音声言語情報処理　4－15
(1994. 12. 16)

韻律モデルを用いたF_0クラスタリングに基づくアクセント句境界検出

中井 満 †(††)　　シンガー ハラルド ‡　　匂坂 芳典 ‡　　下平 博 ††

†東北大学 工学部
〒980 仙台市青葉区荒巻字青葉
‡ATR 音声翻訳通信研究所
〒619–02 京都府相楽郡精華町光台 2-2
††北陸先端科学技術大学院大学
〒923–12 石川県能美郡辰口町旭台 15

あらまし　連続音声の認識や理解は非常に困難であり、認識精度、処理効率を上げるためには句境界検出等の支援が不可欠である。本稿ではピッチ情報を用いて連続音声中のアクセント句境界を自動検出する手法について述べる。システムの学習時には視察で与えたアクセント句を藤崎らの提案するピッチパタンのモデルによるパラメータで表現し、クラスタリングの手法によって代表パタン (テンプレート) を作成する。句境界検出時には未知入力音声のピッチパタンとテンプレートを DP 連続整合させることにより、N-best 句境界を検出する。ATR の連続音声データベースを用いた実験では不特定話者に対して、視察句境界のおよそ 90%が正しく検出された。

和文キーワード 句境界検出、ピッチ、韻律、連続音声認識

Accent Phrase Segmentation by F_0 Clustering Using Prosodic Model

Mitsuru NAKAI †(††)　　Harald SINGER ‡
Yoshinori SAGISAKA ‡　　Hiroshi SHIMODAIRA ††

†Faculty of Engineering, Tohoku University
Aramaki - Aoba, Aoba-ku, Sendai-shi, 980 JAPAN
‡ATR Interpreting Telecommunications Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto-fu, 619–02 JAPAN
††Japan Advanced Institute of Science and Technology
15, Asahidai, Tatsunokuchi, Ishikawa, 923–12 JAPAN

Abstract　Continuous speech recognition and understanding is a difficult task and it is indispensable to use phrase boundary information for raising the recognition accuracy. In this paper, we propose an automatic method for detecting accent phrase boundaries in continuous speech by using F_0 information. In the training phase, hand labeled accent patterns are parameterized according to a superpositional model proposed by Fujisaki, and assigned to some clusters by a clustering method, in which accent templates are calculated as centroid of the each cluster. In the segmentation phase, automatic N-best extraction of boundaries is performed by One-Stage DP matching between the reference templates and the target F_0 contour. About 90% of accent phrase boundaries were correctly detected in speaker independent experiments with the ATR continuous speech database.

英文 key words　Segmentation, Pitch, Prosody, Continuous speech recognition

1 はじめに

連続音声の認識や理解は非常に困難であり、膨大な処理時間とメモリを費しているのが現状である。これらの認識精度、処理効率の向上のためには韻律句境界情報等の支援が不可欠であると考えられ、韻律特徴量から境界位置の推定を行なうことは重要は課題となっている。

アクセント句を単位とした場合、句境界がピッチパタン (基本周波数パタン) 上の谷間の構造として比較的明瞭に現れ易いという理由から、これまでにピッチパタンを用いた様々なアプローチが試みられている。例えば、局所的な特徴 (ピッチパタンの谷間、ピッチパタンの局所変化、境界の時間的な間隔など) にスコアを与えて直接的に句境界を推定する手法 [1][2] やピッチパタンの生成モデルに基づく分析合成手法 [3] などが提案されている。

一方、我々はアクセント句境界の検出をアクセントパタン列の認識に置き換えた間接的な検出法である「ピッチパタン連続整合法」を提案した [4][5]。この手法は、アクセント句のピッチパタンの形状は少数個のクラスタに分類できるという仮定、並びに一つの発話はクラスタの代表パタン (テンプレート) の連接で表現されるという仮定に基礎をおいている。類似した手法にアクセント型別に HMM で表現する高橋らの手法 [6] があり、これはアクセント型に関する知識を分類基準として与える手法である。また、クラスタ毎に HMM で学習し、句境界らしさを数値化する手法 [7] が花沢らによって報告されている。以上の手法はピッチパタンに関するモデルをほとんど必要としないという特徴がある。しかし、アクセントを的確に把えたピッチパタンのモデルであれば、アクセント句境界検出においても有効な情報となると考えられる。そこで、本報告では、新たにピッチパタンの生成モデルという枠組を考慮した句境界検出について検討する。

ピッチパタンのモデルについては藤崎ら [8] によって提案されているモデルが広く認められており、アクセント成分、フレーズ成分などのモデルのパラメータの推定に関する研究が数多く行なわれている [9][10]。しかし、モデルは指令のタイミングと大きさを与えるものであり、パラメータからアクセント句境界を直接推定することはできないという問題がある。AbS に基づくパラメータの推定を補助的に用いた今野らの句境界検出法 [11] も句境界の仮説を立てた上で指令を発生させ、ピッチの誤差から仮説を検証する手法であって、句境界とモデルの指令の位置的な関係について扱ってるものではない。そこで、本稿ではアクセント句とパラメータの位置的な関係をモデル化したテンプレートによるパタン連続整合法を提案する。この手法は、従来のパタン連続整合法に比べて、1 つのアクセント句をより低次元のベクトルで表現できるため、ピッチ推定エラーなどの影響を受けにくいこと、また生成モデルを考慮したパタン連続整合ができることなどの特徴がある。

2 句境界検出法の概略

図 1に句境界検出システムの構成を示す。ポーズは入力音声パワーの閾値を設定して検出し、入力文章はポーズ毎に分割されて句境界検出の処理に送られる。また、ピッチ抽出には lag-window 法 [12] を使用する。

システムの学習時には半自動的に抽出されたピチパタンモデルのパラメータによって視察アクセント句をモデル化し、クラスタリングの手法でアクセント句を分類することによって複数のテンプレートを作成する。

認識時には、入力された連続音声の F_0 パタンとテンプレートとの One Stage DP 整合を行い、入力音声区間全体における最小二乗誤差基準による N-best テンプレート系列を求める。得られたテンプレート系列の接続境界に対応する個所が未知入力音声のアクセント句境界として検出される。

3 テンプレートの学習

3.1 アクセント句のモデル化

藤崎らの提案するピッチパタンのモデルでは、ピッチパタンは文頭から文末に向かって緩やかに下降するフレーズ成分と、局所的に起伏するアクセント成分との和で表現され、そのモデルにおける対数基本周波数 ($\ln F_0$) は時刻 t の関数として

$$\begin{aligned}\ln F_0(t) &= \ln F_{\min} + \sum_{i=1}^{I} A_{p_i} G_{p_i}(t - T_{0i}) \\ &+ \sum_{j=1}^{J} A_{a_j} \left\{ G_{a_j}(t - T_{1j}) - G_{a_j}(t - T_{2j}) \right\}\end{aligned} \tag{1}$$

により与えられる。ここで $F_{\min}$ は声帯振動が可能な最低周波数、I, J は 1 つの発話におけるフレーズ数およびアクセント数、A_{p_i}, A_{a_j} は i 番目のフレーズおよび j 番目のアクセントの大きさ、T_{0i} は i 番目のフレーズの開始点、T_{1j}, T_{2j} は j 番目のアクセントの開始点及び終了点である。また $G_{p_i}(t), G_{a_j}(t)$ はそれぞれフレーズ制御機構のインパルス応答関数、アクセント制御機構のステップ応答関数であり、α_i, β_j をそれぞれの固有角周波数とすれば

$$G_{p_i}(t) = \alpha_i t e^{-\alpha_i t} \tag{2}$$

$$G_{a_j}(t) = \min[1 - (1 + \beta_j t) e^{-\beta_j t}, \theta] \tag{3}$$

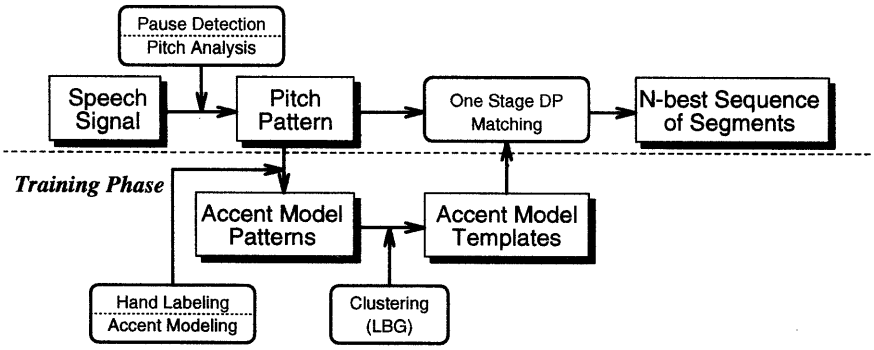
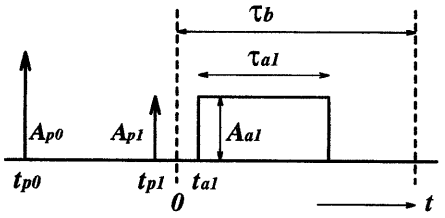


図 1: 句境界検出システムの概略



current accent command	
τ_b:	duration of accent pattern
A_{a1}:	amplitude of accent command
t_{a1}:	onset of accent command
τ_{a1}:	duration of accent command
previous phrase command	
A_{p0}:	magnitude
t_{p0}:	instant of occurrence
current phrase command	
A_{p1}:	magnitude
t_{p1}:	instant of occurrence

図 2: アクセントモデルのパラメータ

である。ただし、$t \leq 0$ ではともに 0 であり、θ は $G_{a_j}(t)$ の上限値 (およそ 0.9) である。

我々はこのアクセント指令、フレーズ指令のパラメータを用いて、1 つのアクセント句に対し、図 2のようなモデル化を行なう。ここでは、着目している当該アクセント句に影響を及ぼすパラメータは当該アクセント句のアクセント指令 (a_1) と直前のフレーズ指令 (p_1)、および先行アクセント句の直前のフレーズ指令 (p_0) の大きさとタイミングのみを考えている。アクセント成分は正と負の等しい大きさのステップ応答によって打ち消し合うことから、後続のピッチ周波数値にあまり影響を与えないと考えられるため、先行アクセント句のアクセント指令は当該アクセント句のモデルの要素に加えない。また、当該アクセント句内で後続アクセント句のアクセント成分が立ち上がることがあるが、後述の句境界検出法の性質上、参照テンプレートの終端のタイミングが一定では無いので、これも除外する。さらに、固有角周波数 α, β については文献 [10] の値を使用して、それぞれ 3.0, 20.0 に固定した。これらの値は話者、発話様式の違いによる差 [13] が他のパラメータに比べて小さく、モデル化に関してはほとんど影響がないと思われる。

3.2 アクセントモデルのクラスタリング

モデル化したアクセントパタンをクラスタリングの手法 (LBG 法) を用いて分類し、参照テンプレートを作成する。分類に用いる特徴量として、モデルのパラメータを要素としたベクトルを用いる場合と、モデルから生成された対数ピッチ周波数値のベクトルを用いる場合が考えられる。ここでは、句境界検出時のピッチパタン整合において対数ピッチパタンの二乗誤差基準を使用することから、後者の特徴量を用いる。

まず、式 (1) のピッチパタンモデルに基づいて、個々のモデル化アクセント句を対数 F_0 パタン $P_n = (p_{n1}, \cdots, p_{nL})$ に変換する。ここで、p_{ni} は対数 F_0 値の時系列であり、L はパタン長を揃えるための固定値である。このとき、2 つのパタン P_1, P_2 間の距離を

$$D(P_1, P_2) = \sum_{i=1}^{L} (p_{1i} - p_{2i})^2 \qquad (4)$$

で定義する。この距離尺度を基準にクラスタリングを行ない、クラスタが収束する毎に各クラスタに属して

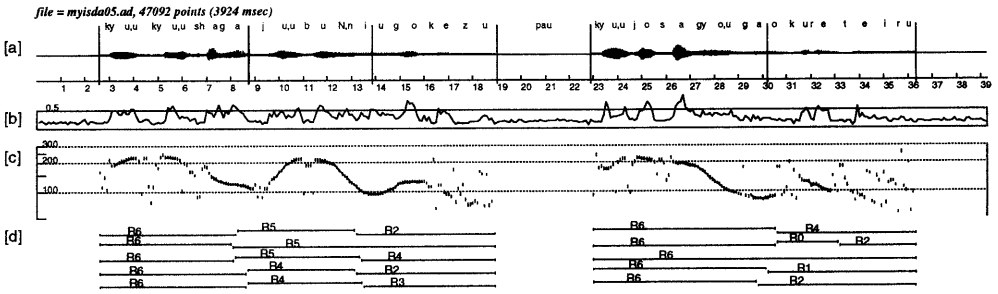


図 4: 句境界検出結果の例

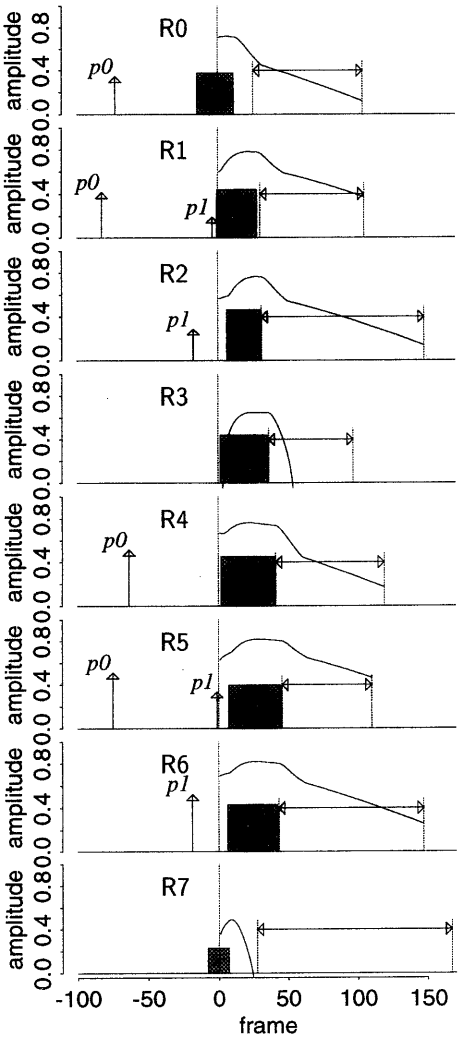


図 3: クラスタリング結果の例

いるアクセントモデルの平均パラメータを計算し、当該クラスタを代表するアクセント句モデル (以下、代表モデル) を求める。この代表モデルから生成される F_0 パタンをテンプレートと呼ぶことにする。

図3はクラスタ数が8の場合の分類結果の例である。横軸は時間軸 (1 frame = 10ms) であり、アクセント句の始点を 0 で示している。また、縦軸は指令の大きさであり、フレーズの指令を矢印 (↑) で、アクセント指令を長方形でタイミングと大きさを表している。代表モデルの後方に示された矢印の区間 (↔) はテンプレートの遷移可能な区間 (4.1節参照) を表している。

なお、2 つのフレーズ指令 p_0, p_1 の有無による組み合わせで 4 タイプのモデルが存在するため、クラスタリングにおいて、それぞれのタイプが分類し易いようにフレーズ指令の有無に重みを付けるという操作を加えている。

4 アクセント句境界の自動検出

図4を参照して句境界検出の流れを簡単に説明する。まず入力音声信号 ([a]) からピッチ抽出を行ない、ピッチパタン ([c]) を推定する。このとき同時に自己相関関数のピークの高さ ([b]) を利用してピッチの信頼度とする。学習で得られたテンプレートとピッチパタンを連続整合することにより、N-best 句境界候補 ([d]) が検出される。

図中、[a] の波形を分割している縦線は視察で与えたアクセント句境界である。また横軸の目盛は分析の 10 フレーム単位で刻まれており、1 目盛は 100ms (1 フレーム=10ms) である。[d] 中の R が添えられている横線のそれぞれが 1 つのテンプレートと整合していることを表し、R0 ~ R7 は図3のテンプレート番号に対応している。

4.1 テンプレートの連続整合

未知音声の入力ピッチパタンと代表モデルから生

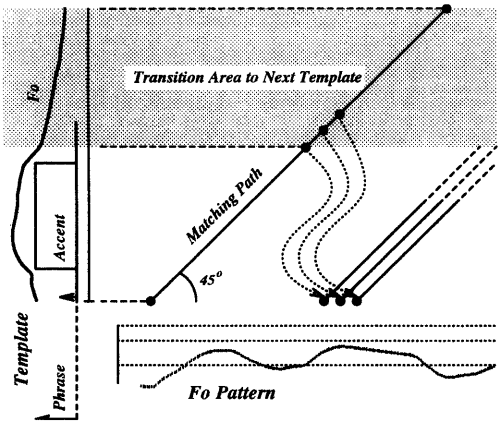


図 5: テンプレートの整合パス

成した F_0 パタンテンプレートとの連続整合は One Stage DP [14] によって行う。このため、整合テンプレート列はフレーム同期で探索され、アクセント句境界も高速かつ安定に求まる。パタン間の距離は全て対数ピッチ周波数の二乗誤差を基準とする。また従来のパタン連続整合法 [4][5] のテンプレートの場合と異なり、整合には図 5のような非線形の伸縮を許さないパス制限を与える。これはアクセントモデルで生成される F_0 パタン上の任意の値が各指令の大きさと指令発生からの経過時間によって一意に定まるためであり、新たな指令が発生しない限り規則的に変化するためである。更に、式 (2)、(3) における固有角周波数 α、β の値を固定しているため、各指令による対数ピッチ値の増加、減衰速度も等しく、傾きが 1 ($= 45°$) の場合のみを考えればよい。

このとき、テンプレートの長さを固定長にして終端フレームからの遷移しか許さないものとすると、テンプレート系列の全パタン長と入力パタン長が一致しないことによる整合エラーが生じる。したがって、テンプレートの終端のタイミングを何らかの統計的な手法で制御して自由度を高くする必要がある。本システムでは個々のテンプレートに遷移可能領域を設定して、その領域内から他のテンプレートに遷移できるようにする。それぞれのテンプレートについて遷移が可能なフレームは①クラスタに属するアクセントパタンの最小長、②クラスタに属するアクセントパタンの平均長の 1/2、③代表モデルのアクセント指令の終了時、の 3 つを超える時点から始まり、クラスタに属するアクセントパタンの最大長に至るまでの範囲とする。図 3 の例では、この区間を矢印 ($\leftrightarrow$) で示してある。

アルゴリズム (1-best)

未知入力パタンのフレーム： $i = 1, \cdots, N$
ピッチテンプート番号： $k = 1, \cdots, K$
ピッチテンプレート kのフレーム： $j = 1, \cdots, J_k$
対数ピッチ周波数値： $P(i)$
テンプレート kにおける対数ピッチ： $T_k(j)$
入力フレーム i におけるピッチの信頼度： $r(i)$
(i,j,k) におけるフレーム間距離：
$d(i,j,k) = r(i)(P(i) - (T_k(j) + F_{\min}))^2$
(i,j,k) における累積距離： $D(i,j,k)$
テンプレート k の終端可能領域： E_k
テンプレート k^* - k の接続コスト： $C(k^*,k)$

Step1 initialize
for $k := 1$ to K do
$D(1,1,k) = 0$
for $j := 2$ to J_k do
$D(1,j,k) = \infty$

Step2 (a) for $i := 2$ to N do steps (b) - (e)
(b) for $k := 1$ to K do steps (c) - (e)
(c) $(j^*,k^*) =$
$\arg\min_{j' \in E_{k'}, k' \in K} [D(i-1,j',k')]$
$D(i,1,k) = D(i-1,j^*,k^*)$
$+ d(i,1,k) + C(k^*,k)$
(d) for $j := 2$ to J_k do step (e)
(e) $D(i,j,k) = D(i-1,j-1,k) + d(i,j,k)$

Step3 Trace back the best path

4.2 N-best 句境界の検出

N-best 法 [15] を使用した句境界検出法は文献 [5] で報告されており、本手法でも N-best 句境界検出を行なう。ただし、N-best の対象はテンプレートの系列である。実際には異なるテンプレート系列であっても、境界候補としては全く等しくなる場合もあり得るし、またテンプレート系列と最適 (最小二乗誤差) に整合しなければならないという条件を除けば、同一系列に対しても複数の候補が存在するはずである。従って句境界候補の観点からすれば N-best ではないが、この条件によって One Stage DP 上での実装が容易になり、高速に N 候補を検出できることが可能となる。

4.3 遷移確率による接続制御

図 3 のテンプレートはフレーズ指令の有無や大きさによって、発声開始時には現れないパタン、特定のパタンにしか接続し得ないパタンなどが生じる。例えば、発声開始時に現れるパタンはフレーズ指令のうちの p_1 のみを持つ R2 と R6 のいずれかであるし、また、p_0 を持つパタンは p_1 を持つパタンに接続するはずである。このようなテンプレートの接続頻度を学習データについて統計的に調査したものが図 6であり、当該パタン (縦軸) と後続パタン (横軸) の接続頻度を面積で表している。ポーズ直後のアクセント句が R0, R3,

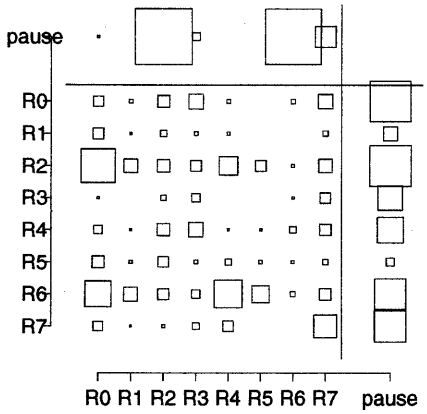


図 6: テンプレートの接続頻度

表 1: ATR 連続音声資料

名称・分類	ATR 日本語音声データベース 連続音声データ
テキスト	音韻バランス 503 文 内分け：10 グループ (A～J) A～I 各 50 文章、J 53 文章

R7 に誤分類されるなどの例も見られるが、おおよそ接続傾向がフレーズ指令の位置関係に従っている。

本システムではこのテンプレートの bigram (#template = 8, perplexity = 3.64) を使用してテンプレートの接続を制御する。距離計算は全て対数尺度の加算で行なっているので、接続コストを

$$C(k^*, k) = -\gamma * \ln(P(k \mid k^*)) \quad (5)$$

として加算する。ここで $P(k \mid k^*)$ はテンプレート k^* から k への遷移確率であり、γ は bigram 制約の強さを表す変数である。

5 アクセント句境界検出実験

5.1 音声資料

ATR の日本語連続音声データベース (表 1) を用いて句境界検出実験を行なう。ピッチパタンの推定は文献 [5] のような補間、スムージング等の処理は行なわず、信頼度を与える。視察アクセント句境界はデータベース付属の韻律情報を使用する。また、アクセント指令、フレーズ指令のパラメータは最小二乗誤差を基準とする手法 [10] で抽出したものを使用する。

表 2: 実験条件

ピッチ抽出部	
分析窓長	512 point (42.7ms)
分析シフト (1 frame)	120 point (10.0ms)
ピッチ探索範囲	(男) 50 ～ 300 Hz
	(女) 100 ～ 500 Hz
抽出法	lag-window 法
句境界検出部	
テンプレート数	8 個
N-best 候補数	10 位

学習データ

男性話者の MHT、MSH、MTK の発話音声 No.51 ～ 503 のうち、モデルパラメータの推定が良好なもの計 565 文章を学習に用いる。

実験データ

男性話者 MHO, MYI、女性話者 FKN、FKS を対象に発声内容が学習データと重複しない No.1 ～ 50 を用いて句境界検出を行う。

5.2 句境界検出実験

実験に使用したパラメータを表 2に示す。また句境界検出時の $F_{\min}$ 値の自動調節は困難なので、今回は $F_{\min}$ が既知の学習用話者と実験対象話者の平均ピッチ周波数値の差から類推した。使用した値はそれぞれ、60Hz(MHO)、80Hz(MYI)、160Hz(FKN)、120Hz(FKS) である。

句境界検出結果は、

$$\text{句境界検出率} = \frac{\text{正解検出数}}{\text{視察句境界の総数}} \quad (6)$$

$$\text{句境界挿入誤り率} = \frac{\text{不正解検出数}}{\text{視察句境界の総数}} \quad (7)$$

によって評価する。ここで正解検出とは視察による句境界の±100ms 内に自動検出されたことを指す。原則として、視察ラベルでポーズとなっている区間と発声区間との境界は視察句境界数に含まないが、ポーズの自動検出で検出できないような短時間のポーズ区間は 1 つの句境界としてカウントする。また N-best 候補に対しては、平均句境界検出率、平均句境界挿入誤り率、N-best (累積) 句境界検出率によって評価する。

5.3 結果と考察

5.3.1 テンプレートの接続コストの強さについて

図 7は話者 MYI を実験対象に、テンプレートの接続コストのパラメータである γ の値を変化させた時の

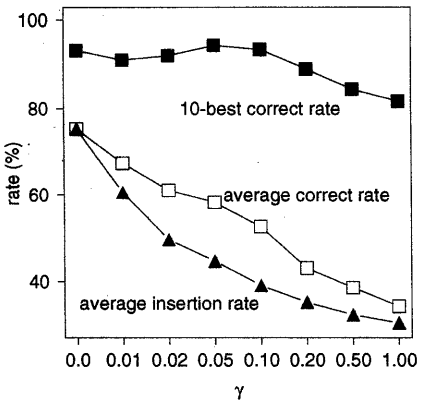


図 7: 接続コストの強さによる句境界検出精度 (MYI)

平均句境界検出率、平均挿入誤り率、および 10-best 句境界検出率についてプロットしたものである。$\gamma = 0.0$ はテンプレートの bigram の制約が全く無い場合であり、大きくなるにつれて制約が強くなる場合である。ただし、距離尺度の正規化を行なっていないので、変数 γ の値自体に意味は無く、正の任意の値で設定している。実験では、値を 0.0 から 1.0 まで対数尺度でほぼ等間隔になるように適当に値を変えて行なった。概して制約が強くなるにしたがって 1 候補あたりの句境界検出率、挿入誤り率、ともに減少する傾向にあるが、10-best 句境界検出率については上位候補で脱落した句境界が下位候補で検出されるこによって補われるので、急激に減少することはない。γ を 0.0 ～ 0.1 まで変化させた場合、10-best 句境界検出率はほとんど変化せず、90% 以上得られているのに対して、挿入誤り率は 30% 以上抑制できることが分かる。したがって、bigram の使用は効率的な句境界挿入誤りの抑制に有効である言える。

5.3.2 N-best 句境界検出率、挿入誤り率について

表 3は話者別に 10-best 句境界候補のうちの 1 位のみ、あるいは、～ 3 位、～ 5 位、～ 10 位までの候補をマージした場合の句境界検出率と 10-best の平均句境界検出率、平均挿入誤り率をまとめたものである。$\gamma = 0.0$ の 1 位検出率がこれまでに報告した手法の結果 [5] に比べて低くなっているが、これはシステムの処理速度を考慮してピッチパタンの補間処理を行なわなかったので、ピッチパタンの精度による影響と思われる。まず、$\gamma = 0.0$ の場合の 1 位検出率が 10-best 平均の検出率に比べてほぼ等しいことに対し、$\gamma = 0.1$ では 1 位候補の検出率が 10-best 平均検出率よりもかなり低いことが分かる。これは、接続コストの影響

表 3: 句境界検出率 (%)

話者	句境界検出率 (%)				10-best 平均	
	1 位	～ 3	～ 5	～ 10	検出	挿入
$\gamma = 0.1$ の場合						
MYI	45.2	69.2	81.2	93.3	52.9	38.6
MHO	55.5	79.2	86.1	94.1	58.5	60.0
FKN	30.6	74.8	67.7	83.6	41.0	84.7
FKS	39.4	64.6	72.1	88.0	47.8	51.4
$\gamma = 0.0$ の場合						
MYI	77.5	86.8	88.2	93.1	75.3	72.0

表 4: 境界検出エラーのうちわけ (MYI)

句境界がピッチパタンの谷間でない	6 個
抽出困難 ··· (～が, も) ある	3 個
ポーズ検出精度の影響	4 個
ピッチ抽出精度の影響	2 個
計	15 個

で、上位候補は比較的少ないテンプレート数で整合する傾向があり、下位候補の方が句境界検出数が多く、句境界検出率、挿入誤り率、ともに高くなる。そのため、上位候補では大局的に見て明瞭な句境界が検出され、下位候補で局所的な境界を補う形となる。また、文献 [5] で報告したピッチのモデルを仮定しないテンプレートによる句境界検出法に比べると、アクセント句の立ち上がりの構造がテンプレートの接続で良好に近似でき、N-best 候補による検出句境界位置も比較的安定している。

挿入誤り率については話者によって大きな差が見られる。これについては、$F_{\min}$ の推定誤差の影響と考えられ、特に女性の FKN では推定誤差が大きいようである。この $F_{\min}$ の推定値によってテンプレートの整合結果が大きく変わるという結果は予備調査で得られいる。

なお、従来のテンプレートの句境界検出実験を話者 MYI について同じ条件のもとで行なった結果、10-best 句境界検出率 87.5%、平均挿入誤り率 67.9% であった。つまり、従来法では視察句境界のうち検出できなかったものが 12.5% であったのに対して、本手法では約半分の 6.7% まで減らすことができた。また、挿入誤り率においても 38.6% まで大幅に削減することができた。

5.3.3 句境界の脱落について

話者 MYI の句境界検出実験で検出されなかった句境界の総数は 15 個 (/223 個) であった。本システムで

抽出困難だったものは「(～が, も) ある」のような文末の短い句、およびピッチパタン上に句境界らしいと思える谷間の構造が見られない句境界である。また、ピッチ推定において信頼度が低いところはN-best 句境界も不安定となる。そのほか、ポーズの検出エラー(MYI の実験でのポーズ検出率は87.2% ポーズ挿入誤り率は7.9%) による影響もあった。これらのうちわけは表4の通りである。その他、句境界としてピッチパタン上に明瞭に表われているものに対しては確実に検出できた。

6 まとめ

本稿では藤崎らによって提案されているピッチパタンのモデルを用いてアクセント句のモデル化を行ない、モデルベースのテンプレートを作成した。未知入力音声のピッチパタンとの二乗誤差最小を基準とした句境界検出では、モデルを仮定したことによるテンプレートの整合規則によりDP パスを簡略化 (45° の直線パス) することができ処理速度の面でも改善できた。また、N-best 句境界候補が安定に検出でき、不特定話者で90%前後の句境界検出率を挙げることができた。しかし、句境界検出の自動化のためには整合時の $F_{\min}$ の自動推定が必要であり、この点については今後の課題としておきたい。

また、将来的にはこれらの句境界情報を連続音声認識に統合していく予定であり、どのようなシステムに対して、どのような形の句境界情報が必要であるのか検討していきたい。

謝辞

研究の機会を与えて戴いたATR 音声翻訳通信研究所 山崎泰弘 社長に感謝致します。また、熱心に討論して戴いた同研究所の 樋口宣男 第2研究室室長、平井俊男 研究員ならびに同研究所の皆様に感謝致します。

参考文献

[1] 浮田輝彦, 中川聖一, 坂井利之, "日本語算術文の音声認識におけるピッチパターンの利用", 信学論 (D), Vol. **J63-D**, pp. 954–961, (1980-11).

[2] 鈴木良弥, 関口芳廣, 重永実, "日本語連続音声認識のための韻律情報を利用した句境界の抽出", 信学論 (D-II), Vol. **J72-DII**, pp. 1609–1614, (1989-10).

[3] H. Fujisaki, K. Hirose, and H. Lei, "Prosody and Syntax in Spoken Sentences of Standard Chinese", In *ICSLP-92*, pp. 433–436, (1992).

[4] 下平博, 木村正行, 嵯峨山茂樹, "ピッチパタン連続整合による連続音声のセメンテーション", 信学技報, **SP90**-72, (1990).

[5] M. Nakai and H. Shimodaira, "Accent Phrase Segmentation by Finding N-best Sequences of Pitch Pattern Templates", In *ICSLP-94*, (1994–09).

[6] 高橋敏, 松永昭一, "統計的韻律モデルによる連続音声の句境界検出", 信学技報, **SP90**-71, (1990).

[7] 花沢利行, 阿部芳春, 中島邦男, "意味主導型音声理解システムのための文節スポッティングの検討", 平6 音講論 I, 1–Q–14, pp. 169–170, (1994-10).

[8] 藤崎博也, 広瀬啓吉, 高橋登, 横尾真, "連続音声中のアクセント成分の実現", 音声研資, **S84**-36, (1984-07).

[9] E. Geoffrois, "Estimation of Prosodic Events from Japanese F_0 Contours", 信学技報, **SP93**-24, pp. 1–8, (1993-06).

[10] 平井俊男, 岩橋直人, Hélène Valbret, 樋口宣男, 匂坂芳典, "統計的手法による基本周波数パタンの制御", 平5 秋音講論 I, 2–8–3, pp. 225–226, (1993–10).

[11] 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).

[12] 嵯峨山, 古井, "ラグ窓を用いたピッチの抽出の一方法", 昭53 信学総全大 1235, (1978–03).

[13] 藤崎博也, 廣瀬啓吉, 高橋登, "共通語のイントネーションの音響音声学的特徴と方言の影響", 音声研資 **S83**-36, pp. 277–284, (1983–10).

[14] Hermann Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition", *IEEE Trans. Acoust., Speech & Signal Process.*, Vol. **ASSP-32**, 2, pp. 263–271, (1984–04).

[15] R.Schwartz and Y.-L. Chow, "The N-best Algorithm: an efficient and extract procedure for finding the N most likely sentence hypotheses", In *ICASSP-90*, pp. 81–84, (1990).