係り受けと F_0 の関係に着目した波形接続型音声合成における単位選択の改良の検討

藤井慶 [†1]　柏岡秀紀 [†1,†2]　ニック・キャンベル [†1,†3,†4]

波形接続型音声合成の単位選択コストの精度向上は合成音の品質向上につながる．韻律に関するターゲットコストは，ターゲット韻律と音声単位候補の韻律の差から求められる．このとき参照するターゲット韻律は，音韻の種類やアクセント型，係り受けなど様々な言語的要因を説明変数として求められたものである．本稿では合成音の品質向上を目的とし，係り受けと F_0 の関係 (直後の句に係るか否かで後続句の F_0 が上昇/下降するという傾向) をコストに取り入れることを考える．まずその動機について述べ，コストの設計，評価について述べる．2 つの評価実験の結果，(1) 係り受けの伝達は従来法のみで十分な精度を持つ，(2) 提案コスト導入で自然性がより向上するとの結果を得た．

Unit Selection Based on the Relation Between Dependency Structure and F_0 for Concatenative Speech Synthesis

Kei FUJII[†1]　Hideki KASHIOKA[†1,†2]　Nick CAMPBELL[†1,†3,†4]

We propose a more precise cost function for concatenative speech synthesis to improve the quality of synthetic speech. The F_0 target cost is one of the most important unit-selection costs for Japanese speech synthesis. It is calculated as the distortion between the F_0 of a candidate unit and that of the target which is generated by the prediction model from various linguistic factors. In this paper we propose use of the relation between the F_0 and the phrase dependency structure as an improved cost function. Two kinds of evaluation experiment showed that (a) both the conventional cost and the proposed cost functions result in synthetic speech with prosody correctly expressing the dependency structure (using a 40-minute source corpus), but (b) the proposed cost results in higher naturalness.

1　はじめに

音声言語は言語的情報以外にパラ言語的情報や非言語的情報を含んでおり，文字上は同一の文でも様々な意図を付与して発話し聞き手へ伝達することが出来る [1]．韻律的特徴はこのような発話の実現において大きな役割を果たしており，円滑な対人あるいは対機械コミュニケーションのために韻律情報の利用は重要であると考えられている．

韻律の制御と言語的情報は深く関わっており，これまでに F_0 および音韻セグメント長に影響を与える言語的要因は [2] で，パワーに影響を与え得る要因は [3] で整理されており，これらの知見はテキスト音声合成の韻律予測モデルに応用されている．

信号処理を用いない波形接続型音声合成 [4] は，話者の音声波形をある単位で分割し合成の際にそれらの波形を接続して出力することで音声の自然性や話者性を保存する合成方式であり，一般に，言語解析，韻律予測，単位選択，波形接続の順に処理を行う．

韻律予測部では，言語解析部で得られた情報を韻律予測モデルに入力し，[2][3] で述べられる種々の要因を説明変数として合成音のターゲット韻律を生成する．

単位選択部ではコーパス内を探策し，コスト最小となる基本音声単位列を探す．コストはターゲット

[†1] 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Sciense, Nara Institute of Science and Technology
[†2] ATR 音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Laboratories
[†3] ATR 人間情報科学研究所
ATR Human Information Science Laboratories
[†4] 科学技術振興機構/戦略的基礎研究推進事業
JST/CREST Expressive Speech Processing

コストと接続コストに大別され, ターゲットコストではターゲット情報に対する歪みの度合いを, 接続コストでは接続音声単位間の歪みの度合いを定量化している. 韻律に関する従来のターゲットコストは 3 種のサブコスト (ターゲットに対する F_0 差, デュレーション差, パワー差) に分けて計算される. このとき参照するターゲット韻律は前段で生成したものである (図 1).

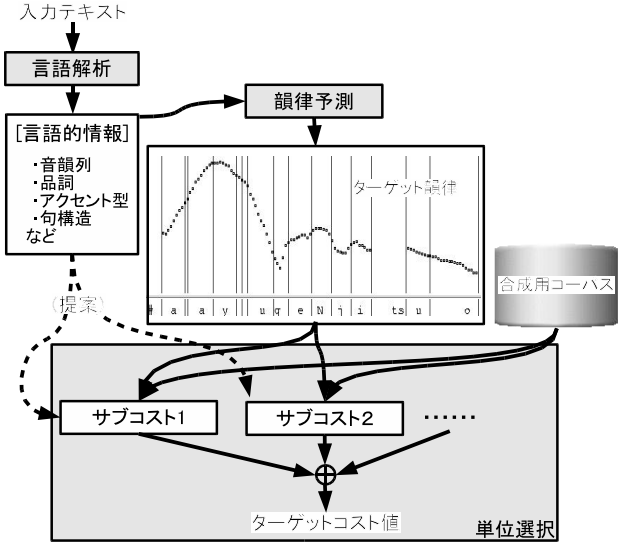

図 1: 韻律に関するターゲットコスト計算

図より, ターゲット韻律は種々の言語的情報の影響を統合したものであり, 従来用いている (韻律に関する) ターゲットコスト値は伝達したい言語的情報を統合した値に対する距離と考えられる.

単位選択に関する改良のアプローチとして従来の韻律ターゲットコストを, 伝達したい情報毎に分解するやり方が考えられる. これは従来法の情報の流れ (図 1 の実線) に対してさらに破線で示した流れを追加し, 破線と実線を通して得られる情報を用いたコストを設計し利用することとなる. 例えば [5][6] では種々の言語的情報のうちアクセント型と F_0 の関係をコスト化することで品質向上を試みているが, これは言語解析で得られたターゲット情報のうちアクセント型をコスト計算に直接反映させる試みである. 伝達したい情報毎にコストを分けて計算するメリットとして以下のことが挙げられる.

1. 音質の向上が期待出来る.
   情報毎にコストを分けることで, 品質劣化のふるまいにより忠実なコスト関数を設定出来, その結果合成音の音質が向上する.
2. 情報毎に重み付けが出来るようになる.
   テキストのみでは複数の意味解釈 (係り受け先の違いなど) や意図解釈 (平叙/疑問など) が可能な文を合成する場合には, 聞き手に誤解が生じないよう, それらに関するコストを優先して単位選択を行うことが出来る. またそれに関する処理や規則化をより直感的に行えるようになると考えられる.

そこで本報告では種々の言語的要因のうち句の係り受け構造と F_0 の関係を新たにコスト化することで合成音の品質向上を目指し, コストを設計, 評価した結果について述べる.

2 係り受けと F_0 に関するコスト

本節では係り受けと F_0 の関係のコスト化について説明する. まず今回着目した係り受けと F_0 の関係について説明し, 次に設計したコストについて説明する. なお本稿での単位選択はビームサーチを併用した動的計画法を用いている.

2.1 着眼点

係り受けの情報は句全体の F_0 に影響しており, 多くの場合において, 先行句が隣接する後続句に係る場合には後続句の F_0 が下降し, 係らない場合には上昇することが知られている [2]. 文献 [8] では決定木を用いて隣接句間の修飾関係の有無を判別している. この手法では言語的情報を説明変数に用いず韻律情報のみ (句の平均 F_0 とポーズ長) を用いているが, 朗読音声においては約 75 %, また同一文字表記で意味の異なる文の朗読において約 92 % の精度を得ている.

これらのことから隣接句間の係り受け有無と F_0 の関係を単位選択のコストに取り入れることを考えた. すなわち当該句が直前の句から修飾されている (いない) 場合には, ある基準より高い (低い) F_0 の単位候補に より大きな値を課すようなコストである.

このようなコストを導入する利点の単純な例として, 先行句が直後の句を修飾する場合の単位選択例を図 2 に挙げる. なお図のパス候補 1 と 2 の各音声単位は, F_0 以外の特徴はほぼ同じ値を持つものとする.

図より, パス候補 1 はパス候補 2 に比べ総じて

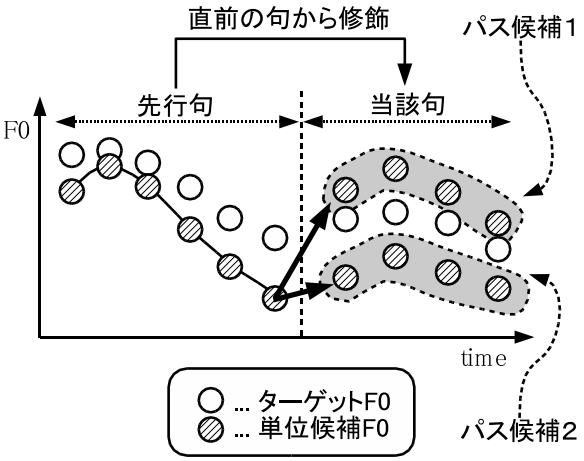

図 2: 従来の F_0 ターゲットコストのみでは問題が生じる例

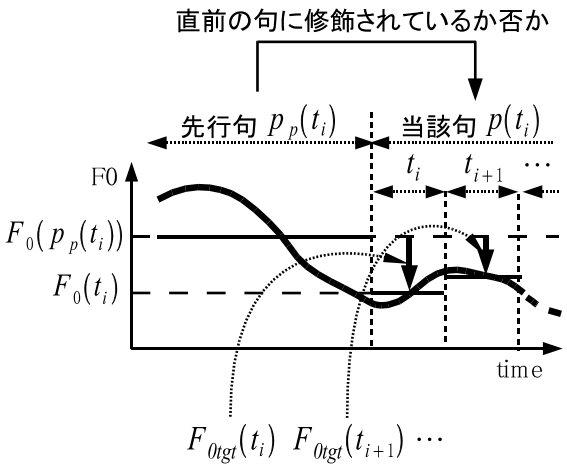

図 3: ターゲット特徴量の抽出

ターゲット F_0 に近いため, 従来の F_0 ターゲットコストのみではパス候補 1 の音声単位列が選択される. しかしながら, パス候補 1 の F_0 はターゲット F_0 よりおしなべて高く, 先行句の単位候補はターゲット F_0 よりも低いため, パス候補 1 を選択すると, 先行句から当該句にかけて F_0 の立て直しのある合成音となり, 不自然なイントネーションとなる恐れがある. またもしこの合成ターゲットが, テキストのみでは係り受けを一意に決められない文であった場合, パス候補 1 の音を出力すると聞き手を誤解させてしまう可能性がある. 一方提案コストはパス候補 1 により大きなコスト値を与えるため, 提案コストを導入することでパス候補 2 を選択することが出来る. これは [6][7] の論旨 (ある予測ターゲットに最も近い単位列が必ずしも最適ではなく, 若干離れた単位列の方がより望ましい場合があるという主張) と同様の観点である.

なお係り受け判別にはポーズ挿入位置も大きな役割を果たすが, それは今後の課題とし, 今回は F_0 のみを考慮して設計, 評価を行った.

2.2 提案コストの計算法

ここでは本稿で提案する係り受けと F_0 に関するコストの計算手順を説明する. まずターゲットに関して, i 番目のターゲット音声単位を $t_i(1 \leq i \leq N_i)$ とし, t_i の平均 F_0 を $F_0(t_i)$, t_i が属する句を $p(t_i)$, $p(t_i)$ の直前の句を $p_p(t_i)$ とおく (図 3).

単位選択を行う前に, 各 t_i について次のターゲット情報を用意する.

- $p_p(t_i)$ からの係り受けの有無
- 先行句の平均 F_0 ($F_0(p_p(t_i))$ とおく) に対する t_i の平均 F_0 の相対的な高さ:
  $F_{0tgt}(t_i) = F_0(t_i) - F_0(p_p(t_i))$

単位選択における, i 番目ターゲット t_i に対するある音声単位候補 u_{ij} ($0 \leq j <$ 候補数) のコスト計算は次のように行う.

1. 直前の句に属する音声単位候補列の F_0 の平均 ($F_0(p_p(u_{ij}))$) を計算
2. 先行句に対する u_{ij} の相対的な高さを計算:
   $F_{0cand}(u_{ij}) = F_0(u_{ij}) - F_0(p_p(u_{ij}))$
3. コスト計算:
   - 直前の句から修飾されている場合:
     $$Cost(u_{ij}) = \left\{ \begin{array}{l} F_{0cand}(u_{ij}) - F_{0tgt}(t_i) \\ \quad \text{if } F_{0cand}(u_{ij}) > F_{0tgt}(t_i) \\ 0 \quad \text{otherwise} \end{array} \right\} \quad (1)$$
   - 直前の句から修飾されていない場合:
     $$Cost(u_{ij}) = \left\{ \begin{array}{l} F_{0tgt}(t_i) - F_{0cand}(u_{ij}) \\ \quad \text{if } F_{0cand}(u_{ij}) < F_{0tgt}(t_i) \\ 0 \quad \text{otherwise} \end{array} \right\} \quad (2)$$

ここで $F_{0tgt}(t_i)$ と $F_{0cand}(u_{ij})$ はいずれも先行句の平均 F_0 を基準とした当該音声単位 F_0 の相対的な高さを表している. ただし $F_{0cand}(u_{ij})$ では, u_{ij} に至る音声単位列から, 基準となる先行句の平均 F_0

$(F_0(p_p(u_{ij})))$ を改めて算出する．このことからこのコストのみの下では，句の F_0 の相対的な上下関係を保つ限りにおいて，合成音の F_0 値はターゲット F_0 値に必ずしも一致しなくても良いことになる．なお，t_i あるいは u_{ij} が無声の場合，および $p(t_i)$ が先行句を持たない場合は $Cost(u_{ij}) = 0$ とした．

コスト計算例として図 4 に示すような，ある候補単位列のパスに u_{i0} を追加する場合および u_{i1} を追加する場合を考える．図では直前の句が当該句に係っているが，u_{i0} については $F_{0tgt}(t_i)$ ほど F_0 が下降していない $(F_{0cand}(u_{i0}) > F_{0tgt}(t_i))$ ため，その差分が提案コスト値となり，単位選択の総コストに加算される．一方 u_{i1} は $F_{0cand}(u_{i1}) < F_{0tgt}(t_i)$ なので提案コスト値は 0 となる．よって $Cost(u_{i0}) > Cost(u_{i1})$ となり，提案法を導入すると u_{i1} の方がより選ばれやすくなる．

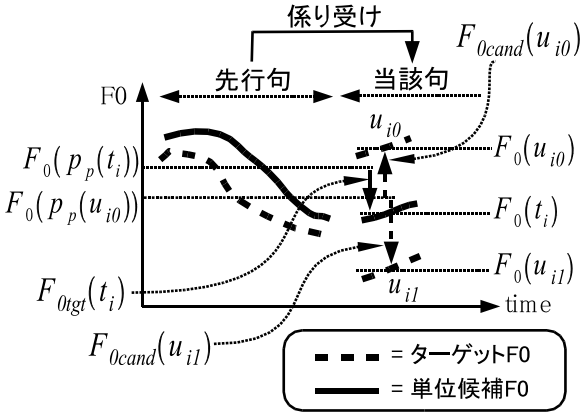

図 4: コスト計算例

今回考案したコストでは音声単位候補毎にコストを発生させている．だが [8] では先行句の平均 F_0 と当該句の平均 F_0 の差を判別に用いており，このことを忠実にコスト化するならば句毎にコストを発生させるべきである．筆者らも当初は，句境界の音声単位候補でのみ先行句と当該句それぞれに属する音声単位候補列の平均 F_0 を求め，コストを発生させようと考えた．しかし従来法を用いた幾つかの合成結果の句境界上で，総コストの低い順に 50 位までの単位候補列の傾向を調べたところ，候補列間で異なるのは数個の単位のみで，その他は同じ単位候補が選択される場合が多かった．このことから，句境界でのみコストを発生させるようにすると，例えば図 2 でのパス候補 2 は句境界に至る前に単位選択のビーム幅の外へ出てしまう恐れがある (本稿の単位選択ではビームサーチを用いているため) と考え，今回は上記のようなコストを設計した．

3 評価実験

2.2 節で述べたコストの有用性を確かめるため，合成音の聴取による評価実験を 2 種類行った．まず係り受けの判別率に関する実験について述べ，次に合成音の自然性に関する実験について説明する．

3.1 実験 1 (係り受けの判別率)

3.1.1 実験内容

提案コストを導入することで係り受けに関してより誤解のない合成音が得られるかを評価するため聴取実験を行った．

まず合成ターゲットとして，同一の文字からは複数の意味解釈が出来る文を 4 文作成した (表 1)．例えば「大きな靴の穴」では，靴が大きいという解釈 (図 5 (a)) と穴が大きいという解釈 (図 5 (b)) が出来る．これらの文を男性話者 1 名が意味毎に発声し，各々に対して WaveSurfer[11] を用いて F_0，パワーを抽出し，人手でラベリングを行った．またこの文の話者と合成用コーパスの話者は異なるため，両者の F_0，デュレーション，パワーの平均を求め，合成用コーパスの平均に一致するように文の話者の各値をシフトしたものを合成ターゲットとした．

表 1: 実験に用いた文

1. 大きな靴の穴
2. 西野と 3 回も優勝した山岡
3. 私はあわてて走る子供を追った
4. 私は山田と飲むのが好きな鈴木に会った

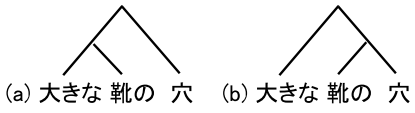

図 5: 係り受けの例

次に従来法，提案法それぞれで単位選択を行い，合成音を作成した．今回用いた合成用音声コーパスは男性話者による ATR503 文の朗読音声であり，半音素を基本音声単位としてコーパスを構築した．音声単位数は約 58,000 である．ここで従来法で用いたコストは，音韻環境，ターゲットコスト : (F_0, F_0 傾き [9], デュレーション, パワー), 接続コスト : (F_0, パワー, MFCC) であり，rms コスト [10] 化したものである．この従来法に 2.2 節で述べたコストを追

加したものが提案法である.

実験では合成テキストと解釈可能な意味 (2 通り) を被検者に提示した上で音声を聞いてもらい, どちらに解釈したかを回答してもらった. 受聴する音声は原音声, 従来法, 提案法の 3 種であり, よって試料数は 24 文である. 試料の順序は文毎にランダムに並び換えており, 被検者は判定を下すまで任意に聞き直せるものとした. なお web を通して実験を実施したため, 再生機器や場所は統一されていない.

3.1.2 実験結果

実験結果を表 2 に示す. 表より, 従来法の判別率は 81.3%, 提案法は 84.4% であり今回の実験では手法間の差がほとんどなく, またいずれも原音声とほぼ同じ判別率を得ていることが分かる. 従来法と提案法で同じ音声単位列が選択されているか確認したところ, どのターゲットでも手法間で異なる単位列が選ばれていた. これらのことから今回用いた合成用コーパスは, 係り受けの正確な伝達に関しては既に十分なサイズであった可能性が考えられる. よって係り受け判別精度の評価の観点からは, 今後コーパスサイズの削減や合成試料数の増加を行った再実験, あるいはコスト値と判別率の相関を調べることが考えられる.

3.2 実験 2 (合成音の自然性)

3.2.1 実験内容

提案コストを導入することで音質向上が得られるかを評価するため聴取実験を行った.

合成ターゲットとして, 合成用コーパス中にある自然音声の音韻および韻律, 係り受け情報を取り出して用い, 従来法および提案法で単位選択を行った. その際にターゲットとなる音声単位自身は選択候補から除外した, 実験に用いたコーパス, コスト等は 3.1 節の実験と同様である.

従来法と提案法の単位選択結果が同一の文を除いた上で, 20 文の合成音声 40 個を実験試料とした. 被検者には従来法・提案法の合成音を聞いてもらい, より人の発声に近い合成音を選んでもらった. 各合成音の順序はランダムに入れ替えており, 判定を下すまで任意に聞き直せるものとした. また本実験も 3.1 節の実験同様 web ページを用いて実験を実施したため, 再生機器や場所は統一されていない.

表 2: 被検者毎の選択率

文	意味	音	被検者毎の回答 A	B	C	D	正解率 (%)
1	1	従来法	○	○	○	○	100
		提案法	○	○	○	○	100
		原音声	○	○	○	○	100
	2	従来法	○	○	○	○	100
		提案法	○	○	○	○	100
		原音声	○	○	○	○	100
2	1	従来法	○	○	×	×	50
		提案法	×	○	○	×	50
		原音声	×	○	○	○	75
	2	従来法	○	×	○	×	50
		提案法	○	○	×	○	75
		原音声	×	○	○	○	75
3	1	従来法	○	×	○	○	75
		提案法	○	×	○	○	75
		原音声	○	×	○	○	75
	2	従来法	○	○	○	×	75
		提案法	×	○	○	○	75
		原音声	○	○	×	×	50
4	1	従来法	○	○	○	○	100
		提案法	○	○	○	○	100
		原音声	○	×	○	○	75
	2	従来法	○	○	○	○	100
		提案法	○	○	○	○	100
		原音声	○	○	○	○	100
合計		従来法	100	75	87.5	62.5	81.3
		提案法	75	87.5	87.5	87.5	84.4
		原音声	75	75	87.5	87.5	81.3

表 3: 被検者毎の選択率

選択肢	被検者毎の選択率 (%) A	B	C	D	E	F	G	全体 (%)
従来法	30	55	25	20	30	10	30	28.6
提案法	60	45	75	40	60	50	50	54.3
優劣無し	10	0	0	30	10	40	20	15.7

3.2.2 実験結果

実験結果を表 3 に示す. 全体では 28.6% の割合で従来法が, 54.3% の割合で提案法が選択されており, 25.7% 多く提案法が選択された. また被検者毎に見ると被検者 B 以外は全員提案法の選択率が従来法を上回っている.

4 考察

3.2 節の実験は特に F_0 のみの評価として行ったのではなく総合評価として行った. これは, 今回の提案コストは F_0 に関するものであるが, コスト導入による音質の変化は F_0 のみとは限らないと考え

たためである. すなわち単位選択は各種コストの全体の最小化に基づくものであり, あるコストのふるまいを変えることで別の音声単位列が選択されれば, 他のコストの値についても違ったものが得られるためである.

また判別率についての評価 (3.1 節) では原音声, 従来法, 提案法それぞれに差がほとんどなく, 今回用いた合成用コーパスサイズでも係り受けを正確に伝達させることが可能であるとの結果を得た. これら 2 つの実験結果より, 今回用いたコーパスでは係り受け伝達に関しての手法間の差はないものの, 提案法の方がより自然性の高い合成音を得られると考えられる.

今後の課題としては以下のことが挙げられる.

1. 係り受けと F_0 のより詳細な関係の利用
   多くの場合, 先行句が直後の句に係る場合には後続句の F_0 が下降し, 係らない場合には上昇する傾向を今回利用したが, この傾向には例外がある. [12] ではより詳細に係り受けと F_0 の関係を調べ, 句間が並列関係の場合には F_0 が上昇すること, 右枝分かれ (直前の句から修飾されない) が連続した場合に F_0 が下降することを示している. このことをコストに反映させることで精度向上が期待出来る.

2. 小コーパスでの係り受け判定
   今回の係り受け判別実験では手法間に差が出なかったが, より小規模なコーパスでは差が出る可能性がある. そこでコーパスサイズの削減, 合成ターゲット文の増加などを行った上での再実験が望まれる.

5 まとめ

本稿では言語的情報を単位選択のコストに直接反映させる取り組みの一つとして係り受けと F_0 の関係に着目し, コストを設計し, 2 つの聴取実験で評価した. その結果, テキストが同一で複数の意味を持ち得る 4 文の合成音の係り受けの判別率に関しては, 従来法, 提案法共に原音声とほぼ同じ判別率が得られ差が出なかったが, 全体の自然性に関しては提案法が従来法に対して 25.7% 多く好ましいとの結果を得た.

謝辞 本研究の一部は科学技術振興機構戦略的基礎研究推進事業 (JST/CREST) の援助により行われた.

参考文献

[1] 藤崎博也, “韻律研究の諸側面とその課題”, 音響学会講演論文集, 2-5-11, pp.287–290 (1994).

[2] 匂坂芳典, “日本語音声の韻律的特徴とその計算モデル”, 音響学会講演論文集, 2-5-13, pp.295–298 (1994).

[3] 三村克彦, 海木延佳, 匂坂芳典, “統計的手法を用いた音声パワーの分析と制御”, 日本音響学会誌, Vol.49, No.4, pp.253–259 (1993).

[4] 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”, 電子情報通信学会音声研究会資料, SP96-7, pp.45–52 (1996).

[5] 藤澤謙, 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”, 音響学会講演論文集, 2-3-1, pp.227–228 (1999).

[6] 藤井慶, ニック・キャンベル, “アクセント型を考慮したモーラ単位選択の検討”, 音響学会講演論文集, 1-P-18, pp.387–388 (2002).

[7] Toshio HIRAI, Seiichi TENPAKU, Kiyohiro SHIKANO, “Speech Unit Selection Based on Target Values Driven by Speech Data in Concatenative Speech Synthesis”, Proc. IEEE 2002 Workshop on Speech Synthesis (2002).

[8] 関口芳廣, 鈴木良祢, 菊川智之, 高橋安子, 重永実, “韻律情報を利用した連続音声中の隣接句間の修飾関係有無の判定”, 電子情報通信学会論文誌, D-II, Vol.J78-D-II, No.11, pp.1581–1588 (1995).

[9] 藤澤謙, 平井俊男, 樋口宣男, “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基準の改良”, 音響学会講演論文集, 2-7-2, pp.219–220 (1997).

[10] 戸田智基, 河井恒, 津崎実, 鹿野清宏, “波形接続型テキスト音声合成における素片選択コストの知覚的評価”, 電子情報通信学会音声研究会資料, SP2002-69, pp.19–24 (2002).

[11] http://www.speech.kth.se/wavesurfer/

[12] 海木延佳, 匂坂芳典, “局所的句構造に基づく F_0 制御”, 電子情報通信学会論文誌, D-II, Vol.J83-D-II, No.9, pp.1853–1860 (2000).