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

2003－NL－155 (12)
2003－SLP－46 (1)
2003／5／27

韻律の局所的特徴による文の構造の推定

大須賀智子　　堀内靖雄　　市川熹

千葉大学大学院自然科学研究科
〒 263-8522 千葉市稲毛区弥生町 1-33
Email: osuga@graduate.chiba-u.jp

本研究では、音声の韻律情報のみを用いた文の構造の推定手法について検討した。推定に用いる韻律パラメータとして、今回は新たに、先行するアクセント句末 1 モーラにおける局所的な韻律的特徴を用いて文の木構造の生成を試みた。ATR503 文を対象として実験を行った結果、部分木のレベルで約 76%の推定精度を得ることができた。これは従来の、後続音声区間にまたがる、より大局的なパラメータを用いた場合に対し、約 4%の低下にとどまり、ほぼ遜色のない結果が得られた。すなわち、先行する音声区間の局所的な韻律情報のみから、後続の音声区間への係り受け関係がある程度推定可能であることが確かめられた。この結果から、局所的特徴も文構造の理解へ貢献しており、韻律情報が我々人間の実時間および実環境での発話理解を支えるために、頑健な構造となっている可能性が示唆されたといえる。

Estimating Syntactic Structure from Local Prosodic Features

Tomoko Ohsuga　Yasuo Horiuchi　Akira Ichikawa

Graduate School of Science and Technology, Chiba University
1-33, Yayoi-cho, Inage-ku, Chiba, 263-8522 Japan
Email: osuga@graduate.chiba-u.jp

In this study, we introduce a method of estimating the syntactic tree structure of Japanese speech from the F0 contour and time duration. We formed the hypothesis that we can infer a syntactic relation with the following part by listening only to the leading part of speech, and we proposed an estimating method which uses only the local prosodic features of the final part of the leading phrase. We applied the method to the ATR 503 speech database. The experimental results indicated an estimation accuracy of 76% for the branching judgment for each sequence of three leaves. We consider this result to be fairly good for the difficult task of estimating a syntactic structure that includes a future part by using only local prosodic features in the past, and also consider prosodic information to be very effective in real-time communication with speech.

1　はじめに

本研究では、音声の韻律に関する研究のひとつとして、韻律情報を用いた文構造の推定に着目する。我々人間が書き言葉よりも話し言葉の方が円滑に話を進められるのは、音声特有の韻律情報が加わるためであり、意味の区切りを明確にするなど言語情報の理解を手助けするとともに、意図や強調、感情を表現したり、相手との対話を制御したりなど、多くの機能を有していると言われている。中でも統語的な情報は音韻情報に負うところが大きいものの、音韻情報のみでは発話内容を実時間で正しく理解することは困難であり、特別な意図を含まない通常の読み上げにおいても、その文法的な理解に対して話し手は無意識のうちに韻律情報を有効利用していると考えられる。

近年、この韻律情報と統語情報との関わりを音声認識や合成、自然言語処理などの分野においても積極的に活用しようという動きがあり、認識の際に韻律的特徴から統語境界を抽出するのに利用したり、統語構造を合成時の韻律制御に利用したりしている。また、言語的な係り受け規則とさまざまな韻律パラメータから係り受け解析を行ったり [1, 2]、隣接句間の修飾関係有無の判定に韻律情報を利用している研究 [3] などもあり、韻律的特徴量を係り受け解析に用いることの有用性が示されている。これらの研究の成果は、将来の対話システムにおける意味理解などへの貢献のみならず、話し手が文構造に対してどのような韻律的特徴を付加しているのか、また聞き手がどのような韻律的特徴を用いて構造の理解を行っているのかなど、人間の認知的な音声情報処理の仕組みについても有用な知見を与えられるものと期待される。

我々もこれまで、以上のような観点から、韻律情報を用いた文構造の推定に関する研究を行っている [4, 5, 6]。本稿では、文の構造理解に対する韻律情報についてさらに分析するため、先行する音声区間の局所的な特徴のみを用いた文の木構造の生成を試みた。これは、文献 [5, 6] において小松らによる先行研究 [7] の改良を行った文構造推定手法では、アクセント句レベルでの韻律的特徴（時間長、後続句とのポーズ長や、基本周波数の差）を用いたのに対し、先行するアクセント句末 1 モーラにおける局所的な韻律的特徴（基本周波数の傾き、時間長など）を新たに定義して用いたものである。つまり、先行する音声区間の韻律情報のみから、後続の音声区間への係り受け関係を推定することになる。もしこのような先行する局所的特徴の効果が高ければ、揮発的な音声の理解において、韻律情報がリアルタイムでの情報処理を支援していることへのひとつの裏付けにもなると考えられる。

2　文構造推定手法

我々はすでに小松らの先行研究 [7] に基づいた文構造推定手法を提案している [5, 6]。これは韻律情報のみから生成する発話文の木構造（これを「韻律木」と呼ぶことにする）を、構文木に近くなるように学習したものである。本節ではこの従来手法について簡単に述べるとともに、今回の改良点について詳述する。

2.1　韻律木の生成

はじめに、韻律木の最下層の葉に相当する韻律単位を PU（Prosodic Unit）と定義し、入力音声から PU 成分を決定する。PU は基本周波数形状の極小値点、ならびにポーズにより分割されるものとするが、ここでは理想的な状態として、アクセント句 [8] を最小の PU として用いることにする。

一般に構文木はいくつかの部分木により構成されているものと考えることができる。ここで部分木を連続する 3 つの PU からなるものとすると、部分木のタイプは図 1 に示すとおり、左枝分かれと右枝分かれの 2 タイプに分類することができる。

(a) 左枝分かれ　　(b) 右枝分かれ

図 1: 部分木のタイプ

構文木から求められる部分木のタイプを正解とし、大量の学習データから判別分析により部分木の判別関数を求める。入力された音声より、葉から順に部分木を作成していき、得られた判別関数を用いて各部分木ごとにタイプの判別を行う。その判別結果に基づいて、適宜 PU の結合を行い、結合された PU を新たにひとつの PU として同様の処理を繰り返していくことで、最終的に文全体についての韻律木を生成することができる。このような手順とすることで、韻律木は準実時間での生成が可能である。

2.2 韻律パラメータ

2.2.1 従来手法におけるパラメータ

従来手法では、2 タイプの部分木の判別には、各 PU における F0 形状と PU 間のポーズ長などに関する 4 種類パラメータを用いた。以下、PU 成分は各 PU 区切りごとの F0 の最小二乗近似直線としてある。ただし、F0 はすべて対数 F0 として扱うものとする。これらのパラメータの詳細を以下に示す。

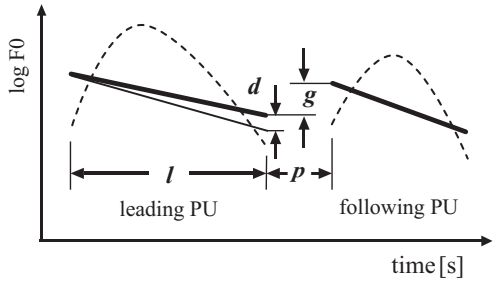

図 2: 従来手法におけるパラメータ

g: 先行 PU の終点と後続 PU の始点との基本周波数の差

l: 先行 PU の時間長（秒）

d: 先行 PU の傾きと標準的な PU の傾き（$-25Hz$/秒）との差の絶対値 [1]

p: PU 間のポーズの時間長（秒）

図 2 では 2PU 間のパラメータを示しているが、実際には部分木は 3PU からなるため、第 1、第 2PU 間におけるパラメータを $g1, l1, d1, p1$ とし、第 2、第 3PU 間におけるパラメータを $g2, l2, d2, p2$ として、計 8 個のパラメータを用いることにする。

2.2.2 局所パラメータ

上述の 8 つのパラメータは、ポーズに関するパラメータ p を除き、PU 全体に関わる値であり、大局的なパラメータであるとも言える。この捉え方は、リアルタイムの音声情報処理において、人間はそれほど複雑な処理はしておらず、感情や強調などを含まない標準的な発声においては、韻律情報は文構造の大枠的理解に貢献しているのではないか、という従来からの我々の仮説に沿うものである。

しかし一方で、リアルタイムで考えながら次々と音を算出していくという発話行為を考えると、ダイナミックに情報を変化させることも可能であり、その瞬間、瞬間に新たに情報をのせている可能性は大きいと思われる。また、人間の知覚－認識メカニズムでの逐次学習という観点から、聞き手にとっても局所的特徴が有効に働いている可能性がある。さらに、先行する発話により後続の構造が予測できれば、リアルタイムでの発話理解が効率的に行えることになる。

このような観点から、我々は先行する部分の局所的な韻律情報が、文構造の推定にも効果を与えているのではないかという仮説を立て、この仮説を検証するために、従来の文構造推定手法に対して新たに局所的なパラメータを導入し、その効果を評価することにした。局所パラメータをとる対象としては、韻律木の最下層の葉に相当する韻律単位 PU（ここではアクセント句）の末尾 1 モーラに注目する。その理由としては、パラメータをとる区間を明確にできること、モーラ単位での比較を行うことで時間的な伸縮傾向も見られることが挙げられる。また古くから “final lengthening” と呼ばれる、強い句境界の前でモーラの伸張がみられるという現象も考慮したものである。

以下に、新たに定義した局所パラメータの詳細を示す。従来と同様、F0 についてはすべて対数 F0 として扱うことにする。また、末尾 1 モーラにおける F0 近似直線を母音部だけとしたのは、子音におけるマイクロプロソディの影響を除くためである。

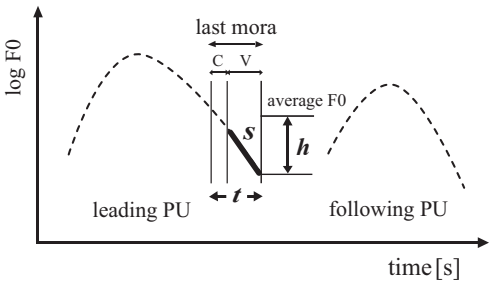

図 3: 新たに定義した局所パラメータ

s: PU 末 1 モーラの母音部における F0 近似直線の勾配

h: PU 末 1 モーラの母音部における F0 近似直線の末端値と、その話者の F0 平均値との差

[1] ただし、小松らによる従来手法では F0 を実空間で扱っていたが、ここでは F0 はすべて対数 F0 として扱うことにし、-25[Hz/s] という値も対数に変換したものを用いた [6]。

t: PU 末 1 モーラ長の、当該アクセント句における末尾以前の平均モーラ長からの伸縮率

なお、従来のパラメータ同様、3PU 間の関係を表すために、第 1PU 末におけるパラメータを $s1, h1, t1$ とし、第 2PU 末におけるパラメータを $s2, h2, t2$ として、計 6 個のパラメータとする。

3 実験

第 2.2.2 節で新たに定義した局所パラメータを用いて、文構造の推定実験を行った。なお比較のため、従来のパラメータを用いた場合、また両者のパラメータを組み合わせた場合も同様に実験を行った。

3.1 音声試料

音声資料としては、ATR「研究用日本語連続音声データベース」[9] の音素バランス文 503 文のうち、セット A の 50 文を全 10 名分（男性 6 名、女性 4 名）用いた。従って、総発話数は $50 \times 10 = 500$ 文である。ただし、話者によって 2 アクセント句だけで発声された 4 文については部分木がひとつも作成できないことから評価から除いてある。

基本周波数データに関しては、データベースには 4 話者分しか存在しないため、全話者分について音声分析ツール ESPS の基本周波数抽出プログラム get_f0 を用いて抽出し、手修正を行ったのち、底を 2 とする対数に変換したものを用いた。各発話文における構文木の情報は、データベースに付随する言語・韻律情報データを利用した。最小の韻律単位 PU とするアクセント句の区切りに関しては、話者 1 名分に付与されていたデータに基づいて、残りの 9 話者分については聴取により取得した。最終的に得られたアクセント句数は 3267 である。また、各モーラ区切りについては、全話者分についてデータベースに付随している音素ラベルデータを利用して求め [2]、各アクセント句の開始時刻および終了時刻も同様に音素ラベルデータから求めた。

[2]モーラ境界の音素が融合化していることによりその境界時刻が与えられていない場合には、便宜上融合化した音素数で等分化して境界時刻とした。

3.2 評価方法

判別分析においては、データベースより得られた構文木を正解の木構造として、提案手法により生成した韻律木の正解率を求めることにする。

オープン実験では、音声データを各話者 10 文、計 100 文ずつの 5 セットに分割し、4 セットを学習、残り 1 セットを評価に用い、この組み合わせを入れ替えて 5 回実験を行い、その平均値により正解率を算出した。表 1 に、最終的な判別件数（部分木数）を示す。ただし、局所パラメータを用いる条件においては、PU 末 1 モーラの母音部に連続 20[ms] 未満の F0 値しか観測されなかった場合には、そのモーラの F0 勾配は取得不可能であるとし、そのような PU が含まれる文は評価対象から除いた。

表 1: 判別件数（部分木数）

左枝分かれ	右枝分かれ	合計
1686	1794	3480

判別分析にはマハラノビスの汎距離を用い、説明変数の選択にステップワイズ法（変数編入基準・除去基準とも 5%）を採用し、統計的に各韻律パラメータの効果をはかることにした。

4 結果と考察

4.1 変数選択の結果

まず、判別分析に先立ち行った変数選択の結果について述べる。表 2 は、全話者データのデータで変数選択を行った際の、編入パラメータ、除去パラメータを列挙したものである。上段が従来の大局的パラメータ（8 個）のみを用いた場合、中段が今回提案した局所パラメータ（6 個）のみを用いた場合、下段が両パラメータ（計 14 個）を組み合わせた場合の結果である。なお、編入パラメータについては編入された順、すなわち判別関数への寄与（F 値）の高い順に、除去パラメータについては変数選択の終了時点において F 値の高かった順に、それぞれ並べてある。

表 2 から、従来のパラメータのみ、今回の局所パラメータのみの場合には、それぞれ全パラメータが部分木タイプの判別に有効であると認められた。一方両パラメータを組み合わせた場合には、g と h の相関が

表 2: 判別分析における変数選択の結果

	編入パラメータ	除去パラメータ
従来のみ	$p1, p2, g1, g2, d1, d2, l2, l1$	
局所のみ	$t1, h1, h2, s1, s2, t2$	
両方	$p1, p2, g1, h2, t1, d1, t2, d2, s2, l2$	$l1, s1, h1, g2$

高かったこともあり 10 個に絞られたが、その中に局所パラメータも含まれていることが分かる。具体的には、ポーズに関するパラメータ p の寄与がもっとも高く、次いで PU 境界における F0 値（先行 PU の末端値 h もしくは後続 PU との差分 g）の寄与が高い。これは、先行研究 [1] により係り受け解析にもっとも有効な韻律的特徴はポーズであるという知見とも一致している。

4.2 部分木レベルの判別正解率

続いて、実際に文構造の推定を行った結果について示す。表 3 は、部分木レベルでのクローズド、オープン実験それぞれの判別正解率について、パラメータの条件ごとにまとめたものである。

表 3: 部分木の判別正解率

パラメータ	クローズド	オープン
従来のみ	80.0%	79.3%
局所のみ	76.3%	75.5%
両方	81.8%	81.4%

表 3 から、局所パラメータのみで 76%と高い精度が得られており、従来の大局的なパラメータでの結果と比較して 4%程度しか低下していない。これは、句末の情報のみ、しかもポーズ情報も用いていない条件であることを考慮すると、高く評価される結果であると思われる。また別の観点からすれば、先行する音声区間の情報のみから後続の音声区間への係り受け関係を推定することになる今回の手法において、後続の情報も用いる従来の方法と比較してほぼ遜色のない結果となっており、韻律情報が文構造のリアルタイムでの理解に効果的に利用されていることを示唆していると考えられる。

一方、表 3 の下段より、従来の大局的パラメータに局所パラメータを追加することにより、2%程度判別正解率が上昇していることが分かる。この結果より、韻律情報のみでの推定に限界も見られるものの、局所パラメータを導入することによる効果は認められると考えられる。

また、各条件においてクローズド実験、オープン実験での差があまり見られないことから、得られた判別関数は一般的な韻律木の構造をよく反映できているといえよう。話者別の正解率においても、話者間で数%の差は生じたものの、話者依存・非依存の結果にはほとんど差が見られなかった。

4.3 文レベルの判別正解率

表 4 に文レベルでのクローズド、オープン実験それぞれの判別正解率を示す。文レベルでの正解率とは、1 文中のすべての部分木が正しく判別された文の割合を表している。

表 4: 文正解率

パラメータ	クローズド	オープン
従来のみ	25.8%	24.6%
局所のみ	19.7%	19.5%
両方	27.8%	26.3%

表から、文レベルでの正解率は 30%を下回っているが、この値は 1 文中の部分木のうち 1 つでも誤判別があれば不正解と扱われてしまうこと、1 文あたり平均して 7 部分木からなることを考慮すると、妥当な数値であるといえる。特に今回の音声資料のような長い文章では、構文構造が複雑になってしまうことも文正解率に影響を与えていると思われるが、本手法では言語情報をまったく用いていないことから、韻律情報のみでこれだけの推定が可能であるということは、韻律が文構造の理解に少なからず貢献していると言えるだろう。ただし、今回の実験に用いた音声はプロ（アナウンサ）による丁寧な読み上げであり、一般人の発声と

比較して整った構造をしていることも指摘されることから、今後さまざまな音声資料に対して本手法を適用し、評価を行う必要性があると思われる。

4.4 誤判別の例

誤判別の例として典型的なものに、ポーズの影響が挙げられる。これは従来手法の際にも取り上げたものであるが、名詞句の並列において間の助詞が省略されている場合などに、その助詞の代わりにポーズを挿入する例が見られる。このようにポーズを挿入することで意味の区切りを捉えやすくしていると解釈される。しかし、構文的には近い関係にあるものの、一般にポーズは句の境界が深いことを示す役割の方が圧倒的に大きいことから、構文木と韻律木とが異なる構造となってしまう。また文全体の発声のリズムを整えるために、特に長い名詞句などが含まれる場合に、ポーズを利用して F0 を立て直す例なども見られた。今回新たに導入した局所パラメータでは直接ポーズの情報は用いていないものの、ポーズを挿入することで前後の韻律的特徴も大きく影響を受けてしまうことから、このような例での判別率は若干改善されたにとどまった。すなわち、この例のようなポーズの役割を韻律のみで正しく解釈することはほぼ不可能であり、言語情報などと合わせた判別が必要になると考えられる。

その他に、強調などのパラ言語情報の影響と考えられる例なども見受けられたが、これら誤判別の箇所については、得られた韻律木がどのような構造になっているのか、今後より詳細な分析を行いたいと考えている。

5 まとめと今後の課題

本稿では、韻律情報のみからの文構造推定に関して、新たに先行句末の局所的パラメータ（句末 1 モーラの F0 勾配、高さ、時間長）を導入した手法について提案を行った。3PU からなる部分木が左枝分かれと右枝分かれのどちらであるかを、大量データを用いて判別分析により評価した結果、局所パラメータのみを用いた場合に 76%、局所パラメータと従来のパラメータを組み合わせた場合で 82%の正解率を得ることができた。これらの結果は、クローズドおよびオープン実験、話者依存および非依存での実験でも結果に大きな差がなかったことから、文構造の推定に対して一般性を持つ判別関数が得られていると考えられる。また局所的パラメータのみでもある程度推定が可能であったことから、局所的特徴も文構造の理解へ貢献していることが確かめられ、韻律情報が我々人間の実時間および実環境での発話理解を支えるために、頑健な構造となっている可能性が示唆されたとも捉えられる。

今後は、第 4.4 節で述べたような本手法では推定が困難な例の詳細な分析と合わせ、より多様な音声資料への適用を考えたい。また今回用いたデータのようなプロの発声だけでなく、一般人の発声などでの実験を通し、韻律木に関してより詳細な分析を行いたいと考えている。

参考文献

[1] 江口徳博, 尾関和彦. 韻律情報を利用した係り受け解析. 日本音響学会誌, Vol. 52, No. 12, pp. 973–978, 1996.

[2] 広瀬幸由, 尾関和彦, 高木一幸. 日本語読み上げ文の係り受け解析における韻律的特徴量の有効性. 自然言語処理, Vol. 8, No. 4, pp. 71–89, 2001-10.

[3] 関口芳廣, 鈴木良弥, 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. 1581–11588, 1995-11.

[4] 大須賀智子, 鈴木則夫, 堀内靖雄, 市川熹. 基本周波数とポーズによる構文構造の推定. 人工知能学会研究会資料 SIG-SLUD-A201-7, pp. 41–46, 2002-6.

[5] Yasuo Horiuchi, Tomoko Ohsuga, and Akira Ichikawa. Estimating Syntactic Structure from F0 Contour and Pause Duration in Japanese Speech. In *Proceedings of International Conference on Spoken Language Processing*, pp. 1177–1180, 2002.

[6] Tomoko Ohsuga, Yasuo Horiuchi, and Akira Ichikawa. Estimating Syntactic Structure from Prosody in Japanese Speech. *The Transactions of the Institute of Electronics, Information and Communication Engineers*, Vol. E86-D, No. 3, pp. 558–564, 1996.

[7] 小松昭男, 大平栄二, 市川熹. 韻律情報を利用した構文推定およびワードスポッティングによる会話音声理解方式. 電子情報通信学会論文誌 D, Vol. J71-D, No. 7, pp. 1218–1228, 1988-7.

[8] J. J. Venditti. Japanese ToBI Labelling Guidelines. In *Papers from the Linguistics Laboratory, Ohio State University Working Papers in Linguistics*, Vol. 50, pp. 127–162, 1997.

[9] 阿部匡伸, 匂坂芳典, 梅田哲夫, 桑原尚夫. 研究用日本語音声データベース利用解説書（連続音声データ編）. ATR 自動翻訳電話研究所, 1990.