日本語文音声における韻律情報と焦点の検討

北川 敏　ニック キャンベル
ATR 音声翻訳通信研究所
〒 619-0288 京都府相楽郡精華町光台 2-2
satoshi@itl.atr.co.jp, nick@itl.atr.co.jp

あらまし
現在、TTS(Text To Speech) による合成音声の韻律は一意に決まっていることが多い。そこで、自由に韻律を付与できる音声合成システムの構築をめざしているが、そのためには、韻律情報の利用が不可欠だと考えられる。韻律情報によって表されるものには様々なものがあるが、本稿では意味的強調 (以下、焦点と表現) に着目した。音声合成の際に、音声の任意の箇所に焦点の情報を付与する際に必要な基準は確立されていない。また、合成時に焦点の情報を利用するためには、焦点の情報のデータベース化が必要だと思われる。本稿では、このデータベース作成のための第一段階として、焦点を抽出することを試みた。その抽出は、焦点を含まない音声と含んだ音声を比較し、その違いをみるという方法を用いた。韻律情報は、基本周波数・パワー・音素継続時間の 3 つの要素からなる。本稿では、基本周波数・音素継続時間の情報を用いて、焦点の抽出を試みた。(1) 自然音声どおしで、(2) 合成音声と自然音声の 2 通りの組合せで抽出を行い、今回の方法の有効性を調べた。また、2 つの韻律情報を組み合わせて用いた場合の焦点の抽出も試みた。その結果、合成音声も自然音声と同様に韻律情報の基準としての使用が可能であることがわかった。また、焦点に対して基本周波数と同様に音素継続時間も何らかの影響をもつということがわかった。

キーワード ● 基本周波数 ● 音素継続時間 ● 焦点抽出 ● 自動抽出

The relation of prosodic characteristics to focal prominence in Japanese read speech

Satoshi KITAGAWA and Nick CAMPBELL
ATR Interpreting Telecommunications Research Laboratories

2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
satoshi@itl.atr.co.jp, nick@itl.atr.co.jp

Abstract
In this paper we present results of an analysis of the prosodic characteristics of focus in spoken Japanese. We show first that focal prominence can be detected from the fundamental frequency contour, and second that it can also be detected from durational characteristics of the utterance. In each case we perform the detection by comparison with a reference utterance, either natural or synthesised. Because the finding for duration was unexpected, we also present a further analysis that shows the durational differences to be independent of the changes in fundamental frequency. The current model is limited to phrase-level granularity,but results show that the phrase carrying focal prominence can be reliably detected in the majority of cases. We leave for future work the detection of prominence on units smaller than the syntactic phrase.

Key words ● pitch ● duration ● focus extraction ● auto extraction

1 はじめに

現在、TTS(Text To Speech)による音声合成は身近なものになってきている。しかし、そこで得られる合成音声の韻律は一意に決まっていることが多い。そこで、自由に韻律を付与できる音声合成システムの構築をめざした。

そのためには、韻律情報の利用が不可欠だと考えられる。韻律情報によって表されるものには様々なものがあるが、本稿では意味的強調(以下、焦点と表現)に着目した。この焦点がわかると話者の意図を明確に表現できると考えられるからである。音声合成の際に、音声の任意の箇所に焦点の情報を付与するには、聞き手が焦点を判断することができるために基準となるものが必要となるが、それは確立されていない。また、合成時に焦点の情報を利用するためには、焦点の情報のデータベース化は不可欠である。

本稿では、このデータベース作成のための第一段階として、焦点を抽出することを試みた。その抽出は、焦点を含まない音声と含んだ音声を比較し、その違いをみるという方法を用いた。

韻律情報は、基本周波数・パワー・音素継続時間の3つの要素からなる。[1]これまで焦点を抽出するために、音声波形のパワーと音素継続時間の情報を用いて分析を行い、その有効性が確かめられている。[2][3]また、強調と発話速度について、一定の傾向があるという報告も行われている。[4]

本稿では、基本周波数・音素継続時間の情報を用いて、焦点の抽出を試みた。ただし、焦点は1文につき1ヶ所含まれているものとしている。(1)自然音声どおしで、(2)合成音声と自然音声の2通りの組合せで抽出を行い、今回の方法の有効性を調べた。また、2つの韻律情報を組み合わせて用いた場合の焦点の抽出も試みた。

2 焦点抽出

前章で述べたように、本稿における焦点の判定は2つの音声を比較することにより行う。韻律情報としては基本周波数と音素継続時間の2つを用いるが、いずれの場合も音素単位で比較できるように、正規化を行った後、音素ごとの代表値を求めている。以下、それぞれの場合における焦点の抽出法を示す。

2.1 基本周波数による焦点の抽出

基本周波数による焦点の抽出は図1の手順で行った。

まず、焦点を含まない音声(基準音声)と含んだ音声のそれぞれに対して、基本周波数曲線から各音素の平均基本周波数を計算し、各音声の平均基本周波数からの移動量を求めた。そして、その移動量の差を求めた。焦点化すると、つまり強調を行った箇所は強調を行っていない場合に比べて基本周波数が高くなる傾向にある。そこで、その差がもっとも大きい音素を含む文節に焦点があると判断することにした。

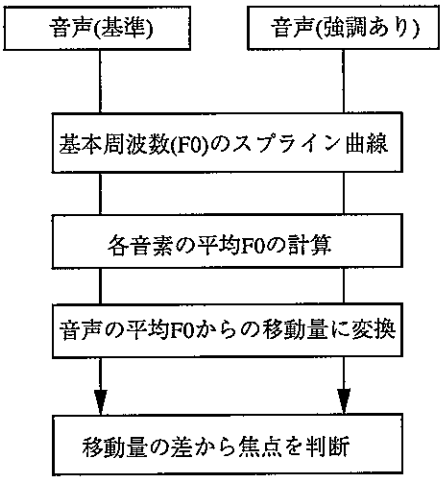

図 1: 移動量の算出手順

ここで述べた移動量は次の手順で求めた。

1. 各音声から有声区間の基本周波数を10ms単位で求める。
2. 求めた基本周波数列からスプライン関数によるスムージングを行い、得られたスプライン曲線をその音声の基本周波数のスプライン曲線とする。(その音声の連続したスムーズな基本周波数曲線を得る)
3. このスプライン曲線から各音素の基本周波数の平均値を求める。(各音素ごとに基本周波数の代表値を求める)
4. 上で求めた各音素の基本周波数の平均値の、文音声全体の平均値に対する移動量を求める。(基本周波数の変化量を比較する)

2.2 音素継続時間による焦点の抽出

音素継続時間による焦点の抽出方法を図2に示す。

はじめに、焦点を含まない音声(基準音声)と含んだ音声のそれぞれのスペクトル時系列から、DTW(時間軸伸縮マッチング)による時間伸縮関数を求める。

そして、この時間伸縮関数に対して回帰直線を求める。このことにより、2つの音声の平均話速の違いの影響を除くことができる。

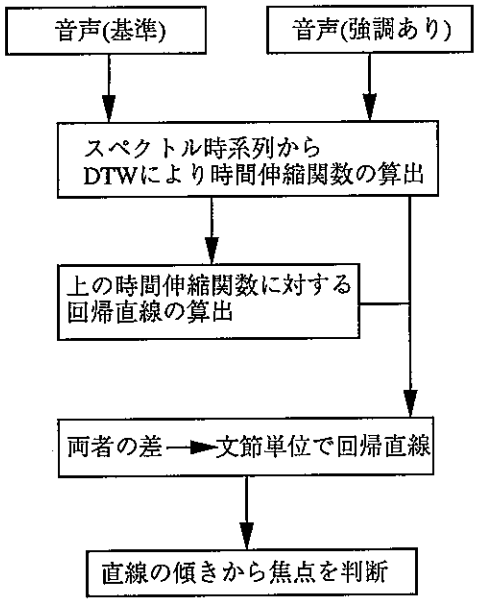

図 2: 音素継続時間による焦点の抽出手順

次に時間伸縮関数と回帰直線の差を求める。この差が大きくなる方に変化した場合、だんだんゆっくりと話していることを示している。

こうして得られた差に対して、文節単位で回帰直線を求める。時間伸縮関数と回帰直線の差が大きくなる方に変化すると、この回帰直線の傾きは大きくなる。音声試料を調べてみると、焦点化された文節の終りの方の音素がより長くなる傾向がみられた。このような場合に、この回帰直線の傾きは大きくなる。そこで、上の回帰直線の傾きがもっとも大きい文節に焦点があるという基準を用いた。

3 自然音声どおしの比較

はじめに今回用いようとしている方法が有効であるかどうかを確認するために、基準音声にも自然音声を用いて、比較を行うことにした。

3.1 音声試料

実験で使用した音声は、1名の男性話者による会議登録の対話文を読んだもので全部で89文である。

この対話文は、

1. 今回は 割引を おこなっておりません。(基準音声)
2. 今回は 割引を おこなっておりません。(焦点 1)
3. 今回は 割引を おこなっておりません。(焦点 2)

のように同一の文字列から構成されており、焦点の位置だけが異なるものである。このうち、基準音声の数は25文である。

3.2 基本周波数

ある文(「今回は割引をおこなっておりません。」)について、基本周波数の平均値からの移動量の差の変化を図3に示す。図の横軸は各音素、縦軸は平均基本周波数からの移動量の差を表す。縦軸の値が大きいほど、焦点の度合いが大きいことが期待される。図の実線は「今回は」、破線は「おこなっておりません」の部分に焦点がある音声との組み合わせの場合である。

実線の変化パターンを見ると、文頭から7番目の音素、すなわち「koNkaiwa」の /w/ でこのパターンにおける最大値となっている。この /w/ は「今回は」に含まれるため、この文の焦点のある部分は「今回は」であると判定される事になる。また破線についても、「おこなっておりません」が焦点と判定されることになる。よって、以上の2つの場合は、正しく焦点が抽出されたことになる。

焦点の判定を、64通りの組み合わせに対して行った。その結果を表1に示す。表に示しているように、高い精度で焦点が抽出されていることがわかる。

3.3 音素継続時間

ある文(「では お名前と 人数を お願いいたします。」)について、時間伸縮関数とその回帰直線の差の変化パターンを図4に示す。

図の横軸は各音素、縦軸は時間伸縮関数とその回

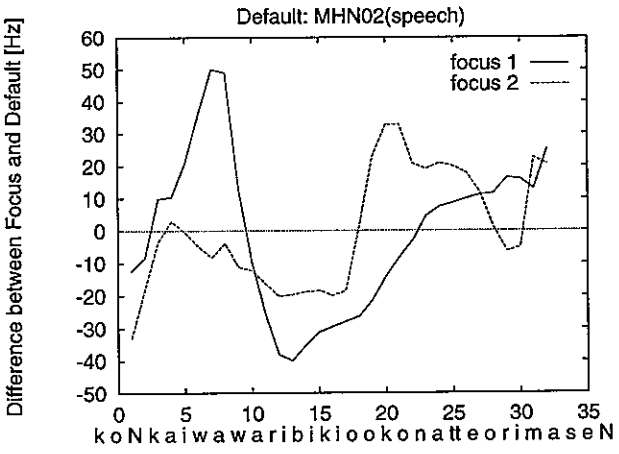

図 3: 平均基本周波数からの移動量の差の変化パターン(自然音声)

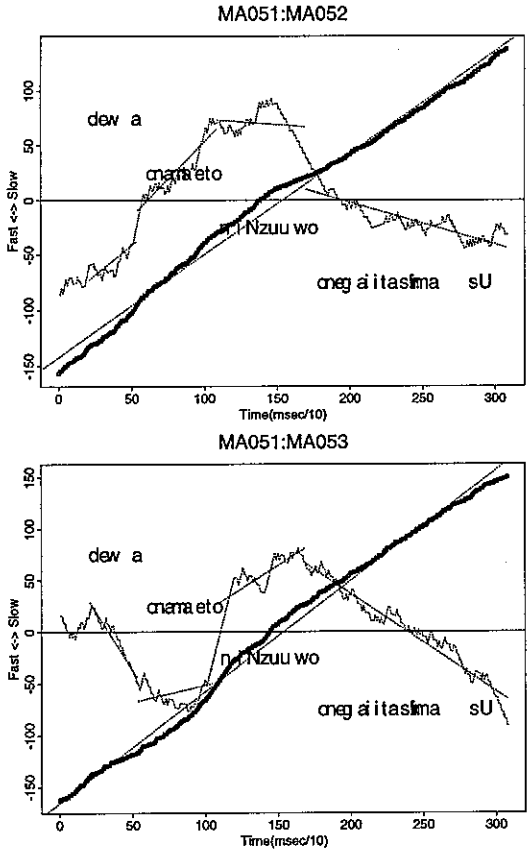

図 4: 時間伸縮関数とその回帰直線の差の変化パターン（「では お名前と 人数を お願いいたします。」）［自然音声］

帰直線の差を表す。文節ごとに求めた回帰直線の傾きが大きいほど、焦点の度合いが大きいことが期待される。上図は「お名前と」の部分を、下図は「人数を」の部分を焦点化した音声との組み合わせの場合である。

上図の文節ごとの回帰直線の傾きを見ると、「お名前と」の部分の傾きが最も大きくなっている。このことから、この文の焦点は「お名前と」であると判定される事になる。

また下図についても、「人数を」が焦点と判定されることになる。よって、以上の 2 つの場合についても、正しく焦点が抽出されたことになる。

焦点の判定を、64 通りの組み合わせに対して行った。その結果を表 1 に示す。表に示しているように、基本周波数による抽出結果には及ばないものの、チャンスレベルよりは高い正解率が得られた。

表 1: 焦点抽出の正解率 (自然音声)

	Correct	False	正解率
F0	52	12	81.3%(=52/64)
Duration	38	25	59.3%(=39/64)

3.4 考察

焦点化の手段としては、基本周波数を変化させることが基本となるため、表 1 における基本周波数の結果は、比較的高い数値となったと考えられる。それに対して、音素継続時間の結果は予想より高いものであった。

焦点と音素継続時間に 1 対 1 の依存性がないとすると、抽出の結果は無作為抽出の場合と同様になるはずである。つまり、今回用いた音声試料の大半は 4 文節から構成されるため、25% 前後の値になると期待される。

しかし、表 1 からもわかるように 60% 近い結果となった。このことから、音素継続時間と焦点になんらかの関係があると考えられる。

4 CHATR による合成音声との比較

前章では、基準音声として自然音声を用いた。しかし、基準音声を自然音声とした場合、焦点抽出を行うことができる音声が限られてしまう。ここで、基準音声として合成音声に対しても、自然音声の場合と同様に焦点抽出をおこなうことができれば、さまざまな入力音声に対処できると予想される。

そこで、本章では基準音声を合成音声として、前章と同様の方法により焦点の抽出を試みた。

4.1 音声試料

使用した音声資料は、前章と同一で 1 名の男性話者による会議登録の対話文を読んだもので 89 文である。焦点についても同様で 1 文につき 1 ヶ所含まれているものとしている。ただし、基準音声として合成音声を使用しており、この合成音声は波形接続型音声合成システム CHATR[5] により作成したものを用いた。

4.2 基本周波数

「今回は割引をおこなっておりません。」について、平均基本周波数からの移動量の差の変化を図 5 に示す。図 3 と同様に、図の実線は「今回は」、破線は「おこなっておりません」の部分を強調した音声である。自然音声の場合と同じように、実線については「今回は」に、破線については「おこなっており

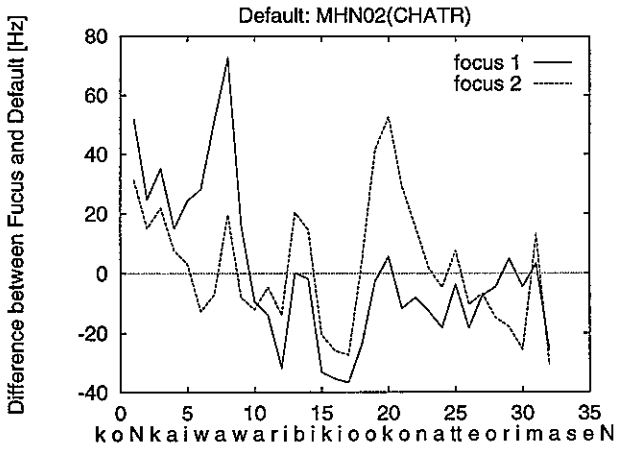

図 5: 平均基本周波数からの移動量の差の変化パターン (CHATR)

ません」を抽出することができた。また、合成音声の場合、micro-prosody 情報 (音素の違いによる韻律の違い、音素片の接続歪みなど) がないために、パターンの変化が自然音声の場合と比べて激しくなっていると考えられる。

4.3 音素継続時間

ある文 (「会議の 間に 市内観光が あるそうです。」) について、時間伸縮関数とその回帰直線の差の変化パターンを図 6 に示す。

前節と同様、図の横軸は各音素、縦軸は時間伸縮関数とその回帰直線の差を表す。文節ごとに求めた回帰直線の傾きが大きいほど、焦点の度合いが大きいことが期待される。上図は「会議の」の部分を、下図は「間に」の部分を焦点化した音声と比較した場合である。

4.4 焦点抽出

64 通りの組み合わせに対して焦点の判定を行った結果を表 2 に示す。焦点の抽出の精度は、それぞれ 75.0%、50.0% と自然音声の場合とほぼおなじ精度の抽出ができた。

この結果から、合成音声に対しても自然音声の場合と同じレベルで焦点の抽出ができたといえる。しかし、自動的に焦点を抽出しようとするには十分とはいえないため、より多くのデータによる分析を行う必要があると思われる。

表 2: 焦点抽出の正解率 (CHATR)

	Correct	False	正解率
F0	48	16	75.0%(=48/64)
Duration	32	32	50.0%(=32/64)

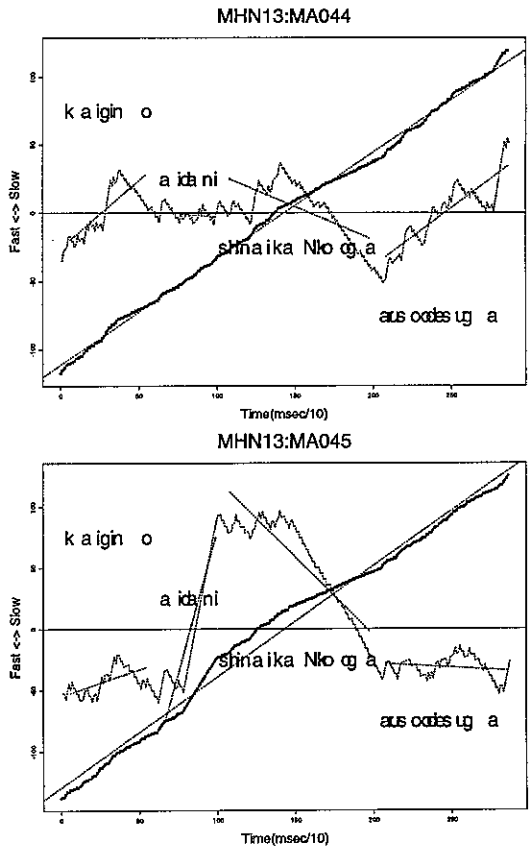

図 6: 時間伸縮関数とその回帰直線の差の変化パターン (「会議の 間に 市内観光が あるそうです。」) [CHATR]

5 複数の韻律情報による判定

5.1 音素継続時間による焦点の判定方法の再検討

前章までは、1 つの韻律情報のみを使用して焦点抽出をおこなったが、組み合わせることにより抽出の精度を向上することは可能であると想像される。

音素継続時間による判定方法を変更して、基本周波数と同様の処理を行うことにする。前章では DTW によるマッチングを利用したが、本章では各音素ごとの継続時間の z-score を求め、その差の変化パターンから判定を試みている。

そこで、音素継続時間の差の最大値 (peak) と焦点

表 3: duration の z-score の差の peak と焦点

焦点と同一文節にある	26
直前の文節の最終モーラ	16
その他	24
計	66

の位置関係を調べてみた。表 5 に示すように、peak は、焦点と同一文節に最も多く存在することがわかるが、その直前の最終モーラにもあることがわかる。

そこで、音素継続時間については「peak が F0 の peak と同一文節にある」と「peak が焦点の直前の文節の最終モーラにある」を判定基準として用い、音素継続時間のみの場合と基本周波数と組み合わせた場合の 2 通りで抽出を試みた。

その結果、正解率は、それぞれ 63.6%(=42/66)、54.5%(=36/66) となり、基本周波数のみを用いた場合の結果には及ばなかった。また、音素継続時間について別の焦点の判定方法を用いたにもかかわらず、結果としては同程度であった。このことからも音素継続時間と焦点には関係があると考えられる。

5.2 基本周波数と音素継続時間の関係

基本周波数と音素継続時間については相関があるという可能性も考えられる。そこで両者の関係について調べてみた。

差の最大値 (peak) となる箇所で、基本周波数の場合はもっとも高く、音素継続時間の場合はもっとも長くなる。この両者の音素が一致すると両者には相関があると考えられる。

文節を単位として、焦点の有無に対する分布を表 6 に示す。表からわかるように、基本周波数の peak と音素継続時間の peak は全て一致していない。このことから、基本周波数と音素継続時間の間に 1 対 1 の依存性はなく、異なる情報を持つと考えられる。

表 4: F0 と duration の peak の分布

	焦点あり	焦点なし
一致	26 (38.8%)	3 (1.2%)
F0 のみ	30 (44.8%)	9 (3.7%)
Dur. のみ	1 (1.5%)	38 (15.4%)
peak なし	10 (14.9%)	196 (79.7%)
計	67 (100%)	246 (100%)

一致 = F0 と duration の peak が一致 (前後 1 音素)
F0 のみ = F0 のみ peak あり
Dur. のみ = duration のみ peak あり

6 考察

本稿では、焦点 の無い文と焦点 を 1 ヶ所もつ文との比較を行っているが、焦点 の無い文、あるいは 2 ヶ所以上もつ文に対する実験は行っていない。もし、これらの文も抽出の対象にするためには、焦点の検出限を確定する必要がある。

また、文節単位で抽出を行ったが、実際には「では お名前 と 人数をお願いいたします。」という焦点 もありえるので、文節より小さい焦点 の範囲の確立も必要であると思われる。

7 まとめ

今回、自然音声と合成音声を焦点のない音声として用いたものを基準として、焦点を 1 箇所含んだ音声の基本周波数・音素継続時間の違いから音声の焦点の抽出を行うことにより、各韻律情報と焦点の関係について調べた。

その結果、合成音声も自然音声と同様に韻律情報の基準としての使用が可能であることがわかった。また、焦点に対して基本周波数と同様に音素継続時間も何らかの影響をもつということがわかった。このことから、リズムを変化させることにより焦点を示すことも可能だと思われる。

今後の課題として、

- 1 つの文中に複数の焦点 をもつ場合への対応
- 文節より小さい単位の焦点 の抽出法の確立
- 複数の韻律情報を用いた場合の焦点 の判定法の確立

などが考えられる。

参考文献

[1] 杉藤美代子 編 : “日本語音声 2 アクセント・イントネーション・リズムとポーズ”, 三省堂, (1997.7).

[2] W.N.Campbell : “PROSODIC ENCODING OF ENGLISH SPEECH”, Proc. ICSLP'92, pp.663–666 (1992).

[3] 音声文法研究会 : 文法と音声, (くろしお出版, 東京, 1997), pp62–64.

[4] 大野, 藤崎, 高橋 : “日本語音声における強調の韻律的特徴に与える影響について”, 音講論, pp.201–202 (1998.9).

[5] W.N.Campbell, A.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, (1996,5).