

TR-IT-0264

CHATR におけるポーズと F_0 下降の関係 を利用した音素選択

森本 康夫 藤澤 謙 Nick Campbell

1998.5.29

概要

自然音声波形接続型任意音声合成システムである CHATR の合成音は、音素の接続性は安定しつつあるがイントネーションの乱れが大きい。この対策としてポーズと F_0 下降の現象を利用することで、ポーズを一種の F_0 下降指令とみなし音素選択時の音素列に組み込むことで、全体の音素の接続性を低下させずに部分的に F_0 波形を矯正できるか検証した。

©ATR Interpreting Telecommunications
Research Laboratory.

©ATR 音声翻訳通信研究所

もくじ

1 はじめに	1
1.1 音声合成システム CHATR の概要と現状	1
1.2 目的	3
2 ポーズと F_0 下降の関係について	4
2.1 ポーズと F_0 下降	4
2.2 ポーズ記号による音素選択の制約	4
3 アクセント句境界でのポーズ記号制約実験	6
3.1 目的	6
3.2 実験	6
3.2.1 条件	6
3.2.2 方法	7
3.3 結果	8
3.4 考察	11
3.5 課題	11
4 韻律句境界 (F_0 下降位置) でのポーズ記号制約実験	12
4.1 目的	12
4.2 実験	12
4.2.1 条件	12
4.2.2 方法	13
4.3 結果	14

4.4	考察	16
4.5	課題	16
5	韻律不適切部分でのポーズ記号制約実験	17
5.1	目的	17
5.2	実験	17
5.2.1	条件	17
5.2.2	方法	17
5.3	結果	19
5.4	考察	22
5.4.1	修正場所がアクセント句境界の場合について	22
5.4.2	総合評価が悪化した合成音声について	22
5.4.3	修正場所がアクセント句境界以外の場合について	22
6	まとめ	24
6.1	結果のまとめ	24
6.2	今後の課題	24
	謝 辞	25
	参考文献	26
A	研究資料の格納場所	27
A.1	第3章関係	27
A.2	第4章関係	27
A.3	第5章関係	28

第 1 章

はじめに

昨今では、音声に関するシステムへの関心が高まりつつある。そういった中で、音声合成とはどのような役割を担うことができるのであろうか。

発声させたい内容やその文章量が増大してくると、単純な録音再生方式では賄いきれないのは、明白であり、何らかの情報変換や信号処理を行いたい場合も、その要求を満たすことは難しい。

音声合成システムを用いた場合では、多少の音声品質の低下を代償とする代わりに、大規模かつ多種多様な発話内容に対して、音声再生コストを引き下げることができ、またその音声に対して付加価値を付随できるなど、単なる音声再生システムとは、一線を画する面も持つ。

例えば、音声翻訳の音声出力では、翻訳後も翻訳前と同一人物の音声で発声させたり、音声再生時に聞き取りやすい音声に微細に調節したり、声質の変更も容易に可能である。

1.1 音声合成システム CHATR の概要と現状

音声合成システムである CHATR では、予め録音された音声をデータベースとして、音声合成時にデータベースから最適な音素を選びだし、それを滑らかに接続して音声を合成するという手法を取っている。この方式を自然音声波形接続型任意音声合成システムと呼ぶ。[1]

このシステムでの利点は、実際に人が発声した音声波形をそのまま加工せず用いるので、信号処理のものよりも自然な音声を合成できることである。また、データベースを変えることで、多言語に対応することができ、話者の特徴をある程度再現することも可能である。

欠点は合成音の品質が、データベースによって左右されるということである。データベースは、前もって録音し、音声合成時に音素を選びだすために必要なラベル付けをしておかねばならない。また、録音時間が1時間を下回ると、合成音声に不自然さが目立つようになる。

現段階では、ラベル付けに関しては、ある程度自動化されており、録音時間については、発声内容を吟味することで、短縮することが見込まれる。

現在、CHATR の日本語合成で問題となっているのは、イントネーションである。現状では、音素選択の最適解の導出は、図 1.1 のようになっている。

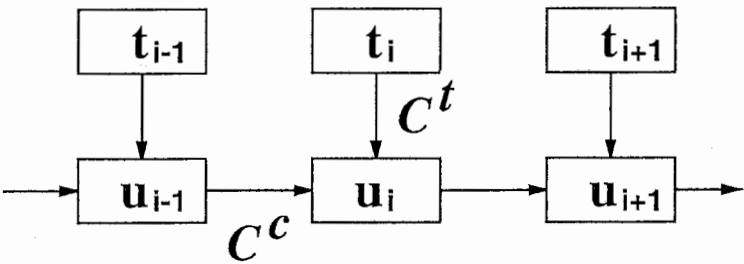


図 1.1: CHATR の音素選択

ただし、 t_i は i 番目のターゲット音素、 u_i は i 番目のユニット音素、 C^c は音素の接続コスト、 C^t は音素のターゲットコストである。

ターゲットコストと接続コストを合算し、そのうちで最適音素候補を幾つか選択しておく。次にそれらの候補について、実際の発話音素系列のパスを計算し、最適パスの候補を幾つか保持して、次の音素選択計算に移る。全ての発話音素系列の計算が終了した後、最適パスを選択する。この様子を図 1.2 に示す。「赤」という音素列のパスを選択する場合について示している。

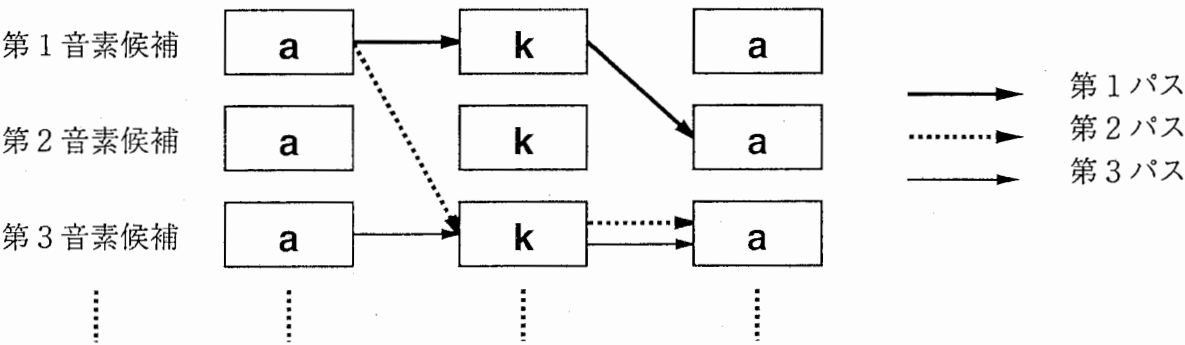


図 1.2: CHATR のパス選択

接続コストが重視されると、滑らかな合成音が得られるが、イントネーションが不自然になることが多い。

一方、音素のつながりが悪いと、聞きづらい合成音となり、機械的で不自然な音声になってしまう

う。

ターゲットコストと接続コストの合算値から選択された音素の候補のうち、第一候補以外の音素であっても、イントネーション的にはよりよい物であることもある。

1.2 目的

現在の CHATR での主な問題点は、韻律の乱れが目立つことである。特に、イントネーションが有らぬ部分で上昇、または下降することで、不自然な音声となっている。

現状では、音素自体のコストより、接続のコストを重視しているため、発話は非常に滑らかではあるが、この品質を大きく低下させずに、少しでもイントネーションを向上させることがねらいである。

第 2 章

ポーズと F_0 下降の関係について

2.1 ポーズと F_0 下降

人間の発声では、無制限に F_0 が上下するわけではなく、一定の周波数の範囲内で変化している。

発声時においては、徐々に呼気が減少するなどして、発声が一旦沈静化して、また再び始めの状態から繰り返される傾向が支配的である。このような現象は、藤崎モデルによって、2 次の線形微分方程式で表すことができる。この沈静化については、息継ぎのように、構造的なものからくるとされる面もあるが、発声者が主としてポーズを挿入することで、発声をあえて沈静化させ意味を持たせる場合が中心的役割となっている。[2][3]

この場合、同時に F_0 概形も下降することがほとんどである。このポーズ位置の前で、 F_0 概形が下降するという現象を音声合成の際にも取り入れる必要がある。そうすることにより、 F_0 の概形を自然なものにすることができる。

ただし、 F_0 の場合は、ポーズ位置周辺以外で F_0 が下降する場合もあり、また F_0 概形の下降の仕方も一通りではない。

2.2 ポーズ記号による音素選択の制約

CHATR では、音素の選択時には、選択する音素の前 2 音素と後 2 音素が一致するような音素を優先的に選択する。このように前と後ろに音素列としての制約を掛けている。前後の重みの比は、現状で前から順に 1:8:4:1 であり、1 つ前の音素が最重要視されている。

ポーズ自体は音素ではないが、CHATR ではポーズを便宜上音素列として扱っている。このポーズ記号は、本来は音素以外の呼気や無音区間を表すために用いられている。ここでは、この呼気や無音を表す役割とは別に、実際にポーズの直前に発声されるような音素を選択するための音素列の制約として利用する。よって、ここで言うポーズ記号制約では実際のポーズ区間は必要ない。本報告では時間的制約により、一旦ポーズ記号を用いて音声を合成した後、人手によってポーズ

自体を編集削除して音声を作成した。今後は音素列の制約に用いるポーズ記号をポーズ記号制約と呼んで区別する。

第 3 章

アクセント句境界でのポーズ記号制約実験

一般的にポーズは発話にいくつかあるアクセント句境界のどこかに挿入される場合が多いということが知られている。またこの場合 F_0 が下降することが多い。実際に音声合成する場合は、このような現象を取り入れることで、合成音の自然性が増す。

そこで、予め句構造を参考にポーズ記号の挿入による音素選択制約を導入して音声合成を行った。これにより合成音声の韻律が矯正されるかどうかを調べた。

3.1 目的

CHATR の音素選択で、実際にアクセント句境界でポーズ記号制約によって F_0 下降音素が選択されるか検証する。

3.2 実験

3.2.1 条件

CHATR の test_seg を利用した音声合成では、韻律予測を行わずに、データベース中の音声パラメータをそのまま利用する。そのとき元となるデータベース中の音声ファイルを排除して、合成を行う。test_seg によって韻律予測の良し悪しを考慮する必要は無くなり、音素選択だけを考慮することができる。

既にラベル付けされている、構文解析によって得られたアクセント句境界に基づいて全てのアクセント句毎にポーズによる制約を加えて、音声合成した。

実験で合成に使用した文章は 5 文である。CHATR のバージョンは、0.94alpha(7 May 1998)、データベースは ATR503 文、比較対象は test_seg の合成音声、話者は FTK とした。

合成音声の F_0 波形と test_seg による合成音声の F_0 波形、データベース音声の F_0 波形を比較し

た。

F_0 値の計算は、対象とする音素の継続時間に対する F_0 値の平均である。

3.2.2 方法

/dept2/work25/pi/data/chatr_dbs/FTK/chatr/seg/FTK_SD_A*.seg.ch ファイルを元に、nus_params ヘッダーに

```
(set excludes (list 'exclude_list "FTK_SD_A*" ))
```

```
(set nus_params (cons excludes nus_params))
```

を加えたファイルを作成し、(* は文章番号で、01 から 05 とする。)

```
chatr>(speaker_FTK)
```

```
chatr>(Say(Synth(load "FTK_SD_A*.seg")))
```

として、音声を作成した。ただし、「chatr>」は CHATR のプロンプトである。

この際、ポーズ記号制約位置に

```
( # 1 2.740000 ((150 0)))
```

をダミーポーズとして挿入した。

音声合成の後、

```
chatr>(Save Wave "FTK_SD_A*")
```

として合成波形を保存し、xwaves+ を用いてポーズ記号制約によって挿入された実際のポーズを手作業により削除した。この時、編集後の音声ができるだけ滑らかになるように注意した。

作成した音声は

/home/as52/xymorimo/Work/First/EDIT/WAV/

に ESPTS 形式で保存してある。

3.3 結果

ポーズ記号の挿入によって、ポーズ記号位置の前の1音素の F_0 値とその音素を含めた F_0 の概形が test_seg による合成の同音素と比較してどう変化したか表 3.1、表 3.2に示す。

表 3.1: ポーズ記号制約による F_0 の変化

	F_0 値減少	変化無し	F_0 値増加	不明	全体
ポーズ記号挿入後	17(68%)	1(4%)	5(20%)	2(8%)	25(100%)

ポーズ記号挿入後 F_0 の値に変化の無い1例を調べるとポーズ記号挿入位置は文末であった。

不明の2例は、ともに文末で母音が無声化している場合である。

ポーズ記号挿入後 F_0 の値が上昇した5例について調べると、そのうち4例がポーズ記号挿入位置の2つ前の音素と1つ前の音素の F_0 下降差が増大していた。つまり test_seg の時と比較すると、ポーズ記号を挿入することで F_0 概形の傾斜が下降方向に大きくなった。

表 3.2: ポーズ記号制約による F_0 概形の変化

	F_0 概形傾斜下降	変化無し	F_0 概形傾斜上昇	不明	全体
ポーズ記号挿入後	21(84%)	1(4%)	1(4%)	2(8%)	25(100%)

1例のみ、test_seg と比較して、ポーズ記号挿入位置前の音素の F_0 値と前の数音素を含めた F_0 概形の傾斜の両方とも上昇したものがあった。

次に F_0 概形の例を図 3.1、図 3.2に示す。

図 3.1の発話内容は、

「あらゆる現実をすべて自分の方へねじ曲げたのだ。」

図 3.2の発話内容は、

「物価の変動を考慮して給付水準を決める必要がある。」

であり、全てのアクセント句境界にポーズ記号制約を行っている。

図 3.1では、test_seg の合成音声に比べてポーズ記号制約をした合成音声は「方へ」の部分で全体的に F_0 概形自体は少し上昇しているが、ポーズ記号の前の傾斜は下向きに大きくなっており、 F_0 概形が下降調となった。

図 3.2では、test_seg の合成音声に比べてポーズ記号制約をした合成音声はアクセント句毎に F_0 概形が下降調に変化している。

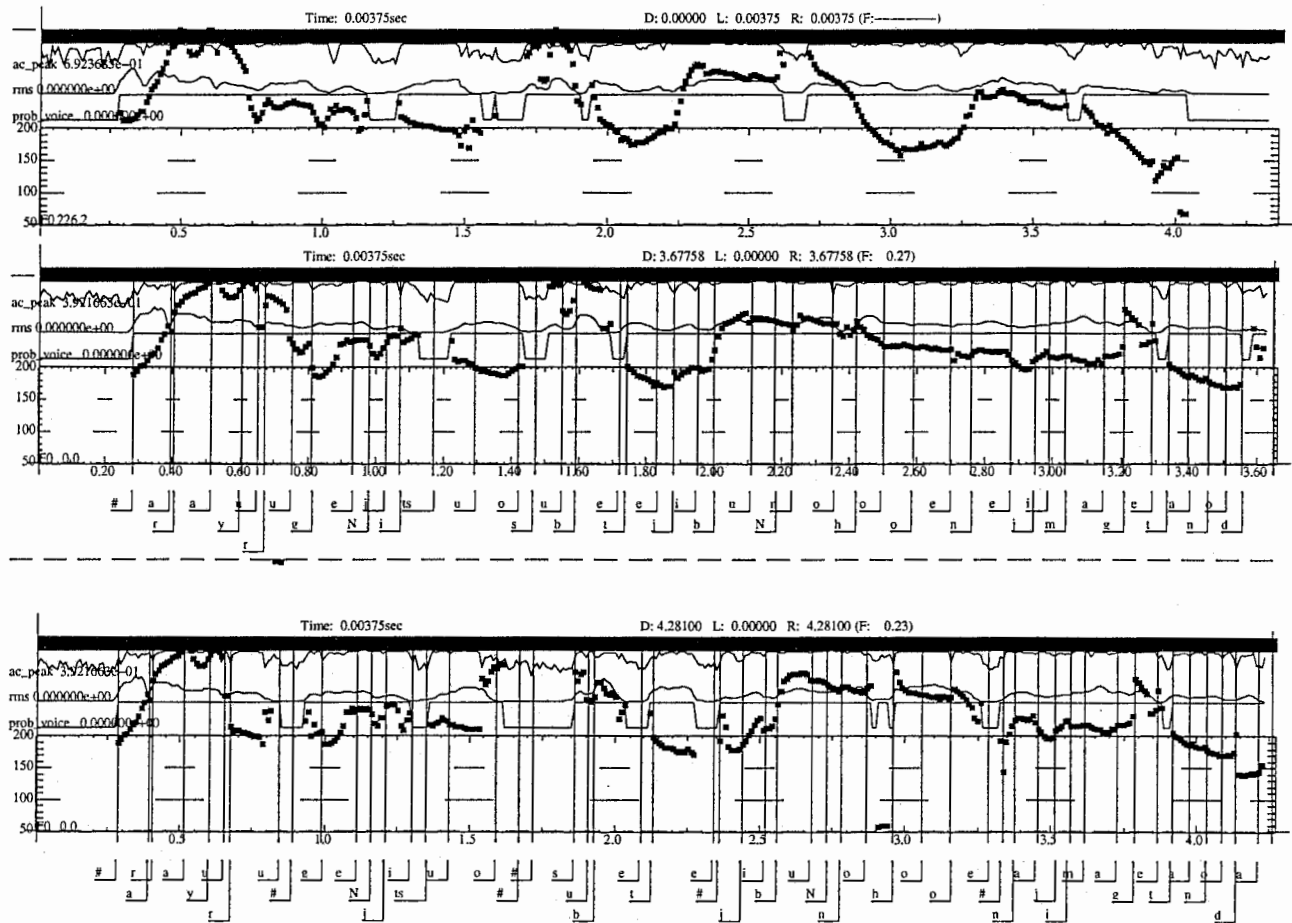


図 3.1: FTK_SD_A01 の F_0 上: データベース 中: test_seg 下: ポーズ記号制約

「あらゆる現実をすべて自分の方へねじ曲げたのだ。」

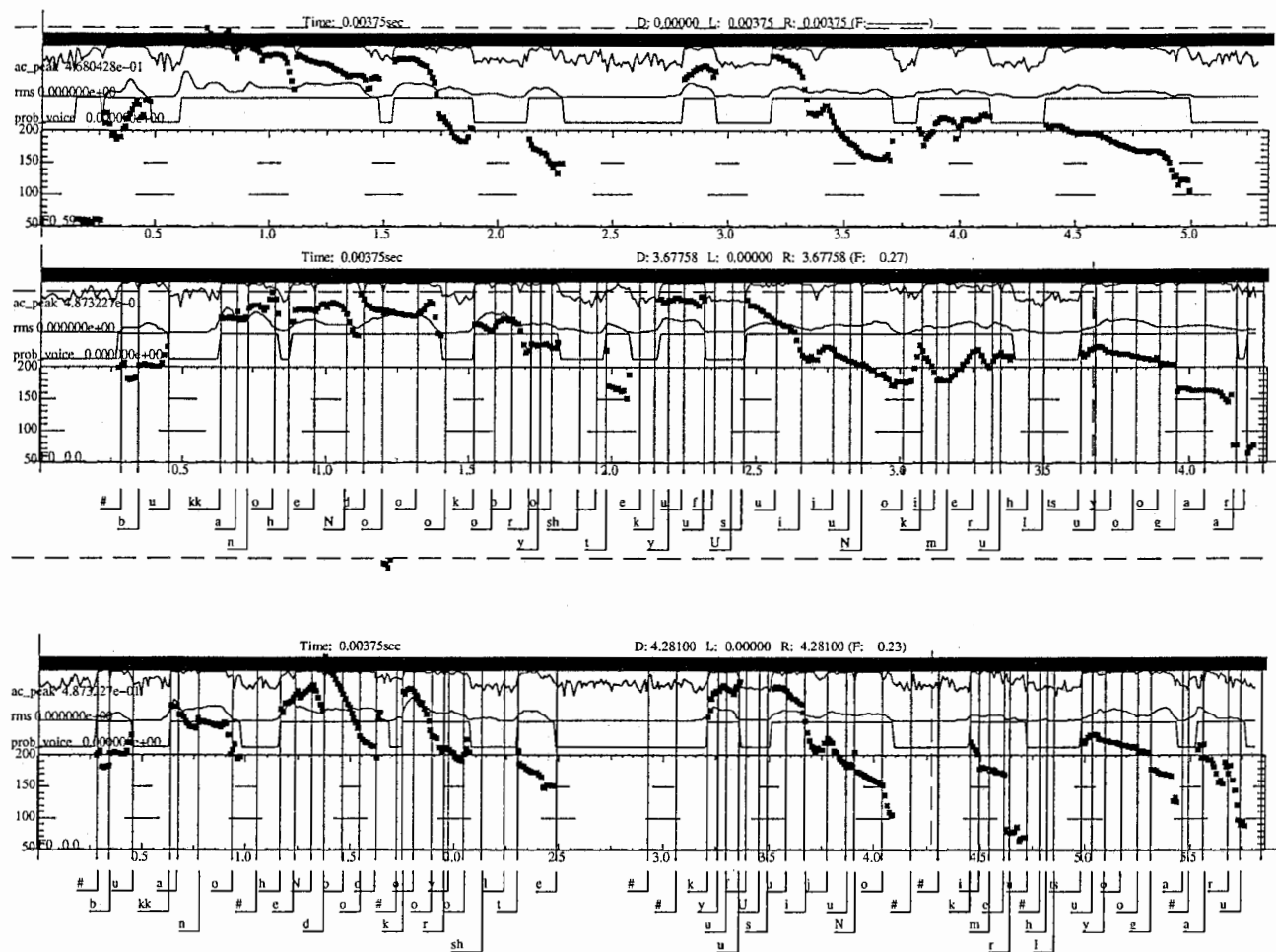


図 3.2: FTK_SD_A04 の F_0 上: データベース 中: test_seg 下: ポーズ記号制約

「物価の変動を考慮して給付水準を決める必要がある。」

3.4 考察

今回の結果から、アクセント句境界でポーズ記号制約を行うと、ほとんどの場合で F_0 の概形が下降調となることが確認できる。これは自然音声では一般的とされるポーズ位置での F_0 下降を反映していると思われる。

CHATR では自然音声波形をデータベースとし、その音声を利用しているため、当然データベース中には発話の一般的特徴、さらに個人的特徴が含まれている。

一般的特徴の場合は、例えば今回の様にポーズの前では F_0 概形が下降するといった、発話の特徴を引き出すことが可能と思われる。

3.5 課題

合成音声の自然性向上のためには、発話の特徴を引き出すだけではなく、どのアクセント境界にポーズ記号制約を適用するかという課題がある。これについては、次の F_0 下降位置でのポーズ記号制約で検討した。

第 4 章

韻律句境界 (F_0 下降位置) でのポーズ記号制約実験

自然音声でポーズ位置の F_0 概形が下降することが多いということは、一般的な事象となっている。そこで、立場を変えて F_0 概形の下降部分は、どのような位置となることが多いかを調査した。これにより、ポーズ位置中心ではなく、 F_0 下降位置を中心にポーズ記号制約位置を決定して制約を適用できる可能性があり、イントネーションの矯正という目的に、より合致する。

4.1 目的

データベースの自然音声から、 F_0 値が下降する部分を検出し、その検出位置がアクセント句境界に該当するか検証する。

4.2 実験

4.2.1 条件

CHATR の `test_seg` を利用した音声合成では、韻律予測を行わずに、データベース中の音声パラメータを目的とする音素の韻律目標値とする。そのうちの F_0 の目標値を用いて F_0 の下降位置を検出する。

F_0 の下降位置を次の条件を満たす場所とした。

- 母音と撥音が連続して 3 回以上かつ 10 ヘルツ以上の下降があるとき
- 母音と撥音が連続して 3 回以上かつ 20 ヘルツ以上の下降があるとき

検出は、音素の並びのうち、有声母音と撥音だけを音素列の順に取り上げて、1 つ前の有声母音もしくは撥音と F_0 値を比較する。 F_0 値が既定値より減少している場合が連続 3 回以上続いたとき、この条件を満たす最後の有声母音もしくは撥音の後を F_0 下降位置とする。

実験で使った文章は、FTK_SD_A01 から 50 までの 50 文である。CHATR のバージョンは、0.94alpha(7 May 1998)、データベースは ATR503 文、比較対象は test_seg の合成音声、話者は FTK とした。

4.2.2 方法

CHATR を起動し、

```
chatr>(speaker_FTK)
```

```
chatr>(Say(test_seg "FTK_SD_A*"))
```

```
chatr>(Save UdbselectInfo "FTK_SD_A*")
```

として、sel ファイルを作成する。sel ファイルには、データベース自然音声から抽出された F_0 の値が書き込まれている。通常の test_seg を用いない音声合成では、韻律予測した F_0 値となる。

この sel ファイルを F_0 下降検出プログラムで処理する。

F_0 下降検出プログラムは awk で書かれており、2 つの処理からなる。このプログラムは

```
/home/as52/xymorimo/Work/Second/Prog/f0_delta.awk
```

```
/home/as52/xymorimo/Work/Second/Prog/f0_down.awk
```

として保存してある。

f0_delta.awk は前の音素との F_0 値との差分をとり、f0_down.awk は F_0 の下降回数をチェックしている。

利用法は、

```
/home/as52/xymorimo/Work/Second/Prog/f0_check.csh
```

を起動すれば良い。ただしディレクトリ名などは適時変更する必要がある。

4.3 結果

データベースの自然音声から抽出された F_0 値の下降部分の検出位置の結果を表 4.1 に示す。

アクセント句境界の位置は、データベース中の構文構造のデータを利用した。

1 モーラずれは、アクセント句境界から前後 1 モーラの位置を検出したものである。

文末無声化母音は文末の母音が無声化したために、1 つ前の母音または撥音が検出位置となった場合である。

表 4.1: 自然音声の F_0 下降検出位置の分類結果

	アクセント句境界	1 モーラずれ	文末無声化母音	その他の位置	全体
10Hz 下降	75(71%)	24(23%)	4(4%)	2(2%)	105(100%)
20Hz 下降	38(75%)	11(22%)	0(0%)	2(4%)	51(100%)

表 4.1 の 1 モーラずれの内訳を表 4.2 に示す。

判別不能は F_0 下降検出位置の前後 1 モーラの位置両方にアクセント句境界がある場合である。つまり 2 モーラのアクセント句の真中を境界として検出した。

表 4.2: 自然音声の F_0 下降検出位置の分類結果

	前	後	判別不能	全体
10Hz 下降	17(71%)	6(25%)	1(4%)	24(100%)
20Hz 下降	9(82%)	2(18%)	0(0%)	11(100%)

データベース自然音声の F_0 波形の例を図 4.1、図 4.2、図 4.3 に示す。

発話内容は、順に

「叔父さんは岬の一軒家に、独りぼっちで住んでいた。」

「普通、中距離トラックのドライバーは中年の人が多い。」

「十進法は両手の十本の指を数えることから起こった。」

である。

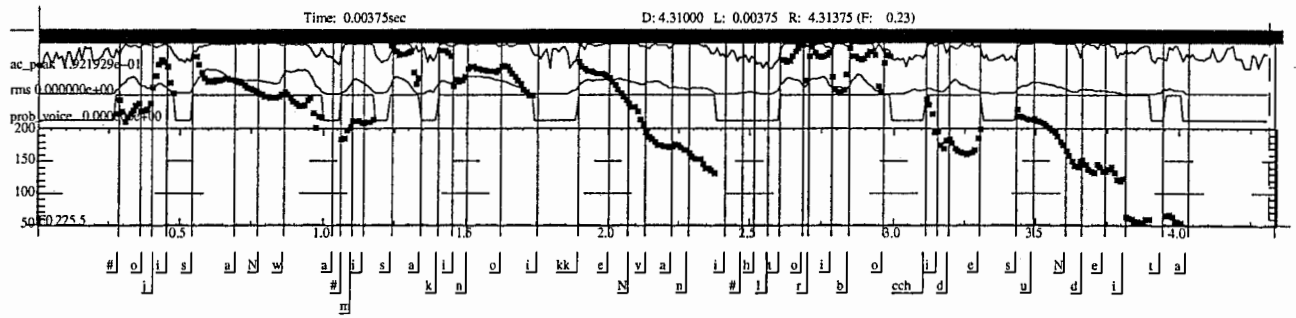


図 4.1: データベース音声 FTK_SD_A18 の F_0

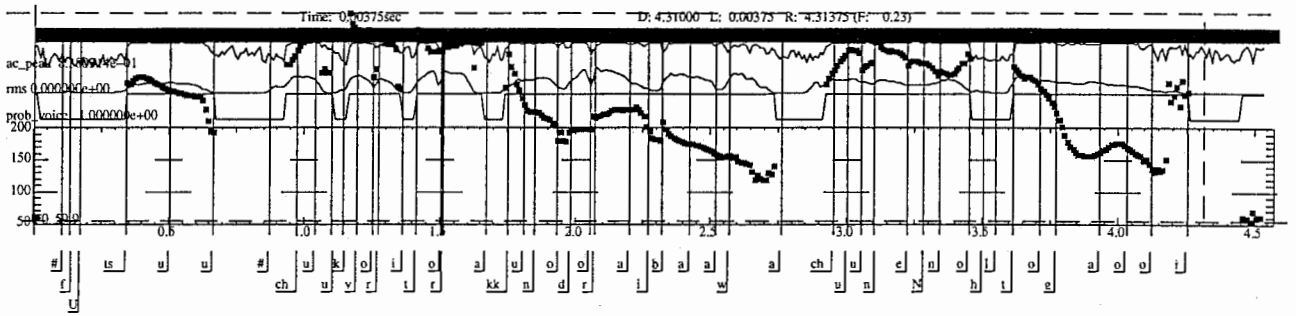


図 4.2: データベース音声 FTK_SD_A38 の F_0

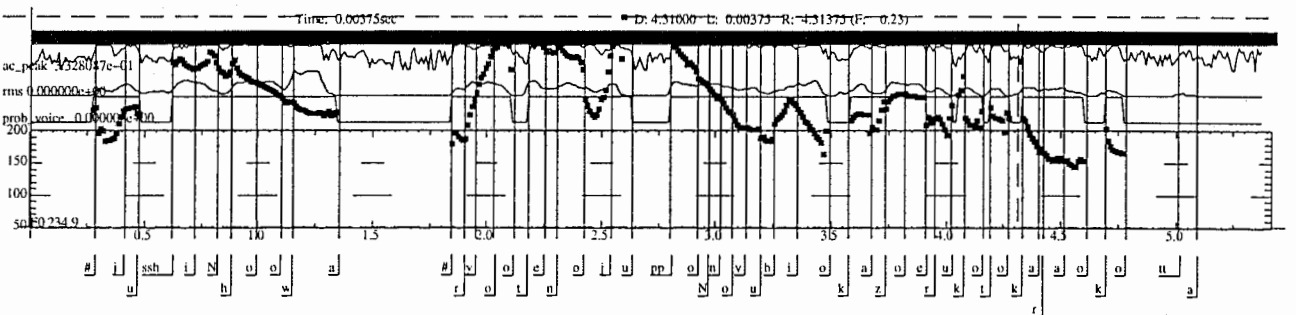


図 4.3: データベース音声 FTK_SD_A44 の F_0

4.4 考察

F_0 下降位置として検出された個所は、アクセント句境界から1モーラずれを含めるならば、ほとんどの検出位置が、文中のアクセント句境界に位置することになる。複数あるアクセント境界のうち、どこにポーズ記号制約を行うかは、音声合成時のターゲットの F_0 下降位置を調べることで決定することができる。

1モーラずれを調べると、 F_0 波形で見る限りでは、前にずれたものは2種類に分けられ、後ろにずれたものは、全て同種とみなせる。

1. 前に1モーラずれた場合で周辺の F_0 の下降概形が二段階になっている。(図4.1「一軒家に」の末部参照)
2. 前に1モーラずれた場合で、 F_0 の概形がV字型になっている。(図4.2「ドライバーは」の末部参照)
3. 後ろに1モーラずれた場合で次のアクセント句の始めが含まれる。(図4.3「十本のゆびを」の「のゆ」部分参照)

このうち2、3については、検出結果が境界の後の音素の F_0 に影響を受けていると推察できる。滑らかな発声では、単語毎やアクセント句毎に F_0 が断続せず、連続的に F_0 が変化している。そういったことを考慮すれば、自然に発声した場合、境界位置の前後1モーラを調節して F_0 を滑らかに変化させることは自然な事といえる。

4.5 課題

境界位置以外でのポーズ記号制約は、音声の接続性を低下させる可能性が高い。よって、 F_0 下降位置に一番近いアクセント句境界を選択するというようなポーズ記号制約位置決定をする必要がある。

第 5 章

韻律不適切部分でのポーズ記号制約実験

音素選択の際、アクセント句境界ではポーズ記号制約によって F_0 が低い音素が選択される傾向が、3 章から明らかになった。このことを利用して、アクセント句境界以外の位置でも、合成音声の韻律不適切部分が修正できるか調査した。

5.1 目的

test_seg の合成音声のイントネーション不適部分に任意にポーズ記号制約を施し、イントネーションと音声のつながりを考慮して、境界位置を問わずいかなる位置、場合であれば、CHATR におけるポーズ記号制約が有効か検証する。

5.2 実験

今回作成した合成音声を test_seg の合成音声と一対比較により、聴取実験を行った。

5.2.1 条件

聴取者は 8 名で、内訳はラベラー 4 名、研究員 4 名である。使用した合成音声は、FTK_SD_A01 から 20 までの 20 文でそのうち 4 文は test_seg の合成音声に変更を加えていない。CHATR のバージョンは、0.94alpha(7 May 1998)、データベースは ATR503 文、比較対象は test_seg の合成音声、話者は FTK とした。

5.2.2 方法

文献 TR-IT-0254[4] の聴取結果からその文章中で、不自然だと感じた聴取者が最も多い部分をおおよそ 1 文で 1 箇所選び、手作業でポーズ記号制約位置を決定した。このとき制約位置は 1 モー

ラずつずらしてイントネーションが良くなったと思われる部分とした。

/dept2/work25/pi/data/chatr_dbs/FTK/chatr/seg/FTK_SD_A*.seg.ch

ファイルを元に、nus_params ヘッダーに

```
(set excludes (list 'exclude_list "FTK_SD_A*" ))
```

```
(set nus_params (cons excludes nus_params))
```

を加えたファイルを作成し、(* は文章番号で、01 から 20 とする。)

```
chatr>(speaker_FTK)
```

```
chatr>(Say(Synth(load "FTK_SD_A*.seg")))
```

として、音声を作成した。ただし、「chatr>」は CHATR のプロンプトである。

この際、ポーズ記号制約位置に

```
( # 1 2.740000 ((150 0)))
```

をダミーポーズとして挿入した。

音声合成の後、

```
chatr>(Save Wave FTK_SD_A*)
```

として合成波形を保存し、xwaves+ を用いてポーズ記号制約によって挿入された実際のポーズを手作業により削除した。この時、編集後の音声ができるだけ滑らかになるように注意した。

作成した音声は

/home/as52/xymorimo/Work/Third/WAV/EDIT

に ESPS 形式で保存してある。

5.3 結果

聴取実験の評価は次のようにした。

合成音声に対してイントネーション、音素の接続性、総合評価の3つの視点から品質を判断する。

聴取実験の集計値の最頻値をその合成音声の評価とする。最頻値が同数の場合は、以下のように取り扱った。

- 良い=悪い ⇒ 同等
- 良い=同等 ⇒ 良い
- 悪い=同等 ⇒ 悪い

全体の結果を表5.1に示す。この実験ではイントネーションの矯正に主眼がおかれているため、表5.1の結果をみると、確かにイントネーションでは、おおむね test_seg の合成音声よりポーズ記号制約を用いた合成音声の方が結果が良くなっている。しかしポーズ記号制約を用いると音素の接続性が低下することが読み取れる。

表 5.1: 聴取実験におけるポーズ記号制約を行った合成音声の一対比較評価

	良い	悪い	同等
イントネーション	7	2	7
接続性	2	7	7
総合	6	5	5

各合成音声に対するポーズ記号制約を行った位置とその影響を表5.2に示す。番号抜けになっているものは、ポーズ記号制約を適用しなかったもので、いかなる変更も加えていない合成音声である。

修正場所の句末はアクセント句境界を表す。また、合成音声 No.8 は単語境界であるが、便宜上アクセント句末として処理した。単語境界は品詞の境界位置とする。語中は品詞の境界以外の位置であり、例えば「アフリカ」の「フ」と「リ」の間などである。

F_0 概形の変化は、下降、整形、分断の3つに分類した。下降はポーズ記号制約によってポーズ記号の前の数音素の傾斜が下方方向に大きくなった場合を指す。整形はデータベース自然音声波形の該当部の概形の傾斜に近づいた場合である。この場合は、傾斜の方向は問わない。分断は、ポーズ記号制約の前後の音素の F_0 値が掛け離れて、ポーズ記号制約の前後数音素の傾斜が大きく異なる場合である。ただし、無声音による影響は除外して考える。

表 5.2: ポーズ記号制約を行った位置とその後の F_0 概形の変化

No.	総合評価	修正場所			F_0 概形の変化			音素 N の位置	
		句末	音素 N の後	語中	下降	整形	分断	単語境界	語中
01	同	○			○				
02	良		○		○			○	
03	同		○		○	○		○	
05	悪	○			○		○		
07	同		○			○	○	○	
08	同	○					○		
09	良	○			○	○			
11	悪		○		○				○
12	悪			○	○				
13	同		○			○		○	
14	良	○			○	○			
15	良			○			○		
16	悪			○			○		
18	良	○			○				
19	良			○		○			
20	悪			○	○				

ポーズ記号制約による合成音声の例を図 5.1 に示す。発話内容は

「救急車が十分に動けず救助作業が遅れている。」

である。「十分に」の後がポーズ記号制約位置である。

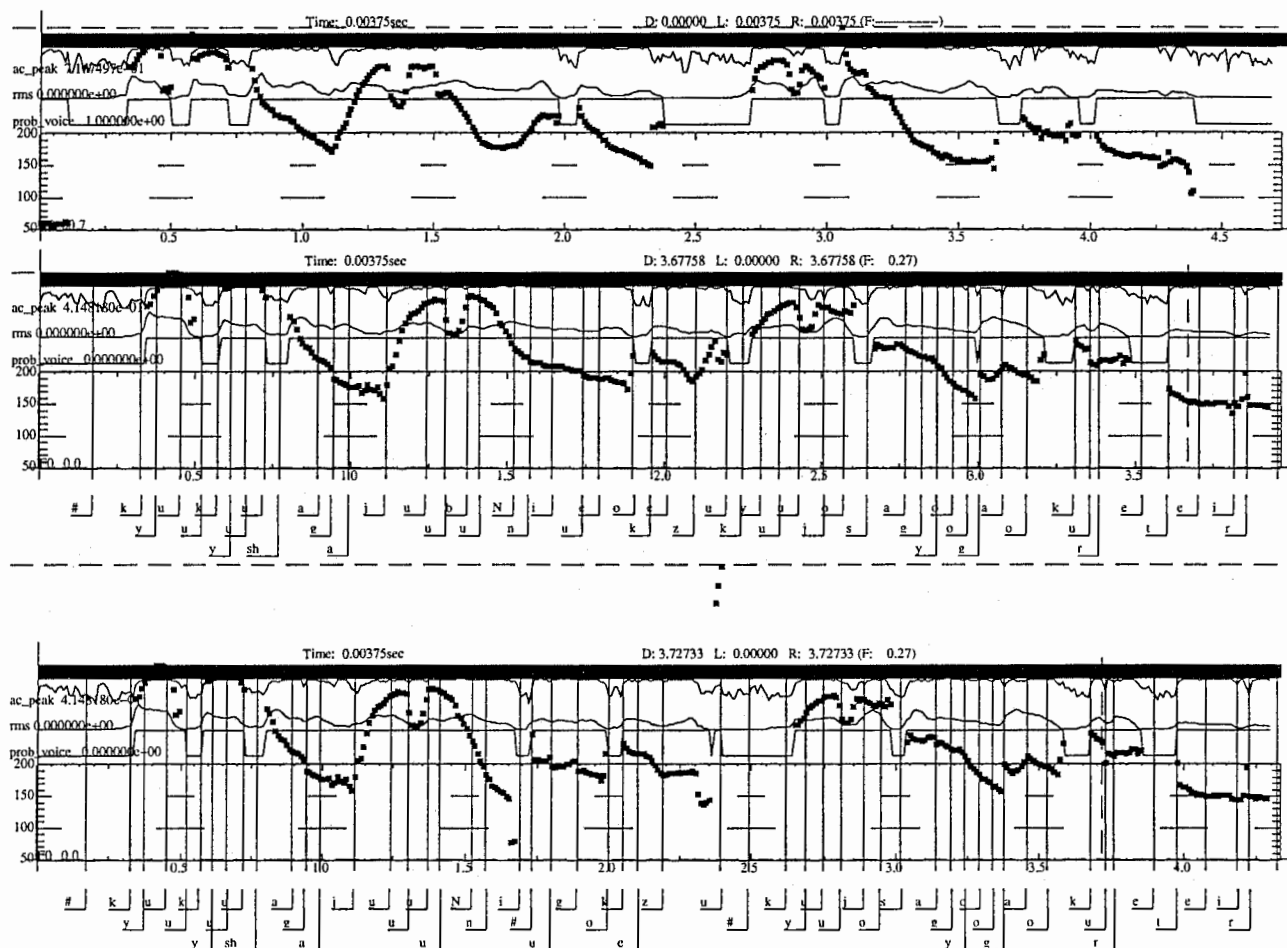


図 5.1: FTK_SD_A05 の F_0 上: データベース 中: test_seg 下: ポーズ記号制約

5.4 考察

5.4.1 修正場所がアクセント句境界の場合について

表 5.2 をみると、No.05 だけは総合評価が悪くなっているがこれについては次の「総合評価が悪化した合成音声について」で述べる。

No.05 を除けば、アクセント句境界でのポーズ記号制約による合成音声は、総合評価が良くなるか、同じである (No.01,08,09,14,18) 。この結果を見る限りでは修正位置がアクセント境界である場合は、ポーズ記号制約を用いても総合評価としては悪くはならないと予想できる。

5.4.2 総合評価が悪化した合成音声について

総合評価の悪くなった 5 例を調べると、修正位置では語中が 4 例 (No.11,12,16,20) 、アクセント句末が 1 例 (No.05) であった。さらに F_0 概形の変化をみると下降型が 3 例 (No.11,12,20) で、1 例 (No.16) が分断型であった。残りの 1 例 (No.05) は下降と分断の両方であった。

語中の F_0 概形下降に注目すると、この様な修正で総合評価の良くなった合成音声は 3 例中 1 例もなく、全て総合評価が悪くなっている。これにより語中の F_0 概形下降型の修正は、合成音声の品質を低下させる傾向があると予想できる。

これは語中の F_0 概形が平板になるべき場合と、 F_0 概形下降となるべき場合がある。また語中の F_0 概形下降は、アクセント句末などの F_0 概形下降とは性質が異なっていることも考えられる。

次に、 F_0 概形分断型の 1 例 (No.16) についてみると語中の F_0 概形分断でありこれは自然音声で考えれば、語中で発声を分断することはあまり無いので聴取者が不自然と感じるのは当然と考えられる。

最後に (No.05) は、アクセント句末での F_0 概形分断型であるが、 F_0 概形をみると、元々のイントネーションの不自然さは、ポーズ記号制約位置の後ろにある (図 5.1 「動けず」部分参照) 。よって、ポーズ記号制約位置の前の F_0 概形は下降型となっているが、後ろのイントネーションの矯正については効果が及ばなかったとみられる。

5.4.3 修正場所がアクセント句境界以外の場合について

全体として考えると、修正場所として音素 N の後が目をはく。さらに 1 例 (No.15) だけ語中の F_0 概形分断型で評価の良くなった合成音声がある。これらについてとりあげる。

まず、修正場所としての音素 N の後について考える。実験結果についてみると、音素 N の後は単語境界に属する限りは悪い評価には繋がっていないが、詳細に音素選択を調べると、ポーズ記号制約が無視されている事がある。

これは、自然音声で音素 N の後にポーズが入るような場合が少ない事と、CHATR では、音素選

択時の制約で指定された音素が無い場合は、条件に合う範囲から音素を選択するためである。

次に、語中の F_0 概形分断型で 1 例 (No.15) だけ、評価がよくなった合成音声がある。これは CHATR の音素選択が、最適パスを選択することに関係があると思われる。

すなわち、ポーズ記号制約を行うことによって、音素列がポーズ記号制約の分だけ増加する。CHATR は、最終的に全体としての音素列の最良パスを選択するので、一部分だけの変化であっても、様々な場所の音素が入れ替わることが有り得る。よってこの影響がこの 1 例に出たのではないかと思われる。

第 6 章

まとめ

6.1 結果のまとめ

人間の発声では、ポーズ位置で F_0 が下降する場合が多いことを利用して、音素選択時にポーズ記号による制約を設けて、合成音声がどのように変化するか調査した。その結果、以下のようなことが判明した。

- アクセント句境界で音素選択にポーズ制約をすると F_0 下降が起こる。
- 自然音声の F_0 下降位置は、ほぼアクセント句境界に位置する。
- アクセント句境界の場合であれば、ポーズ記号制約を用いると接続性は若干悪化するが、総合評価において発話の自然性の向上が見込める。

6.2 今後の課題

音声合成をする際に、あらかじめ自動的にポーズ記号制約を組み込むことで、イントネーションの乱れを部分的に防ぐことができる可能性がある。これを実現するためには、合成音声全体に対するひな形としてポーズ記号制約を検討する必要がある。例えば、始めからおおよそそのポーズ位置や韻律句境界に対応するような F_0 下降位置を決めたテンプレートを用意しておき、発話内容に一番近いテンプレートを利用するといったことが考えられる。

謝 辞

まず始めに、学外実習生として受け入れて下さった山本 誠一 ATR 音声翻訳通信研究所社長に深く感謝致します。

本研究に取り組むにあたり、大変親身に御指導を賜りました Nick Campbell 第二研究室室長、藤澤 謙研究員に厚く御礼を申し上げます。

また、有益な助言を頂きました第二研究室の皆様、聴取実験にご協力下さいました被験者の皆様に厚く御礼を申し上げます。

さらに、論文の解説に快く応じて下さいました匂坂芳典第一研究室室長に深く感謝致します。

ATR 音声翻訳通信研究所でお世話して下さいました関係者の皆様に深く感謝致します。

最後に、学外実習の機会を与えて下さった電気通信大学の樽松 明教授に深く感謝致します。

1998 年 5 月 29 日

森本 康夫

参考文献

- [1] Nick Campbell, Alan W. Black, “CHATR: 自然音声波形接続型任意音声合成システム”, 信学技報, SP96-7, May.1996.
- [2] 海木延佳, 匂坂芳典 “局所的な句構造によるポーズ挿入規則化の検討”, 信学論 (D-II), vol.J79-D-II, no.9, pp.1455-1463, Spet.1996.
- [3] 海木延佳, 匂坂芳典 “局所的句構造に基づく F_0 制御”, 信学技報, SP96-6, May.1992.
- [4] 藤澤 友紀子, 藤澤 謙, Nick Campbell, “CHATR に於ける日本語の韻律評価”, TR-IT0254, Feb.1998.

付録 A

研究資料の格納場所

A.1 第3章関係

- ポーズ記号制約を行った合成音声
/home/as52/xymorimo/Work/First/EDIT/*.wav
- ポーズ記号制約を行った合成音声の F_0 とラベルの eps ファイル
/home/as52/xymorimo/Work/First/EDIT/*.f0.eps
/home/as52/xymorimo/Work/First/EDIT/*.lv.eps
- ポーズ記号制約を行った合成音声の seg ファイル
/home/as52/xymorimo/Work/First/EDIT/*.seg
- test_seg の seg ファイルよりポーズを削除した seg ファイル
/home/as52/xymorimo/Work/First/ORG/*.seg
- test_seg の seg ファイルよりポーズを削除した seg ファイルから音声合成した F_0 とラベルの eps ファイル
/home/as52/xymorimo/Work/First/ORG/*.f0.eps
/home/as52/xymorimo/Work/First/ORG/*.lv.eps

A.2 第4章関係

- F_0 下降位置検出 awk プログラム
/home/as52/xymorimo/Work/Second/Prog/f0_delta.awk
/home/as52/xymorimo/Work/Second/Prog/f0_down.awk

- F_0 下降位置検出実行 c シェルプログラム
/home/as52/xymorimo/Work/Second/Prog/f0_check.csh
- 検出位置でポーズ記号制約を行った合成音声の seg ファイル
/home/as52/xymorimo/Work/Second/10Hz/*.seg
/home/as52/xymorimo/Work/Second/20Hz/*.seg
- 検出位置でポーズ記号制約を行った合成音声の F_0 とラベルの eps ファイル
/home/as52/xymorimo/Work/Second/10Hz/*.f0.eps
/home/as52/xymorimo/Work/Second/10Hz/*.lv.eps
/home/as52/xymorimo/Work/Second/20Hz/*.f0.eps
/home/as52/xymorimo/Work/Second/20Hz/*.lv.eps
- 検出位置でポーズ記号制約を行った合成音声
/home/as52/xymorimo/Work/Second/10Hz/FTK*.wav
- 検出位置でポーズ記号制約を行った後ポーズ削除した合成音声
/home/as52/xymorimo/Work/Second/10Hz/1.wav
/home/as52/xymorimo/Work/Second/10Hz/2.wav
/home/as52/xymorimo/Work/Second/10Hz/3.wav
/home/as52/xymorimo/Work/Second/10Hz/4.wav
/home/as52/xymorimo/Work/Second/10Hz/5.wav
- test_seg の合成音声から抽出した sel ファイル (CHATR 0.93)
/home/as52/xymorimo/Work/Second/Sel/old-ftk*.sel
- test_seg の合成音声から抽出した sel ファイル (CHATR 0.94)
/home/as52/xymorimo/Work/Second/Sel/ftk*.sel

A.3 第 5 章関係

- ポーズ記号制約を行った合成音声
/home/as52/xymorimo/Work/Third/WAV/PLAY/*.wav
- ポーズ記号制約を行った後ポーズを削除した合成音声
/home/as52/xymorimo/Work/Third/WAV/EDIT/*.wav.d

- test_seg による合成音声
/home/as52/xymorimo/Work/Third/WAV/ORG-tseg/*.wav.d
- ポーズ記号制約を行った合成音声の F_0 とラベルの eps ファイル
/home/as52/xymorimo/Work/Third/F0/*.f0.eps
/home/as52/xymorimo/Work/Third/F0/*.lv.eps
- test_seg 用 seg ファイルにヘッダー付加した seg ファイル
/home/as52/xymorimo/Work/Third/SEG/EDIT/*.seg.ch
- ポーズ記号制約を行った合成音声のヘッダー付 seg ファイル
/home/as52/xymorimo/Work/Third/SEG/EDIT/*.seg.ch
- ポーズ記号制約を行った合成音声の聴取実験結果
/home/as52/xymorimo/Work/Third/SEG/RES/total_result
- ポーズ記号制約を行った合成音声の聴取実験用音声ファイル
/home/as52/xymorimo/TEST/