社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（17）
2007／12／20

STRAIGHTを用いたF0モデルパラメータの変換・再合成ツールの開発

佐藤 翔太† 木村 太郎† 堀内 靖雄‡
西田 昌史‡ 黒岩 眞吾‡ 市川 熹‡
† ‡千葉大学
〒263-8522 千葉市稲毛区弥生町１－３３
E-mail: † {sato_s, kimta }@graduate.chiba-u.jp, ‡ {nishida, hory, kuroiwa, ichikawa }@faculty.chiba-u.jp

あらまし F0モデルパラメータに基づいて音声の変換・再合成を行うツールを開発した．先行研究において，発話の平均モーラ長やパワー，F0 モデルパラメータなどの韻律情報から係り受け構造，話者交代/継続の予測がある程度の精度で可能であるという結果が得られていた．これらの結果を実際のシステムに反映するためには，音声の聴取による心理実験が必要である．本研究では遺伝的アルゴリズムによる A-b-S を利用して推定された F0 モデルのパラメータを変更し，STRAIGHT によって再合成を行うツールを開発し，心理実験に使用する音声を快適に作成できる GUI 環境を構築した．

キーワード F0モデル，遺伝的アルゴリズム，STRAIGHT，GUI

Development of Transform and Re-synthesis Tool of F0 Model Parameter using STRAIGHT

Shota SATO† Taro KIMURA‡ Yasuo HORIUCHI†
Masafumi NISHIDA‡ Shingo KUROIWA† and Akira ICHIKAWA‡
† ‡ Chiba University
1-33 Yayoicho, Inage-ku, Chiba, 263-8522 Japan
E-mail: † {sato_s, kimta }@graduate.chiba-u.jp, ‡ {nishida, hory, kuroiwa, ichikawa }@faculty.chiba-u.jp

Abstract We have been developing F0 modification and re-synthesis tool of speech based on F0 model. In the preceding research, syntactic structure and turn-taking were able to be predicted by prosodic information such as average mora duration, power and F0 model parameters. To evaluate the effectiveness of this idea in actual applications, we need to perform psychological listening experiments. In this research, to realize the environment that can easily make speech samples used for listening experiments, we have been developing a tool which can freely change F0 model parameters which were automatically estimated by the genetic algorithm and can re-synthesize the speech data with changed F0 model parameters by using STRAIGHT technology.

Keyword F0 model, genetic algorithm, STRAIGHT, GUI

1. はじめに

近年，音声対話型のシステムは音声認識技術，音声合成技術の著しい発展により，アナウンサが話すような丁寧な書き言葉的発話については実用化に至る段階となった．しかしながら，話し言葉のような自然な発話に対しては，依然として問題点が多く実用化にはいたっていない．その主な原因として，韻律による情報を活かしきれていないということが考えられる．

大須賀[1] [2]らは発話の韻律情報に注目し，平均モーラ長やパワー，基本周波数（以下 F0）等の韻律パラメータと係り受け構造，話者交代/継続との関連性についてそれぞれ分析を行った．この結果から，係り受け構造，話者交代/継続が韻律パラメータからある程度予測が可能であることを示した．しかしそこでの F0 のパラメータは F0 パターンを直線近似することによって算出されており，語末にくる単語のアクセント型などが悪影響を及ぼしてしまい精度にやや難があることや，分析により得られた結果を音声合成等の他技術へ応用することが困難である等の問題があった．そこで，木村ら[3]は従来の直線近似による F0 パラメータに代え

て，F0 パターンの生成過程モデル（以下 F0 モデル）を用いて F0 パラメータを算出した．F0 モデルは藤崎らによって提案された[4]，F0 パターンをフレーズ成分とアクセント成分の２つの成分の線形和として表現するモデルであり，多くの言語に対して高い近似性能を持っている．実測された音声の F0 パターンから F0 モデルのパラメータを推定・抽出する手法としては，遺伝的アルゴリズムによる A-b-S を利用して行い，自動的にパラメータを算出できるようにした[5]．現在，この推定･抽出手法より求められたパラメータを用いて，話者交代/継続の判別を行っており，直線近似によるパラメータを用いた場合と比較して判別精度の向上が見られている[5]．

これらの分析結果を実際のシステムに反映するためには，音声の聴取による心理実験を行い，韻律パラメータと知覚との関係を明らかにする必要がある．河原らによって音声知覚の研究用に開発された STRAIGHT[6] [7]は自然音声を知覚的に意味のあるパラメータとして分析し，変換を行った後に再合成する分析変換合成型のシステムであり，高品質の合成音声を作成できることが知られている．

本研究においては，韻律パラメータの心理実験に使用する音声の作成を支援するため，遺伝的アルゴリズムによる A-b-S を利用して自動的に推定された F0 モデルパラメータを変更し，STRAIGHT を用いて再合成を行うツールの開発を行う．

2. F0 モデル

F0 モデルは藤崎らによって提案された F0 パターンを数学的に説明するモデルであり[4][8]，フレーズ成分とアクセント成分という 2 つの成分の線形和によって構成されている．フレーズ成分は発話頭から発話末にかけて緩やかに減衰する成分であり，インパルス応答の形で記述される．またアクセント成分は局所的に上昇下降する成分であり，ステップ応答の形で記述されている．F0 モデルは(1)式のように記述される．

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\left\{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\right\} \quad (1)$$

ここで，F_b は F0 の基底値であり，話者ごとのベースとなる F0 値を示す．A_p はフレーズ指令（インパルス）の大きさ，A_a はアクセント指令（ステップ）の大きさであり，T_{0i} は i 番目のフレーズ指令の生起時点，T_{1j} は j 番目のアクセント指令の始点，T_{2j} は j 番目のアクセント指令の終点である．また，(1)式内の G_p，G_a はそれぞれフレーズ制御機構，アクセント制御機構の関数であり(2)式，(3)式によって記述される．

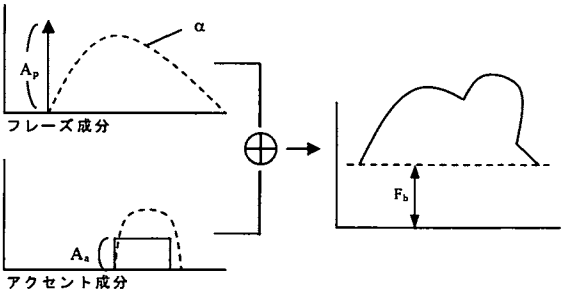

図 1　F0 モデル

$$G_{pi}(t) = \begin{cases} \alpha_i^{\,2} t e^{-\alpha_i t} & : t \geq 0 \\ 0 & : t < 0 \end{cases} \quad (2)$$

$$G_{aj}(t) = \begin{cases} \min\left[1 - (1 + \beta_j t)e^{-\beta_j t}, \gamma\right] & : t \geq 0 \\ 0 & : t < 0 \end{cases} \quad (3)$$

ここでαはフレーズ制御機構の固有角周波数でありフレーズ成分の減衰の速さを，βはアクセント制御機構の固有角周波数でありアクセントの上昇下降の早さを決定するパラメータである．

3. 遺伝的アルゴリズムによる F0 モデルパラメータの推定[5]

木村らによって提案された F0 モデルパラメータの推定手法について説明する．F0 モデルパラメータであるαとβは，発話ごとの変動が比較的小さいとされており，α=3.0，β=20.0 という値を用いることが多い．しかし，木村らはその変動にも注目し，αおよびβの値を変数として推定を行った．F0 モデルパラメータの推定を行う場合，非常に多くのパラメータを有した最適化問題を解くことが要求される．F0 モデル内の各パラメータはそれぞれ相互に影響を与え合うため，全てのパラメータを同時に変動させるような最適化法が適している．そこで木村らは問題の最適化法として，遺伝的アルゴリズムによる A-b-S を提案した．

3. 1 遺伝的アルゴリズムによる A-b-S

遺伝的アルゴリズムではパラメータをビット列化したものを遺伝子とみなし，この遺伝子を進化させていくことによってパラメータの最適化を行う．実音声の F0 パターン[$F_0(t)$]を環境とし，遺伝子の持つパラメータ情報から F0 モデルによって計算される F0 パターン[$F_m(t)$]との平均二乗誤差の逆数をその遺伝子の適応度とする（(4) 式）．より最適解に近いパラメータ情報を有する遺伝子ほど平均二乗誤差が小さくなるため，適応度は高くなる．

$$adp=\left[\frac{1}{N}\sum_{t}^{N}\left(F_0(t)-F_m(t)\right)^2+1\right]^{-1} \quad (4)$$

遺伝的アルゴリズムは選択・交叉・突然変異・エリート戦略の4ステップによって構成される.

(1) 選択

全遺伝子の中から, 適応度に比例した確率で2本の遺伝子を選択する. このステップによって環境に適応しない遺伝子は淘汰されるため, 最適解に近づく方向に進化を進めることが可能である.

(2) 交叉

選択された2本の遺伝子をある一点で切断し, 各々の前後を交換し, 新たな2本の遺伝子を生成する. 両親のパラメータ情報を継承して, なおかつ新しいパラメータ情報を持つ遺伝子が生成される.

$$\begin{matrix}00001|111\\01010|101\end{matrix}\Rightarrow\begin{matrix}00001101\\01010111\end{matrix}$$

(3) 突然変異

交叉後の遺伝子内のビットに対して, 低確率で0→1, 1→0の変化を起こす.

(4) エリート戦略

各世代において最高適応度をもつ遺伝子を無条件に次世代に継承する. これにより, その時点で最良の個体は交叉や、突然変異によって破壊されない.

第1世代N本の遺伝子に対して以上の4ステップをN/2回繰り返すことにより, 新たにN本の遺伝子を生成する. それらを第2世代の遺伝子N本として再びステップを繰り返すことにより, 第3世代の遺伝子N本を生成する. 以下, 第M世代まで進化させることによってパラメータの最適化を図る.

3. 2. フレーズ・アクセント数の決定

自発音声を分析対象とする場合, 同じテキストの発話であっても話者の癖や対話状況などによって F0 パターンが一意であるとは言いがたく, 言語情報からフレーズ・アクセント成分の数を判断することは難しい.

そこで, 木村らはパラメータ推定時に初期値として定められていたフレーズ・アクセント成分の数を変動させ繰り返し推定を行うことによって, 言語情報を用いず最適な各成分の数を検索する手法をとった.

A-b-S によるパラメータ推定法では, 発話の時間長に応じてフレーズ・アクセント成分の数を初期値として設定しパラメータを推定する. 推定したフレーズ・アクセント成分が無音区間にある場合, 最小二乗誤差が計算されず適応度に影響を与えない. そのため, フレーズ・アクセント成分の存在の有無はパラメータの遺伝ステップに影響を与えず, 不要な成分が無音区間に動かされるように進化することがある. 各フレーズ・アクセント成分のパラメータの中で上記のような進化をしたパラメータが含まれている場合, 該当するフレーズ・アクセント成分を「不要」と判断する. フレーズ・アクセント数を決定する具体的な手順を以下に示す.

(1) 発話の時間長に応じてフレーズ・アクセント成分の数を初期値として設定しパラメータを推定する.

(2) 不要なフレーズ・アクセント成分を判定し, 不要とされるフレーズ成分が無くなるまで除外していく.

(3)(2)で不要な成分が無いとされた場合, フレーズ, アクセント成分の数を増加して再び推定を行う.

(4)(3)の結果より, 増加した分の成分が不要だと判定された場合は, 余分な成分も足りない成分も無いと判断して成分の数を決定し最終推定に進む. 増加した分の成分が不要と判定されなかった場合には (2) に戻り再度推定を行う.

4. STRAIGHT

STRAIGHT は図2に示すように, スペクトル情報抽出部, 音源情報抽出部, 合成部の3つの機構から成る. 図の左から入力された音声は, 音韻性に関わる成分(スペクトル包絡から構成される時間周波数表現), 韻律に関わる成分 (基本周波数と有声/無声等の情報) に分解され再合成される. この時各パラメータを調整することも可能である. 合成音声の駆動音源には, 通常のパルス列に代えて群遅延を操作することによって合成音特有のバス臭さが軽減されており, 高品質な合成音声を作成できる.

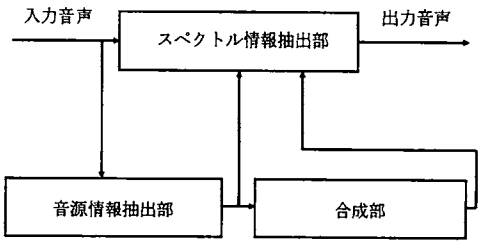

図2 STRAIGHT の構成

5. F0 モデルの変換・再合成ツールの開発

開発するツールについて以下に説明する．木村らによって行われた話者交代/継続の判別では韻律パラメータとして，AP 時間長，平均モーラ長，パワー近似直線勾配，パワー近似直線始端値，パワー最大値，F0 最大値および F0 モデルパラメータが用いられた．判別結果において有効であった韻律パラメータとしては，平均モーラ長，パワー勾配，F0 モデルパラメータが挙げられ判別に与える影響が大きいことが確認されている．

これらの分析結果を音声合成等の技術に応用するには音声の聴取による心理実験を行い，韻律パラメータが知覚に与える影響を明らかにする必要がある．心理実験においては韻律パラメータを変更した合成音声を使用し，知覚との関係を評価する．そのため、実験で使用される音声は聞き手に不自然な印象を与えるものであってはならない．このような要求から今回開発するツールは以下の機能を有するものとする．

・入力音声から F0 モデルパラメータを自動推定できる．

・音声を韻律パラメータに応じて自由に変更できる．

・韻律パラメータを変更して，再合成し，すぐに音声として確認できる．

今回は開発の初期段階として変更を加えることのできるパラメータは F0 モデルパラメータ全てと平均モーラ長（話速）とした．

5. 1. ツールの構成

図 3 にツールの構成を示す．ツールの開発は主に STRAIGHT の動作環境である MATLAB で行うが，F0 モデルパラメータの自動推定プログラムに関しては高速化のため C 言語によって記述している．入力から再合成までは以下のステップによって実現される．

(1) F0 の抽出

STRAIGHT の F0 抽出関数を用いて，入力音声の F0 を抽出する．

(2) F0 モデルパラメータの推定

(1)で抽出した F0 を環境とし，遺伝的アルゴリズムによる A-b-S を用いた F0 モデルパラメータの自動推定プログラムでパラメータを推定する．

(3) F0 モデルパラメータの変更

推定された F0 モデルパラメータに対して変更を行い，そのパラメータより求めた F0 パターンで音声の F0 を置き換える．

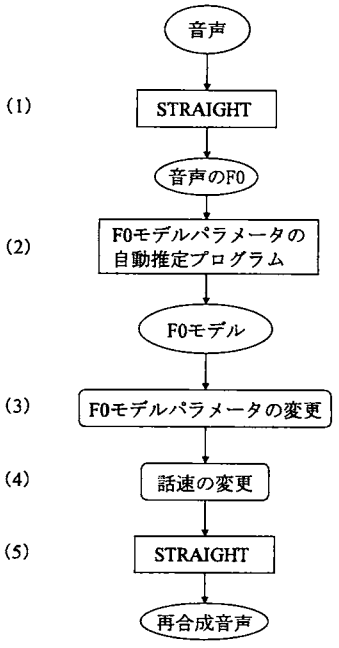

図 3　変換・再合成ツールの構成

(4)話速の変更

STRAIGHT の音声合成関数の既定値を変更することで，時間軸の伸縮を行い，話速を変更する．

(5)音声の再合成

変更した F0 パターン，話速によって STRAIGHT で音声を再合成する．

5. 2. GUI の作成

音声の作成を快適に行うため GUI を作成した．外観を図 4 に，操作パネルを図 5，図 6 に示す．

GUI の操作方法は，まず，open ボタンで入力する音声を選択し，estimate ボタンをクリックすることで，F0 モデルパラメータが推定され画面に F0 モデルより求めた F0 パターンと入力音声の F0 が描画される．ここで F0 パラメータの変更を行う場合は変更したいパラメータをリストボックスから選択し値を入力，transform をクリックすることで F0 モデルが再計算され，変更後の F0 パターンを画面上で確認することができる．また，スライドバーを動かすことで話速を設定できる．resynthesis をクリックすることで変更した F0 パターンで再合成し，play で音声として確認する．ここでは，元の音声も再生でき，作成した音声との印象の違いを比較できる．最後に save をクリックすることで wav 形式のファイルとして作成した音声を保存する．

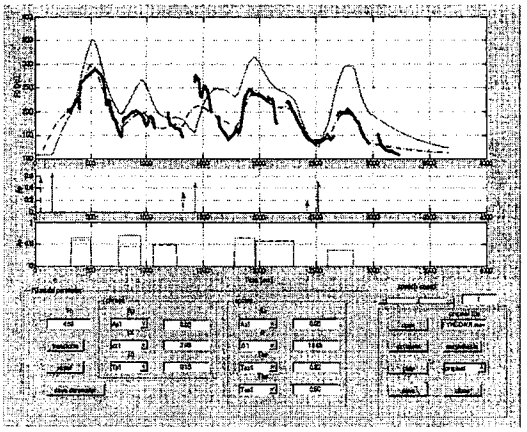

図 4　GUI の外観

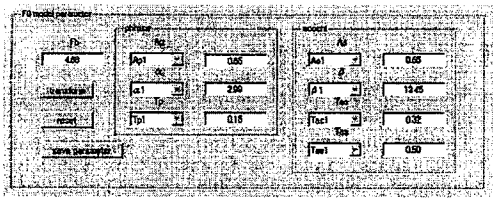

図 5　F0 モデルパラメータの操作パネル

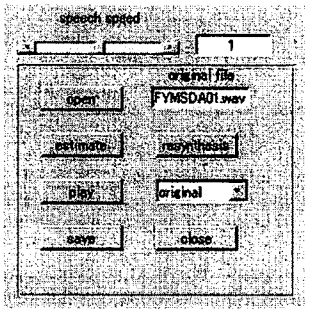

図 6　音声入力・再合成パネル

5. 3. F0 モデルパラメータの推定結果

遺伝的アルゴリズムによる A-b-S を用いた F0 モデルパラメータの自動推定プログラムによる推定結果を図 7, 図 8 に示す. 入力音声は ATR503 文コーパスの男性による 2 発話を用いた. ●が実音声の F0, 実線が推定 F0 パターン, ↑がフレーズ指令, □がアクセント指令を表している.

結果を見ると, 図では有声/無声の変わり目で若干の誤差が見られるが, どちらの場合も実音声の F0 に対して推定 F0 パターンが比較的良く近似されていることが確認できる. これより, STRAIGHT によって抽出した F0 から言語情報を用いず自動で F0 モデルパラメータの推定が可能であることが示された. 入力される音声によっては推定精度の悪い結果になることも考えられるが, 開発したツールを用いてパラメータの変更を加えることで推定結果を修正できる.

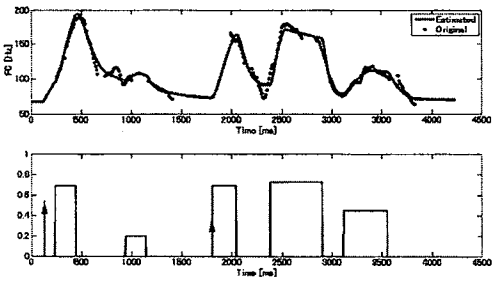

図 7　F0 モデルの推定結果 1
「あらゆる　現実を 全て　自分のほうへ ねじまげたのだ」

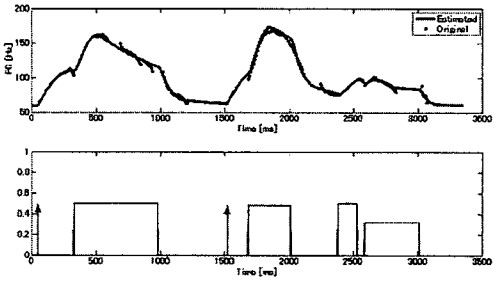

図 8　F0 モデルの推定結果 2
「一週間 ばかり　ニューヨーク を 取材した」

5. 4. F0 モデルパラメータの変更

F0 モデルパラメータの変化に応じて F0 パターンが変形される様子を以下に示す. 元の音声には前節で使用した ATR503 文中の一文「一週間ばかりニューヨークを取材した」を用いた. 図 9, 図 10 にフレーズ指令, アクセント指令のパラメータをそれぞれ変更した結果を示す. ここでは, 点線がオリジナルデータ, 実線が変換後のデータを表す.

・フレーズ指令の変更（図 9）

第 1 フレーズ指令に対しては大きさ A_{p1} を 0.5 から 0.8 と増大させ, 開始時間 T_{01} を 0.05 から 0.2 と遅らせた. また、第 2 フレーズ指令の減衰の速さαを 3.51 から 2 と変更した. 図 9 をみると, A_{p1} の影響で波形の最大値が大きくなっており, フレーズ指令の始まりが遅くなっているのがわかる. また, 第 2 フレーズ指令に対応する F0 パターンを見ると, 減衰が緩やかになっているのがわかる.

・アクセント指令の変更（図 10）

第 1 アクセント指令の大きさ A_{a1} を 0.5 から 0.8 とし，第 2 アクセントの上昇下降の早さ β_2 を 23.41 から 10 とした．第 3 アクセント指令の開始時間 T_{13} を 2.38 から 2.2 と早くしており，第 4 アクセント指令では大きさ A_{a4} を 0.32 から 0.8 と大きくしたうえで，さらに，開始時間 T_{14} を 2.59 から 2.7，終了時間 T_{24} を 3.01 から 2.9 とした．図 10 を見ると，第 1 アクセント指令 A_{a1} を大きくした影響から，アクセント指令と対応する箇所の F0 パターンが大きくなっている．また，第 2 アクセント指令では β_2 を小さくしたので応答が緩やかになっているのが見てとれる．第 3 アクセント指令では開始時間を早めたため，指令区間が広くなっている．第 4 アクセント指令では逆に指令区間は狭くなっており，A_{a4} を変更した影響とあわせて F0 パターンが急激な変化をしているのがわかる．

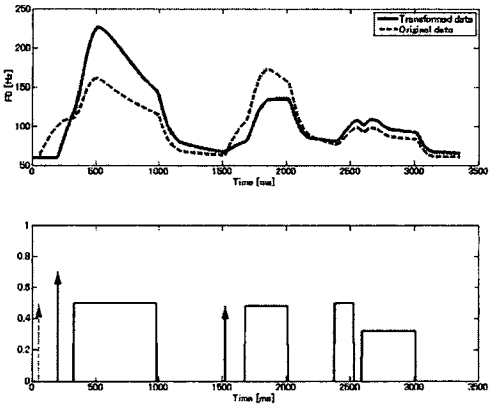

図 9　フレーズ指令の変更例

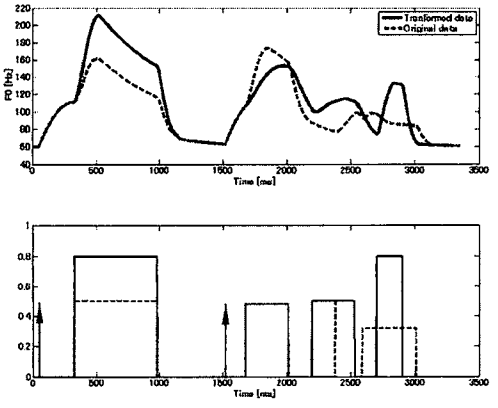

図 10　アクセント指令の変更例

以上のことから，開発したツールが F0 モデルパラメータを自由自在に変更でき，それに応じた音声を再合成し，変更した音声をすぐに確認できるようになった．

6. まとめと今後の予定

遺伝的アルゴリズムによる A-b-S を利用して，自動的に推定された F0 モデルパラメータと時間長を自由自在に変更し，その結果の F0 パターンに基づいて STRAIGHT を用いて再合成可能なツールの開発を行った．

今後は，今回変更するパラメータに加えなかったポーズやパワー等についても変更できるようツールの改良を行いたい．さらには，ツールを用いて韻律パラメータを変換した音声を作成し，人間が実際に係り受け構造，話者交代/継続を予測する際にどのような韻律パラメータが有効であるかについて検討を行いたい．

文　献

[1] 大須賀智子，堀内靖雄，西田昌史，市川熹，“音声対話での話者交替/継続の予測における韻律情報の有効性，” 人工知能学会誌， Vol.21，No.1，pp.1-8，2006.

[2] Tomoko Ohsuga, Yasuo Horiuchi, Akira Ichikawa, “Estimating Syntactic Structure from Prosody in Japanese Speech” IEICE Transactions on Information and Systems,Vol.E86-D,No.3, pp.558-564 2003

[3] 木村 太郎，西田 昌史，堀内 靖雄，市川 熹，“遺伝的アルゴリズムによる F0 モデルパラメータ推定手法と話者交替分析への適用，” 信学技報，SP2006-82，pp.37-42，2006.

[4] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” Jour. Acoust. Soc. Jpn. (E), Vol.5, No.4, pp.233-242, 1984.

[5] 木村 太郎，堀内 靖雄，西田 昌史，市川 熹，“F0 モデルを用いた日本語対話音声における韻律と話者交代の分析，”信学技報，SP2007-75，pp.25-30，2007.

[6] Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveigne,”Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction” Possible role of a reptitive structure in sounds, Speech Communication, 27, pp.187-207 (1999).

[7] 河原英紀：Vocoder のもう一つの可能性を探る--音声分析変換合成システム STRAIGHT の背景と展開--，日本音響学会誌，Vol.63，No.8，pp.442-449 (2007).

[8] 成澤修一，峯松信明，広瀬啓吉，藤崎博也，“基本周波数パターン生成過程モデルのパラメータ自動抽出とその評価，” 信学技報，SP2002-27，pp.19-24，2002.