0zs,delta	0.2363	0.791973
maxcost,meancost	0.2524	0.769653
logslope,f0zs,delta, maxcost,meancost	0.198	0.8287501
slope,f0zs	0.3149	0.715474

図 3.8: 各要素における信頼性

Prediction by Regression Tree *accent-type: 1*

Factor	Residual	Correlation
logslope,f0zs	0.2576	0.7618595
logslope,f0zs,delta	0.2196	0.8013718
maxcost,meancost	0.14	0.8786042
logslope,f0zs,delta, maxcost,meancost	0.1183	0.8985084
slope,f0zs	0.2824	0.7372469

図 3.9: 各要素における信頼性

Prediction by Regression Tree *accent-type: 2*

Factor	Residual	Correlation
logslope,f0zs	0.3756	0.6956823
logslope,f0zs,delta	0.1975	0.8522734
maxcost,meancost	0.2362	0.8202288
logslope,f0zs,delta, maxcost,meancost	0.1764	0.8692456
slope,f0zs	0.3812	0.6901142

図 3.10: 各要素における信頼性

Prediction by Regression Tree *accent-type: 3*

Factor	Residual	Correlation
logslope,f0zs	0.2664	0.7510536
logslope,f0zs,delta	0.2341	0.7853999
maxcost,meancost	0.3191	0.6962063
logslope,f0zs,delta, maxcost,meancost	0.2099	0.8102496
slope,f0zs	0.2801	0.7358906

図 3.11: 各要素における信頼性

Prediction by Regression Tree *accent-type: 4*

Factor	Residual	Correlation
logslope,f0zs	0.1958	0.7912136
logslope,f0zs,delta	0.1751	0.8158242
maxcost,meancost	0.13	0.8736588
logslope,f0zs,delta, maxcost,meancost	0.1739	0.8171969
slope,f0zs	0.1583	0.8394253

図 3.12: 各要素における信頼性

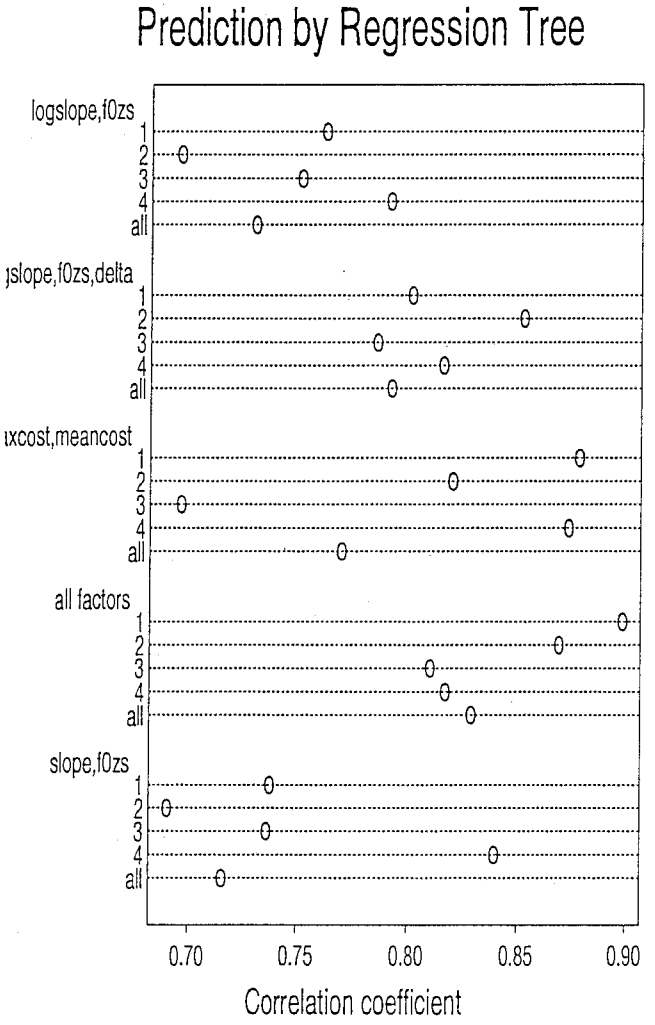


図 3.13: 点グラフによる各アクセントの比較

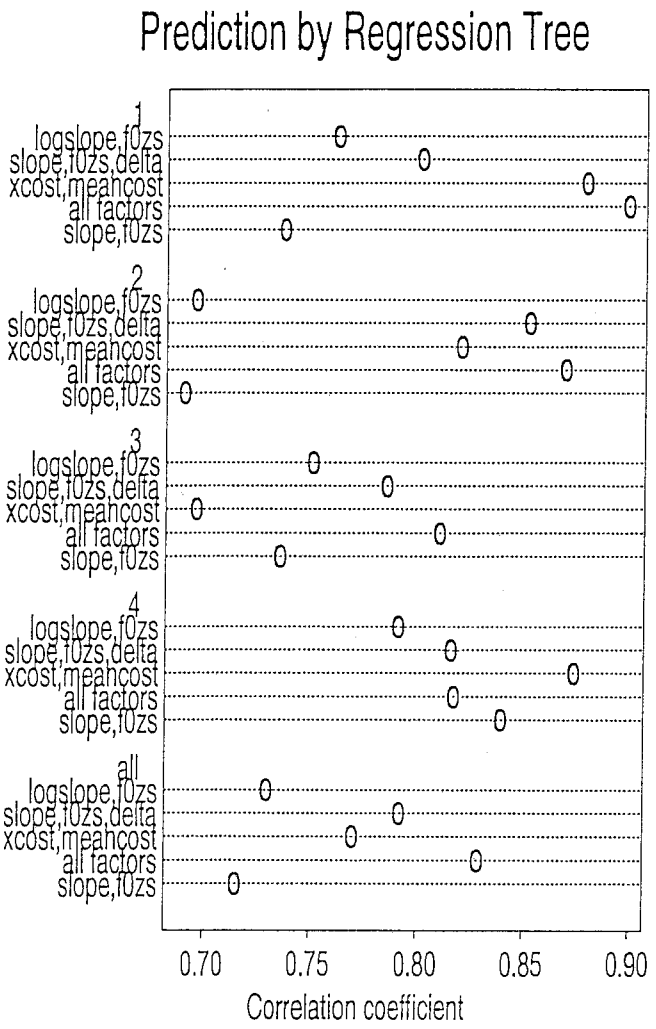


図 3.14: 点グラフによる各要素の比較

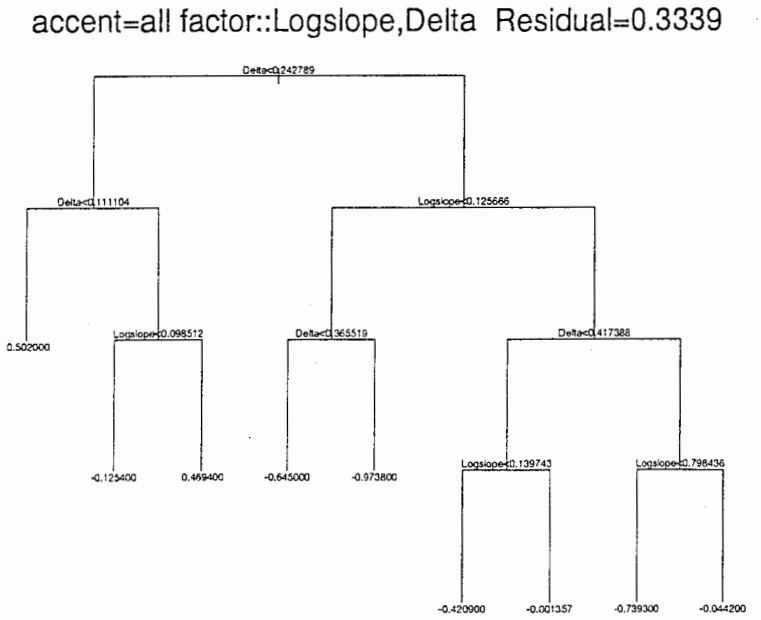


図 3.15: 部分信号処理に用いる決定木

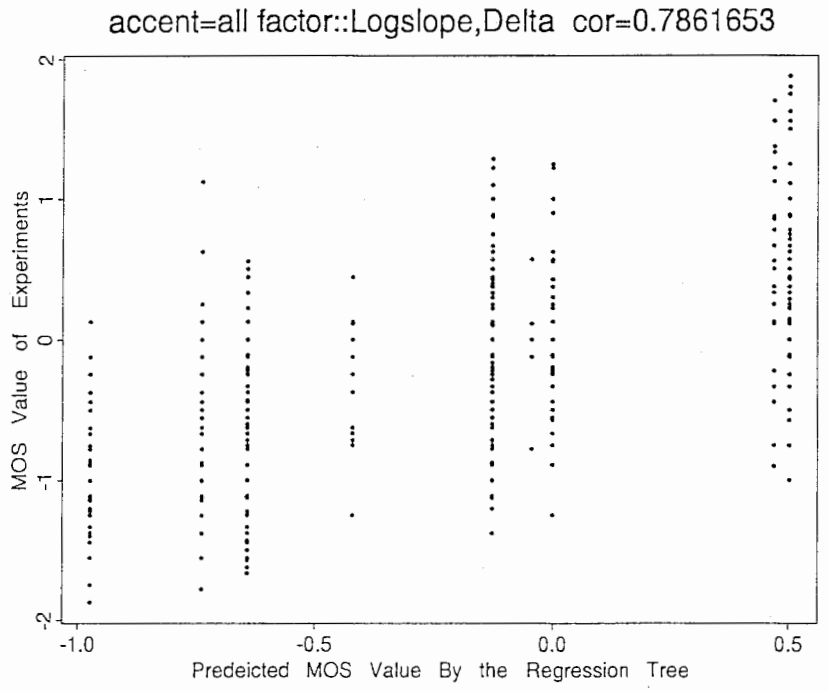


図 3.16: 使用する決定木の信頼性

第 4 章

CHATR への組み込み

聴取実験に基づいたデータから、決定木を構成することができた。次はその決定木を用いて、適所に信号処理を施せるアルゴリズムを考案する。

1 PSOLA(Pitch Synchronous Overlap and Add) 法

韻律改善、すなわちピッチ変更には PSOLA 法を用いている。まず、今回使用する PSOLA 法の概要と特徴を示す。

PSOLA 法とは、音声波形のピッチマーク位置を中心に波形を切り出した“ピッチ素片”を、合成する際の基本周波数に対応するピッチマーク間隔で配置することにより、任意のピッチの音声を得る手法である。この PSOLA 法を用いるためには、予めデータベース中の音声波形のピークの位置などにピッチマーキングをしておく必要がある。

CHATR においてはデータベース作成時にすでにピッチマーキング処理を行っている。[4]

ピッチ素片を切り出す窓関数として基本周期の 2 倍程度の幅をもち、中心から両端に向かってなだらかに減衰する Hanning 窓が用いられる。[2]

切り出したピッチ素片を密に詰めれば基本周波数は高く、粗に間引くと基本周波数は低く変更することができる。

ただし、変更率が高くなると音質の著しい劣化が生じ、文献 [3] によると視察より評価基準が 3.5(”良い”と”普通”の中間) 以上に対応するのは 25% 前後になる。

2 アルゴリズム

概念図を図 4.1に示し、以下に説明する。なお図中では最大ピッチ変更率 x=0.2、モーラ数 M=4、刻み数 N=5 と設定している。

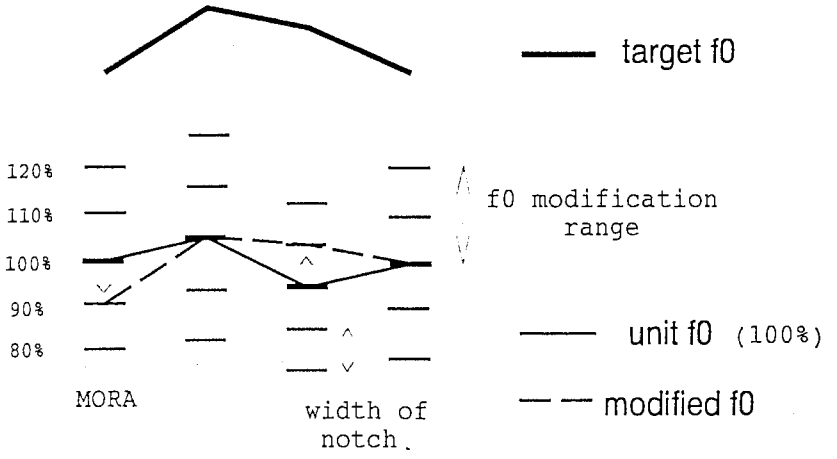


図 4.1: 最適 path 決定アルゴリズム

PSOLA 許容変更値の設定　先にも述べた通り、f0 の最大ピッチ変更率 (f0 modification range) は $x = \pm 0.25$ 程度以内が望ましい。

CHATR で合成されたモーラの母音毎の基本周波数 (unit_f0) に最大変更制限幅 $unit\_f0 \times (1+x)$ を設け、有限数 (N、奇数とする) で区切る。つまり、各モーラに対して N 段階の変更が可能となり、変更幅 (width of notch) は $\frac{2x}{N-1}$ で表される。

決定木による MOS 値の予測　聴取実験より得られたパラメータを要素に決定木を構成するのでそれから予測される値は評価に有効である。

モーラ数を M とすると、N^M 通りの path の組み合わせになるが、各 path ごとに目標とする基本周波数 (target_f0) との傾きの差の 2 乗和平均 (Loglope)、傾きの最大差 (delta) を求め、決定木によって MOS 値を予測する。

音質劣化コストの設定　各 path ごとの PSOLA による音質劣化は文献 [3] の図より、

$$y = ax^2 \tag{4.1}$$

で近似することにする。y_i が音質劣化のコストで、x=x_{max} のとき y=1.0 となるような比例係数 a を初期設定する。

音質劣化のコストの総和をモーラ数 M で割る。

$$psola\_cost = \frac{\sum_{i=1}^{M} y_i}{M} \tag{4.2}$$

当然、unit_f0 は x=0 に相当し、y=psola_cost=0 である。

最適な信号処理位置の決定 以上のように算出した平均音質劣化値 (psola_cost) と予測 MOS 値 (predict_MOS) を式 (4.3) によって各 path ごとのコストを求め、最小になる path を最適 path とする。すなわちどこにどれだけ信号処理を施せば最も劣化を感じにくく韻律補正ができる path を探索する。(式 (4.4))

$$cost = psola\_cost * weight - predict\_MOS \tag{4.3}$$

$$min\_cost = \min_{i=1,N^M} cost \tag{4.4}$$

3 結果考察

今回用いた部分的信号処理による処理前と処理後、目標とする韻律の比較を図 4.2に示す。

weight(4.3式参照) は統計的にみて推測した値であるが、音質劣化も少なく韻律すなわちイントネーションが改善されているのがわかる。(例えば" ワールドシリーズ" の韻律)

weight を変化させていったときの韻律の改善具合いを図 4.3～図 4.4に示す。

weight を 0 であるということは、式 (4.3) より聴取実験のデータから予測される値に依存して最もよくなる path を選んでいることになる。と同時に、f0 変化による音質劣化の影響を無視していくことになる。

図 4.4では *weight*=0.01 と *weight*=0.0001 は同波形のため、このサンプル文においては *weight* が 0.01 以下は 0 に近似したものと同一になる。ただ、図 4.3～図 4.4において、PSOLA ピッチ変更幅はある程度の良い基準以上 (20%) に設定しているので、実際 0.01 という 0 に近似した *weight* でも、信号処理による音質劣化は少ない。

韻律改善で最も望まれるのは、イントネーションが目標とするものと同じになることであるが、psola_cost の影響で音質劣化を重視して無理に韻律を変えないことが多い。よって、20% のピッチ変更率では韻律改善を重視した、*weight*=0.01 を主に利用することにした。

もっと変更率が高い場合には、適切な *weight* を求めることが重要となる。

full search を用いているので、1 フレーズにモーラ数が M、f0 変化刻み数が N あると、最適パスの発見に N^M 通りの探索をしなければならない。1 フレーズごとに計算するので、モーラ数が大きすぎると計算量に膨大な時間がかかり実時間処理できない。

また図 4.2は一般会話文の一部だが、流れがあり区切れが少ないので、処理するのが困難となる。任意にポーズを入れると、target_f0 に影響をもたらし、フレーズ単位の韻律のつなが

りが不自然になりやすい。

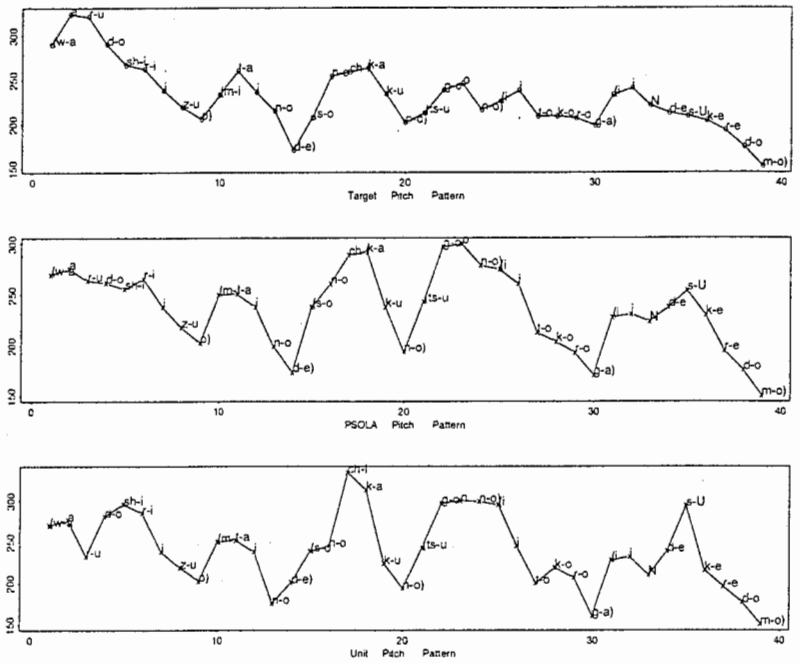


図 4.2: 部分信号処理による韻律の変化

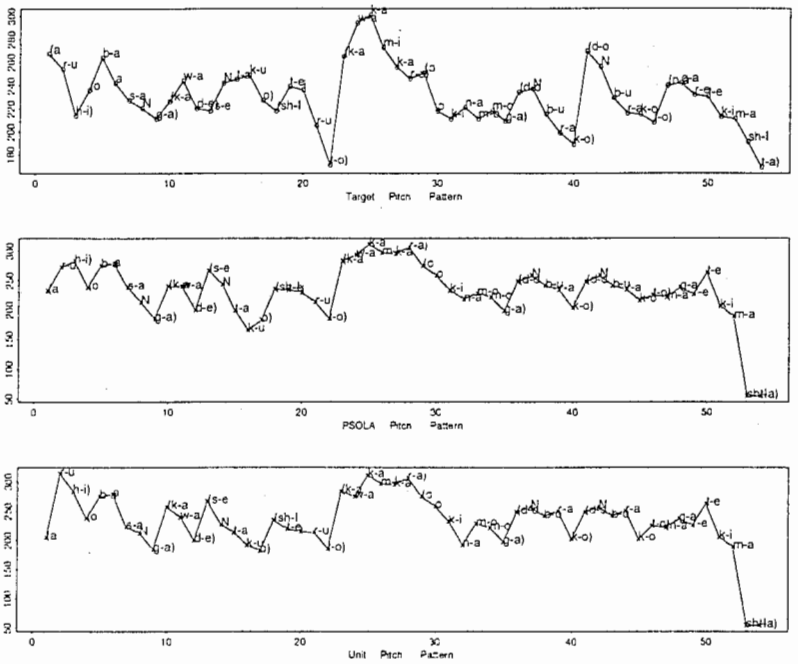


図 4.3: 上から target_f0,$weight$=1 の PSOLA,unit_f0

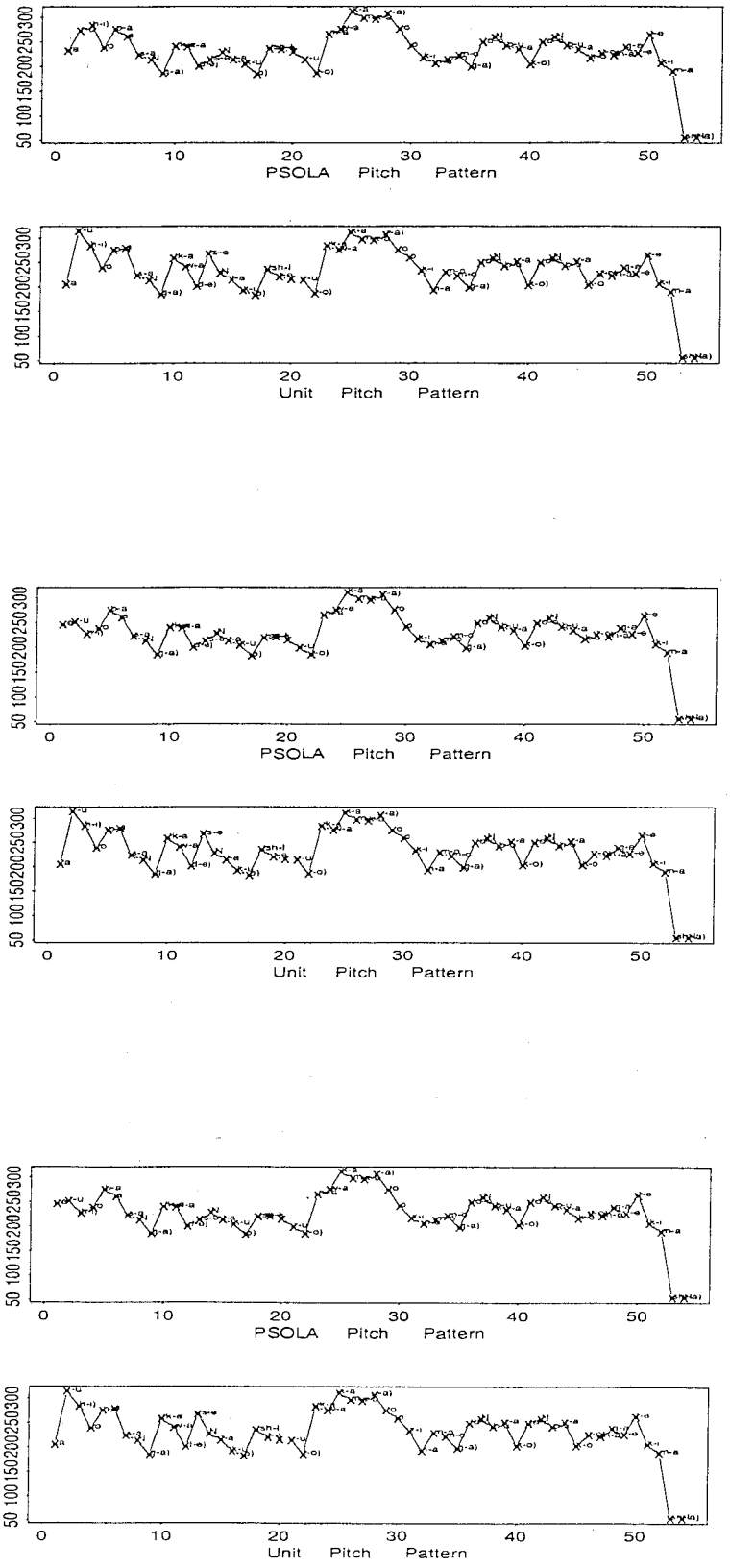


図 4.4: 上から $weight$=0.1,0.01,0.0001 のときの韻律変化

第 5 章

結論

1　まとめ

本論文では自然音声波形接続型音声合成システムにおいて韻律の改善を目指すべく、最適な部分信号処理が施せる方法を提案、検討した。

PSOLA 処理で音質劣化が大きくならない程度の範囲 (±25%) で、もっとも評価の良い path を探索する。その評価は聴取実験から様々なパラメータを得るときの評価値を統計処理 (決定木) で構成し、各 path におけるパラメータから逆に予測したものである。

まず、始めに予測したものと実際の評価値との相関が高いということを示した。これから評価値を決定づけるパラメータとして delta と Logslope — 基本周波数の対数をとった傾きの最大差と差の 2 乗和平均値 — を決定した。

そのパラメータを要素とする韻律改善においては、不必要な信号処理をすることなく韻律を目標に近づけることができた。

2 今後の課題

先にも述べた通り、この最適 path の探索には full search を用いているので、日常会話文などの長いフレーズには大量の計算時間がかかってしまう。不必要と思われる path をはやめに切り取る速い探索アルゴリズムが必要となる。

予測した評価値と信号処理による音質劣化のコストの組合わせ比 ($weight$) は今回は視察により推測したが、今後統計による決定など、適切な値が自動で求まることが望まれる。

本来はピッチの標準化得点 ($f0_{zs}$) も最適韻律を求めるためのパラメータに加える予定であった。そうすることでより正確な予測、信号処理ができることは図 3.14からも分かる。しかし、各 path における $f0_{zs}$ を算出することができなかったので、採用できなかった。

新たなパラメータ、例えば次のピッチが上がるのか下がるのかを判定する "slope threshold" なる要素を加えるなど、効果的な改善が期待される。

謝辞

本研究を進めるにあたりまして、暖かく見守りつつ、惜しみない協力を頂いたATR音声翻訳通信研究所第二研究室の皆様に深く感謝致します。特に数多くの貴重な助言及び御指導を頂きました丁 文 研究員に心から感謝致します。

最後に、実習の機会を与えて下さり、基礎からいろいろ丁寧に教えてくださったNick Campbell室長兼奈良先端大教授及びATR音声翻訳通信研究所の山本誠一社長に感謝致します。

参考文献

[1] ニック・キャンベル, アラン・ブラック: “CHATR：自然音声波形接続型任意音声合成システム”, 電子情報通信学会技術研究報告,SP96-7,pp.45–52(1996-05).

[2] Charpentier,F. and Moulines,E.: “Pitch - Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones”, Proc.Eurospeech’89(1989).

[3] 河井恒, 樋口宜男, 清水徹, 山本誠一: “波形素片接続型音声合成システムの改良と評価”, 電子情報通信学会春季大会報告,SA-5-7,pp.496–497(1994).

[4] 坂野秀樹, ニック・キャンベル: “自然音声波形接続型合成システムにおける波形接続方式の検討”, TR-IT-0197.

付録 A

使用データ

1　聴取実験に使用したサンプル文

文中の添字がついてる区間が、聴覚実験に使用されたサンプル部分で、数字は図 2.1(p.3) における位置に対応している。

- あらゆる$_{1}$現実をすべて$_{2}$自分のほうへねじ曲げたのだ。$_{3}$
- 一週間ばかり$_{1}$ニューヨークを取材した。$_{3}$
- テレビゲーム$_{4}$やパソコンでゲーム$_{2}$をして遊ぶ。$_{3}$
- 救急車$_{4}$が十分に$_{2}$動けず、救助作業が遅れている。$_{3}$
- おごりを$_{1}$捨て、謙虚な$_{4}$姿勢を取り戻さねば、$_{2}$冬は過ごせない。
- しかし、$_{4}$このプロ野球ブームも$_{2}$永久に続くとは限らぬ。$_{3}$
- アフリカ人は、$_{4}$実に$_{2}$巧みにぴゅんとつばを吐く。
- 大昔の$_{1}$フィリピンには$_{2}$豊かな土地があった。
- 旅館や$_{1}$ホテルに$_{4}$着くと、$_{2}$非常口を尋ねる。
- やるべきことは$_{1}$やっており、$_{2}$なんら落ち度はない。
- 私は上着を脱ぎ、$_{4}$石組みの上に$_{2}$両手をついて、うつぶせになった。$_{3}$
- 人間とは$_{1}$微妙で複雑な生き物で$_{2}$ある。$_{3}$
- ここ$_{1}$一か月はほとんど$_{2}$不眠不休の徹夜つづきで目が腫れ上がっている。$_{3}$
- 見上げる$_{1}$フジもいいが露地植え、また鉢植えの花も$_{4}$きれいです。$_{2}$
- 母は$_{4}$脳血栓の$_{2}$後遺症で、老人性痴呆症になり一年前から入院中です。$_{3}$
- 着用中に$_{1}$ダウンや$_{4}$フェザーが$_{2}$飛び出す原因ともなります。

- 普通、(1) 中距離トラックの (2) ドライバーは (3) 中年の人が (1) 多い。
- 「ユーザーにも (4) 責任が (2) ある」との (3) 論理は暴論と言わざるを (2) えません。
- 首相 (1) 自ら (4) 国民 (2) 一人一人 (1) 百ドル、舶来品を買うようにすすめた。
- 日本の (1) エスペラントとして、 (2) やはり 標準語は (3) 必要だ。

2 使用プログラム

2.1 聴取実験データ編集

データ及びプログラム (awk) は

/data1/itlpc06/ding/Percept_Unit_Sele2/

以下の

ch_file/ — 処理するデータ

Result/ — 結果

command/ — 使用プログラム

に納められている。詳しくは Step2.log を参照されたい。

2.2 回帰木、グラフ表示

2.1節同様、 /data1/itlpc06/ding/Percept_Unit_Sele2/

以下の

Result/ — 結果

command/ — 使用プログラム (s 言語)

に納められている。

付録 Bのデータは

/home/as67/xmaru/report/

に post_script ファイルで格納してある。

2.3 PSOLA 組み込み CHATR

コマンド /data1/itlpc06/ding/chatr-0.94/src/main/chatr

で起動する CHATR version 0.94 alpha (28 Aug 1998)

に、最適パス探索かつ PSOLA 信号処理のプログラムを組み込んだ。

/data1/itlpc06/ding/chatr-0.94/src/ruc/ の中に

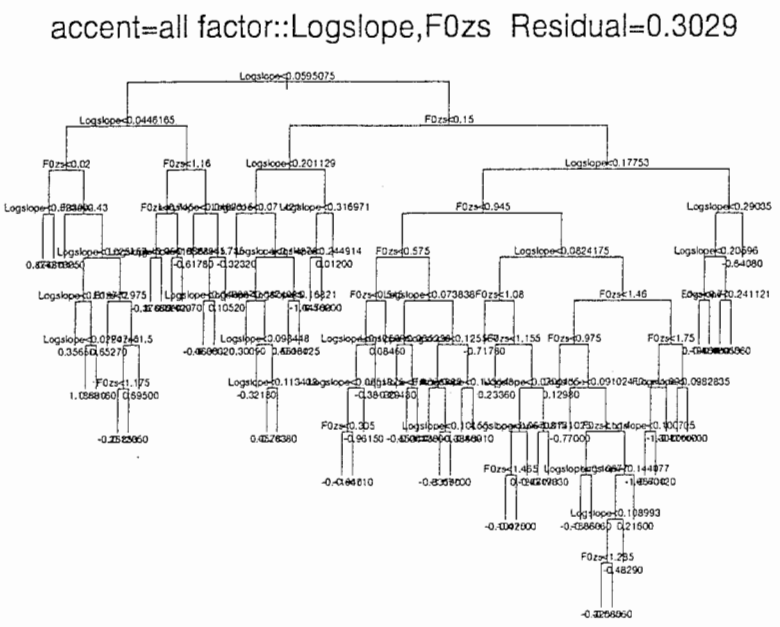
- ding_prosody.c の"psola_min_f0_ratio , psola_max_f0_ratio" の値を変えることができる。
- search_path.h の PSOLAW を変えることで *weight*(式 4.3) の値を変えることができる。
- また search_path.h の NUM で 1 モーラ辺り変更可能なピッチ数 を設定できる。

付録 B

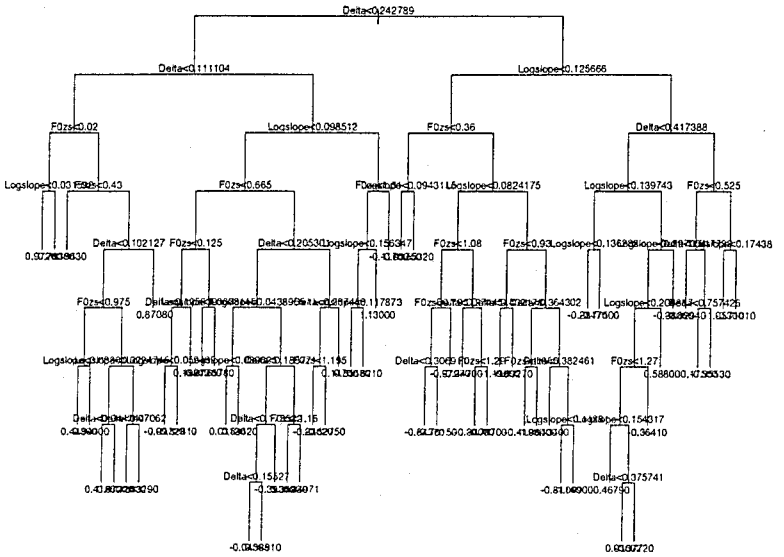
全結果表示

1　決定木

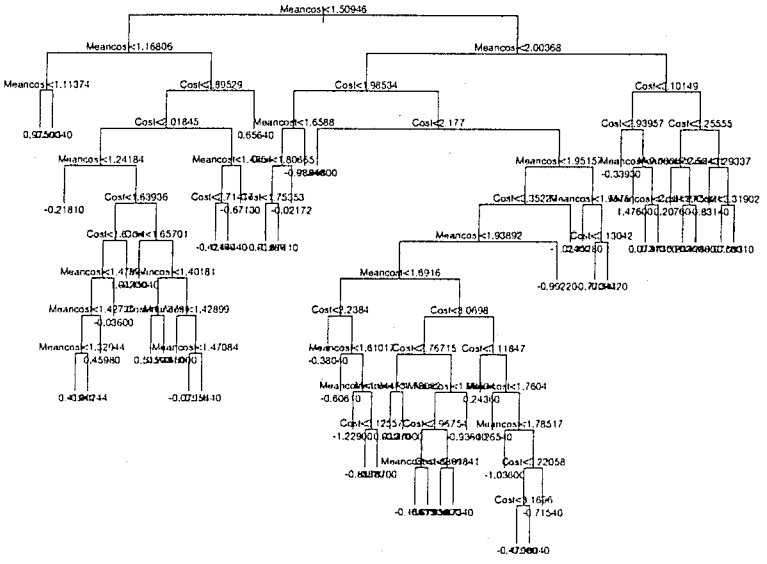
全サンプルから構成した決定木



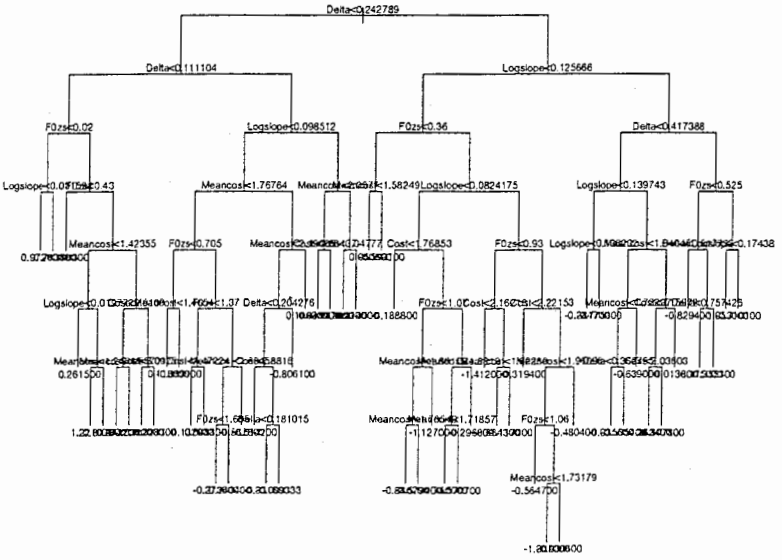
accent=all factor::Logslope,F0zs,Delta Residual=0.2363



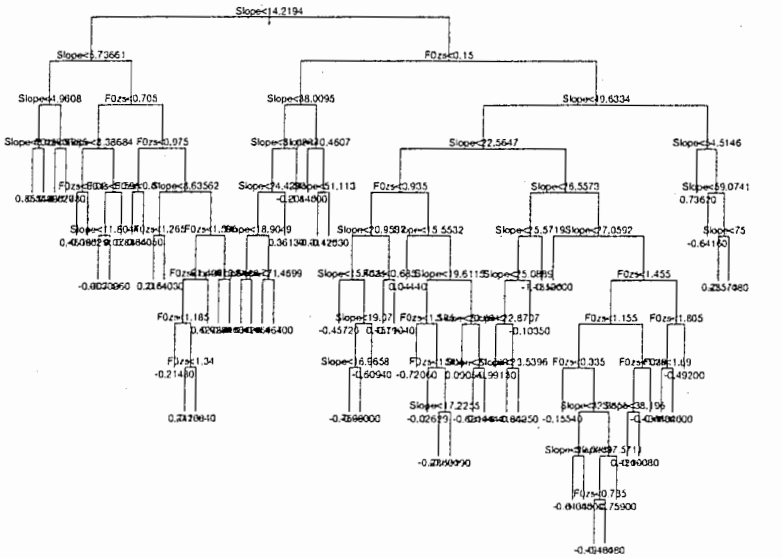
accent=all factor::Maxcost,Meancost Residual=0.2524



accent=all factor::Logslope,F0zs,Maxcost,Meancost,Delta Residual=0.198



accent=all factor::Slope,F0zs Residual=0.3149



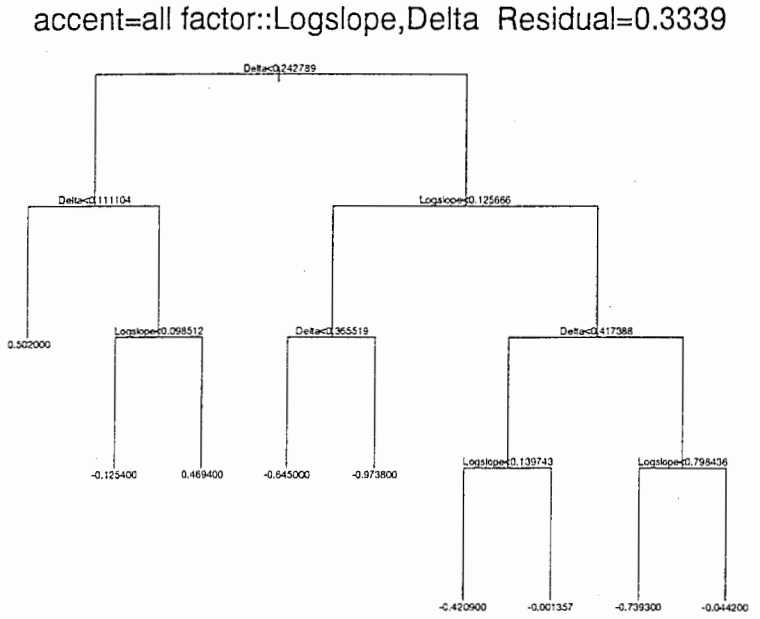
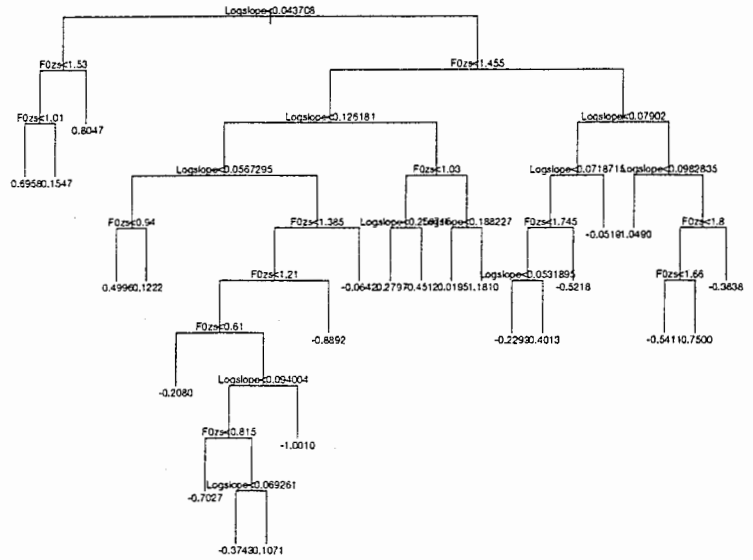


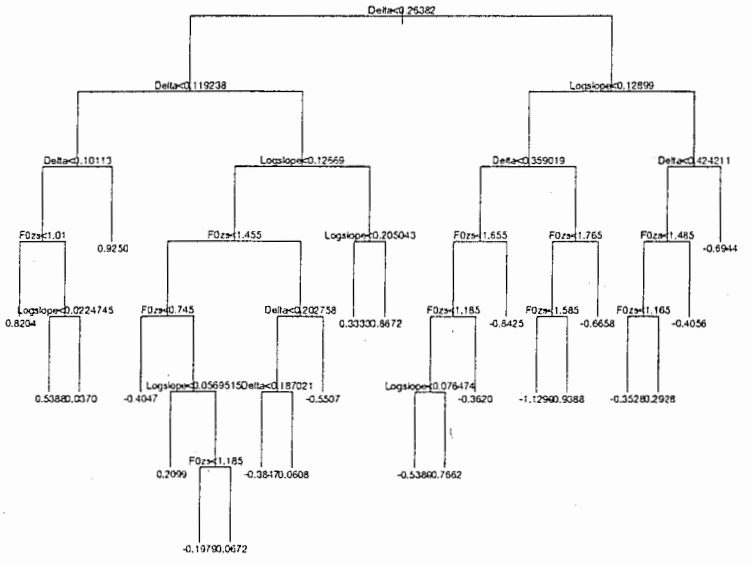
図 B.1: CHATR 使用決定木 (pruning)

サンプルの位置ごとに分類した決定木

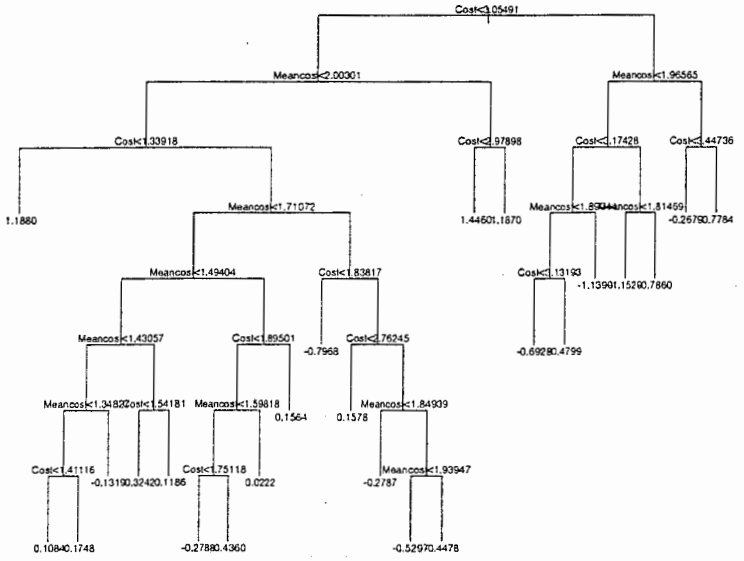
accent=1 factor::Logslope,F0zs Residual=0.2576



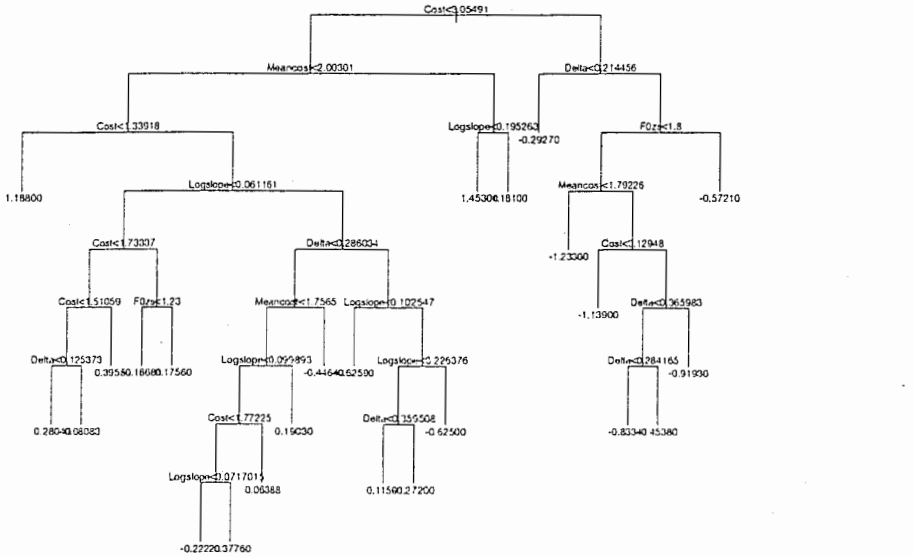
accent=1 factor::Logslope,F0zs,Delta Residual=0.2196



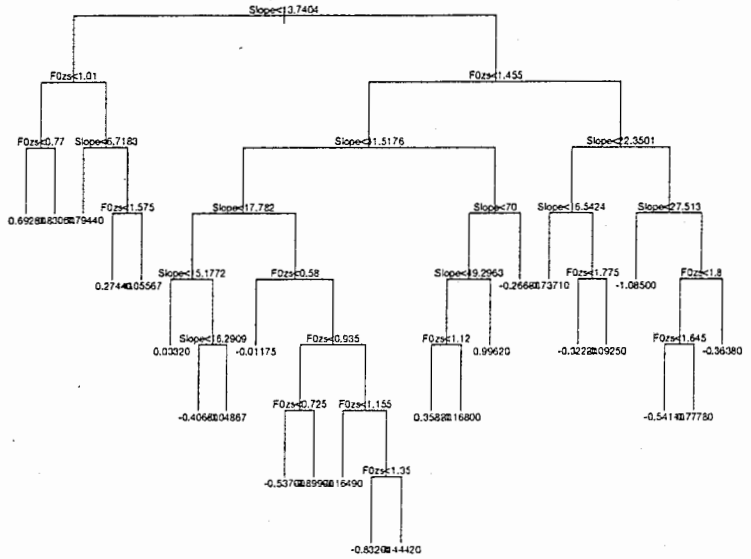
accent=1 factor::Maxcost,Meancost Residual=0.14



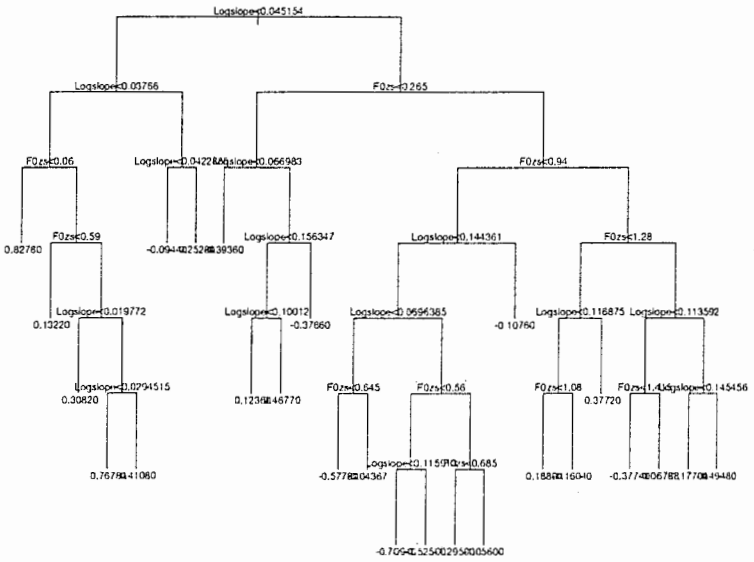
accent=1 factor::Logslope,F0zs,Maxcost,Meancost,Delta Residual=0.1183



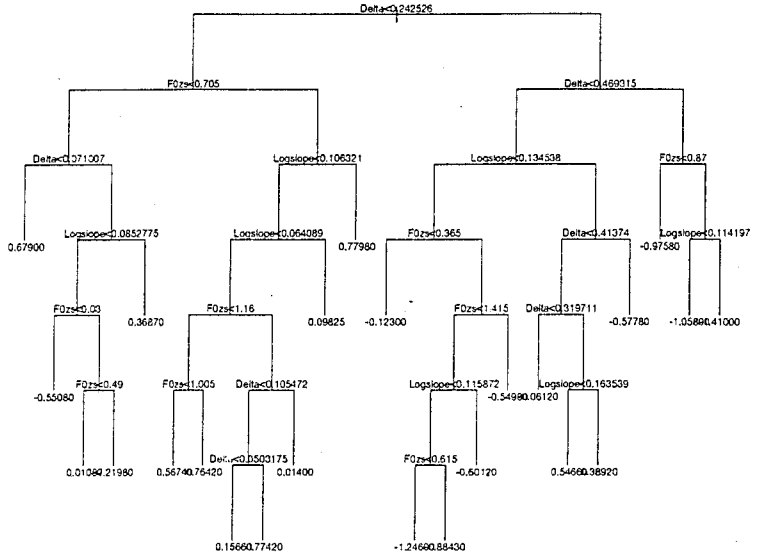
accent=1 factor::Slope,F0zs Residual=0.2824



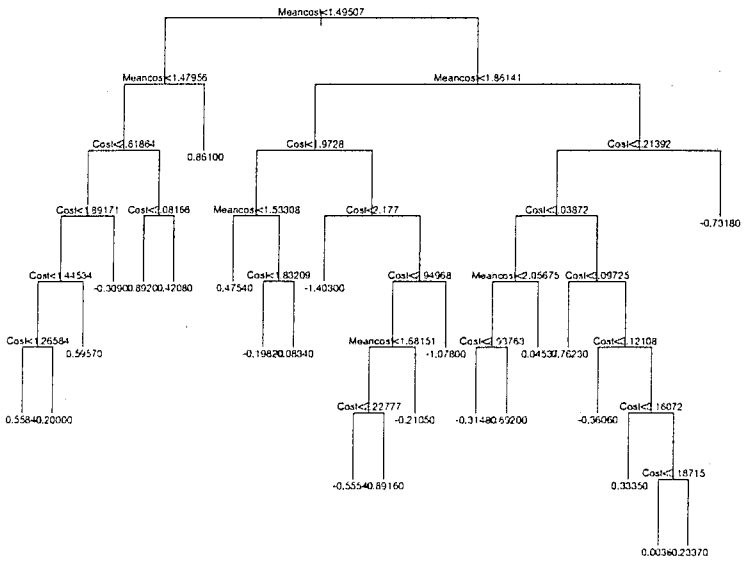
accent=2 factor::Logslope,F0zs Residual=0.3756



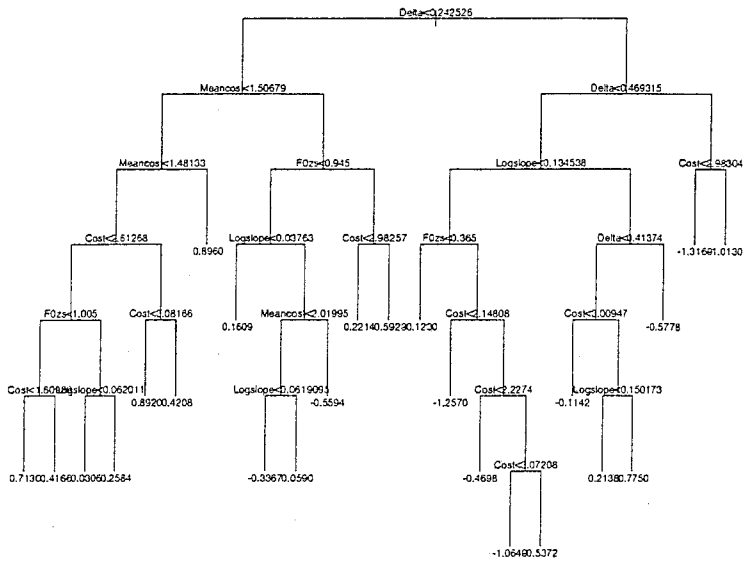
accent=2 factor::Logslope,F0zs,Delta Residual=0.1975



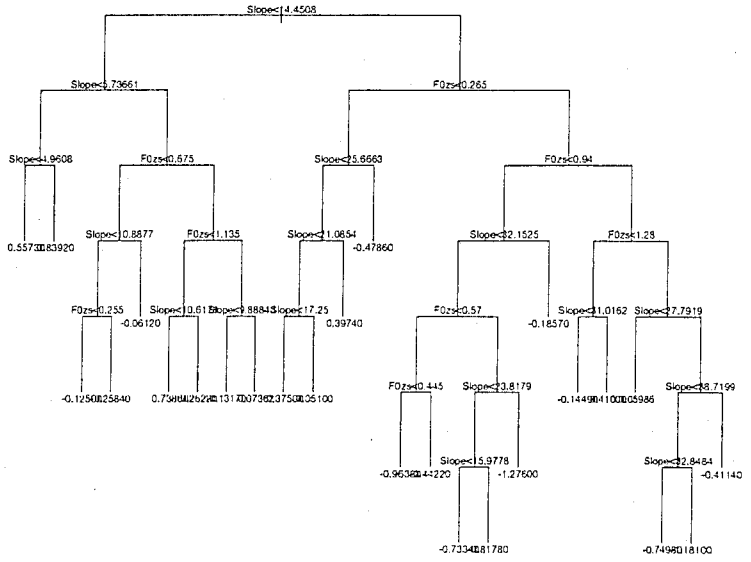
accent=2 factor::Maxcost,Meancost Residual=0.2362



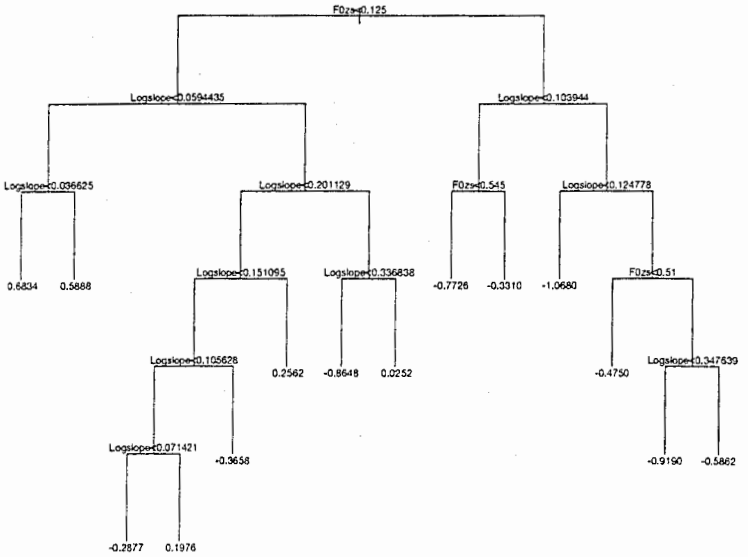
t=2 factor::Logslope,F0zs,Maxcost,Meancost,Delta Residual=



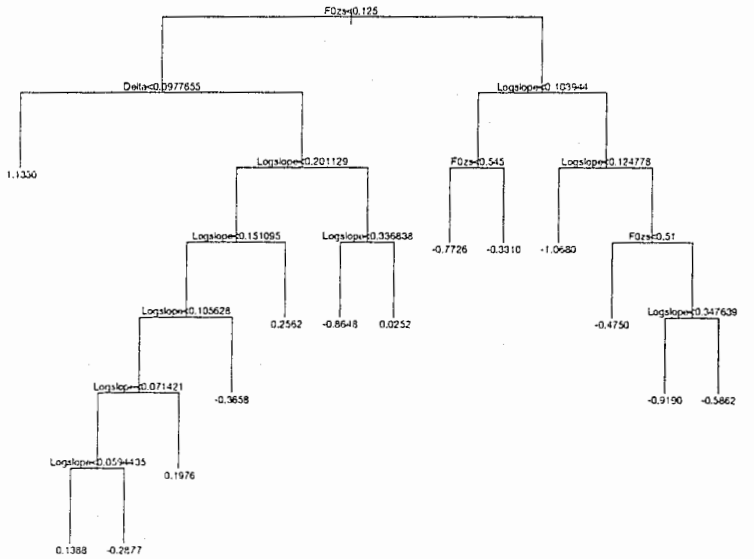
accent=2 factor::Slope,F0zs Residual=0.3812



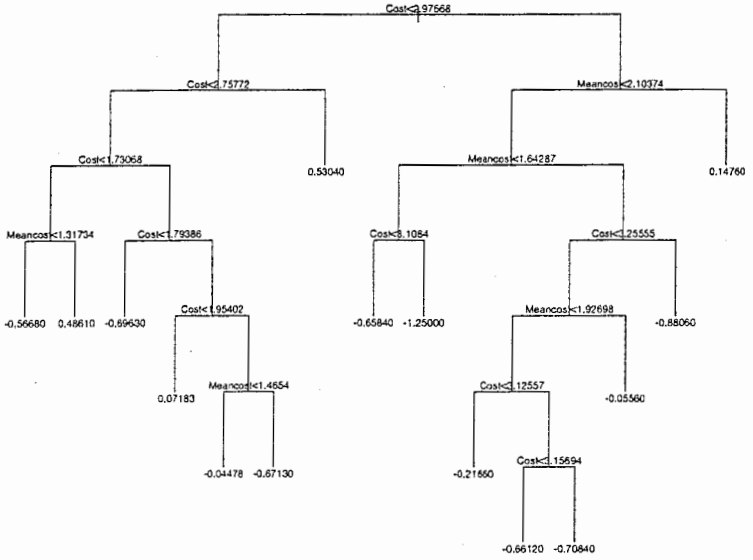
accent=3 factor::Logslope,F0zs Residual=0.2664



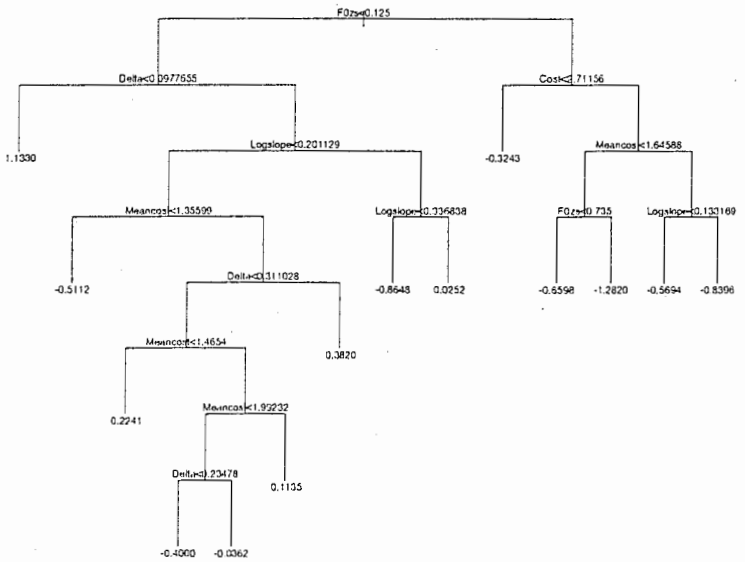
accent=3 factor::Logslope,F0zs,Delta Residual=0.2341



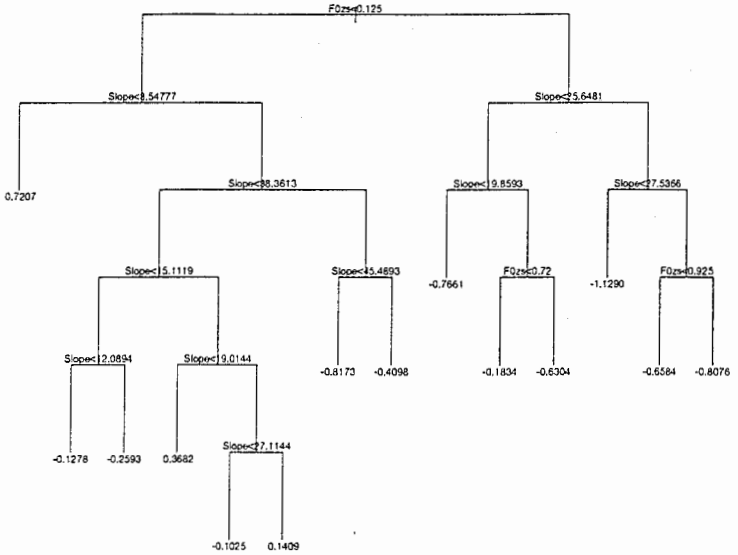
accent=3 factor::Maxcost,Meancost Residual=0.3191



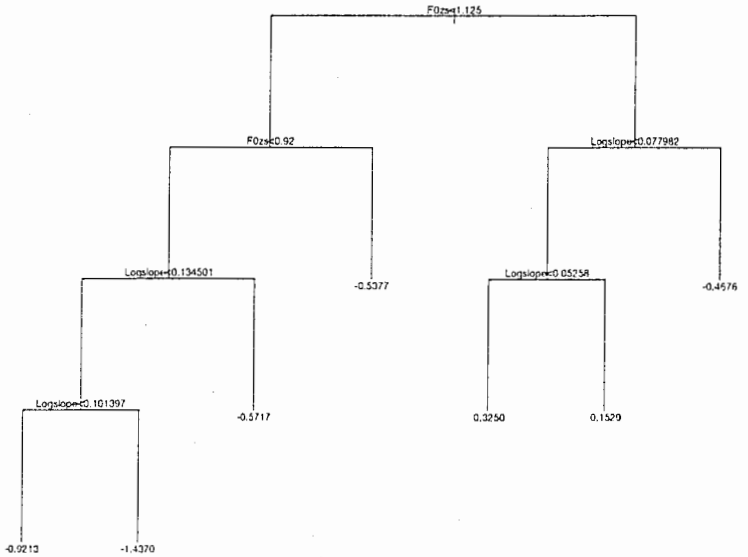
accent=3 factor::Logslope,F0zs,Maxcost,Meancost,Delta Residual=0.2099



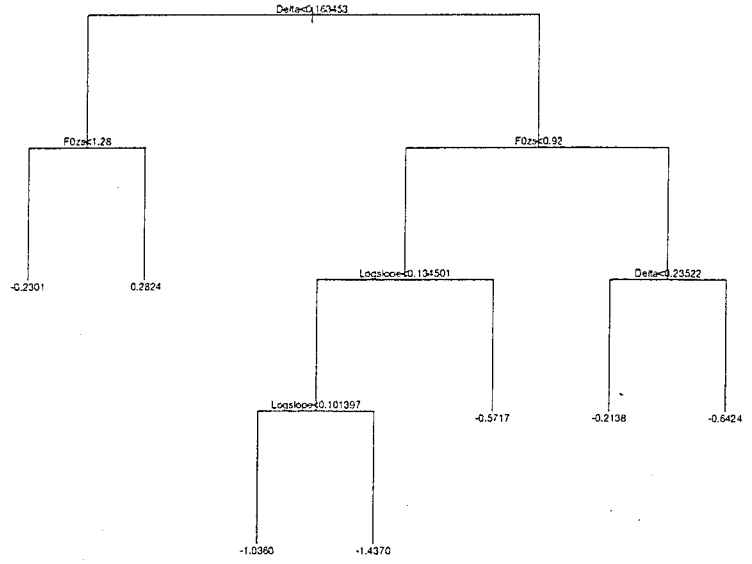
accent=3 factor::Slope,F0zs Residual=0.2801



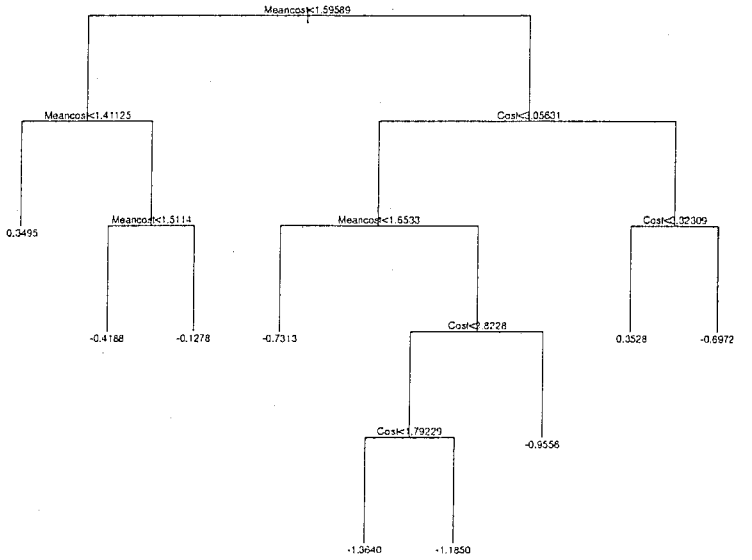
accent=4 factor::Logslope,F0zs Residual=0.1958



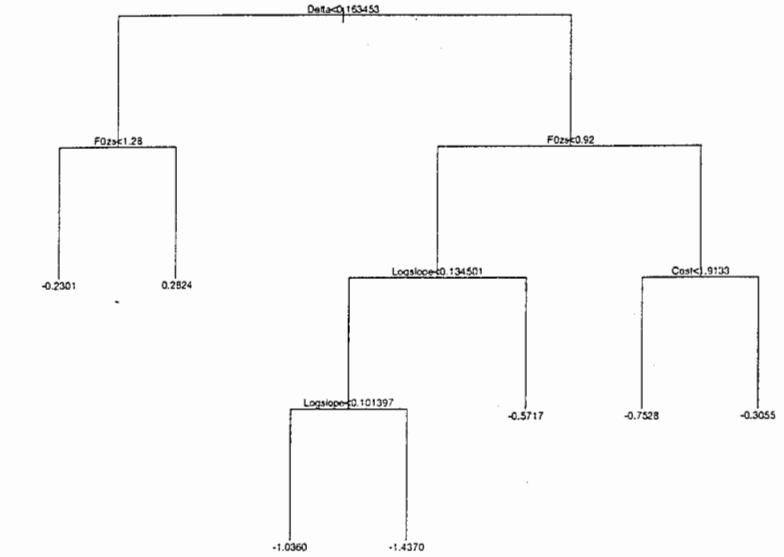
accent=4 factor::Logslope,F0zs,Delta Residual=0.1751



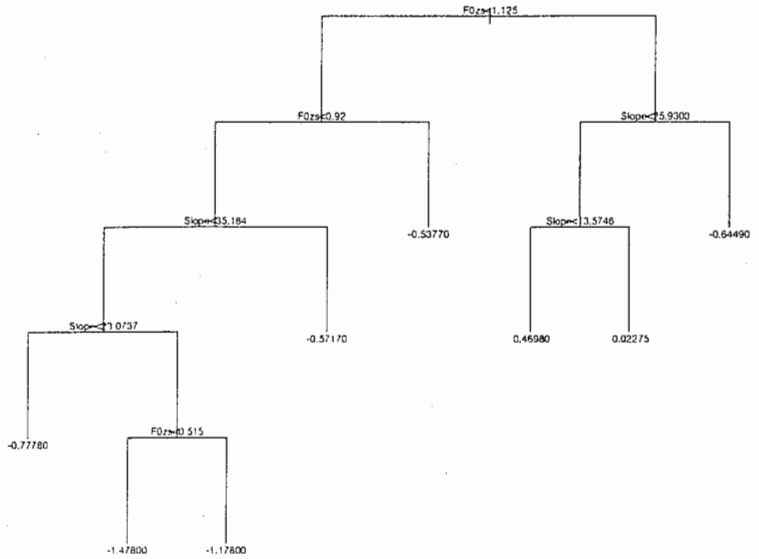
accent=4 factor::Maxcost,Meancost Residual=0.13



accent=4 factor::Logslope,F0zs,Maxcost,Meancost,Delta Residual=0.1739

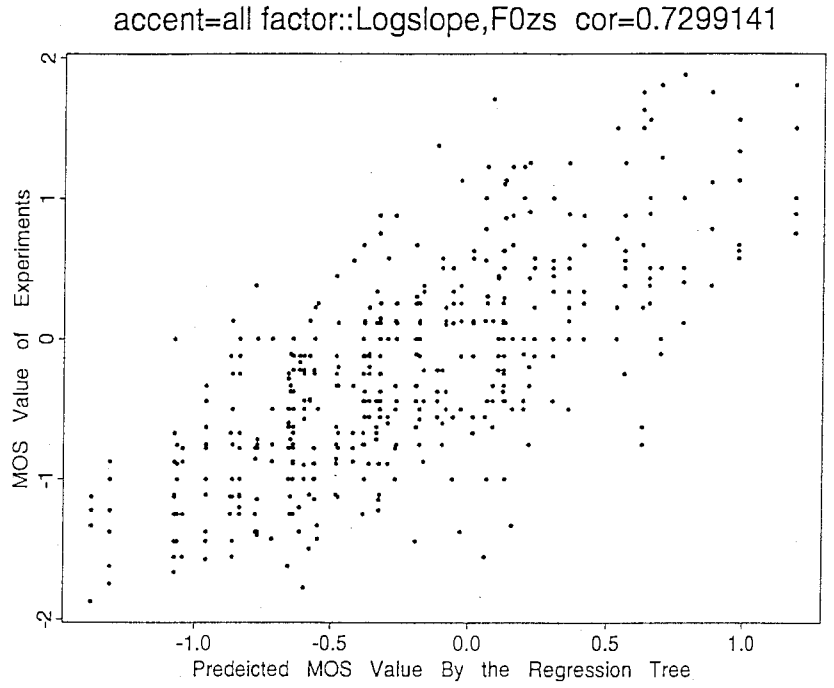


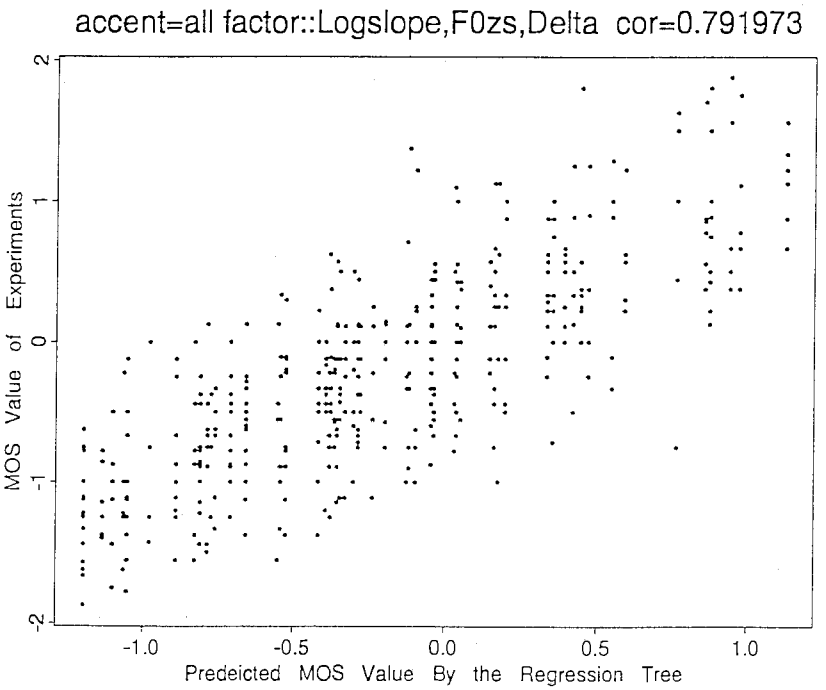
accent=4 factor::Slope,F0zs Residual=0.1583

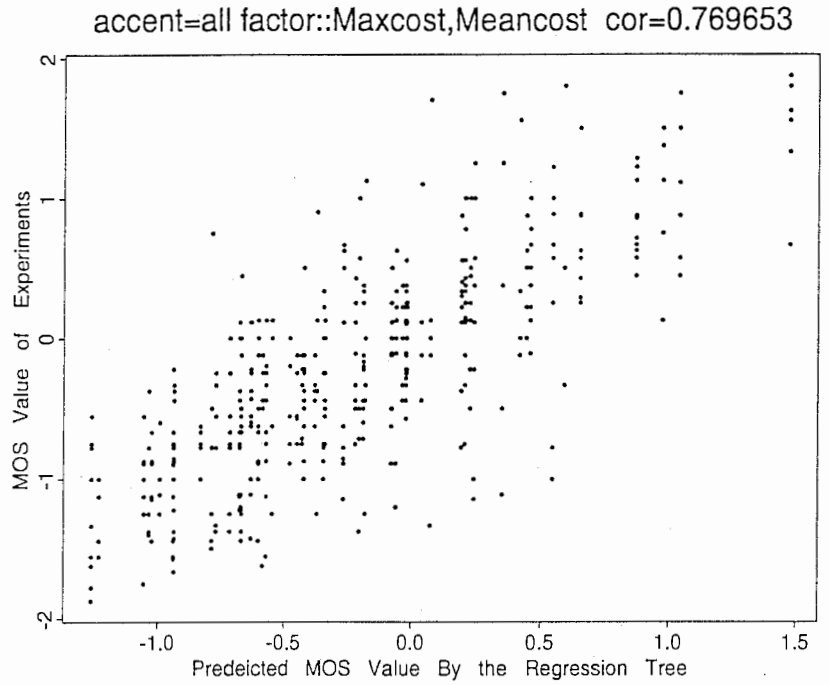


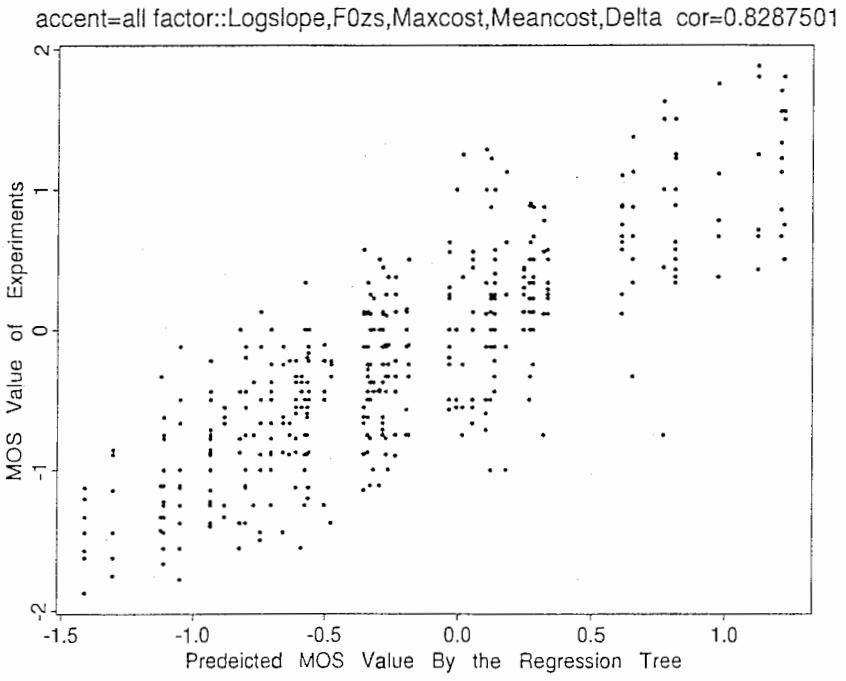
2 決定木による予測値と実データの MOS 値との相関性

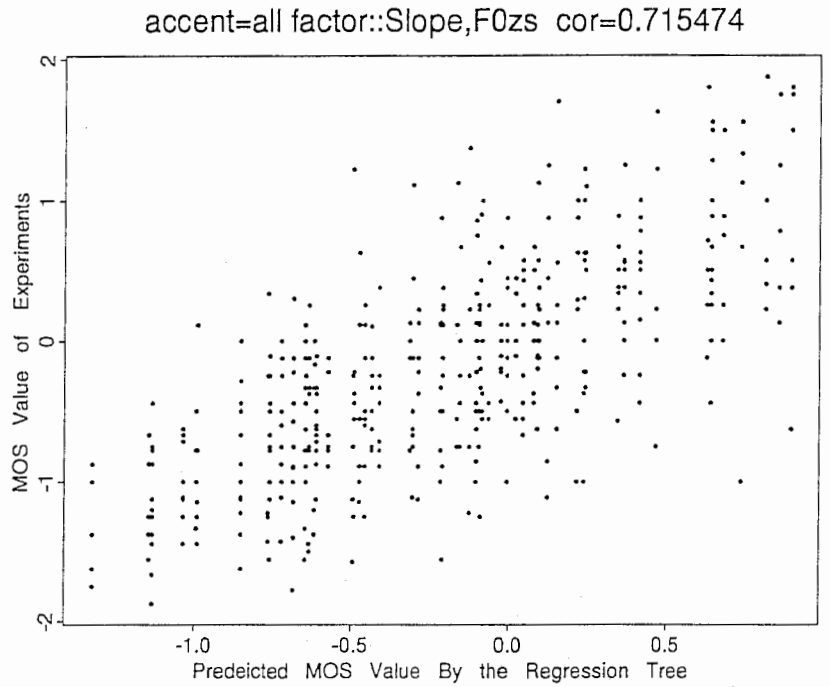
全サンプルのデータで構成、予測、比較






accent=all factor::Maxcost,Meancost cor=0.769653
MOS Value of Experiments
Predeicted MOS Value By the Regression Tree
-2
-1
0
1
2
-1.0
-0.5
0.0
0.5
1.0
1.5


accent=all factor::Logslope,F0zs,Maxcost,Meancost,Delta cor=0.8287501
MOS Value of Experiments
Predeicted MOS Value By the Regression Tree
-2
-1
0
1
2
-1.5
-1.0
-0.5
0.0
0.5
1.0



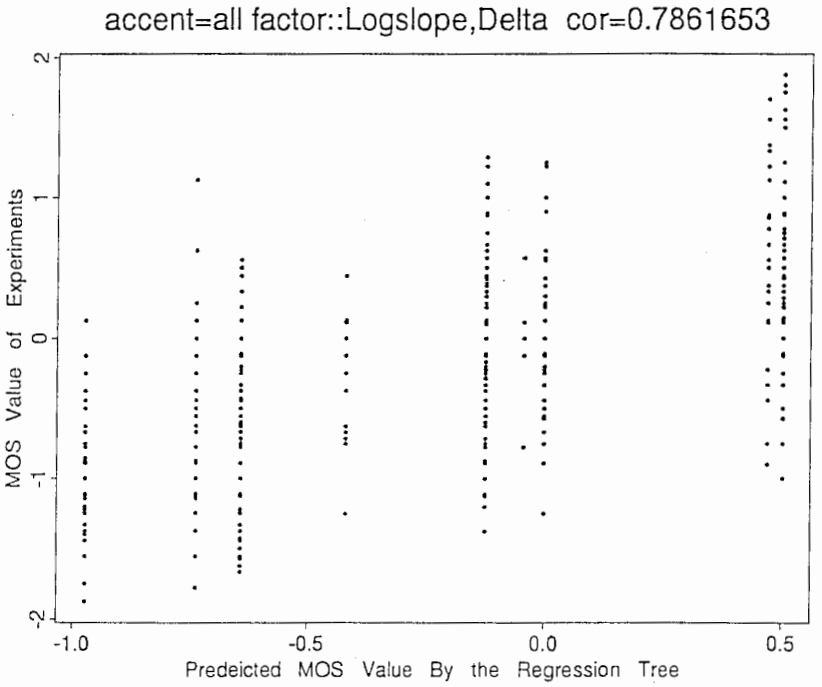
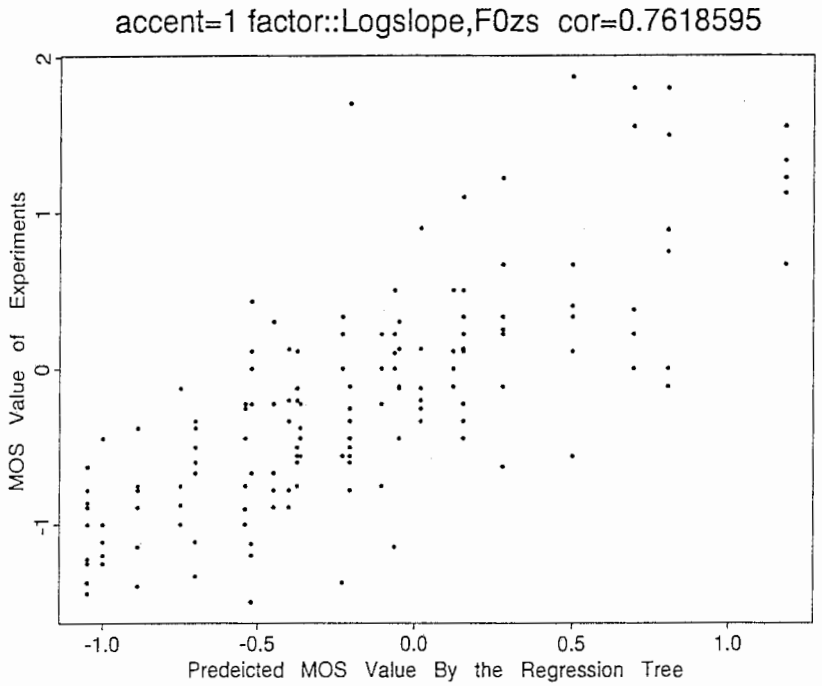
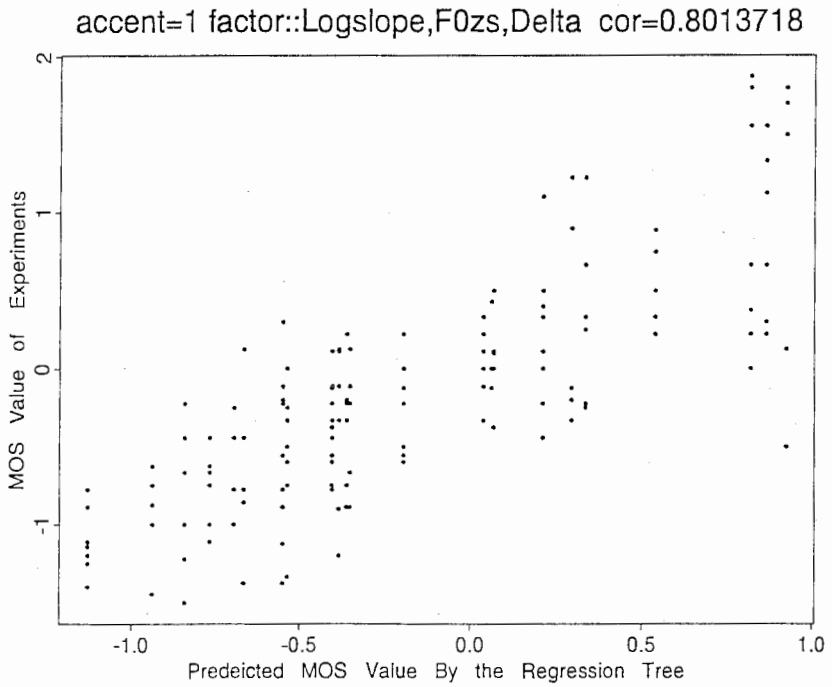
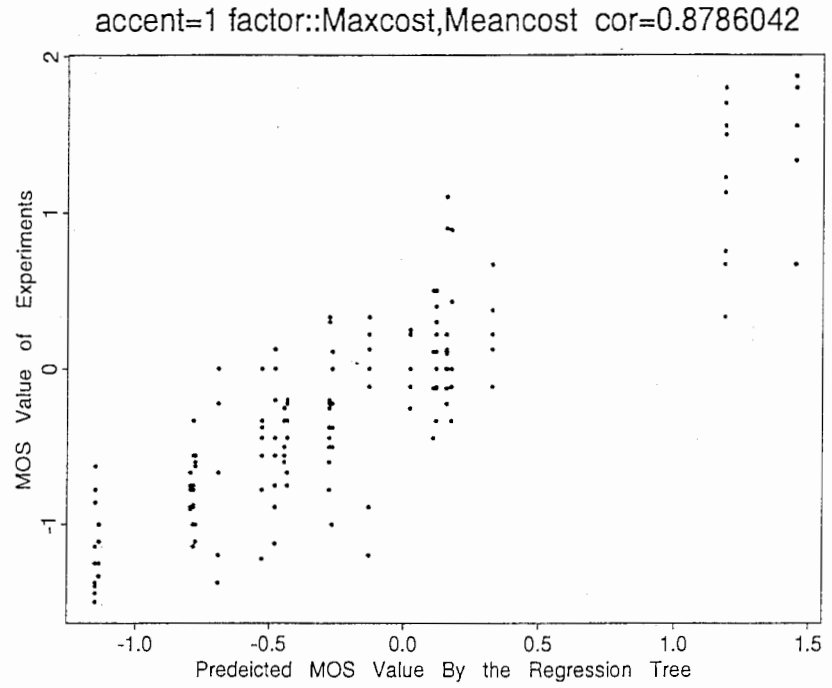


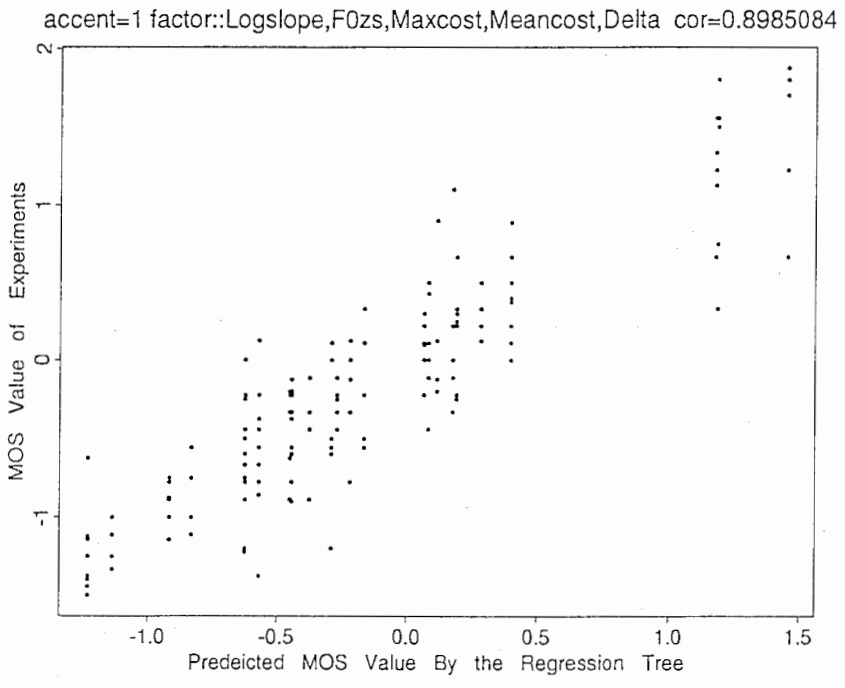
図 B.2: CHATR 使用決定木 (pruning) の評価

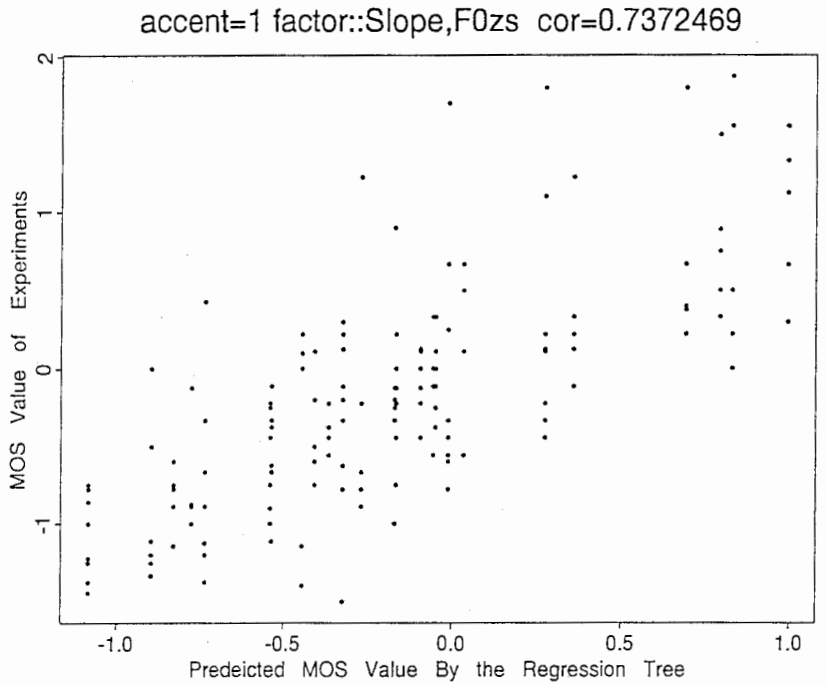
サンプルの位置ごとに構成、予測、比較

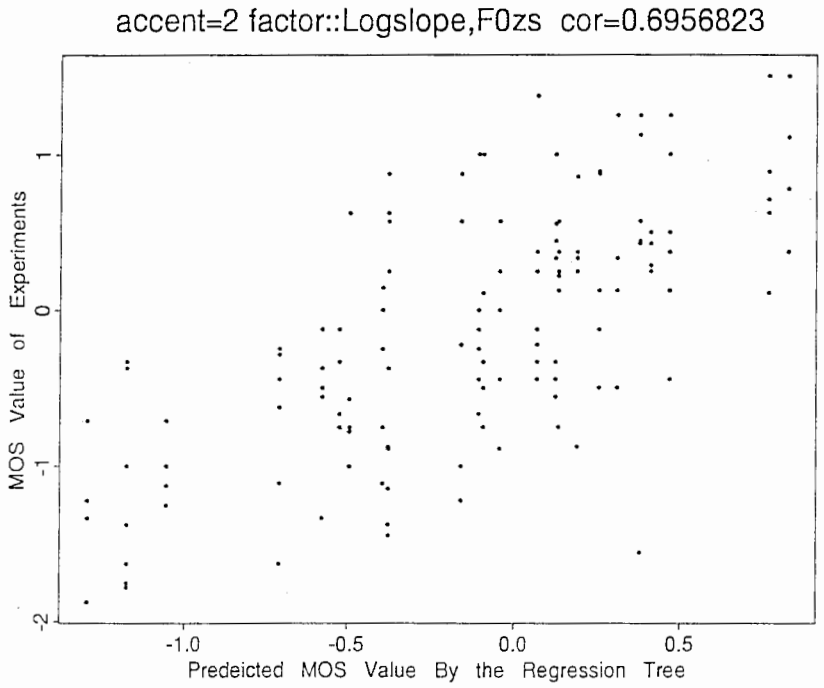


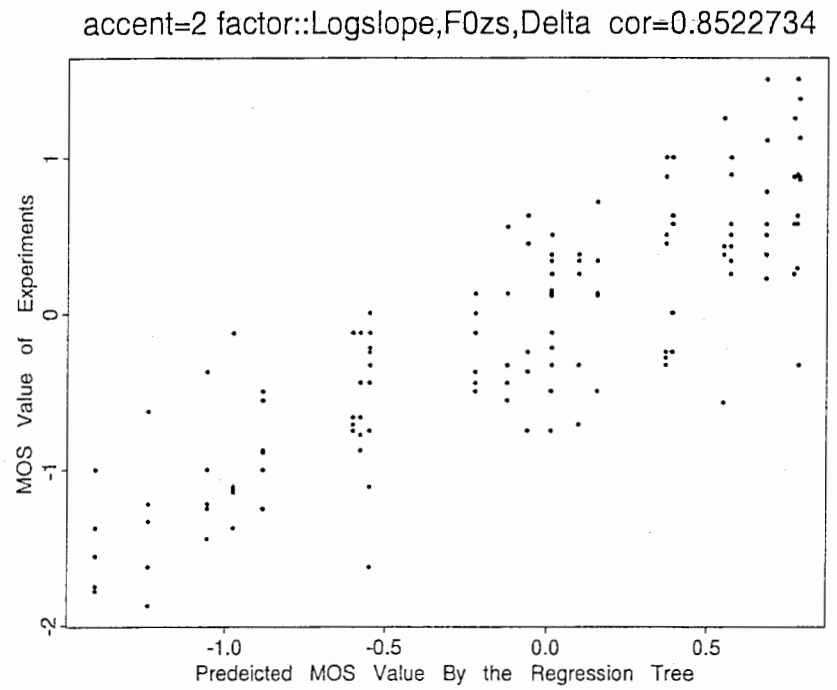


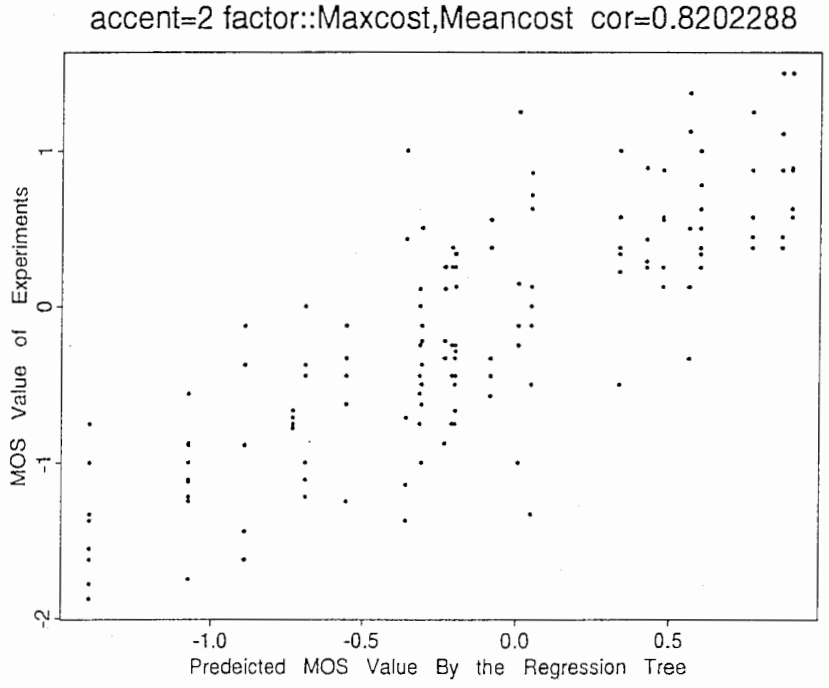

accent=1 factor::Maxcost,Meancost cor=0.8786042
MOS Value of Experiments
Predeicted MOS Value By the Regression Tree
-1
0
1
2
-1.0
-0.5
0.0
0.5
1.0
1.5

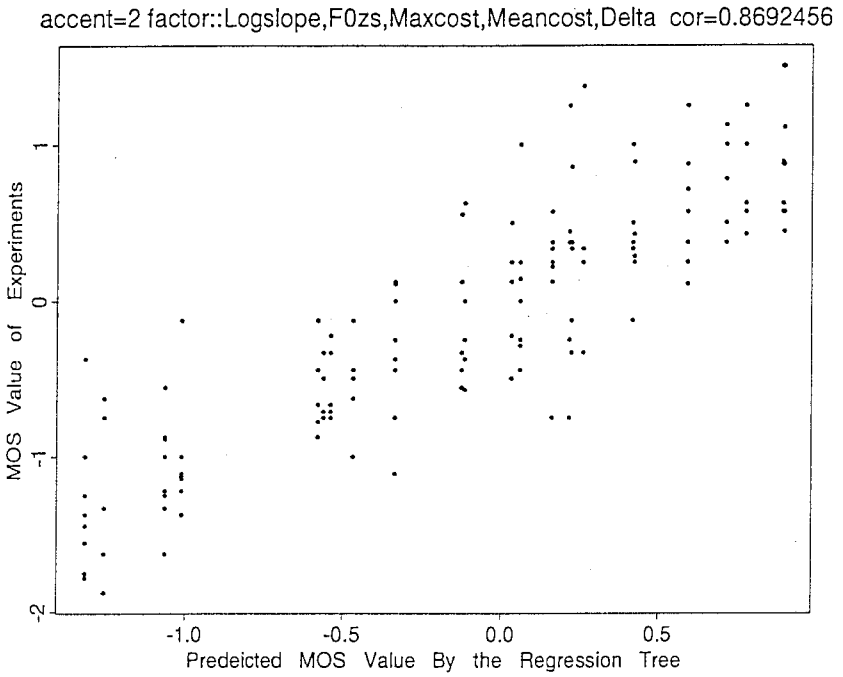

accent=1 factor::Logslope,F0zs,Maxcost,Meancost,Delta cor=0.8985084
MOS Value of Experiments
Predeicted MOS Value By the Regression Tree
-1
0
1
2
-1.0
-0.5
0.0
0.5
1.0
1.5

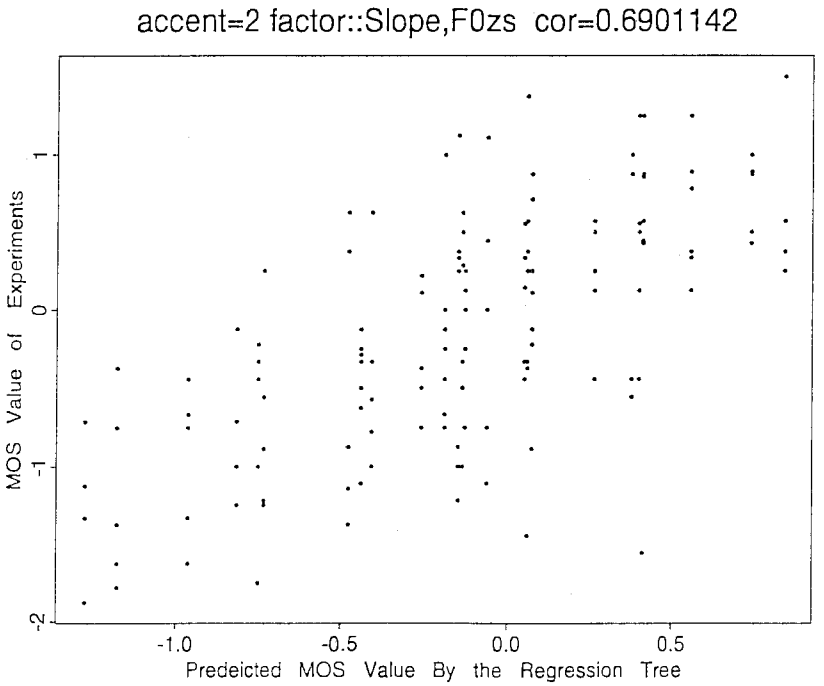

accent=1 factor::Slope,F0zs cor=0.7372469
MOS Value of Experiments
2
1
0
-1
-1.0
-0.5
0.0
0.5
1.0
Predeicted MOS Value By the Regression Tree

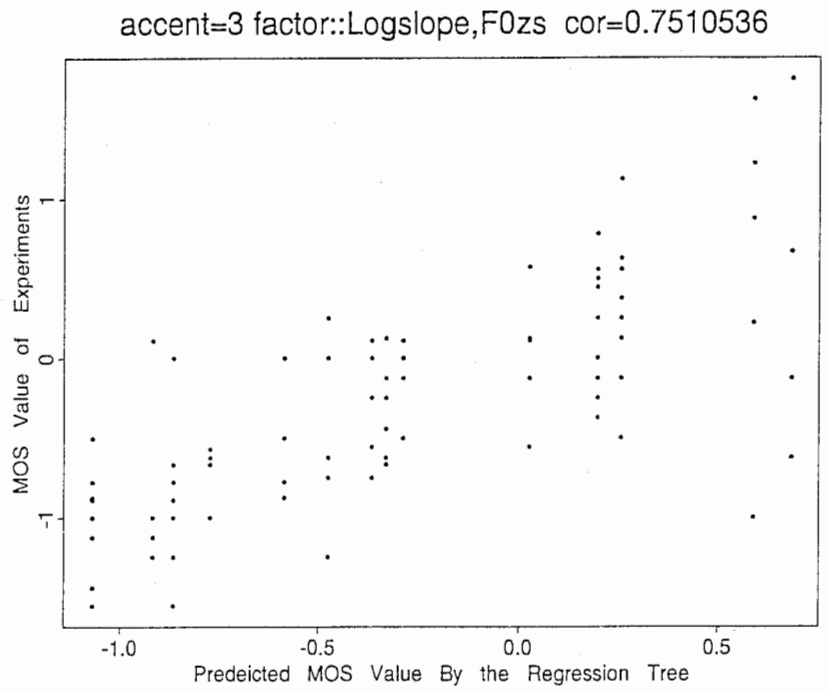

accent=2 factor::Logslope,F0zs cor=0.6956823
MOS Value of Experiments
1
0
-1
-2
-1.0
-0.5
0.0
0.5
Predeicted MOS Value By the Regression Tree

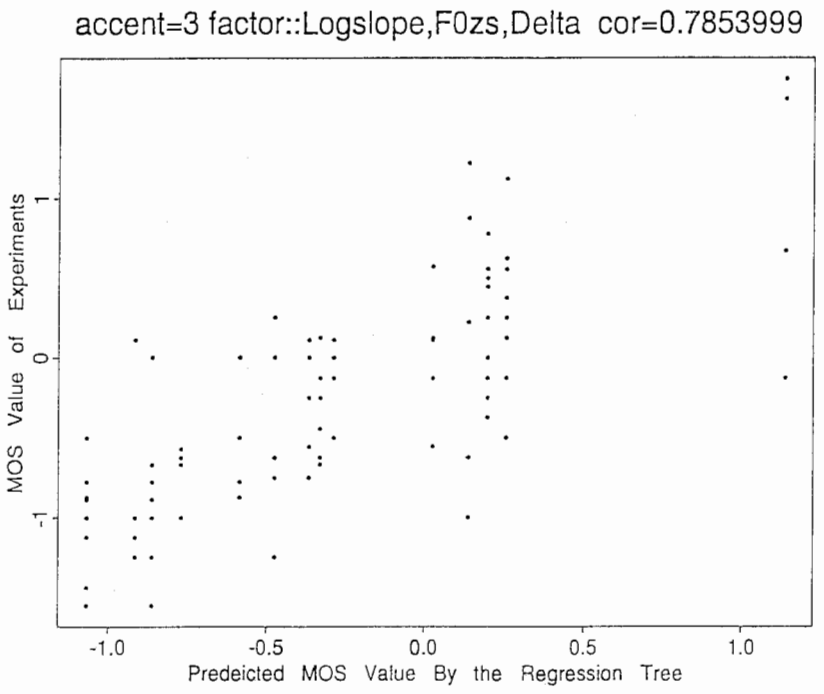

accent=2 factor::Logslope,F0zs,Delta cor=0.8522734
MOS Value of Experiments
-2
-1
0
1
-1.0
-0.5
0.0
0.5
Predeicted MOS Value By the Regression Tree

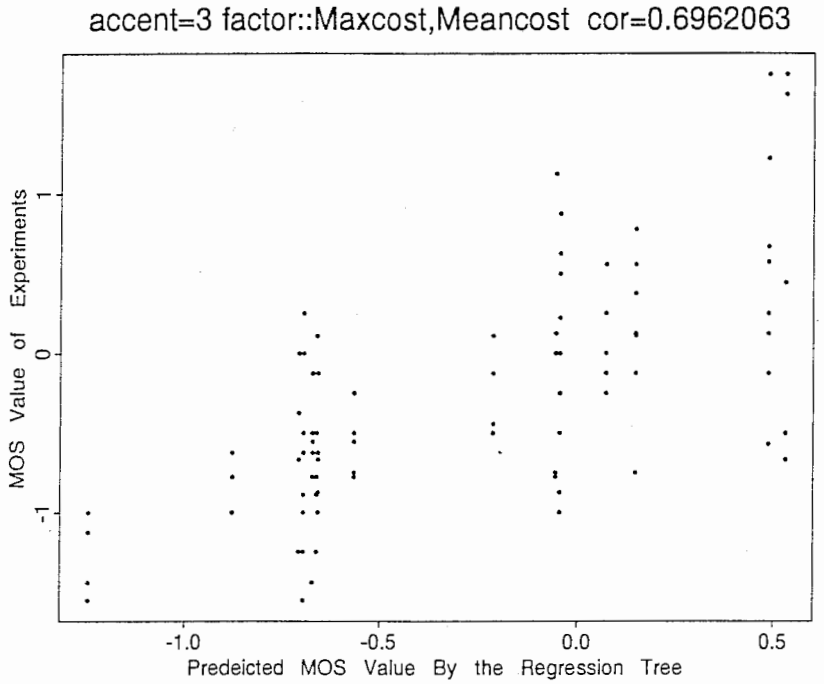

accent=2 factor::Maxcost,Meancost cor=0.8202288
MOS Value of Experiments
-2
-1
0
1
-1.0
-0.5
0.0
0.5
Predeicted MOS Value By the Regression Tree

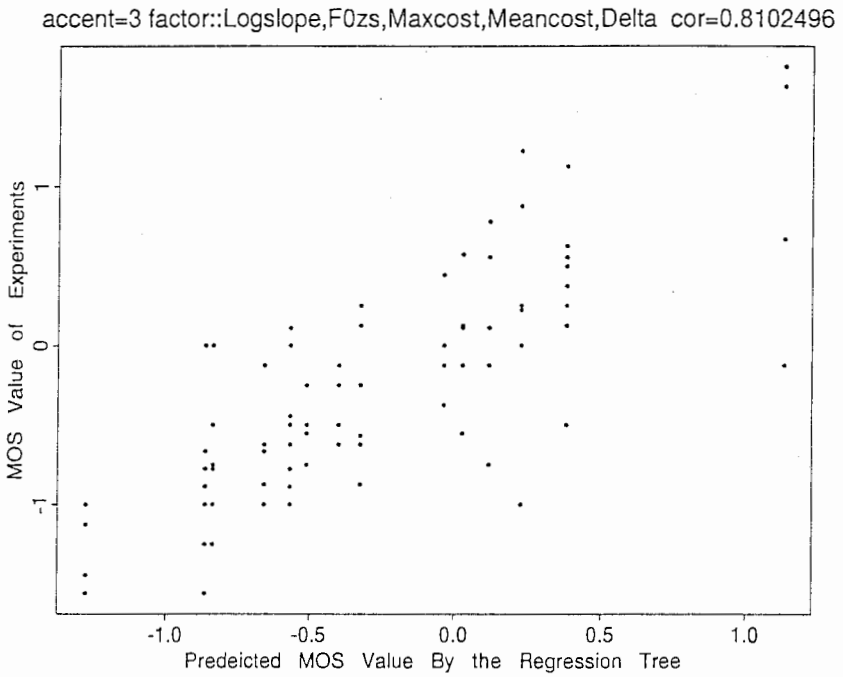

accent=2 factor::Logslope,F0zs,Maxcost,Meancost,Delta cor=0.8692456
MOS Value of Experiments
-2
-1
0
1
-1.0
-0.5
0.0
0.5
Predeicted MOS Value By the Regression Tree

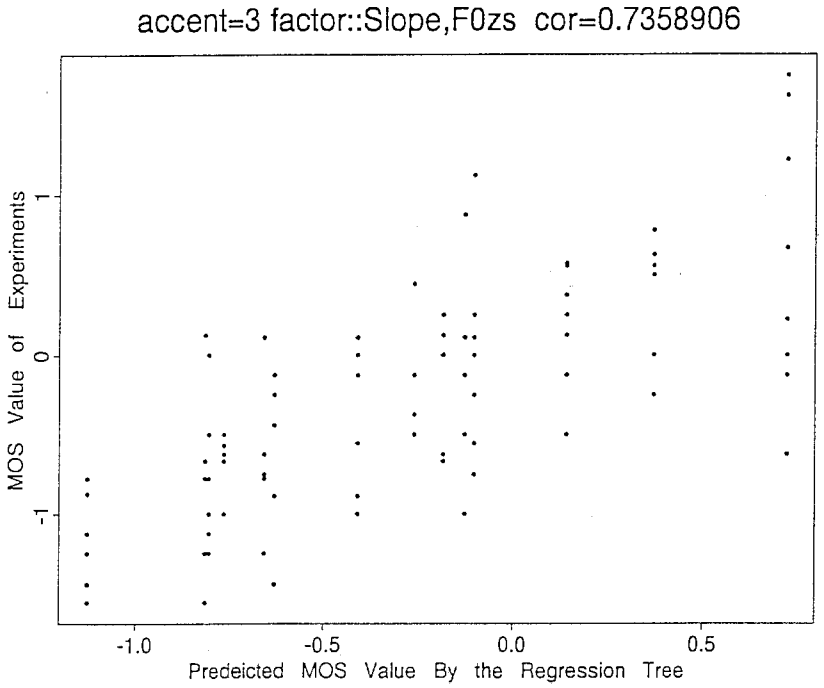

accent=2 factor::Slope,F0zs cor=0.6901142
MOS Value of Experiments
-2
-1
0
1
-1.0
-0.5
0.0
0.5
Predeicted MOS Value By the Regression Tree

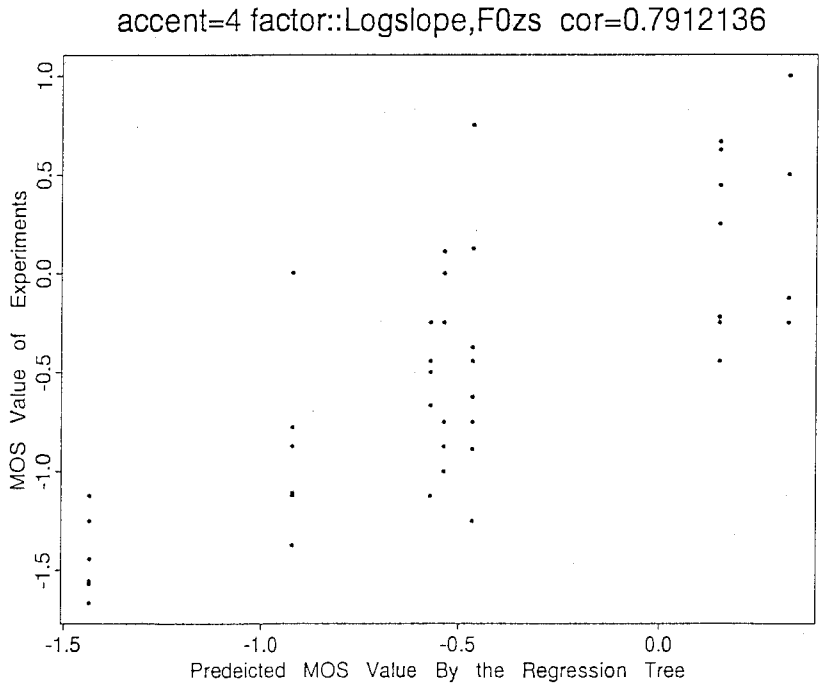

accent=3 factor::Logslope,F0zs cor=0.7510536
MOS Value of Experiments
-1
0
1
-1.0
-0.5
0.0
0.5
Predeicted MOS Value By the Regression Tree

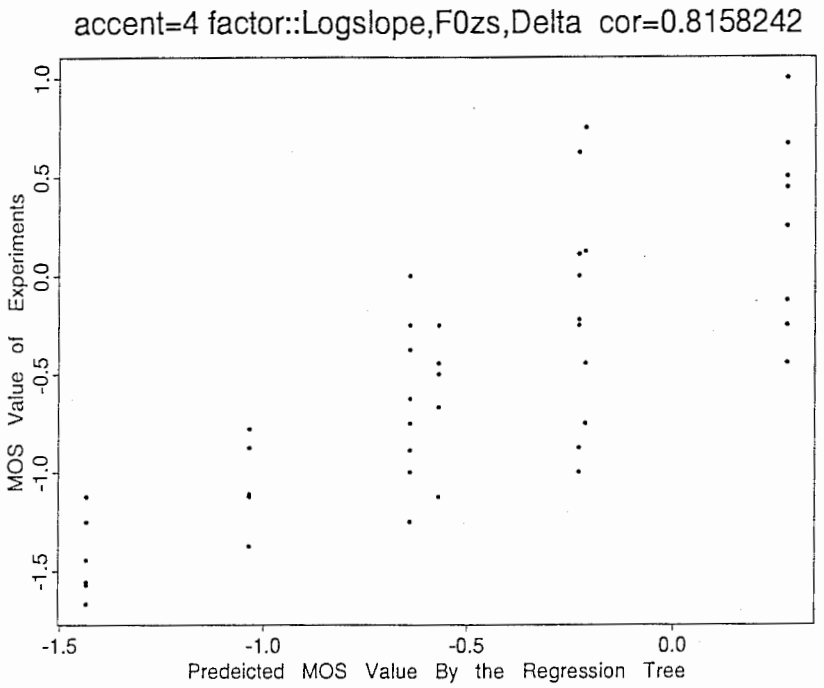

accent=3 factor::Logslope,F0zs,Delta cor=0.7853999
MOS Value of Experiments
-1
0
1
-1.0
-0.5
0.0
0.5
1.0
Predeicted MOS Value By the Regression Tree

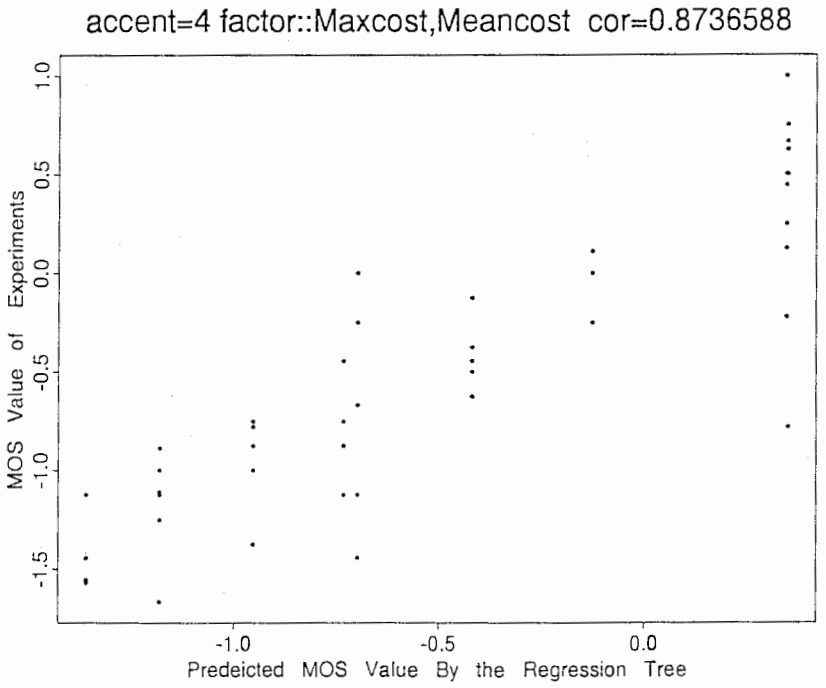

accent=3 factor::Maxcost,Meancost cor=0.6962063
MOS Value of Experiments
-1
0
1
-1.0
-0.5
0.0
0.5
Predeicted MOS Value By the Regression Tree

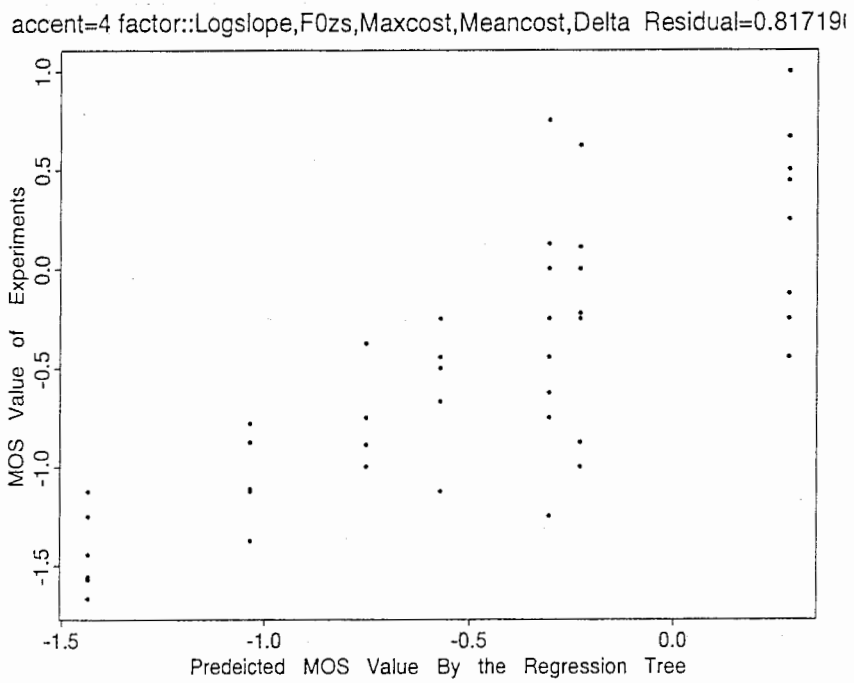

accent=3 factor::Logslope,F0zs,Maxcost,Meancost,Delta cor=0.8102496
MOS Value of Experiments
-1
0
1
-1.0
-0.5
0.0
0.5
1.0
Predeicted MOS Value By the Regression Tree

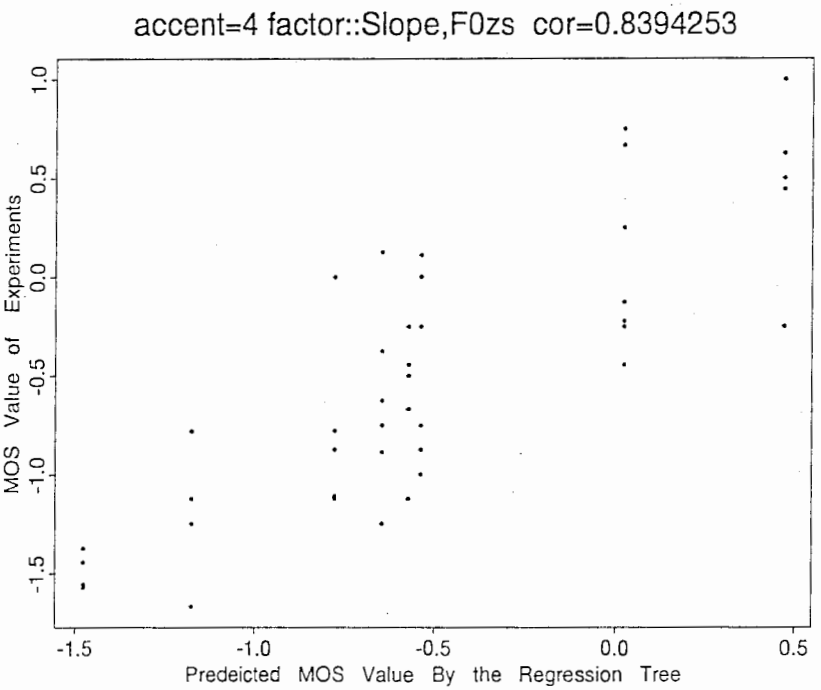

accent=3 factor::Slope,F0zs cor=0.7358906
MOS Value of Experiments
-1
0
1
-1.0
-0.5
0.0
0.5
Predeicted MOS Value By the Regression Tree


accent=4 factor::Logslope,F0zs cor=0.7912136
MOS Value of Experiments
-1.5
-1.0
-0.5
0.0
0.5
1.0
-1.5
-1.0
-0.5
0.0
Predeicted MOS Value By the Regression Tree


accent=4 factor::Logslope,F0zs,Delta cor=0.8158242
MOS Value of Experiments
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
Predeicted MOS Value By the Regression Tree


accent=4 factor::Maxcost,Meancost cor=0.8736588
MOS Value of Experiments
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
Predeicted MOS Value By the Regression Tree


accent=4 factor::Logslope,F0zs,Maxcost,Meancost,Delta Residual=0.81719
MOS Value of Experiments
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
Predeicted MOS Value By the Regression Tree


accent=4 factor::Slope,F0zs cor=0.8394253
MOS Value of Experiments
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
0.5
Predeicted MOS Value By the Regression Tree

3 *weight* の違いによる韻律の比較

weight については式 (4.3)(p.22) 参照。下線部の韻律の変化に注目。

サンプル文：”ある日 おばあさんが 川で 洗濯を していると 川上から 大きな桃が ドンブラコ ドンブラコと 流れてきました”

”a'ruhi 1 oba'asaNga 3 kawa'de 2 seNtakuo 1 shIteiruto 5 kawakamikara 1 o'okina momoga 2 do'Nburako 5 do'Nburakoto 1 naga'rete kimashIta 5”

5 段階、最大変更率 30%

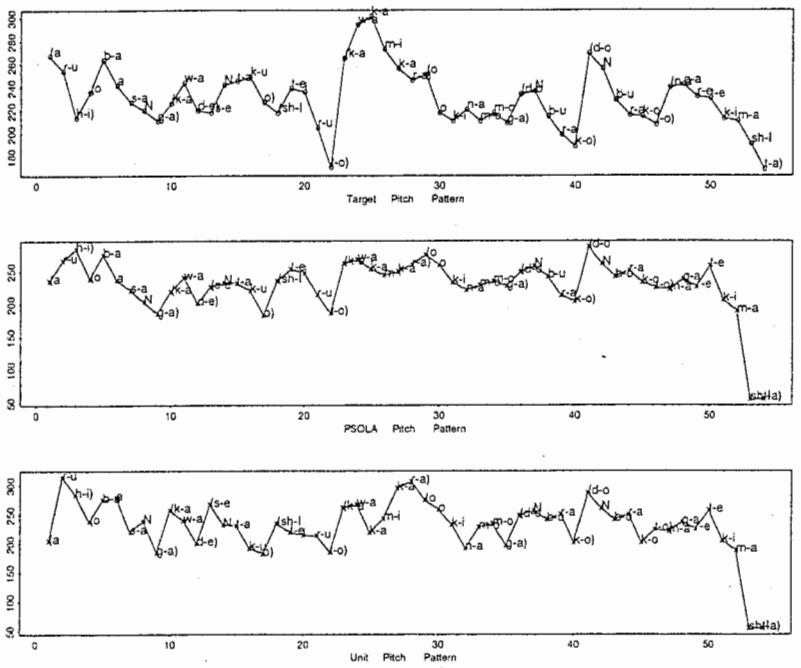


図 B.3: *weight*=1

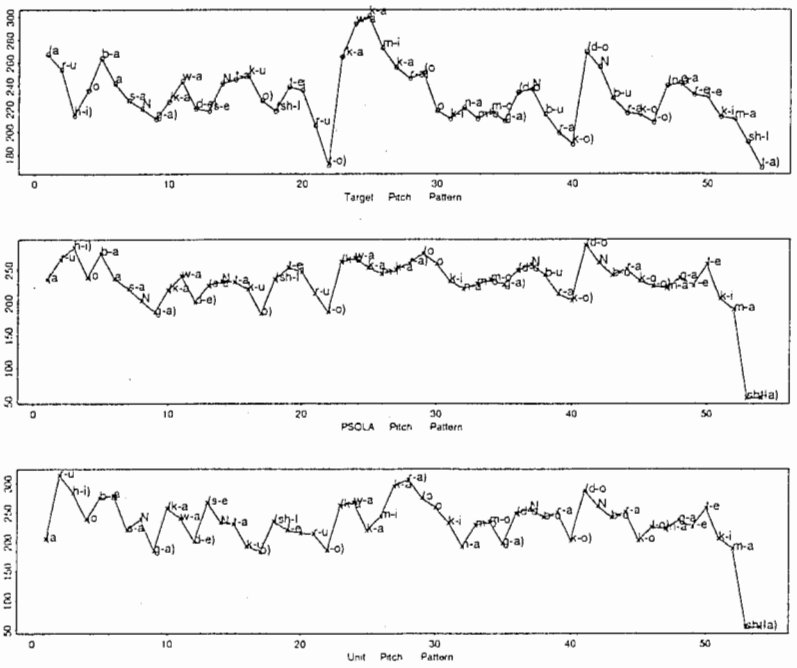


図 B.4: *weight*=0.1

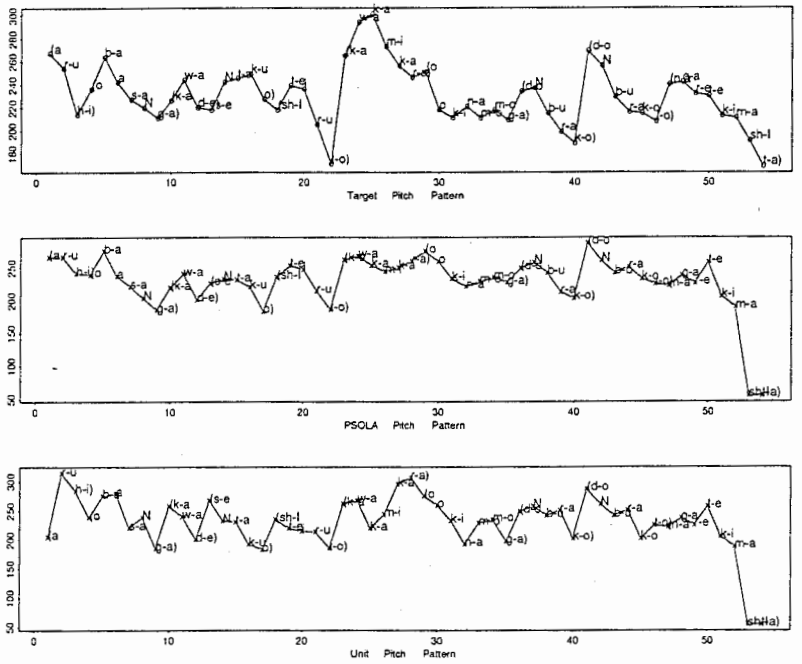


図 B.5: $weight$=0.01

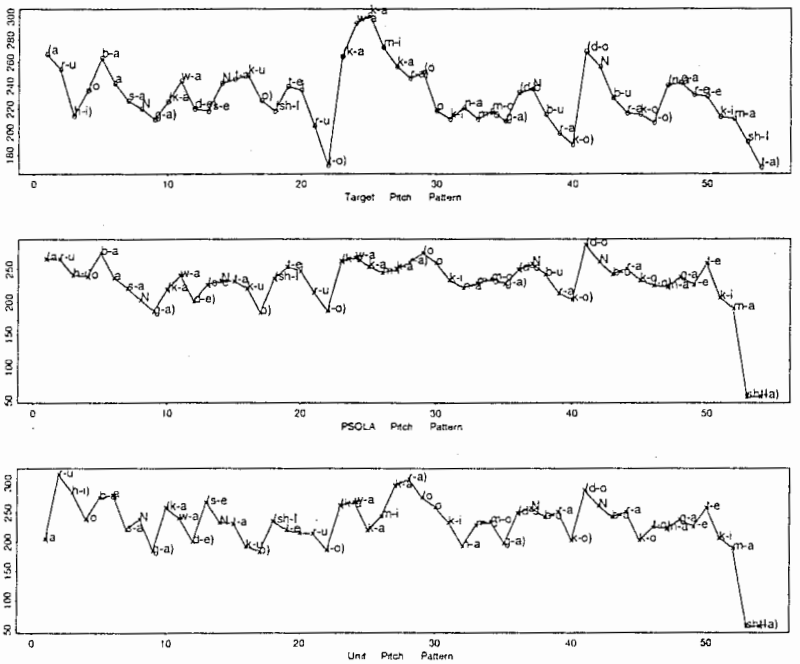


図 B.6: $weight$=0.0001

7 段階、最大変更率 20%

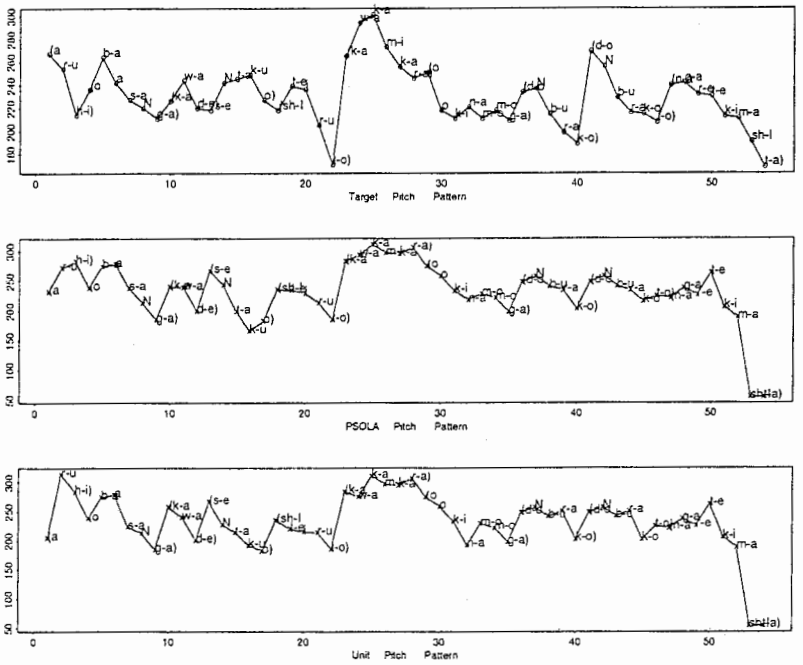


図 B.7: *weight*=1

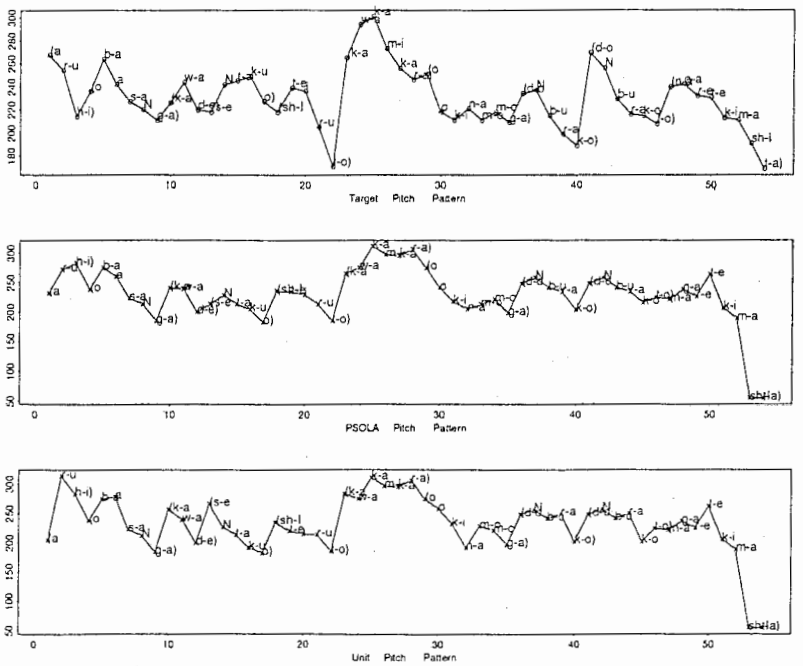


図 B.8: *weight*=0.1

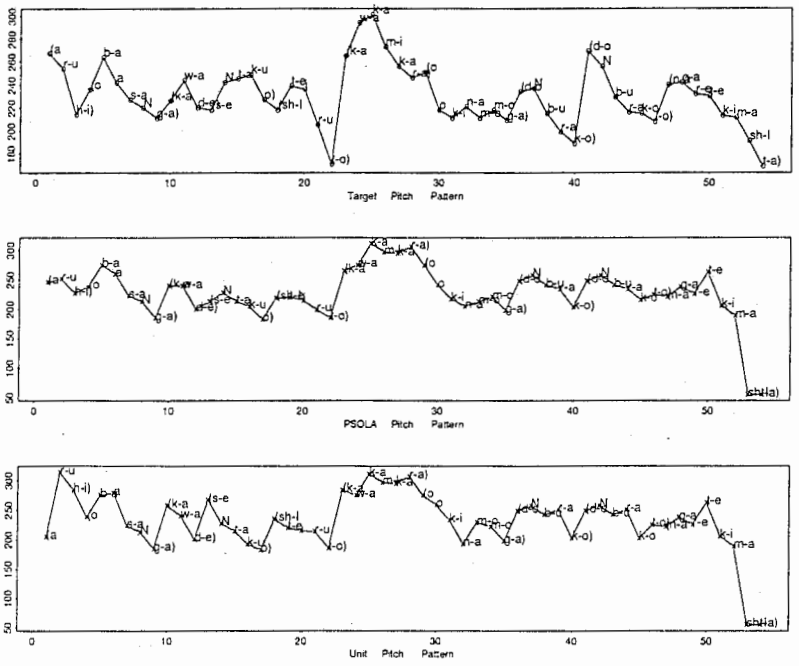


図 B.9: $weight$=0.01

サンプル文："ワールドシリーズを 見たいので、 その近くの
都合の いいところが いいんですけれども。

"(waarudoshiri'izuo 1 mita'inode 1 sono chIka'kuno 1 tsugoono 1 i'i tokoroga 1 i'iNdesUkeredomo 5"

7 段階、最大変更率 20%

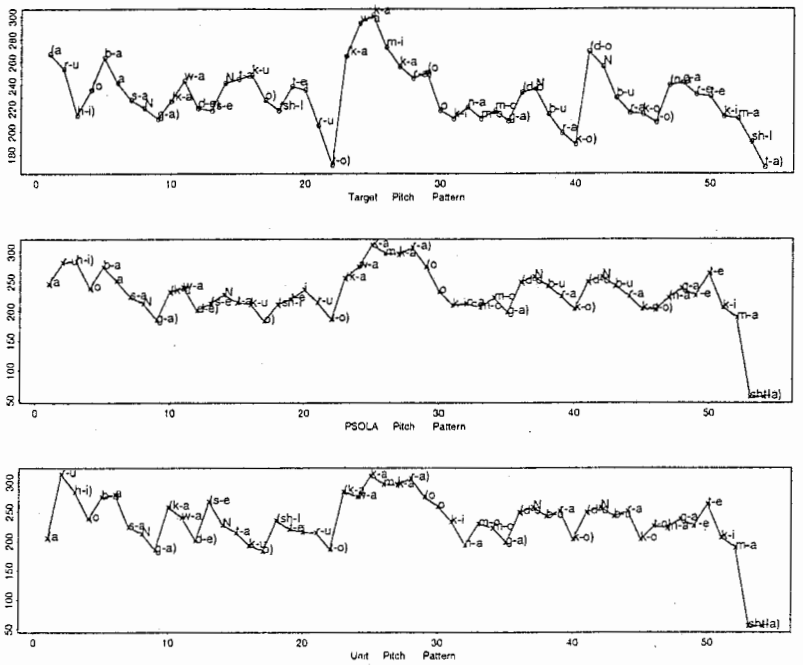


図 B.10: *weight*=1

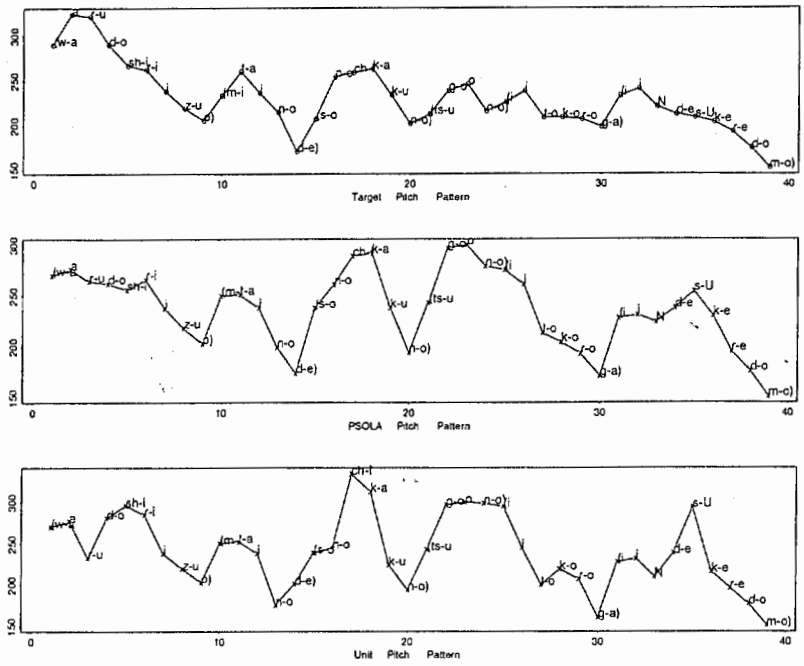


図 B.11: $weight$=0.1

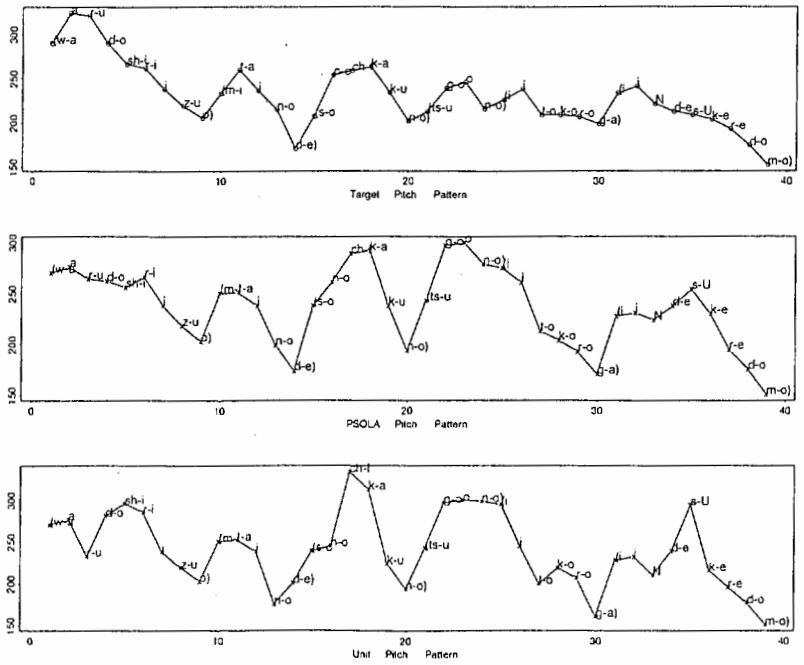


図 B.12: $weight$=0.01