

TR-IT-0297

CHATR の韻律に関する研究

藤澤 謙
Ken Fujisawa

1999.3

概要

本報告では1996年4月から1999年3月までの3年間、筆者がATR音声翻訳通信研究所において行ったCHATRの日本語の韻律に関する研究について、音素単位選択の改良、STRAIGHTによる韻律の修正、自然発話の音声の韻律を用いた音声合成、外国語の音声合成などを記す。

©ATR Interpreting Telecommunications
Research Laboratory.

©ATR 音声翻訳通信研究所

目次

1	はじめに	1
2	CHATR	2
2.1	音素単位選択規準	2
2.2	問題点	3
3	日本語韻律の改良	5
3.1	F0 の傾きを利用した音素単位選択	5
3.1.1	アルゴリズム	5
3.1.2	聴取実験	5
3.2	アクセント型を考慮した音素単位選択	8
3.2.1	音声データベースの分析	8
3.2.2	音素単位選択	9
3.2.3	聴取実験	11
3.3	STRAIGHT を利用した韻律制御	12
3.3.1	STRAIGHT のオフライン処理	12
3.3.2	CHATR の韻律制御	12
3.3.3	評価実験	13
4	入力音声の韻律を用いた音声合成	15
4.1	システム構成	15
4.2	アルゴリズム	15
4.2.1	音素ラベル変換	15
4.2.2	韻律変換	16
5	外国語の音声合成	19
5.1	アルゴリズム	19
5.1.1	話者選択	19

5.1.2	英語 DB による英語音声合成	20
5.1.3	日本語 DB による英語音声合成	20
6	まとめ	22
7	発表文献	23
8	謝辞	25
A	音素バランス	28
A.1	ATR503 文章の CV 分布	28
A.2	ATR503 文章の不足音素／音節	28
A.3	促音の問題 - FTKq	31
B	CHATR の日本語音声合成法	32

第 1 章

はじめに

第二研究室で開発された自然音声波形接続型音声合成システム CHATR [5] は、予め録音された音声データベース中の音素単位の音声波形を、信号処理を行わずに接続して連続音声として出力するため、話者の特徴を失わずに音声を合成することができることを最大の特長としている。しかし、音声データベースのサイズ、不完全な音素単位アルゴリズムなどのために、必ずしも聴覚的に最適な音素が選択されとは限らず、合成音声のイントネーションが概して不自然である。

本レポートでは1996年4月から1999年3月までの3年間、筆者がATR 音声翻訳通信研究所において行った CHATR の日本語の韻律に関する研究について、音素単位選択の改良、STRAIGHT による韻律の修正、自然発話の音声の韻律を用いた音声合成、外国語の音声合成について記述する。

第 2 章

CHATR

2.1 音素単位選択規準

CHATR では、目的とする音声 $t^n = (t_1, \dots, t_n)$ に対して次の 2 つのコストを最小にする音素単位の組合わせ $u^n = (u_1, \dots, u_n)$ を決定する [1] [5].

- (1) 音声データベース中の音素単位 u_i と合成音声として実現したい音素単位 t_i の特徴量の差を表す target cost $C^t(t_i, u_i)$,
- (2) 接続単位 u_{i-1} と u_i との不連続性を表す concatenation(以下 concat) cost $C^c(u_{i-1}, u_i)$

以下、これらについて説明する.

target cost は、実現したい音素単位の特徴ベクトルと、音声データベース中から選ばれた音素単位候補の特徴ベクトルの各要素の差の線形荷重和であり、各 target sub-cost $C_j^t(t_i, u_i)$ の重み w_j^t が与えられた場合、

$$C^t(t_i, u_i) = \sum_{j=1}^p w_j^t C_j^t(t_i, u_i) \quad (2.1)$$

で計算できる. target cost では F_0 の中央値、音韻継続長などを特徴ベクトルの要素としている. なお、 p は特徴ベクトルの次数を表す.

concat cost $C^c(u_{i-1}, u_i)$ は q 個の concat sub-cost $C_j^c(u_{i-1}, u_i)$ の線形荷重和 (重み係数 w_j^c) で表される.

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^q w_j^c C_j^c(u_{i-1}, u_i) \quad (2.2)$$

concat sub-cost としては、接続点における

- ケプストラム距離
- 対数パワーの差の絶対値
- F_0 の差の絶対値

が用いられる.

n 個の音素単位のコストはそれぞれの target cost と concat cost との和となり，最適な音素単位の並びはこの total cost を最小にするような音素単位列 $\bar{u}^n$ で決定される．

$$C(t^n, u^n) = \sum_{i=1}^n C^t(t_i, u_i) + \sum_{i=2}^n C^c(u_{i-1}, u_i) \quad (2.3)$$

$$\bar{u}^n = \operatorname{argmin}_{u_1, \dots, u_n} C(t^n, u^n) \quad (2.4)$$

2.2 問題点

CHATR は，あらかじめ録音された音声データベース中の音素単位の音声波形を信号処理をせずに接続し，音声として出力するため話者の特徴を失わずに音声を合成することを特長としているが，韻律的，音響的に以下の問題点がある．

1 イントネーション

CHATR F_0 ，デュレーション，パワー，前後の音素並びなどを要素として音素単位選択を行うが，各要素の重み付けは経験的に行っているため，必ずしも音響的，韻律的に最適な音素が選択されとは限らず，特に不自然なイントネーションが問題となっている．(図 2.1)

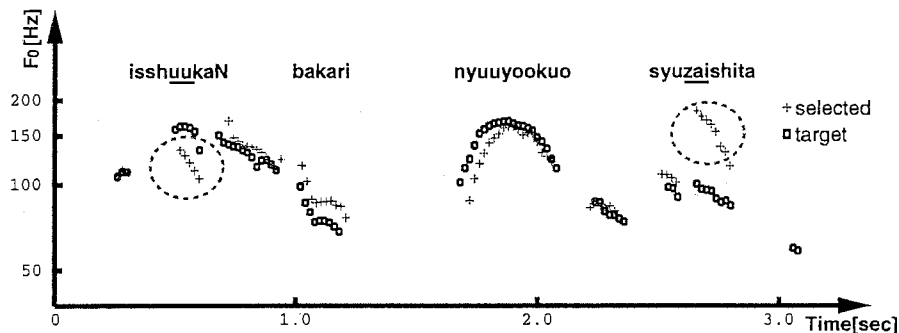


図 2.1: 不自然な F_0

2 デュレーション

前述と同じ理由で，時折ある音素がターゲットとなる目標値の二倍以上や半分以下のデュレーションになることがあり違和感がある．

3 異聴

特に促音において生じやすく，音声データベースに存在しない促音 (/gg/, /jj/, /zz/, /hh/) などで知覚される．

4 音声データベースとの相性

第二研究室における主観評価では男性話者に較べて女性話者の合成音声の評価が高い．原因は不明．同じ性別の話者であっても評価の差は大きい．

5 音声データベースの録音

音声の録音にはどのような発話(朗読, 会話等)を用いるべきか, またどの程度の量が必要なのかが判明していない. いたずらに録音時間を増やすのは, 被録音者の苦痛とディスクスペースの無駄を増やすだけである.

本報告では主に1を改良する方法について記述する.

第 3 章

日本語韻律の改良

3.1 F0 の傾きを利用した音素単位選択

3.1.1 アルゴリズム

CHATR で合成した音声のイントネーションの不自然さは、音素単位間での F_0 のギャップや、アクセント核での不適切な基本周波数パターンを持つ音素単位の選択によるものと考えられる [16][9]. F_0 のギャップは、隣接する音素単位間の F_0 パターンの形状や、大きさの差によって生じるため、これらを考慮する選択規準が必要である。また、適切なアクセントを表現するには音素単位間の相対的な F_0 の大きさを考慮する必要がある。

そこで、音素単位間の F_0 パターンのギャップを減らし、推定された F_0 パターンの形状をより忠実に反映した音素単位が選ばれるよう、 F_0 に関する以下のコスト関数を追加した。

slope 音声データベース中の音素単位の F_0 パターンの傾きを考慮し、実現したい F_0 (以下 target F_0) との傾きとの差を target cost に追加する。 F_0 パターンの傾きは、音声データベース中に十分ある母音に対してのみ考慮し、他の有声音は考慮しないこととする。また、原音声波形からの F_0 抽出誤りの影響を軽減するため、抽出した F_0 をスムージングしてから回帰分析により傾きを計算した。

threshold target cost 中の F_0 中央値の差が、一定の閾値以上であれば penalty cost を追加する。

difference 連続する二つの音素単位の F_0 の差分を、target F_0 の差分に近づけるため、以下を concat cost に追加する。

$$C_j^c(u_{i-1}, u_i) = |u'_{f_0 i} - u_{f_0 i}| \quad (3.1)$$

ここで

$$u'_{f_0 i} = u_{f_0 i-1} + t_{f_0 i} - t_{f_0 i-1}. \quad (3.2)$$

とする。 $t_{f_0 i-1}$, $t_{f_0 i}$ はそれぞれ $i-1$, i 番目の音素の target F_0 , $u_{f_0 i-1}$, $u_{f_0 i}$ はそれぞれ $i-1$, i 番目の音素単位の F_0 を表す。(図 3.1)

3.1.2 聴取実験

無作為に選んだ新聞記事 50 文章から CHATR で合成音声を作成し、被験者に提示した。アクセント付与は半自動的に行った。被験者には合成音声の自然性について、(a) イントネーション

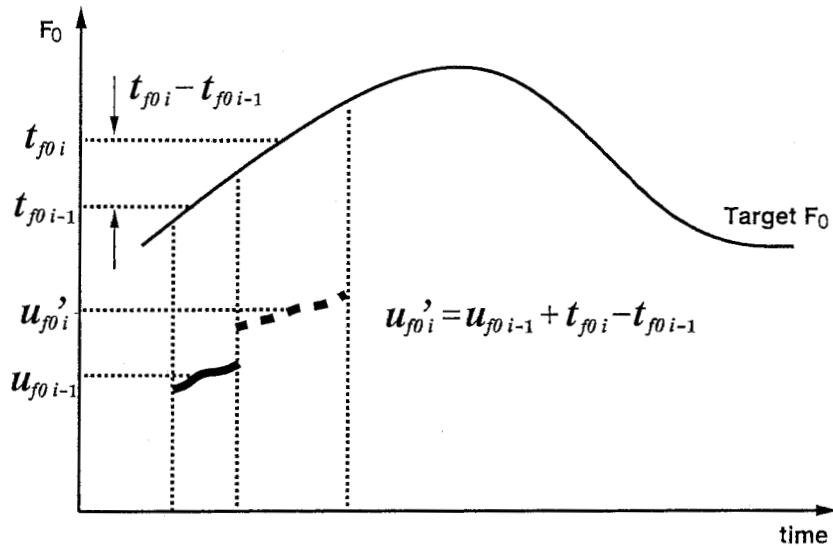


図 3.1: Difference cost.

と (b) 連続性 / 明瞭性の評価を, 「極めて良い」から「極めて悪い」の 5 段階で評価させた. 連続性 / 明瞭性の評価の際には, イントネーションを無視するように被験者に指示した. 合成音声は, (1) オリジナルの CHATR での合成 (以下 original), (2) slope cost を導入して合成, (3) threshold cost を導入して合成 (予備的検討より threshold は 20 Hz に設定), (4) difference cost を導入して合成, (5) 上記の (2)–(4) を同時に使って合成 (以下 all features) の 5 種類を使用した. 被験者は, 合成音声を聞きなれていない 6 名である. また, CHATR での F_0 パターンの予測には J_ToBI [7] を用いた. イントネーションの評価実験の結果を図 3.2 に示す. 図 3.2 から, 今回提案した各コストを個別に導入することにより, イントネーションが悪い / 極めて悪いという評価が 20 % 程度減少し, 逆に良い / 極めてよいという評価が 10 % 程度増加している. 更に, これらのコストを同時に導入した all features では, 悪い / 極めて悪いという評価が original に較べて約半分となっており, 提案した選択規準の有用性が確認できた.

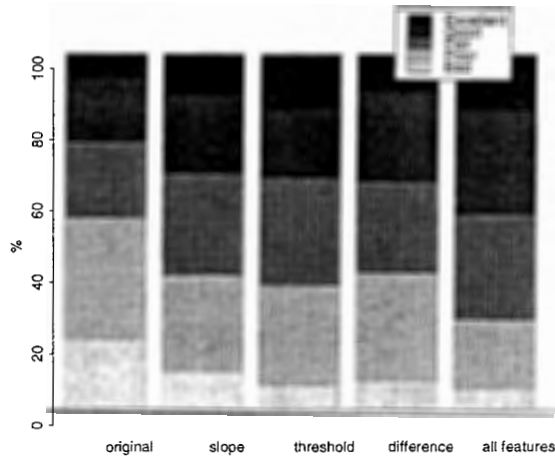


図 3.2: イントネーションの評価

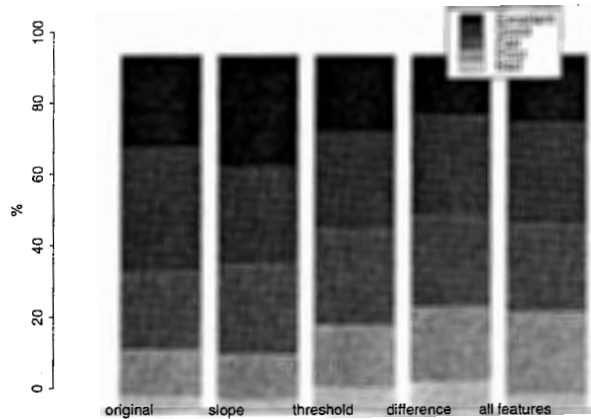


図 3.3: 連続性 / 明瞭性の評価

次に、連続性 / 明瞭性の評価実験の結果を図 3.3 に示す。threshold cost, difference cost を個別に導入した場合、およびすべてのコストを導入した場合は original と較べて若干評価が低かった。この原因としては、これらのコストを追加することにより、original と較べてケプストラムや音韻継続長のコストの影響が相対的に小さくなり、音素単位の接続個所での不連続感や不適当な音韻継続長を持つ音素単位の選択が生じたためであると考えられる。これに対して slope cost を導入した場合、連続性 / 明瞭性は original と較べてあまり劣化がみられなかった。

これらのことから、イントネーションの自然性向上に寄与し、連続性 / 明瞭性に関しても original とほぼ同じ評価を得た slope cost のみを導入することが適切であると考えられる (図 3.3)。

関連 URL:

<http://www.itl.atr.co.jp/~fujisawa/DEMO/pitch1.html>

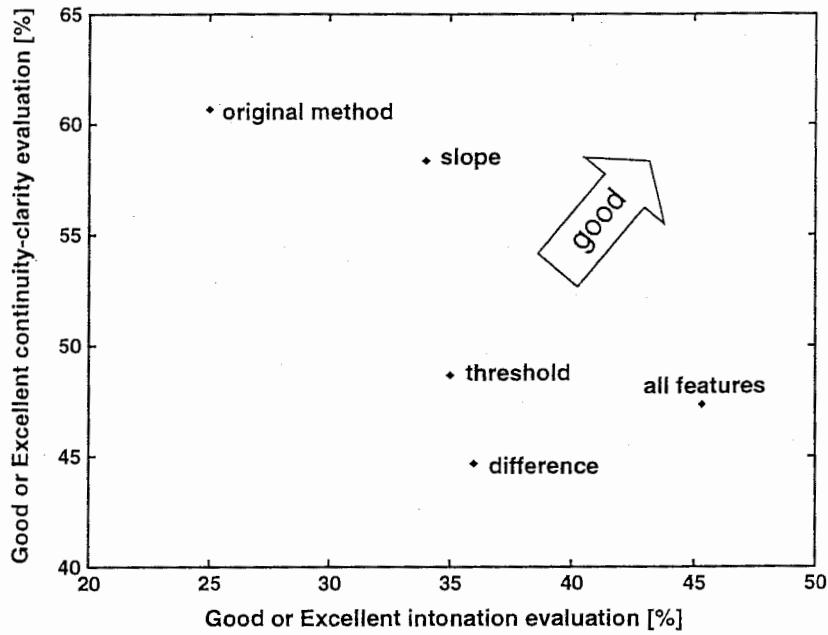


図 3.4: イントネーションと連続性・明瞭性で‘good’以上と評価された割合

3.2 アクセント型を考慮した音素単位選択

予備実験 [19] より通常とは異なるアクセント型の合成音声やアクセント句内で F_0 が複数回上昇する音声で標準語と異なった不自然な韻律として知覚されやすいことを確認したため、正しいアクセント型で合成できるような音素単位選択について検討した [15][8].

3.2.1 音声データベースの分析

女性アナウンサー (FTK) による ATR-503 文章の朗読音声の F_0 を ToBI ラベル [2] を用いてアクセント句単位で以下の要因について分類した.

1. モーラ長
2. アクセント型
3. 左右の BI

アクセント句境界およびアクセント型はそれぞれ ToBI ラベルの ‘L%’ と ‘H*+L’ などから求めた. 図 3.5 に FTK の ATR-503 文朗読発話のアクセント句単位のモーラ長の分布を示す.

アクセント句のそれぞれのカテゴリに対して以下をデータベース化 (以下アクセントデータベース) した.

- 各モーラの平均 F_0 ($m_{l,t,p,n,i}$),
- 連続するモーラの F_0 の差分の平均 ($d_{l,t,p,n,i}$),

- $d_{l,t,p,n,i}$ の標準偏差 ($s_{l,t,p,n,i}$)

ここで l はアクセント句のモーラ長, t はアクセント型, p と n は対象とするアクセント句の左右の BI, i はアクセント句内の i 番目のモーラを表す. ただし, 無声化した母音や促音をもつアクセント句はアクセントデータベースの計算から除いた.

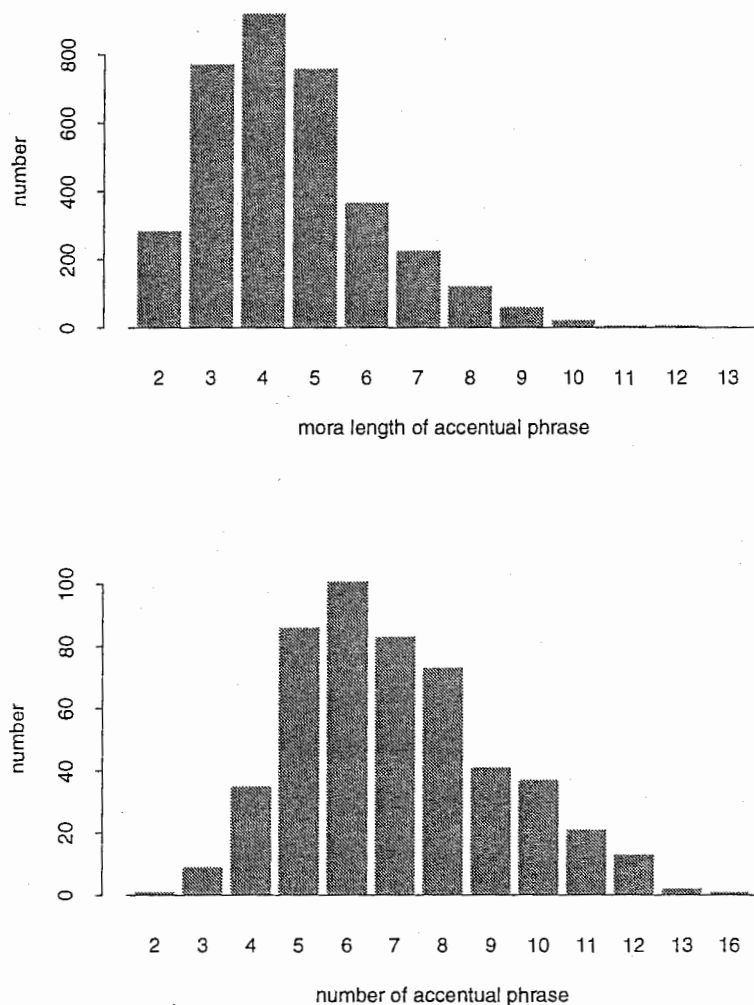


図 3.5: FTK のアクセント句長とモーラ長の分布

3.2.2 音素単位選択

CHATR は目的とする音声に対して予測した韻律的特徴との距離を表わすターゲットコスト C^t と連続する音素間の不連続性を表わす接続コスト C^c の和を最小にするように音素単位を選択する [1]. 本稿ではアクセントデータベース中に対応するアクセント句の情報がある場合, ターゲットとする F_0 として予測値の代わりに $m_{l,t,p,n,i}$ を用いた.

また、以下に示すコスト関数を接続コストに追加し、選択される音素に F_0 の制限を設けた。

$$C_j^c(u_{i-1}, u_i) = \begin{cases} 0.0 & \text{if } u_i - u_{i-1} \leq d_{l,t,p,n,i} \pm \alpha s_{l,t,p,n,i} \\ \text{const.} & \text{otherwise.} \end{cases} \quad (3.3)$$

ここで u_{i-1} と u_i はそれぞれ $(i-1)$ 番目と i 番目の音素の F_0 を示す。 α は F_0 の許容範囲を設定する変数で通常 2.0 として使用する。 Const. は音素候補を合成に使用されない程度に十分大きな値とした。 本コスト関数は母音と撥音についてのみ適用した。 実際の計算では F_0 レンジとして、話者固有性を除くために F_0 の z-スコアを用いた。

$$x_i^z = \frac{x_i - \bar{x}}{\sigma_x} \quad (3.4)$$

ここで σ_x はアクセントデータベースを作成した話者 FTK の音素毎の F_0 の標準偏差、 $\bar{x}$ は FTK の平均 F_0 を示す。 z-スコアを用いることにより、FTK から求めたアクセントデータベースを他の話者に対しても適用できる。 図 3.6 に本コスト関数を適用したときの F_0 の制限の概要を示す。 合成内容は「海が見える」 (u'miga 3 mie'ru) である。 ここで、 “ ’ ” はアクセント、“3” は BI を示す。

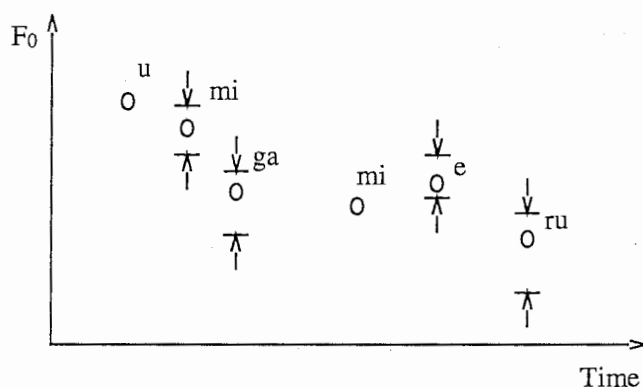


図 3.6: 音素単位選択時の F_0 の制限

ところでアクセントデータベースは ATR-503 文章をベースとして作成しているため、10 モーラを越えるような長いアクセント句については十分なデータがないため、任意の入力に対しては十分とはいえない。予備的見地によると CHATR による合成音声の不自然性は間違ったアクセント型やアクセント句内での複数回の F_0 の上昇で多く知覚されている。一方、東京方言ではアクセント句内で F_0 は複数回上昇しないことが知られている。そこで、モーラ長、アクセント型などの一致するアクセント句がアクセントデータベースに存在しない場合、音素単位選択に以下の条件を課した。

- 頭高型以外のアクセント句の場合、2 モーラ目まで F_0 は上昇する
- アクセント核で F_0 は大きく下降する
- アクセント句以降のモーラの F_0 は緩やかに下降する。

3.2.3 聴取実験

新しく導入した音素単位選択規準の有効性を調べるため、聴取実験を行った。50 文章を BI で区切られたアクセント付きローマ字に変換し CHATR により合成した。アクセントと BI は入力文章から自動的に作成し、目視により人手で修正した。音声の合成には女性話者 (FTK) と男性話者 (MHN) の音声データベースを用いた。被験者は 12 人で、従来法による CHATR の合成音声と提案法による合成音声を聞き、イントネーションが自然に聞こえる方を選ばせた。被験者には好きな回数だけ音声を聞かせた。

実験結果は発話単位で評価し、8 人以上が良いと評価した音声を他方の音声よりも良いとした。5 人から 7 人が一方を良いと評価した場合、すなわち評価に 4 人以上の差が無かった場合は同じとした。図 3.7 に評価結果を示す。図 3.7 から提案法によりイントネーションの自然性が改善できたことを確認できた。アクセントデータベースは FTK から作成されたものを用いているにも関わらず、話者 MHN の改善率は話者 FTK より高い。これは元来話者 MHN による合成音声のイントネーションの評価が FTK のものに較べて低かったことが原因と考えられる [6]。

表 3.1: 聴取実験結果

Speaker	Better	Worse	Same
FTKnew	64 %	8 %	28 %
MHNnew	70 %	12 %	18 %

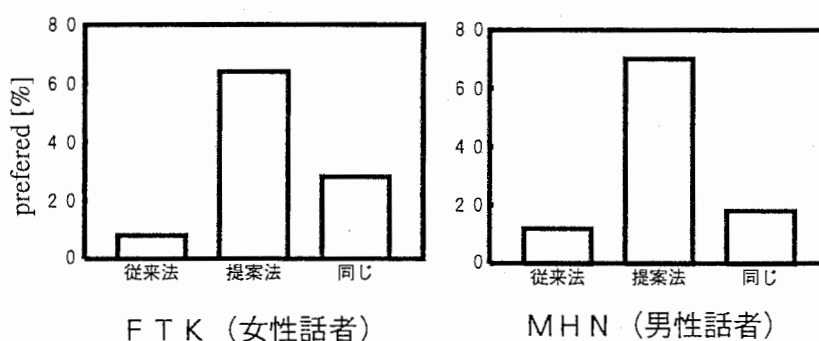


図 3.7: イントネーションの評価結果

関連 URL:

http://www.itl.atr.co.jp/~fujisawa/CHATR/accent_type/
http://www.itl.atr.co.jp/~fujisawa/DEMO/openhouse98/improve_jp/

3.3 STRAIGHT を利用した韻律制御

高品質な音声分析変換合成方法として STRAIGHT が提案されている [14]. 本章では STRAIGHT を用いてオフライン処理で合成用のインパルス応答素片を作り, その素片を編集することで任意のピッチ及びデュレーションの音声を合成する方法について述べると共に, 音声波形接続型合成システム CHATR の韻律制御への応用及び評価実験について述べる [18].

3.3.1 STRAIGHT のオフライン処理

音声信号の周期性を積極的に利用して時間周波数領域でのスペクトログラムの適応的補間により信号の時間周波数特性を表す曲面を求める方法を用いた音声分析, 変換, 合成方法 STRAIGHT が提案されている.

本方式で, 変換及び合成に必要なパラメータは最適補間平滑化スペクトル及び音源情報であり, これらのパラメータを用いて合成用のインパルス応答波形素片が作られる.

本方式は合成時の計算時間を短縮するために一部の処理を予めオフラインで行っておくことができる. 図 3.8 の上段がオフライン処理を, 下段が合成時の処理を表す. CHATR のデータベース音声に対しては, 予めある一定間隔のインパルス応答の素片を STRAIGHT で分析し, ピッチ, 有声/無声情報と一緒にディスクに蓄積しておけば, 合成音声を作る時に, これらの素片を波形上で編集するだけで, 望ましいピッチ及びデュレーションに変換した音声合成できる.

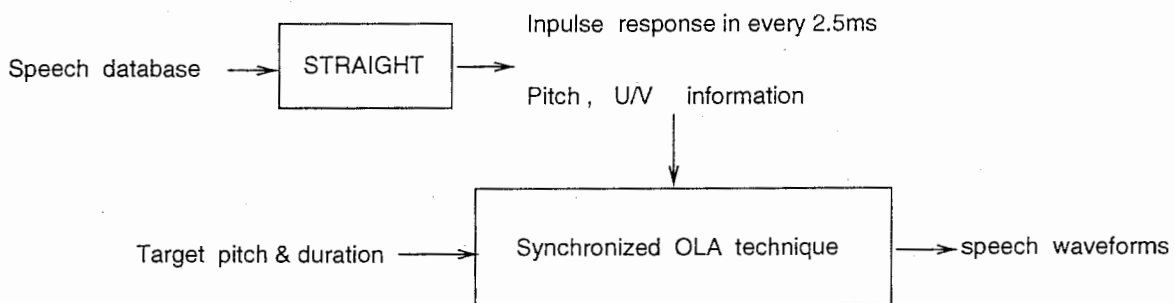


図 3.8: STRAIGHT のオフライン処理とインパルス応答波形素片のオーバーラップアド法による韻律の修正

3.3.2 CHATR の韻律制御

音声波形接続型合成システム CHATR では目標の韻律特徴にもっとも近い音素単位の波形をデータベースから選択して, それらを信号処理せずに接続し, 音声を生成する. その際, データベースから選ばれた音素の中には予測した韻律 (ピッチ, デュレーション) との誤差の大きい音素列が存在する可能性があり, これによって合成音声の一部分の韻律が不自然になることがある. 本報では, CHATR の韻律を改善するために STRAIGHT を用いてピッチ及びデュレーションのモディフィケーションを行った. 具体的には, 以下の二つの方法について検討した:

1. CHATR で選ばれた音素単位の音声のピッチ及びデュレーションをそれに対応する予測値にぴったり合わせる方法と

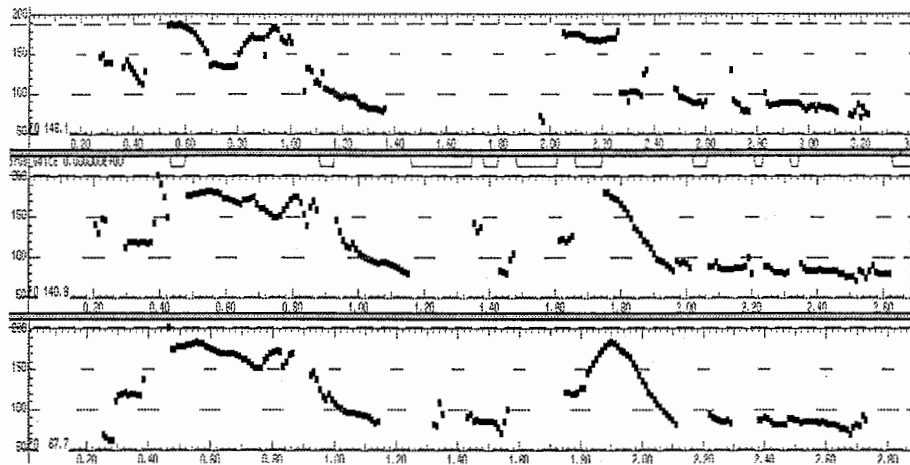


図 3.9: CHATR の出力音声の韻律を STRAIGHT により修正した例. 発話内容: “旅館やホテルに着くと, 非常口を訪ねる.” 上段より CHATR の出力 (上段), STRAIGHT により修正した CHATR の出力 (中段), ターゲット F_0 (下段)

2. 音質のダメージを最小限に抑えるために, 選ばれた音素の韻律と予測韻律の差が大きい部分のみに対して, モディフィケーションを行い, 他の部分を音素素片波形の接続のままとする方法

である. 図 3.9 に示すのは方法 1 により合成されたピッチパターンの一例である.

3.3.3 評価実験

男声 MHN, 女声 FKT のデータベース音声を用いて, 前節で述べた 2 つの手法, およびそれと比較するための 2 つの従来手法で合成した音声の聴取実験を行った. 比較の対象としては, (a) 信号処理を行わずに波形素片を接続した合成音声 (以下 CHATR), 及び (b) 波形接続時に PSOLA を適用した音声 (以下 PSOLA) を用いた. ただし, ここで用いた PSOLA 法は, ターゲットの韻律を忠実に実現する方法ではなく, セグメント単位でピッチを平行移動する方法をとっている. 合成はすべてデータベース中にある自然音声波形の韻律情報をターゲットとして音素単位を選択する手法を用いた. ただし, 元の文の中の該当する音素単位がそのまま使われることを避けるために, 該当する文に含まれる音声の波形素片は合成に使わないものとした.

実験は ABX 法で行い, 最初に自然音声を提示したのち前述した 4 つの方法で合成した音声のうち 2 つの音声を提示して, どちらの音声が最初の音声に近いかを被験者に評価させた. 提示した音声は 10 発話で, 被験者は 11 名である. 被験者には同じ音声資料をヘッドホンとスピーカの二通りで評価させた. なお, 前節で述べた方法 2 (以下 partial-modification) の評価は MHN 音声のみを使って行った. 実験結果を図 3.10, 図 3.11 に示す. MHN, FKT 共に STRAIGHT の音声ではいわゆる buzzy な音が聞こえるため, 総じて CHATR, PSOLA より評価が低くなっている. ヘッドホンで評価した場合はこの buzzy な部分が更に知覚されやすいため, STRAIGHT の評価が低くなったと思われる. 一方, スピーカで評価した場合はヘッドホンに較べて音質の劣化を知覚しにくいいため原音声に近い韻律の STRAIGHT の評価が相対的に高くなるが, それでも音質の差により CHATR や PSOLA には及ばなかった.

これに対して partial-modification と CHATR の比較ではヘッドホン, スピーカ共に partial-modification の方が良い結果を得ている. つまり, 部分的に STRAIGHT の波形を使えば, 音質

の劣化が知覚されずに CHATR の韻律を向上することが可能であると考えられる。

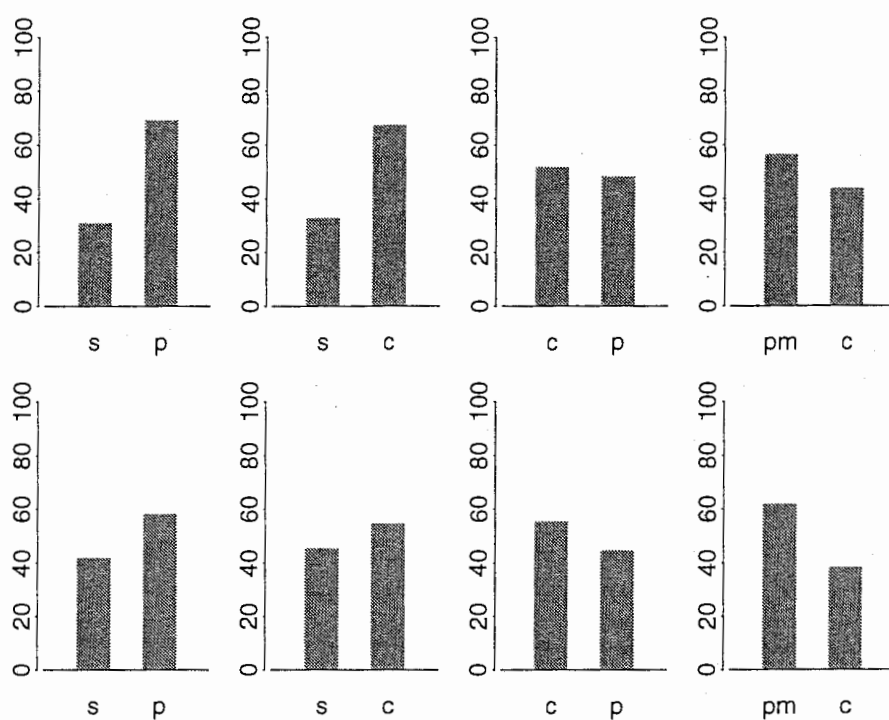


図 3.10: MHN 評価結果

s は STRAIGHT, p は PSOLA, c は CHATR, pm は partial-modification を表す。上段がヘッドホン、下段はスピーカでの結果を表す。

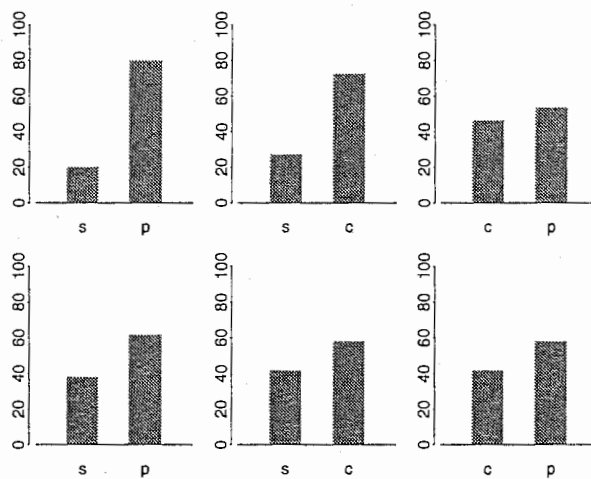


図 3.11: FKT 評価結果

s は STRAIGHT, p は PSOLA, c は CHATR を表す。上段がヘッドホン、下段はスピーカでの結果を表す。

関連 URL:

<http://www.itl.atr.co.jp/~fujisawa/DEMO/Straight/>

第 4 章

入力音声の韻律を用いた音声合成

近年、音声合成技術の向上にともなってカーナビゲーションなどへの応用が検討されており、GUI を用いて合成音声の韻律を制御するシステム [10] [20] が提案されているが、録音再生方式の音声と比較した場合、音声の自然性が不十分である。しかし、録音再生方式のアプリケーションでは発声すべき単語が増えたとき同じ話者で追加録音する必要がある。録音する際のコンディションや前回の録音からの時間的な間隔などによって必ずしも以前と同じような声質の音声で録音できるとは限らず、録音話者と同じ音声を合成で実現できると便利である。

本章ではあらかじめ与えた文章を発声し、その韻律情報を出力話者のレンジに変換したものをターゲットとして音声を合成するシステムについて述べる。韻律情報として実際に発声した音声のものを使用することにより、合成音声の自然性の向上に加え、部分的な強調や韻律の変更など、さまざまな韻律の合成音声を簡便に作成することを目的とする。

4.1 システム構成

システムの実行画面を図 4.1 に、構成を図 4.2 に示す。システムには図 4.3 に示すように合成する文章とその音素表記をあらかじめ与えておく。音声が入力されると与えられた音素ラベルにしたがって Viterbi alignment [12] を行い、音素境界を決定する。また、声門閉鎖点 [11] に基づいて入力音声の F_0 を抽出する。

音声合成には波形接続型音声合成システム CHATR を用いた。 F_0 とパワーはそれぞれ平均、標準偏差より合成音声出力話者のレンジに変換して、音声合成セグメント情報を作成し、CHATR により合成する。

4.2 アルゴリズム

4.2.1 音素ラベル変換

本システムに用いた音声認識と合成システムでは表 4.1 に示す音素ラベルが一対一に対応しない問題があったため、音素ラベルは合成のラベルを与えた。また、有声/無声は入力音声の F_0 より判定し、合成時に音素ラベルの変換を行った。

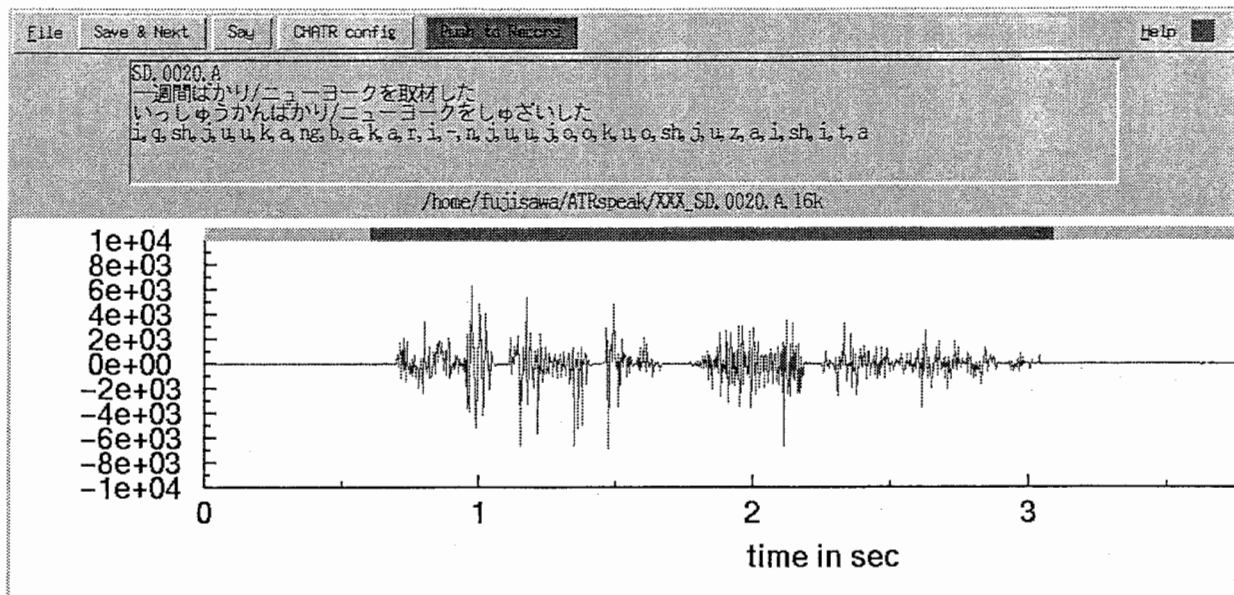


図 4.1: 実行画面

4.2.2 韻律変換

入力が男性の声で女性の合成音声を出力するように，入力音声と合成音声出力話者でピッチレンジが異なる場合，合成器に与える F_0 を出力話者のものに変換しなければならない．本システムでは F_0 を以下の式で変換したものを音声合成時のターゲットとして用いた．時刻 i の入力音声の F_0 を x_i ，出力音声のターゲット F_0 を y_i としたとき，

$$y_i = \bar{y} + w_{f0} \sigma_y x_i^z \quad (4.1)$$

$$x_i^z = \frac{x_i - \bar{x}}{\sigma_x} \quad (4.2)$$

ここで， σ_x は入力音声の F_0 の標準偏差， $\bar{x}$ は入力音声の F_0 の平均値を表わす． σ_y は音声出力話者の F_0 の標準偏差， $\bar{y}$ は音声出力話者の F_0 の平均値を表わし，オフラインであらかじめ出力話者のデータベースより計算しておく． w_{f0} は変換率を表わし，この値を変えることにより合成音声の抑揚の大きさを制御する．通常は 1.0 として使用する．

パワーについては F_0 と同様の変換を行い，音韻継続長は音素境界より得られるものを用いた．

発話内容“あらゆる現実を全て自分の方へねじ曲げたのだ”を男性の声を入力し，女性の話者 (FMP) の F_0 に変換した例を図 4.4(a) に示す． $w_{f0} = 1.0$ として変換した．入力音声の F_0 の平均は 110 [Hz]，標準偏差は 13 [Hz] であり，FMP の F_0 の平均は 228 [Hz]，標準偏差は 44 [Hz] である．FMP の波形データベース中の同発話の F_0 パターンを同図 (b) に示す．

関連 URL:

<http://www.itl.atr.co.jp/~fujisawa/PRODUCTS/voiceln.html>

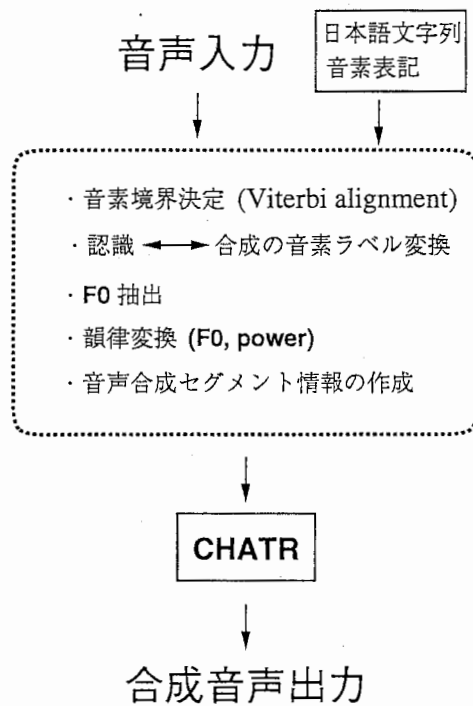


図 4.2: システム構成

SD.0010.A
 あらゆる現実を／すべて自分の方へねじ曲げたのだ
 あらゆるげんじつを／すべてじぶんのほうへねじまげたのだ
 a,r,a,y,u,r,u,g,e,N,j,i,ts,u,o,#,s,u,b,e,t,e,j,i,b,u,N,n,o,h,o,o,e,
 n,e,j,i,m,a,g,e,t,a,n,o,d,a

SD.0020.A
 一週間ばかり／ニューヨークを取材した
 いっしゅうかんばかり／ニューヨークをしゅざいした
 i,ssh,u,u,k,a,N,b,a,k,a,r,i,#,n,y,u,u,y,o,o,k,u,o,sh,u,z,a,i,sh,i,t,a

⋮

図 4.3: 入力テキスト例

表 4.1: 音素ラベルの違い

synthesis	recognition
U	u
I	i
N	ng
j	zh
cch	q, ch
gg	q, g
kk	q, k
pp	q, p
ss	q, s
ssh	q, sh
tt	q, t
tts	q, ts
j	y
h	h
f	h

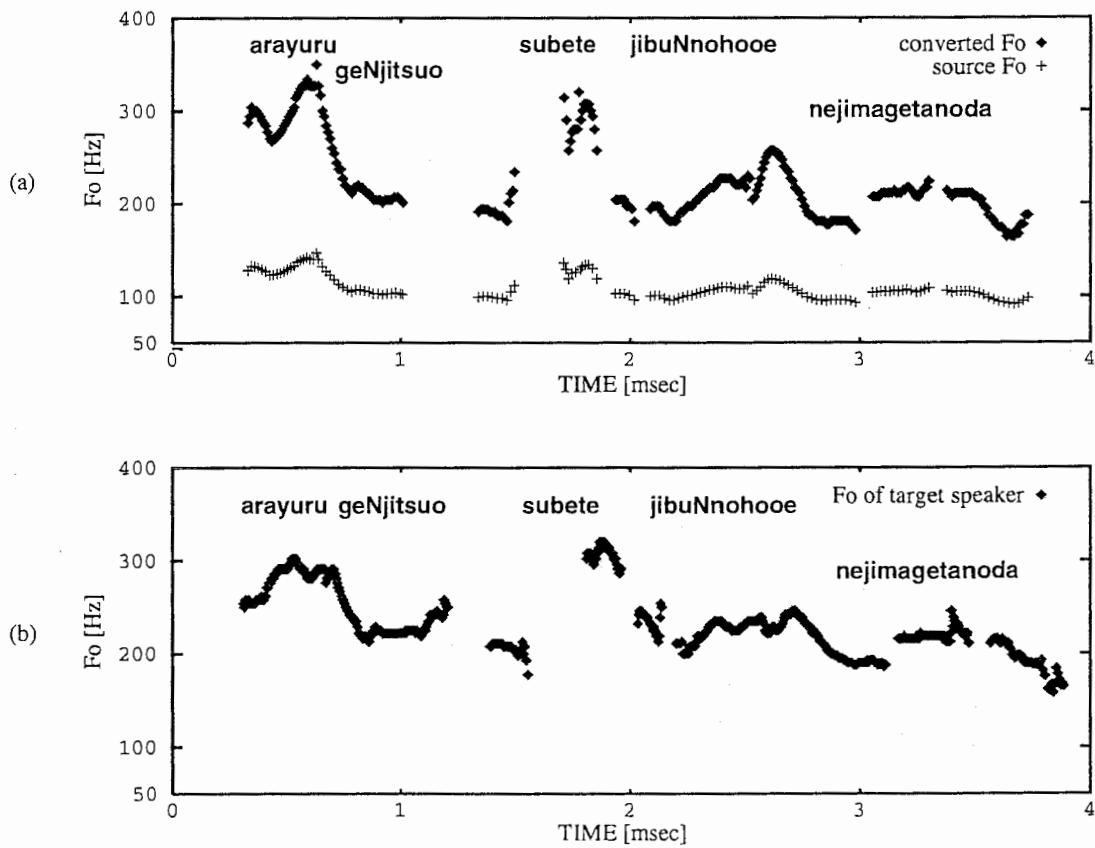


図 4.4: F_0 変換例

第 5 章

外国語の音声合成

多言語翻訳電話で日本語から英語への変換を考えた場合、現状では日本人が日本語で話した言葉を英語に翻訳後、(a) 英語話者すなわち元話者とは全く異なった声の合成音声を出力するか、(b) 元話者の日本語 DB から英語のローマ字読みの合成音声を出力する、しか方法がない。(a) では元話者とは全く異なった音声になり、(b) では元話者の声ではあるものの、いわゆるカタカナ英語の発音となる。ケプストラム情報を音素単位選択に使うことによりこの問題は解決できると考えられるが、音声波形からケプストラムを抽出することはできるものの予測することは困難である。そこで、本章ではある言語の音声 DB から別の言語の音声を合成する際に、一度ネイティブの音声 DB を用いて合成波形を作成し、そのケプストラム情報を元にネイティブ以外の音声 DB から音声を合成する手法について述べる [4][3][17]。原理的に言語の依存性はないため英語、日本語間や多言語間の変換にも応用できる。

5.1 アルゴリズム

例として日本語音声 DB で英語の音声を合成する際のアルゴリズムを図 5.1 に示す。処理の流れは

1. 話者選択
2. 英語音声 DB を用いて英語音声合成
3. 英語の音声波形のケプストラムをターゲットとして日本語音声 DB で英語の音声合成

となる。

5.1.1 話者選択

合成したい日本語音声 DB の声質に近い英語音声 DB を選択する。選択規準として、性別、基本周波数のレンジ、音韻継続長などがあげられる。また登録話者に対して DTW で時間整合した目的話者の学習音声スペクトル時系列と登録話者の学習音声スペクトル時系列との距離を求め、2 乗誤差最小規準により最も距離の小さい登録話者を選択する方法も考えられる。

5.1.2 英語 DB による英語音声合成

2章の方法により音声を合成する. すなわち target cost として実現したい音声単位の基本周波数, 音韻継続長, 対数パワーなどを要素とした特徴ベクトルと, 音声 DB 中から選ばれた候補の音声単位の特徴ベクトルの差を用いる.

5.1.3 日本語 DB による英語音声合成

音素候補

ある言語の音声 DB で他言語の音声を合成する場合, 合成する言語の音素が音声 DB に無い場合がある. そこで, 各言語毎に IPA(the International Phonetic Alphabet) で定義される調音位置や調音様式による分類を用いて音素をクラスタリングし, 各クラスに所属する音素を合成のための音素候補とする. 表 5.1 は日本語の音素を調音様式で分類した例を示す.

表 5.1: 日本語音素分類

調音様式	日本語の音素例
母音	/a/, /i/, /u/, /e/, /o/
半母音	/j/, /w/
破裂音	/p/, /b/, /t/, /d/, /k/, /g/
摩擦音	/f/, /h/, /s/, /sh/, /z/, /r/
破擦音	/ts/
鼻音	/n/, /m/

音素選択

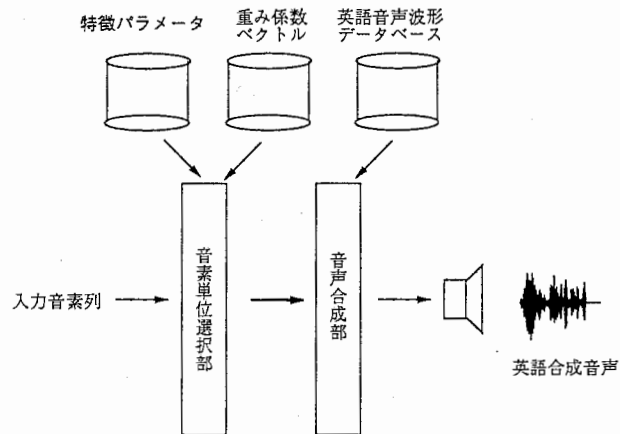
target cost の計算では, ターゲットとして従来の CHATR で使われていた韻律情報ではなく, ?? 節で合成した英語音声のケプストラム情報を用いて音素候補を選択する.

concat cost の計算は従来通りの方法 [2] で行う.

関連 URL:

http://www.itl.atr.co.jp/~fujisawa/DEMO/openhouse98/foreign_synth/

step 1 英語DBによる英語の音声合成



step 2 日本語DBによる英語の音声合成

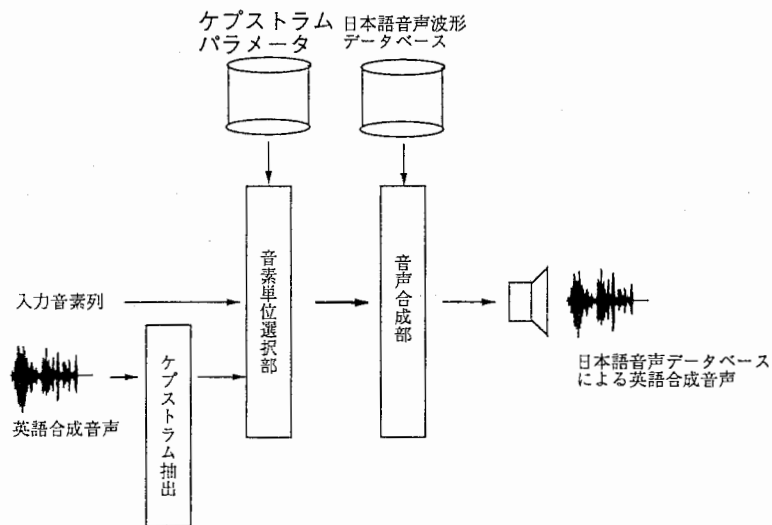


図 5.1: 動作原理

第 6 章

まとめ

本稿では 2 章で波形接続型音声合成システム CHATR のアルゴリズムと問題点について述べ、最も大きな問題点であるイントネーションの自然性の改良について 3 章で、また CHATR を用いたシステムやアプリケーションについて 4 章と 5 章で示した。

3.1, 3.2 節では CHATR の音素単位選択アルゴリズムを改良することにより、イントネーションの自然性を向上できることを示した。ただし、これらの方法はデュレーションや音素単位接続の重みを相対的に小さくするため、音声データベースの充実や、接続歪みを軽減する HNM [13] などの実装が望まれる。

また 3.3 節では部分的に信号処理を適用することにより合成音声のイントネーション、リズムの品質向上を示した。波形全体に STRAIGHT を適用した場合、信号処理を行わずに音素を接続 (DUMB+) した音声に較べて評価が低かった。これは評価実験の方法が原音声と DUMB+, STRAIGHT の合成音声を聞かせて原音声に近い音声を被験者に選択させる方法をとっていたためであり、STRAIGHT による合成音声と DUMB+ による音声を一対比較すると結果が変わる可能性がある。また現在では実験を行った当時より STRAIGHT が改良されており、最新版の STRAIGHT を利用するとどうなるか興味深い。

さまざまな韻律の音声を簡便に合成するため、音声を入力してその韻律を CHATR の音素単位選択に用いるシステムについて 4 章で示した。更に、ある言語の音声データベースから別の言語の音声を合成する方法について 5 章で示した。

第 7 章

発表文献

以下の文献は

<http://www.itl.atr.co.jp/~fujisawa/publication/>

から PostScript ファイルで手にいれることができる。(99.3 現在)

主著

藤澤, 平井, 樋口 “波形接続型音声合成システム CHATR の基本周波数に関する音素単位選択規
準の改良”, 日本音響学会 春季研究発表会 講演論文集, pp.219-220, Mar.1997.

K.Fujisawa, T.Hirai, N.Higuchi. “Use of Pitch Pattern Improvement in the CHATR Speech
Synthesis System”, Proc. Eurospeech97, pp.2671-2674, Sep.1997

藤澤, キャンベル “入力音声の韻律を用いた音声合成”, 日本音響学会 春季研究発表会 講演論文
集, pp.191-192, Mar.1998.

藤澤 謙, Wei Zhang. “VOICE FONTS: 音声合成データベースの作成”, ATR Journal, pp.4-5,
Volume 33, AUTUMN, 1998

K.Fujisawa, N.Campbell. “PROSODY-BASED UNIT-SELECTION FOR JAPANESE SPEECH
SYNTHESIS”, The Third ESCA/COCOSDA Workshop on SPEECH SYNTHESIS, pp.181-
184, Nov.1998

藤澤, キャンベル “波形接続型音声合成システムにおけるアクセント型を考慮した音素単位選択”,
日本音響学会 春季研究発表会 講演論文集, pp.227-228, Mar.1999.

共著

丁文, 藤澤, キャンベル, 樋口, 河原 “STRAIGHT を用いた CHATR の韻律制御”, 日本音響
学会 秋季研究発表会 講演論文集, pp.211-212, Sep.1997.

W.Ding, K.Fujisawa, N.Campbell. “IMPROVING SPEECH SYNTHESIS OF CHATR USING
A PERCEPTUAL DISCONTINUITY FUNCTION AND CONSTRAINTS OF PROSODIC
MODIFICATION”, The Third ESCA/COCOSDA Workshop on SPEECH SYNTHESIS, pp.191-
194, Nov.1998

リープス ベン, 西野 敦士, シンガー ハラルド, 藤澤 謙, 山田 節夫, 菅谷史和, 竹澤寿幸, 横尾 昭男, 山本 誠一. “ATR-MATRIX: A Speech Translation System Between English and Japanese”, 情報処理学会, 1999

共著 (テクニカルレポート)

藤澤 友紀子, 藤澤 謙, Nick Campbell “CHATR に於ける日本語の韻律評価”, TR-IT-0254. Feb.1998.

森本 康夫, 藤澤 謙, Nick Campbell “CHATR におけるポーズと F0 下降の関係を利用した音素選択”, TR-IT-0264. May.1998.

田村 則和, 藤澤 謙, ニック キャンベル “音声合成システム CHATR における日本語話者による英語の音声合成”, TR-IT-0277. Sep.1998.

酒井 慎介, 藤澤 謙, Nick Campbell. “言語フィルタの改良”, TR-IT-0290, Feb.1999

佐渡 詩郎, 藤澤 謙, Nick Campbell. “韻律処理プログラムの改良”, TR-IT-0291, Feb.1999

特許

自然発話音声波形信号接続型音声合成装置 08 音 - 特 29 (平成 09 年度特許願第 048769 号, 特開平 10-247097)

音声合成システム 09 音 - 特 24 (平成 10 年度特許願第 051925 号)

第 8 章

謝辞

ATR 音声翻訳通信研究所での研究の機会を与えてくださった山崎元社長，樋口元室長，出向元の関連部署の皆様に深謝します．研究に関して有益な助言を頂いた，山本社長，キャンベル室長，匂坂室長，第二研究室の研究員の皆様に感謝します．研究に用いるプログラムや研究資料等でサポートしていただいたエンジニアやラベラーの皆様に感謝します．

参考文献

- [1] A. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *Proc. Eurospeech95*, pages 581–584, Apr. 1995.
- [2] N. Campbell. Autolabelling Japanese ToBI. In *Proc. of ICSLP*, pages 2399–2402, Oct 1996.
- [3] N. Campbell. Foreign language speech synthesis. In *The Third ESCA / COCOSDA Workshop on SPEECH SYNTHESIS*, pages 177–180, Nov. 1998.
- [4] N. Campbell. Foreign-language speech synthesis - borrowing voices. In *音響学講論*, pages 179–180, Sep. 1998.
- [5] N. Campbell and A. Black. CHATR: 自然音声波形接続型任意音声合成システム. In *信学技報*, SP96-07, May 1996.
- [6] N. Campbell, Y. Itoh, W. Ding, and N. Higuchi. Factors affecting perceived quality and intelligibility in the CHATR concatenative speech synthesiser. In *Proc. Eurospeech97*, pages 2635–2638, Sep. 1997.
- [7] N. Campbell and J. Venditti. J-ToBI: an intonation labelling system for Japanese. In *Proc. Fall Meeting, Acoust. Soc. Jpn.*, pages 317–318, Sept. 1995.
- [8] K. Fujisawa and N. Campbell. Prosody-based unit-selection for japanese speech synthesis. In *The Third ESCA / COCOSDA Workshop on SPEECH SYNTHESIS*, pages 181–184, Nov. 1998.
- [9] K. Fujisawa, T. Hirai, and N. Higuchi. Use of pitch pattern improvement in the CHATR speech synthesis system. In *Proc. Eurospeech97*, pages 2671–2674, Sep. 1997.
- [10] Kazuo Hakoda, Tomohisa Hirokawa, and Kenzo Itoh. Speech editor based on enhanced user-system interaction for high quality text-to-speech synthesis. In *Proc. ICSLP*, pages 1775–1778, 1994.
- [11] Ashok K. Krishnamurthy and Donald G. Childers. Two-channel speech analysis. *IEEE Trans. on Acoustics, Speech and Signal Proc.*, ASSP-34(4):730–743, Aug. 1986.
- [12] Biing-Hwang Juang Lawrence Rabiner. 音声認識の基礎 (下). NTT アドバンステクノロジー株式会社, 1995.
- [13] Y. Stylianou. Concatenative speech synthesis using a harmonic plus noise model. In *The Third ESCA / COCOSDA Workshop on SPEECH SYNTHESIS*, pages 261–266, Nov. 1998.
- [14] 河原 英紀 and 増田 郁代. 時間周波数領域での補間を用いた音声の変換について. In *信学技報* EA96-28, pages 9–16, Aug. 1996.
- [15] 藤澤 謙 and ニック キャンベル. 波形接続型音声合成システムにおけるアクセント型を考慮した音素単位選択. In *音響学講論*, pages 227–228, Mar. 1999.
- [16] 藤澤 謙, 平井 俊男, and 樋口 宜男. 波形接続型音声合成システム CHATR の基本周波数に関する音素単位選択規準の改良. In *音響学講論*, pages 219–220, Mar. 1997.

- [17] 田村 則和, 藤澤 謙, and ニック キャンベル. 音声合成システム CHATR における日本語話者による英語の音声合成, 1998. TR-IT-0277.
- [18] 丁 文, 藤澤 謙, ニック キャンベル, 樋口 宜男, and 河原 英紀. STRAIGHT を用いた CHATR の韻律制御. In 音響学講論, pages 211-212, Sep. 1997.
- [19] 藤澤 友紀子, 藤澤 謙, and ニック キャンベル. CHATR に於ける日本語の韻律評価, 1998. TR-IT-0254.
- [20] 近藤 玲史, 滝澤 卓也, 若尾 淳, 稲垣 敬子, 吉田 和永, and 三留 幸夫. 合成音声をデザインするシステム - スピーチデザイナの試作 -. In 音響学講論, pages 223-224, Mar. 1997.

付録 A

音素バランス

A.1 ATR503 文章の CV 分布

ATR503 文章の中には /gg/, /jj/, /zz/, /hh/ の音素がない。また音節レベルで見ると幾つか足りない外来語系の音節がある。FTK データベースの ATR503 発話に含まれる音節と、その中に含まれない音節(※で表わす)を表 A.1A.2に示す。

関連 URL:

<http://www.itl.atr.co.jp/~fujisawa/CHATR/db.html>

A.2 ATR503 文章の不足音素／音節

A.1で述べたように ATR503 文章の中には /gg/, /jj/, /zz/, /hh/ の音素がない。また音節レベルで見ると幾つか足りない外来語系の音節がある。そこで後日音声データベースを収録するときのために、WWW から集めた ATR503 文章に含まれない音節を含んだ文章を付す。

関連 URL:

<http://www.itl.atr.co.jp/~fujisawa/CHATR/db.html>
<http://www.itl.atr.co.jp/~fujisawa/CHATR/atr503plus.html>

不足音素 (/gg/, /jj/, /zz/, /hh/)

[gga]

超高速なインクリメンタル・デバッグ環境を提供する。
彼は日本が誇るべきスラッガーだ。

[ggu]

50 年代後期モデルをモチーフに、レッグを全体にスリム化した。
ソルティードッグの塩抜きのことをボガードっていうんだよ。

[jji]

フィルムはパトローネからオールプラスチック製のカートリッジに収納されます。
白黒画像のエッジ抽出を、ラプラシアン・ガウシアンフィルタを用いて行っています。
採点を行うジャッジ団は、9 名のジャッジと 1 名のサブステイテュートジャッジで構成される。

[zzu]

表 A.1: FTK ATR 503 文 CV 分布

あ a (302)	い i (805)	う u (389)	え e (342)	お o (1067)
か ka (366)	き ki (59) ki (173)	く ku (81) ku (255)	け ke (153)	こ ko (251)
さ sa (192)	し shi (147) shi (211)	す su (82) su (142)	せ se (142)	そ so (128)
た ta (337)	ち chi (165)	つ tsu (71) tsu (180)	て te (246)	と to (369)
な na (304)	に ni (354)	ぬ nu (31)	ね ne (84)	の no (504)
は ha (162)	ひ hi (78) hi (68)	ふ fu (70) fu (71)	へ he (37)	ほ ho (93)
ま ma (217)	み mi (157)	む mu (78)	め me (123)	も mo (244)
や ya (149)		ゆ yu (90)		よ yo (175)
ら ra (224)	り ri (202)	る ru (363)	れ re (257)	ろ ro (140)
わ wa (413)				
ん N (900)				
が ga (356)	ぎ gi (74)	ぐ gu (70)	げ ge (102)	ご go (101)
ざ za (45)	じ ji (195)	ず zu (108)	ぜ ze (34)	ぞ zo (28)
だ da (183)			で de (294)	ど do (133)
ば ba (106)	び bi (59)	ぶ bu (96)	べ be (42)	ぼ bo (56)
ぱ pa (19)	ぴ pi (2) pi (11)	ぷ pu (13) pu (3)	ぺ pe (9)	ぽ po (9)
しゃ sha (50)		しゅ shu (64)		しょ sho (88)
ちゃ cha (12)		ちゅ chu (41)		ちょ cho (33)
にゃ nya (3)		にゅ nyu (24)		にょ nyo (1)
みゃ mya (2)		みゅ myu (3)		みょ myo (3)
りゃ rya (※)		りゅ ryu (7)		りょ ryo (34)
きゃ kya (12)		きゅ kyu (46)		きょ kyo (42)
ひゃ hya (3)		ひゅ hyu (1)		ひょ hyo (24)
じゃ ja (16)		じゅ ju (62)		じょ jo (75)
ぎゃ gya (8)		ぎゅ gyu (2)		ぎょ gyo (23)
びゃ bya (1)		びゅ byu (2)		びょ byo (9)
ぴゃ pya (※)		ぴゅ pyu (1)		ぴょ pyo (1)
つか kka (10)	つき kki (12)	つく kku (14)	つけ kke (3)	っこ kko (10)
っさ ssa (3)	っし sshi (13)	っす ssu (※)	っせ sse (5)	っそ sso (1)
った tta (105)	っち cchi (7)	っつ tsu (1)	って tte (116)	っと tto (21)
っが gga (※)		っぐ ggu (※)		
	っじ jji (※)	っず zzu (※)		
				っど ddo (4)
っは hha (※)	っひ hhi (※)	っふ flu (1)	っへ hhe (※)	っほ hho (※)
っぱ ppa (14)	っぴ ppi (7)	っぷ ppu (10)	っぺ ppe (3)	っぽ ppo (26)
っきゃ kky (※)		っきゅ kkyu (1)		っきょ kkyo (5)
っしゅ sshu (4)		っしゅ sshu (12)		っしゅ sshu (4)
っちゅ cchu (4)		っちゅ cchu (※)		っちゅ cchu (1)
っぴゅ ppyu (2)		っぴゅ ppyu (※)		っぴゅ ppyu (3)
			シエ she (※)	
	ティ ti (6)			
			チエ che (1)	
ファ fa (3)	フィ fi (6)		フェ fe (1)	フォ fo (1)
	ウィ wi (1)		ウェ we (※)	ウォ wo (1)
			ジェ je (1)	
		デュ dyu (1)		
	ディ di (2)	ドウ du (※)		
ツァ tsa (※)	ツイ tsi (※)		ツエ tse (※)	ツォ tso (※)

あのタイガー・ウッズがアサヒ飲料の新缶コーヒー「ワンダ」に登場。
オッズの方式には二通りある。

[zzo]

エデュカートリッジシステム対応ゲームソフトの第2弾が発売された。
ギリシャに行ったら一度「ウッズ」を試してください。

[hha]

F1最終戦、シューマッハとビルヌーブ、1ポイント差でむかえた最終戦。
聖トーマス教会でバッハは毎週日曜日の礼拝用のカンタータを作曲し、演奏指導をさせた。

[hhi]

チューリッヒへは日本からノンストップ便の飛行機で12時間かかり、時差は8時間。
今日は、ブッヒー通信のホームページに来ていただき、ありがとうございました。

[hhe]

ノイエ・バッハは戦争と暴力支配の犠牲者を追悼し記念する場所である。
サンフレッチェ戦のテレビ中継で度々ケッヘル監督が映っていました。

表 A.2: ATR 503 文に含まれない低頻度 CV (本当にある?)

	っぎ ggi		っげ gge	っご ggo
っざ zza			っぜ zze	っぞ zzo
っだ dda			っで dde	
っば bba	っび bbi	っぶ bbu	っべ bbe	っぼ bbo
っら rra	っり rri	っる rru	っれ rre	っろ rro
っぎゃ ggya		っぎゅ ggyu		っぎょ ggyo
っじゃ jja		っじゅ jju		っじょ jjo
っひゃ hhya		っひゅ hhyu		っひょ hhyo
っぴゃ ppya		っぴゅ ppyu		っぴょ ppyo
っびゃ bbya		っびゅ bbyu		っびょ bbyo
っりゃ rrya		っりゅ rryu		っりょ rryo
っじゃ jja		っじゅ jju		っじょ jjo
	ッテイ tti			
			ッチェ cche	
ッファ ffa	ッフィ ffi		ッフェ ffe	ッフォ ffo
	ッウイ wwi		ッウエ wwe	ッウオ wwo
			ッジェ jje	
	ッディ ddi	ッデュ ddyu		
		ッドゥ ddu		
	ツツイ ttsi			

[hho]

オランダに生まれたゴッホは、伝道師、画商などの職をへて、1880 年より画家を志した。ゴッホ特有の画風を示すのは、印象派や日本の浮世絵の影響を受けてからである。

不足音節

[rya]

「ひとえに武略、計略、調略」とは元就の合戦哲学です。三十年戦争の軍隊は傭兵が多く、勝てば戦勝のほうびに掠奪を公認された。

[ssu]

線路沿いにまっすぐ、二つ目の止まれの標識を左に 150 メートルです。彼はぞっとして、その日は朝まで一睡も出来ませんでした。

[she]

うちの愛犬「シェクタ」の小犬の頃の写真です。ただ今、ラ・マーシェではダーツクラブのメンバーを募集しています。

[kkya]

ライバルの失脚でチャンスが巡って来たり、不思議と幸運を呼ぶ力を備えています。そこで居酒屋での接客の心得を列挙してみました。

[cchu]

でも十中八九、人生において得るものはなにもないです。しばらく会っていない人やしょっちゅう顔を合わせている人など、いろんな人に会いました。

[ccho]

私って本当におっちょこちょいのコンコンチキです。
桂小五郎や西郷隆盛らによって薩長同盟が調印された。

[we]

海外旅行に行く日本人が年々増えているが、スウェーデンを旅する人は比較的少ない。
バグのないソフトウェアはない。

[du]

アンドゥとは取り消しのことで、リドゥは同一の処理を繰り返すこと。
ボタン国営航空、「ドゥルック・エアー」のお問い合わせは旅行代理店に。

[dda]

受け取ったパケットを、ヘッダ部とデータ部に分離し、パケットの正当性を検査する。
周知の通りブッダとは仏教の開祖である釈迦如来の事である。

[tti]

「フォレスパ」とネーミングされた特製スパゲッティーがお奨めだ。
彼のバッティングには全く無駄がない。

[cche]

神戸駅周辺はグラッチェプレミオとチェリーのチェーンで占めている。

[ffe]

セーヌ左岸に立つエッフェル塔は、1889年にギュスタブ・エッフェルにより建設された。
旅の思い出を語り合う楽しいひとときを支えてくれるのが車内販売とビュッフェです。

[di]

もちろん田口ランディの生原稿もガンガン読める。
いつも、テレホーダイの時間はディアブロやっています。

[she]

限定された地域で産する葡萄を使って造られたものだけをシェリーと呼ぶことができる。
美しいセーシェルにもひとつだけ欠点がある。

A.3 促音の問題 – FTKq

既存の問題点

添付の参考資料に示すように音素バランスをとった ATR503 文には 幾つかの促音の音素が足りない. ex. /gg/, /jj/, /zz/, /hh/,...) これらの音素は XXX_synth.ch: XXX_PhoneSets で定義されていないため現在の CHATR の処理では二つの促音でない音素に分解して処理 されるため促音に必要な無音区間が挿入されないため促音に聞えない という問題がある. 仮に /gg/, /hh/ を音素として登録しても実際に ATR503 文章の DB にはこれらの音素が含まれないため, CHATR の合成処理の途中で無視される.

改良点

促音のラベルを /q/ + / 後続音素 / に変更する (一研の認識ラベルと同じ). 音素不足に効果的 (音素数の少ない小さな DB で特に有効) だが, /ssh/, /ss/ は完全な無音は無いので別扱いする.

問題点

/dd/, /ff/, /hh/, /jj/ などの場合は, 無音区間があるときとないときがある. /qj/, /qf/, /qh/, /qg/ の音も /j/, /f/, /h/, /g/ が強く聞える. /q/ の duration が長すぎることで, それぞれの促音が通常無音区間を含まない可能性がある. また Phoneme 合成のときの入力方式の仕様を決める必要がある.

関連 URL:

http://www.itl.atr.co.jp/~fujisawa/double_consonant/

付録 B

CHATR の日本語音声合成法

Text Analysis

- a kan2rom (Full automatic TTS)
KDD 製の漢字かな混じり文からアクセント付きローマ字 文字列を作成するフィルタプログラムで自動的に作成したアクセント付きローマ字文字列.
- b AUDIOTEX (NTT 情報研)
NTT 情報研の形態素解析+韻律付与プログラム. KDD kan2rom より性能がよいが, Windows 版のバイナリしかないため, ATR-MATRIX では現在使用できない.
- c gfilter
読み, 品詞, アクセント情報を TDMT の生成辞書に付与し, kan2rom のアクセント処理を組み込んだもの. 現在 ATR-MATRIX で使われている.

Prosody Prediction

- A JToBI-lr
現在 CHATR で標準に使われている ToBI ラベルによる F0 予測. LR モデルで予測を行う.
- B JToBI-rule
現在 CHATR で標準に使われている ToBI ラベルによる F0 予測. ルールに基づいて予測するが一部バグあり.
- C Fujisaki model
藤崎モデルによる F0 予測. 現在 MHT のパラメータしかないので, 話者 MHT でしか使えない.
- D test_seg
話者データベース中にある自然音声波形の F0,duration をターゲットとして単位選択する合成方法. ただし, 合成に元波形の波形素片は使わない. 自然音声の韻律をターゲットとするので品質が一番良いが, DB に無い発話は合成できない.

Unit Selection

- i UDB
現在 CHATR で標準に使われている単位選択方法.
- ii SDB
Dave Sann による単位選択方法. 音素単位候補をフルサーチするので合成に時間がかかる.
- iii FDB
ATR503 文章より, アクセント句単位でモーラ長, アクセント型, 前後の境界の深さから各アクセント句を分類し, 各モーラの F0 のレンジを設定したもの.

iv ARX

音源パラメータとフォルマントを単位選択の要素に加えたもの

Concat. Method

i DUMP+

単純に波形素片を接続する方法.

ii (Partial)PSOLA

波形接続時に PSOLA を使って F0 を修正する.

iii CEPLMA

Nuu-talk のケプストラム合成.

iv STRAIGHT

元人間情報の河原室長の開発された STRAIGHT を用いて F0 を修正. CHATR の後プロセスとして処理される. 503 文で 10GB 程度のディスク容量が必要.

v MORPH

画像処理で使われるモーフィングを波形接続時に適用.

関連 URL:

http://www.itl.atr.co.jp/~fujisawa/CHATR/synth_method/
http://www.itl.atr.co.jp/~fujisawa/CHATR/synth_method/samples/