音声言語情報処理　9－2
(1995. 12. 14)

基本周波数パターンの部分 AbS 方式による フレーズ境界の推定に関する検討

桜井 淳宏 広瀬啓吉

東京大学工学部
〒113 東京都文京区本郷 7-3-1 東京大学工学部
atsuhiro@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp

あらまし：韻律的特徴を連続音声認識に利用するために部分 AbS 方式を提案している。この方式は認識結果の各仮説に対して数値モデルにより F_0パターンを生成し、観測される F_0パターンと部分的に比較するものである。ここでは、フレーズ境界がどの程度正確に検出し得るかを調べ、1 モーラの位置ずれを許容して、95%の結果を得た。部分 AbS で得られる観測パターンとモデルで生成されるパターンとの誤差の資料による変動をおさえるために、マイクロプロソディー等による F_0パターンのゆらぎをフィルタリングにより低減する手法を検討した。また、部分 AbS の初期値（アクセント境界の大きさ）を話者によって変える話者適応の検討を行ない、誤差およびその変動を減少し得ることを示した。さらに、フレーズ指令の生起しない、フレーズ境界以外の統語境界について、フレーズ指令の生起を仮定した場合の誤差の傾向を調べ、フレーズ境界との識別の可能性を示した。
キーワード：基本周波数パターン、F_0モデル、部分 AbS、フレーズ境界。

Searching Phrase Boundaries by the Method of Partial AbS of Fundamental Frequency Contours

Atsuhiro Sakurai, Keikichi Hirose
University of Tokyo, Faculty of Engineering
7-3-1 Hongo, Bunkyo-ku, 113 Tokyo, JAPAN
atsuhiro@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp

Abstract: A method was proposed for using prosodic features in continuous speech recognition, and an experiment was realized in which the proposed method, called "partial analysis-by-synthesis", was used to determine the correct position of phrase boundaries. The method consists of comparing the automatically-extracted F_0 contour of an utterance with the contours generated using the recognition hypotheses and a mathematical model. In order to deal with the high variability of the resulting error, even for correct hypotheses, as well as to improve the overall performance of the method, experiments were also carried out under two different conditions: using filtered F_0 contours (the F_0 contours were smoothed in order to minimize microprosodic effects), and adapting the initial conditions of the model (accent command amplitudes) to the specific speech samples of the experiments. At last, the method was evaluated at non-phrasal boundaries, with the objective of detecting different behaviors and qualitative clues that permit the differentiation between phrase boundaries and non-phrasal boundaries.
Key words: Fundamental frequency contours, F_0 model, partial AbS, phrase boundaries.

1 はじめに

近年、連続音声の自動認識に関する研究が盛んに行なわれているが、従来の研究のほとんどでは音声に含まれる音響的特徴のうち、分節的特徴（segmental features）のみが注目され、韻律的特徴（prosodic features）の有効な利用がなされていない。実際、人間同士のコミュニケーションでは分節的特徴のみならず韻律的特徴も情報の伝達に大きな役割を担っていることは良く知られている。従って、連続音声の自動認識の高度化のためには韻律的特徴の利用が不可欠と考えられ、実際、その具体的な方法を開発することが重要な研究課題となりつつある。

韻律情報、特に基本周波数パターンから得られる情報を連続音声認識に用いる方法としていくつかの研究が報告されているが [1] [2]、その多くは音声認識を行なう前に文を韻律語単位に分割し、統語境界さらには文構造を推定しようとするものであり、音声認識にかかわる探索範囲を縮小することを目的としている。しかしながら、韻律的特徴は個人差等の影響を受けやすく、発話毎の変動が大きい上に、必ずしも統語情報を正確に表している訳ではない。そのため、韻律情報のみを用いて統語境界や文構造を推定することには問題がある。

そこで、筆者らは韻律的特徴と分節的特徴を併用して連続音声の自動認識を行なう方式として、基本周波数パターンを用いて従来の分節的特徴による認識結果を部分的に評価する方式（部分 AbS 方式）を提案し、分節的特徴のみからでは曖昧性が生じる若干の文例に対してその有効性を示した [3] [4] [5]。この方式は、数値モデルを用いて認識結果の各候補に対応する基本周波数パターンを生成し、入力音声から直接抽出したパターンとの誤差（対数基本周波数パターン上の距離）を求めることに基づく。

しかし、部分 AbS 方式の結果として得られる誤差の大きさは対象となる部分に強く依存するため、正解・不正解を定める閾値としてそのまま用いることはできない。したがって、正解における誤差のばらつきを縮小することが重要な課題となる。本稿では、誤差のばらつきの縮小に関するいくつかの問題や解決法を検討する。さらに、実験で用いた音声資料により適したアクセント指令の初期値を求めて性能の向上をはかることを試みる。最後に、本方式を用いてフレーズ境界とそれより小さな統語境界（ここでは便宜的にアクセント境界と呼ぶ）の識別を行なうための基本的な実験を実施する。具体的には、フレーズ境界以外の統語境界に仮にフレーズ指令を挿入し、部分 AbS 方式を用いてフレーズ指令の位置を探索する。この実験の結果によってフレーズ境界特有の（アクセント境界と区別可能な）特徴が明らかになれば、フレーズ境界抽出の手助けとして本方式を用いることが可能となる。

2 部分 AbS 方式

部分 AbS（partial analysis-by-synthesis）方式は、複数の認識仮説が存在するとき、入力音声の注目部分の基本周波数パターンと各仮説に対応する基本周波数パターンとを比較して誤差の大小から正しい仮説を選択する方法である。各仮説に対応する基本周波数パターンは数値モデル [6] を用いて生成するが、入力音声のパターンとの比較に際して、発話毎の様々な要因による変動を、AbS 処理によって吸収する。AbS 方式は本来、観測された基本周波数パターンをモデルで表現するための最適パラメータを求める方法であり、観測パターンとモデルで生成されたパターンとの誤差を最小化するようにモデルの各パラメータ値を探索することに基づくが、本稿の部分 AbS 方式は認識を目的としているため、大きな範囲でパラメータ値を最適化することには問題がある。つまり、パラメータ値の最適化をある範囲内（個人差や発話毎の変動として予想される程度の範囲、**Table 1** を参照）で行なう必要がある。

数値モデルのパラメータの初期値は音声合成システム用に開発された基本周波数の合成規則に基づいて決定した [7]。基本周波数パターン上の誤差は、ここでは、フレーズ境界の位置を中心とした 1 秒間の区間（モーラ数で規定することも考えられる）の 1 点あたりの対数基本周波数の平均自乗誤差として求めた。部分 AbS 方式を用いた統語境界推定に基づく連続音声認識システムの構成を **Figure 1** に示す。

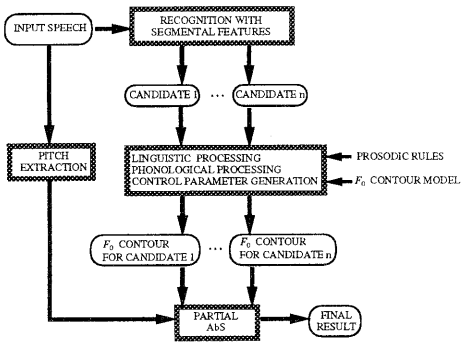

Figure 1. Total configuration of the partial AbS method for searching the correct recognition result from several candidates

3 フレーズ境界推定実験

この実験で、部分AbS方式を用いてフレーズ境界をどの程度の正確さで検出することができるかを調べた。そのために、いくつかの例文に対し、対象としたフレーズ境界を中心として、フレーズ指令の位置についての仮定を、前後に2

Table 1. Restrictions imposed on the parameters of the model for the partial AbS method

Parameter	Allowable Range
Onset Time of the Phrase Command	Up to ± 20 ms from Original Time
Magnitude of the Phrase Command	Up to ± 20% from Original Value
Onset Time of the Accent Command	Up to ± 20 ms from Original Time
End Time of the Accent Command	Up to ± 20 ms from Original Time
Amplitude of the Accent Command	Up to ± 20% from Original Value
Natural Angular Frequency of the Phrase Control Mechanism	Up to ± 20% from Original Value
Natural Angular Frequency of the Accent Control Mechanism	Up to ± 20% from Original Value

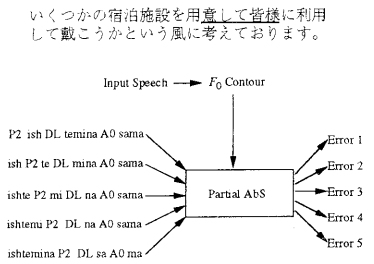

Figure 2. Outline of the experiment of finding the correct position of phrase boundaries

モーラの範囲で変化させて部分AbS処理を行ない、誤差がどのような傾向を示すかを調べた。

実験の概要を **Figure 2** に示す。システムの入力として用いた連続音声はATR連続音声データベース [8] の中から選んだ（男性話者1名）。なお、各例文に対応する基本周波数パターンを生成するため、認識結果を想定したラベルデータの中に、テキストレベルで韻律記号を与えた。**Table 2** はその韻律記号の種類や内容の一覧を示す。

本稿の実験はフレーズ境界に注目するが、文頭や長いポーズ（respiratory pause）の後の境界は比較的容易に抽出することができるため、non-respiratory なフレーズ境界の抽出精度を調べることとした。具体的には、各例文について想定されるP2あるいはP3レベルのフレーズ指令に着目した。

なお、実際の発話においては、フレーズ境界に必ずしもフレーズ指令が現れるとは限らない。従って、実験に先だって通常のAbSによりフレーズ指令の存在を調べ、存在するとされた資料について実験を行なった。

ここで、P2（あるいはP3）の位置を変更するにあたって、いくつかの点に注意する必要がある。まず、フレーズ指令を右（後方）へずらす際、アクセント指令の位置・区間も同時に変化する可能性があることに注意しなければならない。標準日本語の場合、アクセント指令がフレーズ指令と同時に立ち上がるのは1型アクセントのみであり、必ず次のモーラの後で指令は下降する。そのため、本稿では、フレーズ指令を右に移動させたときにフレーズ指令とアクセント指令

Table 2. Command values and positions assigned to the phrase and accent symbols in the prosodic rules. These will serve as the initial parameter values for the process of analysis-by-synthesis.

Command Type	Symbol	Command Value	Position with Respect to Voice Onset (ms)
Phrase Command	P1	0.35	-210
	P2	0.25	-80
	P3	0.15	-80
	P0	(reset)	-80
Accent Command	FH	0.50	-70
	FM	0.25	-70
	FL	0.10	-70
	DH	0.50	-70
	DM	0.35	-70
	DL	0.15	-70
	A0	(reset)	-70

が同時に立ち上がることになった場合、2つの方法で実験を行なった。一つは、アクセント指令の立ち上がり時点をフレーズ指令と同じモーラ数だけ右へずらす方法である（CASE 1）。この際、アクセント指令の立ち下がり時点とかさなる位置までは移動しないようにした。一方、もう一つはフレーズ指令と同時に立ち上がるアクセント指令を次のモーラで下降させる方法である（CASE 2）。もちろん、アクセント指令がもともと次のモーラで下降するようになっていれば、そのままにしておく。この場合、CASE 1はないことにする。

さらに、フレーズ指令がもともとアクセント指令と同時に立ち上がっているときは、アクセント指令もフレーズ指令と同じく左右にずらすことにした。なお、フレーズ指令を左へずらすときは基本的にそのフレーズ指令につづくアクセント指令も同じく左へずらすことにした。フレーズ指令が前のアクセント指令の立ち下がり位置より前に移動する可能性があるが、フレーズ指令はアクセント指令の立ち上がりと立ち下がりの間に位置することがないため、そういう場合はアクセント指令の立ち下がり時点をフレーズ指令と同じ位置まで前方へずらすことにした。また、標準日本語ではフレーズ指令より1モーラ以上遅れてアクセント指令が立ち上がることがないため、フレーズ指令を左へずらすときは基本的にそのフレーズ指令直後のアクセント指令も同じく左へずらすことにした。

4 実験の結果

いくつかの例文において実験を行なった結果、フレーズ指令の位置はだいたい1モーラ程度の正確さで求めることができることが分かった。**Figure 3** は本手法でフレーズ境界が特定できた例（例（a））と特定できなかった例（例（b））を表している。横軸はモデルでのフレーズ指令の位置を表し、縦軸は抽出パターンとの平均自乗誤差を表す。グラフの縦線や文中の矢印はラベルデータによる正確なフレーズ境界の位置を示している。

一般的に、フレーズ境界周辺に促音や無声破裂音等による無声区間がなく、フレーズ境界による基本周波数パターンの立ち上がりが顕著に見られるときは高い精度でフレーズ境界の位置が特定できることが分かった（**Figure 3**(a)）。逆に注目部分内に長い無声区間が存在するときやフレーズ境界に対応するフレーズ指令の大きさが小さく、基本周波数パターンがその部分において平坦に近い場合は誤差がフレーズ境界の位置に依存しなくなる傾向が見られ、フレーズ境界の特定が難しくなる（**Figure 3**(b)）ことが分かった。

このように、合計38個のフレーズ指令を分析し、その結果を **Table 4** にまとめた。**Table 4** から分かるように、最大1モーラのずれでフレーズ境界の位置が検出されたのは95%である。ずれなしでは正解率は約40%にとどまった。この正解率にはまだ改善の余地があるが、ここでは、以下、部分AbS方式で得られる誤差のばらつきを縮小する問題について検討する。

5 フィルタリングの影響

部分AbS方式を用いて統語境界の推定を行なう場合、破裂音に見られる基本周波数の局所的

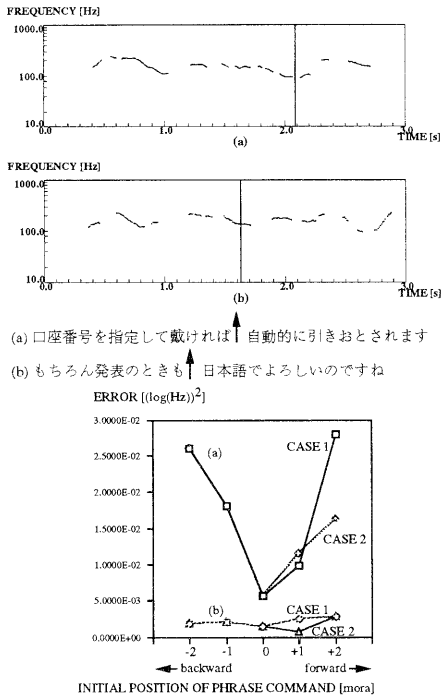

Figure 3. Partial analysis-by-synthesis errors for two speech samples as a function of the position of the phrase command

な起伏などマイクロプロソディによる影響は誤差の値を全体的に大きくするだけでなく、資料毎の変動を大きくし、正解・不正解の基準となる閾値の設定を困難にする。このため、マイクロプロソディの影響を軽減する目的で、自動抽出された基本周波数パターンに適切なスムージング又はフィルタリングを事前に行なうことを考えた。

ここでは、基本周波数パターンを時間軸上の信号として考え、低い遮断周波数（10～30 Hz）のローパスフィルタに通した。フィルタのパラメータは、基本周波数パターンに見られる局所的な起伏を修正し、かつ時間遅れを最小にとどめるように選定した。フィルタリングした基本周波数パターンを用いて実験を行なった結果を **Table 4** に示す。

この結果から分かるように、平均誤差は若干小さくなっているが、フィルタリングによる時間遅れの影響等により正解率がかえって低くなっている。フィルタリングの特性についての検討が必要である。

Table 3. New initial values of accent symbols after adaptation to the speech material

Command Type	Symbol	Command Value	Position with Respect to Voice Onset (ms)
Accent Command	FH	0.62	-70
	FM	0.31	-70
	FL	0.124	-70
	DH	0.62	-70
	DM	0.434	-70
	DL	0.186	-70

6 話者適応に関する課題

部分 AbS を行なって得た F_0パターンを観察した結果、基本周波数のバイアスレベル（以下、F_{min}）の値が全般的に予想より大きくなっていることが分かった。それはアクセント指令の初期値が実験で用いた音声資料に対して小さすぎることによるものと考えられる。そこで、対象とする音声資料のアクセント指令を大雑把に分析し、より適したアクセント指令の大きさで再度実験を行なうことにした。変更後のアクセント指令の初期値を **Table 3** に示す。

このような話者適応を行なった時の部分 AbS の結果を **Table 4** に示す。この結果では、正解率はほとんど変っていないが、誤差のばらつき（標準偏差）が小さくなっていることが分かった。今後はより正確な分析を行なって最適な初期値を求めることを検討する。

7 フレーズ境界以外の統語境界での実験

以上の実験では部分 AbS によりフレーズ境界の抽出がどの程度正確にできるかを調べたが、逆に、フレーズ指令が生起しない統語境界については、指令が生起しないことを検出する必要がある。そこで、フレーズ境界以外の統語境界（アクセント境界）に注目し、フレーズ指令があると仮定して**第 3 節**と同様のフレーズ境界検索実験を行なってからどの様な結果が得られるかを調べた。

実験では、**第 3 節**と同じ音声資料を用いて、10

Table 4. Results of the partial AbS method

	Original Results	With Filtered F_0 Contours	With Speaker Adaptation
Correct Search	16/38 (42%)	9/38 (24%)	14/38 (37%)
Correct Search Admitting Errors of Up to 1 Mora	36/38 (95%)	31/38 (82%)	37/38 (97%)
Mean Error for Correct Hypothesis [(log(Hz)^2]	$5.59\text{x}10^{-3}$	$5.27\text{x}10^{-3}$	$4.86\text{x}10^{-3}$
Standard Deviation for Correct Hypothesis [(log(Hz)^2]	$4.93\text{x}10^{-3}$	$5.14\text{x}10^{-3}$	$3.91\text{x}10^{-3}$

個のアクセント境界を選んでP2（大きさ0.25のフレーズ指令、**Table 2**を参照）を挿入した。それからは以前と同じくその位置をずらして誤差を測定した。

フレーズ指令の位置に対する誤差の分布を見ると、フレーズ指令が実際に存在するときと比べて誤差が大きくなる傾向が見られた他、フレーズ境界で実験を行なったときには見られなかったような誤差の分布パターン（上に凸など）が表れた。しかし、フレーズ指令を挿入することで誤差の値が大きくならなかった場合もあった。これは部分AbS処理の過程でF_{min}の値を最適化しているため、フレーズ指令の挿入による基本周波数パターンの緩やかな上昇がF_{min}の調整によって吸収されるからと考えられる。F_{min}に対する制限が必要である。

8 まとめ

部分AbS方式を用いてフレーズ境界位置の推定を行ない、1モーラ程度の正確さでフレーズ境界を求めることが可能であることを実証した。さらに、マイクロプロソディによる影響をフィルタで除去することを試みた他、パラメータの初期値を音声資料により変更することの実験を行なった。また、フレーズ指令の生起しないアクセント境界での部分AbSの様子を調べ、フレーズ境界とアクセント境界を識別することの基礎的な検討を行なった。

今後、他方式との組合せも考え、統語境界推定アルゴリズムを構築するとともに、認識システムへの組み込みを検討する。

参考文献

[1] 今野博之, 広瀬啓吉, “韻律情報を利用した連続音声認識における句境界の検出”, 平5音講論I,1-8-24, pp.47-48 (1993-10)

[2] 中井満, シンガーハラルド, 匂坂芳典, 下平博, “F_0モデルに基づくアクセントテンプレートの連続整合による句境界検出”, 日本音響学会平成7年度春季研究発表会講演論文集I, pp.21–22 (1995-3)

[3] K.Hirose, A.Sakurai and H.Konno, “Use of prosodic features in the recognition of continuous speech”, *Proc. of ICSLP'94*, Vol.3, pp.1123-1126 (1994-9)

[4] 桜井淳宏, 広瀬啓吉,“部分AbS方式を用いた統語境界の高精度推定”, 日本音響学会講演論文集,2-4-9（1995-3）

[5] 桜井淳宏, 広瀬啓吉,“部分AbS方式による統語境界位置の推定に関する検討”, 日本音響学会講演論文集,2-10-8（1995-9）

[6] H.Fujisaki and K.Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese” *J.Acoust.Soc.Jpn(E)*, Vol.5, No.4, pp.233-242 (1984-10)

[7] K.Hirose and H.Fujisaki, “A system for the synthesis of high-quality speech from texts on general weather conditions”, *IEICE Trans. Fundamentals, Vol.E76-A, No.11, pp.1971-1980* (1993-11)

[8] 武田一哉, 匂坂芳典, 片桐滋, 阿部匡伸, 桑原尚夫, “研究用日本語音声データベース利用解説書”, ATR自動翻訳電話研究所（1988-5）