TR-IT-0068

アクセントモデルを用いた F_0 クラスタリングによる句境界検出

Accent Phrase Segmentation by F_0 Clustering Using Superpositional Modeling

中井 満 Mitsuru Nakai
シンガー ハラルド Harald Singer
匂坂芳典 Yoshinori Sagisaka

1994.09.02

概要

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情報等の支援が不可欠である。しかし、プリプロセッサとして連続音声中の句境界を自動的に検出する手法は未だに確立されておらず、そのため連続音声の認識に膨大な時間とメモリを費しているのが現状である。したがって、入力音声から直接、あるいは簡易処理によって抽出された韻律特徴量から高速に境界位置の推定を行なうことは非常に重要は課題である。

そこで本報告では従来のパタン連続整合による句境界検出法を基に、アクセントパタンのモデルを仮定することによるシステムの実装、およびその成果をまとめる。

この手法は従来のアクセント F_0 テンプレートを使用したパタン連続整合に比べて、句境界検出率、句境界挿入誤り率、処理速度のいずれにおいても良い性能を示している。

目次

図目次

表目次

第 1 章

序論

1 研究の背景・目的

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情報等の支援が不可欠である。しかし、プリプロセッサとして連続音声中の句境界を自動的に検出する手法は未だに確立されておらず、そのため連続音声の認識に膨大な時間とメモリを費しているのが現状である。したがって、入力音声から直接、あるいは簡易処理によって抽出された韻律特徴量から高速に境界位置の推定を行なうことは非常に重要は課題である。

そこで本報告では従来のパタン連続整合による句境界検出法を基に、アクセントパタンのモデルを仮定することによるシステムの実装、およびその成果をまとめる。

2 パタン連続整合による句境界検出法

アクセント句の典型的な F_0 パタンと連続音声の F_0 パタンの連続整合による句境界検出法は下平 (現・北陸先端大) が ATR 滞在中に嵯峨山 (現・NTT ヒューマンインターフェース研) の助言を受けて研究し、提案したものである。この手法はアクセント句のピッチパタンの形状は多様であるが、ランダムではなく、あるクラスを構成しているという仮定、並びに、一つの発話 (ブレスグループ) はそれらのピッチパタンの連接で表現されるという仮定に基礎をおいている。具体的には、クラスタリング学習により予め準備された擬似アクセントパタン (ピッチテンプレート) を参照して全発声区間の F_0 パタンからアクセント句を認識し、その結果、アクセント句の接続境界位置から句境界を推定するものである。この手法の特徴は、ピッチパタンにモデルを仮定せず、抽出されたパタンをそのまま扱っている点で、いわゆるデータ駆動型の手法である点である。したがって、最適なパラメータ設定等の煩わしさが少なく学習が容易で、さらに未知入力音声全体に対する最適アクセントパタン列を求めるので、検出が比較的安定であるという特徴がある。

これまでの研究の流れ[1]

- **ピッチパタン連続整合による連続音声のセグメンテーション** [1]

  ピッチパタンと Δ ピッチパタンの混合パタンによる句境界検出。

- **Accent Phrase Segmentation Using Pitch Pattern Clustering** [2]

  ピッチ抽出エラーによる影響を避けるためこれまでの F_0 contour に対して Pitch Spectrum Pattern を提案。

- **ピッチパタンのクラスタリングに基づく不特定話者連続音声の句境界検出** [3, 4]

  話者に依存したピッチの高さの問題を解消するため、従来の F_0 テンプレートを A 型、ピッチの高さ方向にバイアスを与えることのできる新たな F_0 テンプレートを R 型として提案。句境界検出率は上昇するが挿入誤りの増加も伴なう。なお、これ以降の韻律特徴パタンはピッチパタンのみで、複数ピッチ候補から連続処理したピッチパタンを使用している [5]。

[1] 発表順とは異なる。

- **ピッチパタンのクラスタリングに基づく句境界検出法** [6]

  固定長 F_0 テンプレートの伸縮限界による脱落問題を解消するため、アクセント F_0 パタンのクラスタ毎に異なる長さ (クラスタのアクセントパタンの平均長) の F_0 テンプレートを作成。しかし、短い F_0 テンプレートによる句境界挿入が増加。

- **Accent Phrase Segmentation Using Transition Probabilities between Pitch Pattern Templates**[7]

  F_0 テンプレートの接続を遷移確率によって制御することにより、挿入誤りを減少させる。しかし、テンプレートの遷移にあまり偏りがなく[2]、大きな効果は見られない。また、この時点では 1-best の One Stage DP であったで、遷移確率による重みをかけた場合に最適候補であることが保証されない。

- **N-best 法を用いたアクセント句境界候補の検出** [8, 9]

  F_0 テンプレート系列に対する N-best 候補を求めることにより、N 通りの句境界候補を検出。

3 用語

これまでの研究報告では用語の統一がされていなかったので、本報告で使用する用語について定義しておく。

- **ピッチパタン** ··· F_0 の連続量。 (F_0 contour)
- **F_0 パタン** ··· ピッチパタンに同じ。
- **アクセントモデル** ··· 藤崎らの提案するモデル。
- **アクセントパタン** ··· アクセント句を表すパタン。
- **(アクセント)F_0 パタン** ··· アクセントパタンの一種。F_0 パタンで表現したもの。
- **(アクセント) モデルパタン** ··· アクセントパタンの一種。アクセントモデルによって表現されたもの。モデルパラメータの集合。
- **(アクセント) モデルパラメータ** ··· アクセントモデルの要素。
- **テンプレート** ··· 句境界検出において参照される擬似アクセントパタン。アクセントパタンのクラスタリングによって得られる代表パタン。
- **(アクセント)F_0 テンプレート** ··· テンプレートの一種。F_0 パタンで表現したもの。
- **(アクセント) モデルテンプレート** ··· テンプレートの一種。アクセントモデルによって表現されたもの。

[2] おおよそテンプレート数16で perplexity 12 ~ 13

第 2 章

アクセントモデルに基づく句境界検出システム

1　句境界検出法の概略

図 2.1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[1]。これは、N-best 候補の探索に拡張したときに有効であり、処理の効率においても、句境界候補検出精度においても確実に改善される。また、ピッチ抽出には基本的に lag-window 法 [10] を使用する。

本システムの特徴は参照用のテンプレートとして藤崎ら [11] によってモデル化されているアクセント成分およびフレーズ成分のパラメータを使用していることである。現段階ではこれらのモデルパラメータを自動的に抽出する方法は確立されていないが、これまでに平井ら [12] によって半自動的に抽出する良い手法が提案されている。ここで注意しておきたいことは、我々のシステムにおいてモデルパラメータはテンプレートの学習時にしか使用されないので、完全に自動化されたモデルパラメータの抽出アルゴリズムは必要としないということである。したがってシステムの学習時には半自動的に抽出されたモデルパラメータと視察で与えられたアクセント句を照らし合わせてアクセントパタンをモデル化し、クラスタリングの手法を用いて複数のテンプレートを作成することができる。

認識時には、入力された連続音声の F_0 パタンに対してテンプレートによる 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最適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。

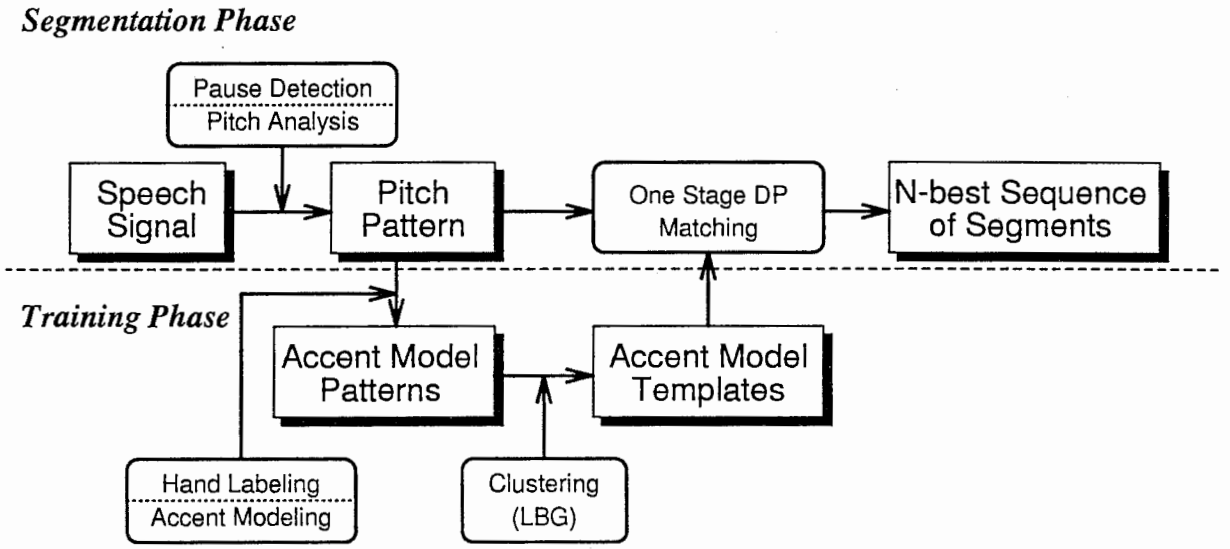


図 2.1: 句境界検出システムの概略

[1] *ICSLP-94* で発表予定の従来法 [8] ではポーズ検出は行なっていたが、句境界検出は文章単位で行なっていた。この点については平 6 秋音講論では改善されている。

2 前処理

2.1 ポーズ検出

ポーズ区間は入力音声のパワーに閾値を設定することで検出する。閾値の決定方法及び、本システム中での使用値を以下に記す。

1. 入力パワーの計算

$$p_n = \sum_{j=0}^{N} \left(H(n,j) * \log \left(\sum_{i=j-r}^{j+r} x_i^2 \right) \right) \tag{2.1}$$

N : 入力フレーム数
p_n : 入力パワー
x_i : 入力音声
r : パワーの有効範囲
$H(n,j)$: スムージング窓(ハミング窓)
(1フレーム10ms、$r = 15$ms、ハミング窓長100msを使用)

2. 閾値の計算

$$p_{\text{high}} = \min \left(\beta \max_n p_n + (1-\beta) \min_n p_n, E(p_n) + \sigma(p_n) \right) \tag{2.2}$$

$$p_{\text{low}} = \max \left((1-\beta) \max_n p_n + \beta \min_n p_n, E(p_n) - \sigma(p_n) \right) \tag{2.3}$$

$$\text{threshold} = \alpha * E(p_n > p_{\text{high}}) + (1-\alpha) * E(p_n < p_{\text{low}}) \tag{2.4}$$

$E(p_n)$: $p_n (1 < n < N)$の平均値
$\sigma(p_n)$: $p_n (1 < n < N)$の標準偏差
($\alpha = 0.15$、$\beta = 0.10$を使用)

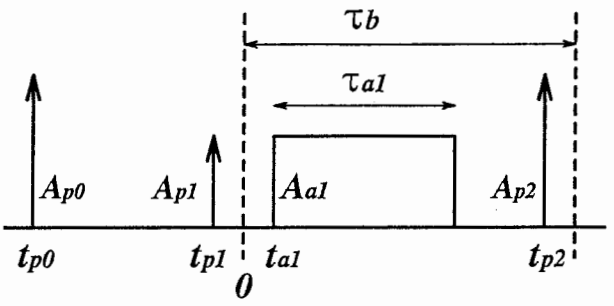
現在、入力は 1 文章毎に行ない、N は入力の全フレーム数をとっている。しかし、リアルタイム処理のためには、この点は改善すべきであり、パワーの閾値を固定にするなどの処置をする必要がある。

2.2 ピッチ抽出

本手法も含め、これまで報告してきたパタン連続整合法を基準とした句境界検出法はピッチパタンとの二乗誤差を基準とするためピッチの抽出精度に大きく依存するところがある。そのため、ピッチの連続性を重視したピッチパタンとして、複数の狭周波数帯域からピッチ候補を求め、DP 法で統合する手法 [5] を提案していたが、実際のところ高周波数帯のピッチ成分の信頼性は低く、また、狭周波数窓幅もデータに依存して調整する必要が生じるであろう[2]。

また処理時間を多く要するばかりではなく、1 フレームのピッチ抽出エラーが前後に及ぼす影響も無視できない。

そこで、本システムでは上述の方法は使わず、 lag-window 法 [10] を使用してピッチ抽出を行なう。ただし、このとき自己相関関数から得られるピークの高さをピッチの信頼度として付与する。



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
succeeding phrase command	
A_{p2}:	magnitude
t_{p2}:	instant of occurrence

図 2.2: アクセントモデルパラメータ

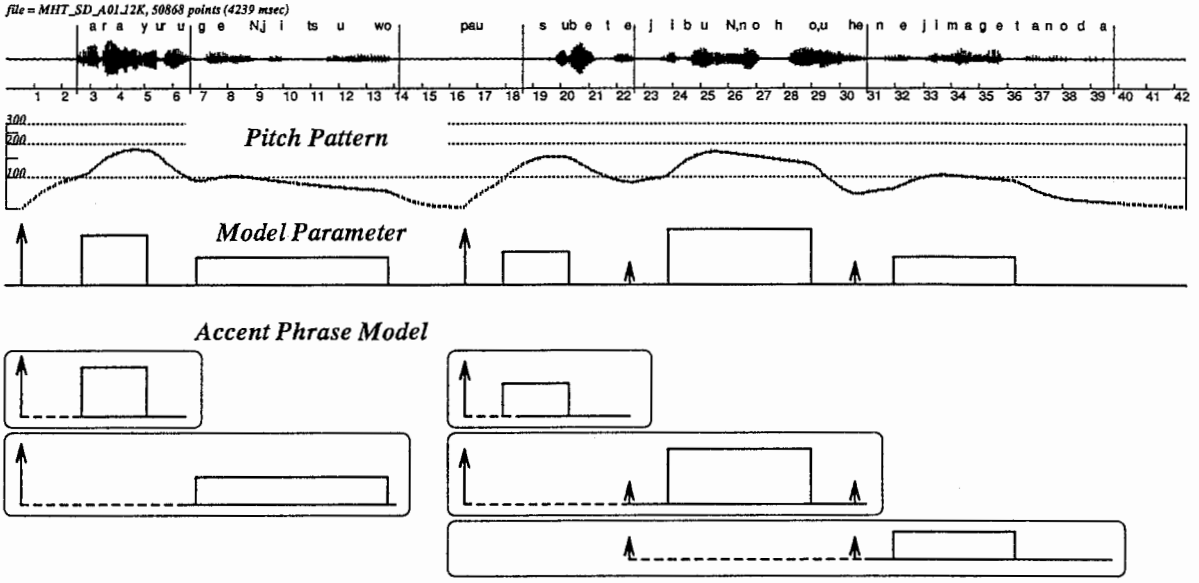


図 2.3: 1 文章から得られる 5 アクセントモデルパタン

3 テンプレートパラメータの学習

3.1 アクセントモデルパラメータ

藤崎らのピッチパタンのモデルによると、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 $\ln F_0$(対数基本周波数) は時刻 t の関数として

$$\ln F_0(t) = \ln F_{\min} + \sum_{i=1}^{I} A_{p_i} G_{p_i}(t - T_{0i}) + \sum_{j=1}^{J} A_{a_j} \left\{ G_{a_j}(t - T_{1j}) - G_{a_j}(t - T_{2j}) \right\} \tag{2.5}$$

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 は一文中での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

$$G_{p_i}(t) = \alpha_i t e^{-\alpha_i t} \tag{2.6}$$

[2]狭周波数窓にすることによってピッチ抽出の精度が上がることは確認されている。

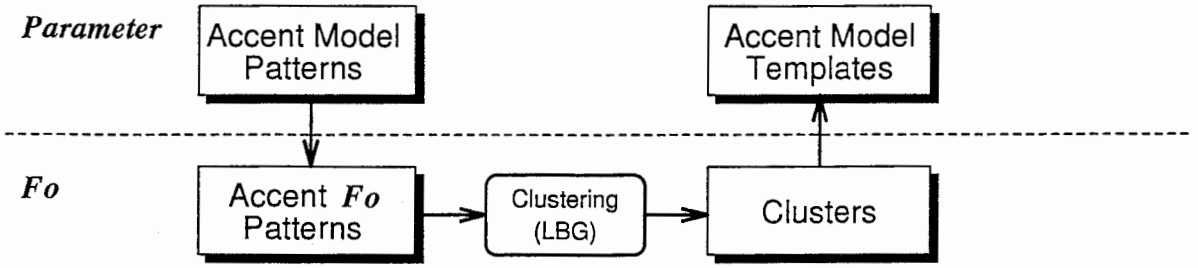


図 2.4: アクセントモデルパタンのクラスタリング (a)

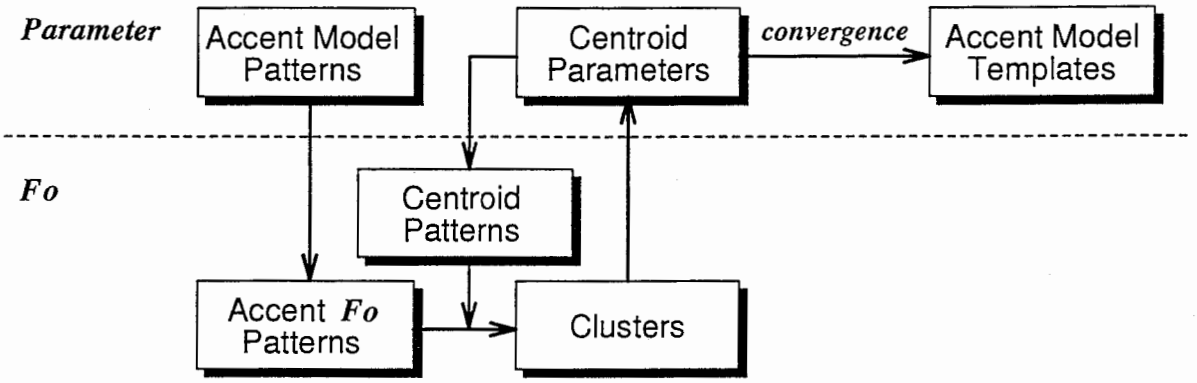


図 2.5: アクセントモデルパタンのクラスタリング (b)

$$G_{a_j}(t) \quad = \quad \min[1-(1+\beta_j t)e^{-\beta_j t}, \theta] \tag{2.7}$$

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成分・フレーズ成分のパラメータを用いて、1 つのアクセント句に対し、図 2.2のようなモデル化を行なう。ここでは、今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そのアクセント句内に発生している各指令と直前のフレーズ指令、および 1 つ前のアクセント句の直前のフレーズ指令のみを考えている。実際には 1 つ前のアクセント成分も少なからず影響しているのであるが、アクセント指令は正と負のステップ応答によって打ち消し合い、後続のアクセント句にあまり影響を与えないことから、ここでは考慮しない。また、アクセント句内で後続のアクセント指令が開始することがあるが、後述のテンプレートの性質上、無視することにする。さらに、α, β についてはそれぞれ 3.0, 20.0 として固定した。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、ましてや、本レポートで扱うデータベースに関してはほとんど差が見られないと予想されるからである。これらについての検討は今後の課題として残しておきたい。

図 2.3は 1 文章中の 5 つのアクセント句についてモデルパタン表現したものである。それぞれアクセント指令については当該アクセント句の要素をそのまま取り出し、フレーズ指令については先行・当該の 2 つのアクセント句に影響を及ぼす 2 つの要素、もしくは、当該アクセント句内で後続のフレーズ指令が発生した場合にはそれを含めて 3 つの要素を抽出している。ただし先行アクセント句に影響を及ぼすフレーズ指令とは 1 つ前のアクセント句の開始直前の指令であって、指令が無い場合には無くても構わず、2 つも 3 つも前のアクセント句まで逆昇することはしない。また図のように途中でポーズが検出されれば、その次のアクセント句を先頭アクセント句として処理する。

3.2　アクセントモデルパタンのクラスタリング

図 2.1の *Training Phase* をもう少し詳しく図示すると図 2.4のようになる。現段階ではアクセントモデルパタンを一度 F_0 パタンに変換して LBG 法 [14] によるクラスタリングを行なった後、各クラスタに属しているアクセントモデルパラメータの平均を計算して、テンプレートにしている。平均を計算する際に注意すべきことは、アクセント句の数とフレーズ指令の数が一致しないということである。つまり、モデル表記上のフレーズ指令 i (P_i) の大きさ $A_{pi}(i=0,1,2)$ が 0.0 のときには、タイミング t_{pi} の値は特定されない。これについては、大きさ A_{pi} はクラスタのメンバー数の平均、タイミング t_{pi} はクラスタのメンバー中の P_i の個数の平均をとることにする。例え

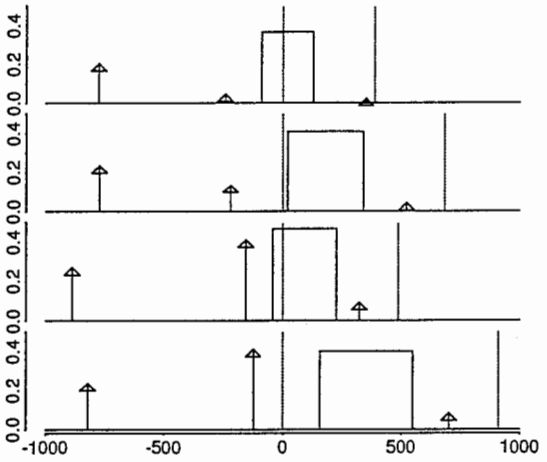


図 2.6: クラスタリング結果の例

ば図 2.3の場合、アクセントの数は 5 であるが、先行 ($p0$)・当該 ($p1$)・後続 ($p2$) フレーズ数はそれぞれ、3、4、1 である。このとき大きさ A_{p0} A_{p1} A_{p2} はアクセント数 5 で割って平均を取り、タイミング t_{p0} t_{p1} t_{p2} はそれぞれのフレーズ数 3、4、1 で割って位置を決定する。この方法ではクラスタ数の小さい時にはフレーズの有無による分類が不十分で、平均をとったためにフレーズ指令が小さくなったテンプレートが続出するであろうが、クラスタ数が増すにつれて解消されると予想される[3]。

また、クラスタリングの手法は従来のものを使い、F_0 表現の領域で計算しているため、実際の F_0 のクラスタ重心とモデルテンプレートのパラメータで生成される F_0 パタンはおそらく大きく異なる。将来的 (図 2.5) にはクラスタを生成した後にパラメータ重心を計算し、再び重心パタンに変換して再度クラスタリングするというように $parameter \leftrightarrow F_0$ 間の変換を密にした手法も検討中であるが、この場合、クラスタの収束が可能かどうかの問題が残される。

この他にも

- アクセント成分とフレーズ成分による separate VQ [4]
- アクセント指令による振幅の差を吸収するような距離尺度
  (あるいはテンプレートの接続ポイントを強調するような距離尺度)

など、改良の余地がたくさん残されている。

図 2.6はテンプレート数が 4 の場合のクラスタリング結果の例である。横軸は時間で、0 がアクセント句の開始時間を示す。縦軸は指令の大きさである。大別して、アクセント指令がアクセント句内で発生しているものとアクセント句の前に発生しているものの 2 種類、直前のフレーズ指令が大きいものと小さいものの 2 種類、それぞれの組み合わせで合計 4 種類のテンプレートになっている様子が分かる。

従来からの距離尺度

異なるアクセントパタン間の距離を簡単に定義するために、ここでは 2 つの距離尺度、1 つはパタンの形状に関する距離、もう 1 つは長さに関する距離、の組合によって定義する。今ここに学習アクセントパタンの集合

$$P_j = (p_{j1}, \cdots, p_{ji}, \cdots, p_{jL_j}) \tag{2.8}$$

がある。ここで p_{ji} は j 番目のアクセントの i フレームにおける対数ピッチ値である。パタンの形状に関する距離を最小二乗誤差基準で簡単に定義するために等しい長さに線形伸縮したパタンを

$$\hat{P}_j = (\hat{p}_{j1}, \cdots, \hat{p}_{ji}, \cdots, \hat{p}_{jL}) \tag{2.9}$$

[3] これについては未確認である。

[4] 簡単な実験ではフレーズ成分のみをクラスタリングした方がテンプレート間の制約が強いようである。

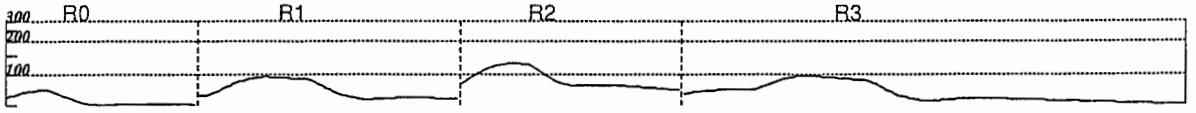


図 2.7: アクセントモデルテンプレート

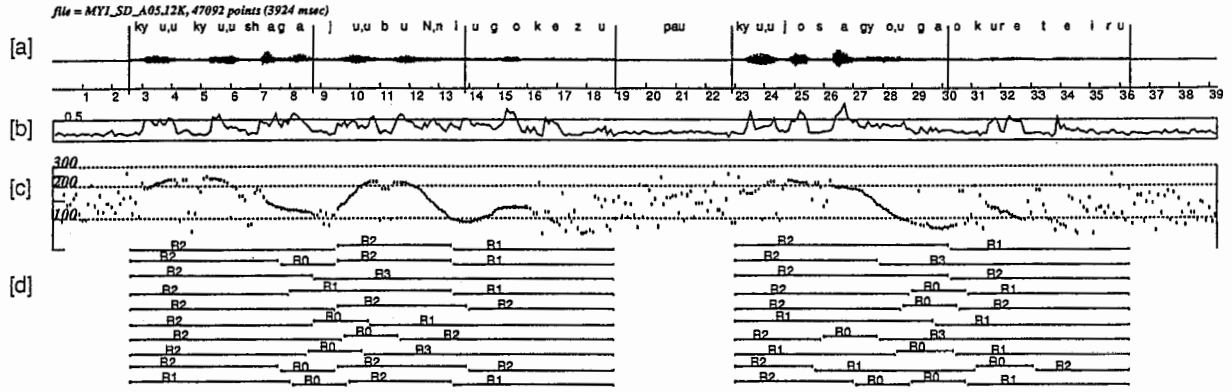


図 2.8: 句境界検出結果の例

とする。このとき、2 つのパタン $\hat{P}_j$ と $\hat{P}_k$ 間の距離は

$$D_S(\hat{P}_j, \hat{P}_k) = \sum_{i=1}^{L} (\hat{p}_{ji} - \hat{p}_{ki} - a)^2 \tag{2.10}$$

で定義される。ここで、a はバイアスであり、従来法でいうところの R 型テプレートであれば $a = \hat{p}_{j1} - \hat{p}_{k1}$ であるが、アクセントモデルにおいては相対的なピッチではなく、ピッチの高さそのものを使用するので $a = 0$ である。一方、長さに関する距離は

$$D_L(\hat{P}_j, \hat{P}_k) = (L_j - L_k)^2 \tag{2.11}$$

として定義する。これらの 2 つの距離尺度を使って、2 つのパタンを

$$D_\lambda(\hat{P}_j, \hat{P}_k) = (1 - \lambda) D_S(\hat{P}_j, \hat{P}_k) + \lambda\gamma D_L(\hat{P}_j, \hat{P}_k) \tag{2.12}$$

と定義する。ここで λ は D_L に対する重み係数であり γ は D_L の正規化係数で、

$$\gamma = \frac{\sum_{\hat{P}_n \in \hat{P}} D_S(\hat{P}_n, \bar{P})}{\sum_{\hat{P}_n \in \hat{P}} D_L(\hat{P}_n, \bar{P})} \tag{2.13}$$

のように表される。$\bar{P}$ は $\hat{P}$ の平均 (クラスタ数 1 の場合の重心) である。

4　アクセント句境界の自動検出

図 2.8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]) を記憶してピッチの信頼度として利用する。図 2.7は学習によって得られた 4 つのテンプレートであり、これらとピッチパタンを連続整合することにより、句境界候補 ([d]) が検出される。

なお図中、[a] の波形を分割している線は視察によって与えたアクセント句境界であり、波形の上の文字列は音韻ラベルである。また横軸の目盛は分析の 10 フレーム単位で刻まれていて、1 目盛は 100ms (1 フレーム =10ms 換算) である。[d] 中の横棒それぞれが 1 つのテンプレートと整合していることを表し、線上に添えられた R で始まる文字が図 2.7のテンプレートと対応している。時間軸方向に見てテンプレートとの整合処理が行なわれていない区間はポーズ検出によって予め除去された区間であって、N-best 候補検出はポーズの前後で別々に処理されている。

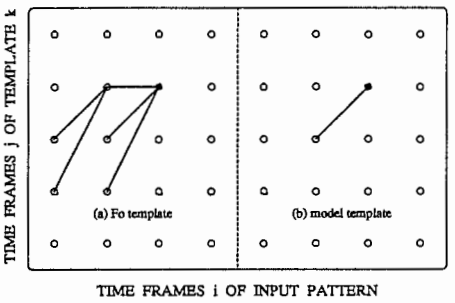


図 2.9: F_0 テンプレートとモデルテンプレートとの整合パスの相違

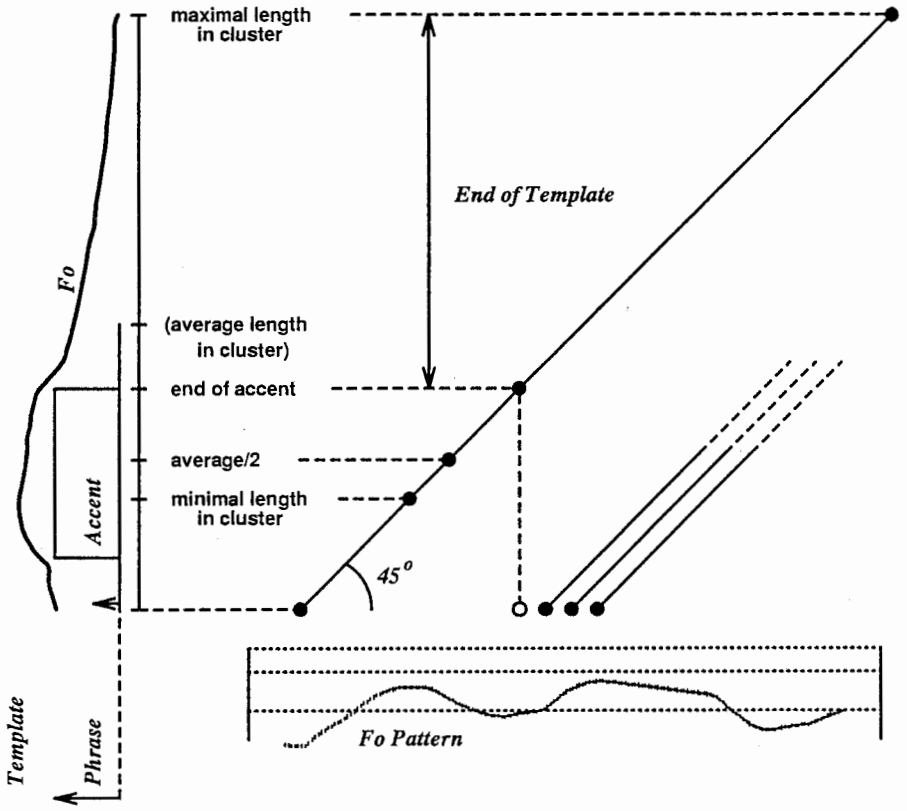


図 2.10: モデルテンプレートの整合パスに関する制約

4.1 テンプレートの連続整合

テンプレートの連続整合は基本的に入力 F_0 パタンとアクセントモデルテンプレートから生成される F_0 パタンの One Stage DP[15] である。パタン間の距離は全て対数尺度を用いて二乗誤差基準で整合する。ただし、F_0 テンプレートでは図 2.9(a) のようなパスを与えていたことに対して、モデルテンプレートでは図 2.9(b) 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 パタン上のあらゆる時間における F_0 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、不規則な変化を考慮する必要がないからである。また、先に述べたように式 (2.6)、(2.7) における固有角周波数 α、β の値を固定にしているため、各指令の増加、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であろう。

この時、問題になるのはテンプレートの終端条件である。従来の F_0 テンプレートと同様にテンプレートの最終端フレームだけでしか次のテンプレートに遷移できないのであれば、テンプレート系列のパタン長と入力パタン長が一致することは極めて稀である。したがって、テンプレートの終端に幅を設けて遷移をある程度自由にしてやらなければならない。本システムでは終端条件として次のような範囲を設定する。

- テンプレート終端の開始点 ··· 以下の全ての条件を見たすとき

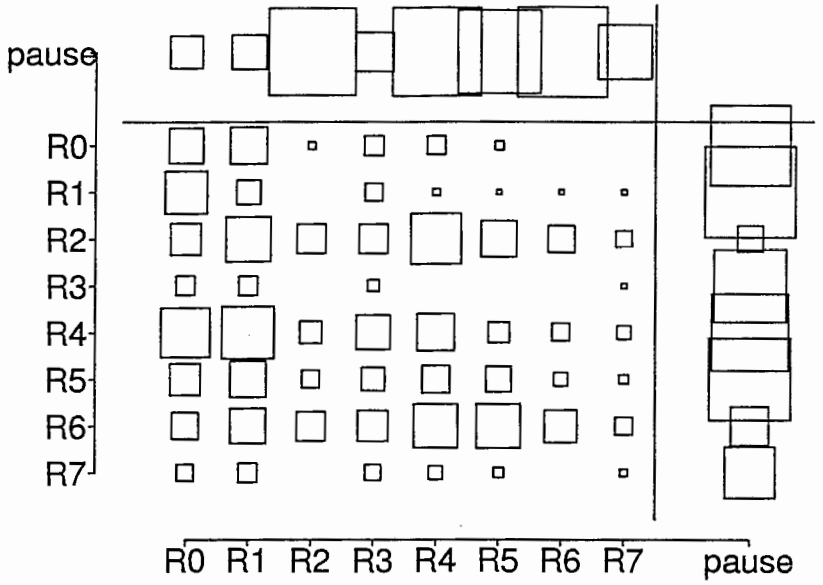


図 2.11: テンプレート間の遷移頻度

1. テンプレートが構成するクラスタに属するアクセントパタンの最小の長さ
2. テンプレートが構成するクラスタに属するアクセントパタンの平均の長さ /2
   (F_0 テンプレートが 1/2 ～ 2 の傾斜で伸縮していることに起因する。)
3. テンプレートのアクセント指令の終わる時間
   (アクセント指令が終了する前に新たなアクセント句が始まることは無いことに起因する。)

- テンプレート終端の終了点[5]

1. テンプレートが構成するクラスタに属するアクセントパタンの最大の長さ

この範囲においてテンプレートは終端することが可能であり、次のテンプレートの先頭フレームに接続することができる。

4.2 遷移確率による接続コスト

モデルテンプレートはフレーズ指令の特徴によって、発声の開始時に現れるパタンと他のアクセント句の後に現れるパタンとに比較的顕著な差がみられる。

図 2.11はテンプレート間の遷移頻度についてまとめたものである。縦軸が遷移前の状態、横軸が遷移後の状態であり、遷移頻度は四角の面積に比例している。pause についてはテンプレートが用意されているわけではなくて、単にテンプレートの遷移の初期状態と終了状態として図示している。ポーズ検出によって分割された句境界検出対象はおよそ平均して 2 ～ 3 個程度のアクセント句で構成されているので統計的にポーズの出現頻度が多くなる。この図から推測できるテンプレートの系列は始めに R2, R4, R5, R6 のいずれかのパタンが現われて、それに R0, R1 のパタンが続き、ポーズになるといったものであろう。これらテンプレートの bigram 情報を用いれば句境界検出の誤りが抑制できると考えられる。距離計算は全て対数値の加算によって行なっているので、このテンプレートの bigram による接続コストを

$$-(\text{scale}) * log_{10}(\text{遷移確率}) \tag{2.14}$$

[5]テンプレート終端の開始点と同様に「テンプレートが構成するクラスタに属するアクセントパタンの平均の長さ × 2」を条件に加えるべきだったかもしれない。

で与える。

現在、終端可能範囲のいずれの点からも等しいコストで遷移が可能なため、接続コストを与えた場合には可能な限り接続回数を少なくしようとする傾向があるし、逆に接続コストを与えない場合にはしばしば終端して新しいテンプレートへと接続しようとする傾向がある。これらの問題を解決するためにアクセントパタンの平均長に対する正規分布的な確率によって遷移をコントロールするなどの方法が考えられるが、これについては今後の課題である。

4.3 アルゴリズム

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：	$j = 1, \cdots, J_k$
(i,j,k) における累積距離：	$D(i,j,k)$
(i,j,k) における高さ方向の対数移動幅：	$O(i,j,k)$
(i,j,k) におけるフレーム間距離：	$d(i,j,k,O)$
対数ピッチ周波数値：	$P(i)$
ピッチテンプレート番号 k における フレーム j の対数ピッチ周波数値：	$T_k(j)$
入力フレーム i におけるピッチの信頼度：	$r(i)$
フレーム間距離：	$d(i,j,k,O) = r(i)(P(i) - (T_k(j) + O))^2$
バイアスの上限：	B
テンプレート k' から k への接続コスト：	$bigram(k',k)$

式 (2.5) における $F_{\min}$ の値は話者に依存してさまざまな値をとるが、テンプレートの $F_{\min}$ の値は学習話者のもので固定されている。このため、従来法で言うところの R 型 (高さ方向に移動可能な) テンプレートの手法と同様にテンプレートに若干の上下移動を与えることにする。B はその時の上限である。

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)

$$(j^*, k^*) = \arg\min_{j',k'} [D(i-1, j', k')]$$

(j'はテンプレートk'における終端可能範囲)

$$O(i,1,k) = \min [P(i) - T_k(1), B]$$

$$D(i,1,k) = D(i-1, j^*, k^*) + d(i,1,k,O(i,1,k)) + bigram(k',k)$$

(d) for $j := 2$ to J_k do step (e)

(e)

$$D(i,j,k) = D(i-1, j-1, k) + d(i,j,k,O(i-1,j-1,k))$$

$$O(i,j,k) = O(i-1,j-1,k)$$

Step3 Trace back the best path

実際には N-best 法 [16] を使用して、N 位までの候補を記憶している。ただし、ここでいう N-best の基準はテンプレートの系列に対してである。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同等な候補となる場合もあり得るし、またテンプレート系列と最適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。従って句境界候補としては 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。

第 3 章

句境界検出実験

1 音声資料

1.1 連続音声データベース

ATR の日本語連続音声データベース [17] を用いて句境界検出実験を行なう。

MHT、MSH、MTK、の三名については平井らの研究 [12] により藤崎らの提案するピッチパタンモデルのパラメータが与えられている。ただし、それぞれ 503 文章中の 200 文程度であり、発話内容については特に統一されていない。

学習データ

男性話者の MHT、MSH、MTK の発話音声 No.51 ～ 503 のうち、アクセントモデルパラメータが与えられている資料を学習に用いる。それぞれの話者の使用した文章番号については付録 A において表でまとめてある。

実験データ

男性話者の MYI、MHO、女性話者の FKN、FKS の発話音声 No.1 ～ 50 を句境界検出の対象とし、話者性、発声内容、ともにオープン実験にする。

表 3.1: 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

1.2 通訳対話音声データベース[1]

4 名 (A,B,C,D) による日本語・英語通訳対話文。このうち話者 A(日本人の発呼者) のデータを自由発話音声として使用する。現在、このデータベースは作成中であり、境界情報が与えられていないものがほとんどであるが、1 部の文章について特別に境界情報を得ることができたので、それらのデータについて句境界検出実験を行なう[2]。境界情報が付与されているものは表 3.2に示されている 103 文のみであるので、これらを句境界検出の対象にする。テンプレートは連続音声資料で学習したものを使用する。

[1]著者が勝手に呼んでいる名称である。

[2]今回は時間の都合により参考程度の実験しか行なっていないので、結果は全て付録 C でまとめる。

表 3.2: 通訳対話音声資料

トピック	旅行・ホテルの部屋の予約 (非対面)	
テキスト	(男性 東京 声優)	TAS12001.A 23 文
		TAS12002.A 20 文
	(女性 大阪 プロナレーター)	TAS22001.A 13 文
		TAS22002.A 16 文
	(女性 大阪 テレビレポーター)	TAS32001.A 18 文
		TAS32002.A 13 文
		計 103 文章

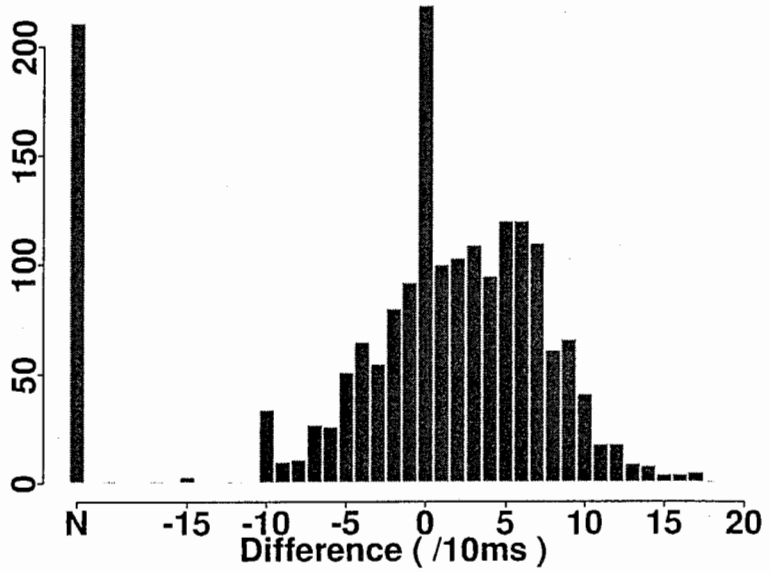


図 3.1: アクセント句境界と F_0 local minima のずれ

2 予備実験

2.1 アクセント句境界と F_0 local minima のずれについて

韻律特徴量から直接的に句境界を検出しようと試みられている研究の多くは F_0 パタンの局所的な谷間の位置を検出することを基本にしている。しかし、局所的な谷間を基準にして検出された句境界は視察あるいは聴取による文節等の意味的なまとまりでラベリングされた句境界と必ずしも一致するとは限らない。具体的にはアクセント句の始まりよりも先に生じるフレーズ指令の影響やアクセント指令の開始の遅れによる影響があり、そのような結果になる。

図 3.1は連続音声の学習データについて句境界位置のずれを調べたものである。(サンプル総数： 1879 句境界) 縦軸は頻度、横軸は視察句境界から見た F_0 local minima の遅れ (+)、進み (−) であり、10ms 間隔 (句境界分析の 1 フレーム相当) で刻んである。"N" で記されているものは視察句境界の前後 200ms 以内に F_0 local minima が見られなかったものであり、全体の 11.2% (211 個) が含まれている。グラフより全体的に視察句境界よりも遅れる傾向があることが分かるが、±100ms 以内に全サンプルの内の 82.7% (1554/1879)、local minima の見られないサンプルを除いたサンプルの内の 93.2% (1554/1668) が含まれている。

表 3.3: テンプレート間の遷移頻度

	R0	R1	R2	R3	R4	R5	R6	R7	pause
pau→	42	46	296	59	307	272	316	115	-
R0→	47	53	2	15	13	3	0	0	249
R1→	70	23	0	12	2	1	1	1	323
R2→	38	78	34	33	99	50	27	10	26
R3→	14	14	0	5	0	0	0	1	204
R4→	94	106	19	46	53	17	12	7	230
R5→	38	50	12	21	29	25	8	3	262
R6→	28	48	32	37	74	76	41	12	57
R7→	11	15	0	10	7	4	0	2	102

表 3.4: テンプレートの bigram による perplexity

	perplexity			
# template	plain F_0 contour	model parameter	phrase parameter	accent parameter
1	1	1	1	1
2	1.929	1.924	1.924	1.932
4	3.567	2.997	3.538	3.249
8	6.340	4.630	5.103	6.779
16	10.412	7.952	7.633	11.940

2.2　テンプレートの遷移の偏りについて

従来の F_0 テンプレートに比べて、モデルパラメータによるテンプレートでは遷移確率の偏りが大きい。これはアクセント句のパラメータ化において、その要素に先行・後続のフレーズ指令を加えたためであって、特にポーズ後の最初のアクセント句と先行する他のアクセント句に接続しているアクセント句とでは大きく異なる。

表 3.3はテンプレート間の遷移頻度であり、図 2.11を数値化したものである。また表 3.4 は F_0 テンプレートとアクセントモデルテンプレートについて bigram の制約のもとでの perplexity をまとめたものである。参考までに、アクセントとフレーズの成分を分離してテンプレートを作成した場合[3]の結果も併記しておく。明らかにアクセントモデルテンプレートの方が perplexity が小さいことが分かる。また、アクセントとフレーズの比較より、これらの傾向がフレーズ成分に依存していることが分かる。これはテンプレートの要素が前述のようにフレーズ成分は前後合わせて 3 つの指令まで考慮しているのに対し、アクセント指令は当該アクセントのみしか要素に入れていないことにも起因しているのであろう。しかし、仮に前後のアクセント成分をモデルパラメータに加えたとしても 3.1節でも述べたように、アクセント成分は正と負の指令によって短時間で打ち消されるので、後続への影響はフレーズ成分に比べてやはり小さいであろうと考えられる。

2.3　処理時間について

F_0 テンプレートにおいては DP パスの傾斜が図 2.9(a) で示される 1/2 ～ 2 であったが、モデルテンプレートでは、傾き 1(45°) の 1 本の経路である。このため、傾き 1 の直線上では N-best 候補の順位が入れ替わることは無く、テンプレート間の遷移の時においてのみ候補の取捨選択をすれば良い。したがって、演算量はかなり削減できる。

図 3.2は入力音声の時間長と N-best 句境界検出に要する時間についてプロットしたものである。対象には話者 MYI の No.01 ～ 50 の 50 文を使用し、テンプレート数 8、N-best 候補数 10、動作環境 HP755/124mips の条件で実験を行なった。なお、同環境においてピッチの分析時間は 入力音声 1 秒あたり 2.42 秒であった。プログラ

[3]ただし、フレーズ成分のクラスタリングについては学習サンプルパタンを固定長に線形伸縮をする操作のかわりに、同一の長さまで、「延長する」という手法を使った。

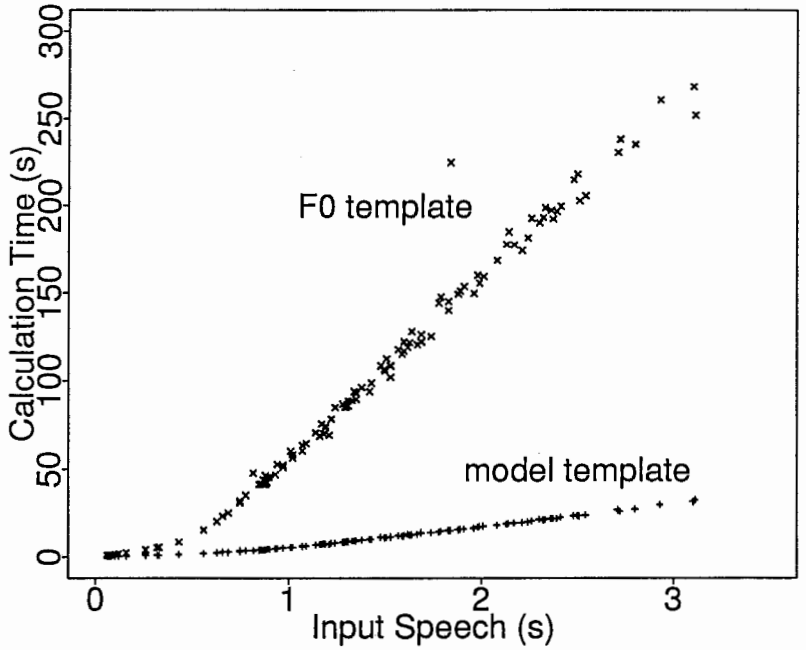


図 3.2: N-best search に要する時間 (HP755/124mips)

ムの実装上では、N-best 候補の選択や候補の履歴比較などが最適なアルゴリズムにはなっていないため、処理の無駄が多く、アクセントモデルテンプレートによる句境界検出は入力実時間に対して約 10 倍の時間を要しているが、それでも F_0 テンプレート に比べれば 7 ～ 8 分の 1 程度の処理で済む。

3 句境界検出実験

実験に使用したパラメータを表 3.5にまとめておく。

表 3.5: 実験条件

ピッチ抽出部	
FFT	512 point (42.7ms)
分析シフト	120 point (10.0ms)
ピッチ抽出時の探索範囲	70 ～ 300 Hz
抽出法	lag-window 法 (自動抽出)
句境界検出部	
ピッチテンプレート数	8 個
N-best 候補数	10 位
バイアスの上限	60Hz
bigram の強さ (scale)	0.1

予め断わっておくが、今回の実験では男性、女性のピッチ探索範囲を変更するのを忘れている。そのため、女性の句境界検出結果は明らかに半ピッチの影響がでているので、結果は参考程度にして戴きたい。

句境界検出評価基準

まず、本稿における句境界とはアクセント句とアクセント句の境界である。ポーズとアクセント句の境界はポーズ境界として定義する[4]。ただし、ポーズ検出処理において検出されなかった文中のポーズは句境界として扱う。このとき、

- 未検出ポーズ時間が長く、ポーズの両端がアクセント句境界として検出された場合は 2 つの句境界に対して 2 つ正解検出されたものとする。
- 未検出ポーズ時間が短く、ポーズ間、もしくはその周辺に 1 つのアクセント句境界が検出された場合は 1 つの句境界に対して 1 つ正解検出されたものとする。

また、句境界検出結果は、

$$\text{句境界検出率} = \frac{\text{正解検出数}}{\text{視察による句境界の総数}} \tag{3.1}$$

$$\text{句境界挿入誤り率} = \frac{\text{不正解検出数}}{\text{視察による句境界の総数}} \tag{3.2}$$

によって評価する。ここで正解検出句境界とは視察による句境界の前後 100ms 内に自動検出されたものを指す。また N-best 候補に対しては、n 位候補までの平均句境界検出率、n 位候補までの平均句境界挿入誤り率、n 位候補までの累積句境界検出率、および n 位候補中の最大句境界検出率を挙げた候補についての句境界検出率と句境界挿入誤り率を評価する。

3.1　アクセント F_0 テンプレートによる句境界検出法

パタン連続整合法による句境界検出法のうち、現在までの最高検出率を挙げたものは平 6 秋音響学会 [9] で発表が予定されている N-best 句境界検出法である。ここでは実験環境を統一して追実験を行なう。また、ピッチ抽出法についても修正を行なう。2.2節でも述べたように、これまで複数のピッチ候補を抽出してその中から F_0 contour の連続性を保証するようなピッチを選択し、100% の信頼度を与えてきたが、本報告ではピッチ候補は 1 つとし、ピッチ抽出時の自己相関関数のピークの高さを信頼度として与える。

話者 MYI の実験結果を表 3.6にまとめる。その他の話者については付録 Bを参照してもらいたい。

3.2　アクセントモデルテンプレートによる句境界検出法

本稿で提案するアクセントモデルテンプレートを使用して以下の 3 種類の実験を行なう。

- bigram をテンプレートの接続コストとして使用しない実験。
  F_0 テンプレートを使用した従来の句境界検出法と比較のための実験である。
- bigram の代わりに等確率でテンプレートが接続すると仮定した実験。
  bigram を使用した実験の対象実験である。
- bigram をテンプレートの接続コストとして与えた実験。

これらについても同様に実験結果をそれぞれ表 3.7、表 3.8、表 3.9にまとめる。

また、通訳対話音声データベースについては、bigram を使用した実験のみ行なったので、付録 Cに結果をまとめておく。

[4] 平 6 秋音講論 [9] までの報告ではポーズ境界もアクセント句境界にカウントしていた。

表 3.6: F_0 テンプレートによる句境界検出精度 (話者 MYI)

MYI	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	71.4	(71.4)	[71.4]	102.3	(102.3)
3	80.5	(72.6)	[81.0]	80.5	(98.6)
5	86.0	(72.8)	[86.5]	71.0	(98.8)
10	89.6	(72.3)	[91.2]	65.8	(100.1)

表 3.7: モデルテンプレートによる句境界検出精度 (話者 MYI)

MYI	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	82.5	(82.5)	[82.5]	77.1	(77.1)
3	90.1	(81.9)	[90.2]	64.1	(83.5)
5	93.3	(81.0)	[93.8]	59.8	(88.9)
10	96.4	(81.7)	[97.4]	49.3	(95.1)

表 3.8: モデルテンプレート (等確率遷移) による句境界検出精度 (話者 MYI)

MYI	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	61.9	(61.9)	[61.9]	42.2	(42.2)
3	80.1	(63.5)	[83.5]	34.4	(52.2)
5	85.9	(64.6)	[90.0]	31.8	(57.3)
10	92.3	(66.6)	[96.1]	28.1	(64.8)

表 3.9: モデルテンプレート (Bigram) による句境界検出精度 (話者 MYI)

MYI	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	59.4	(59.4)	[59.4]	41.0	(41.0)
3	76.5	(61.0)	[81.2]	30.0	(45.6)
5	81.2	(62.1)	[87.1]	28.4	(52.0)
10	91.4	(64.9)	[94.7]	25.5	(59.4)

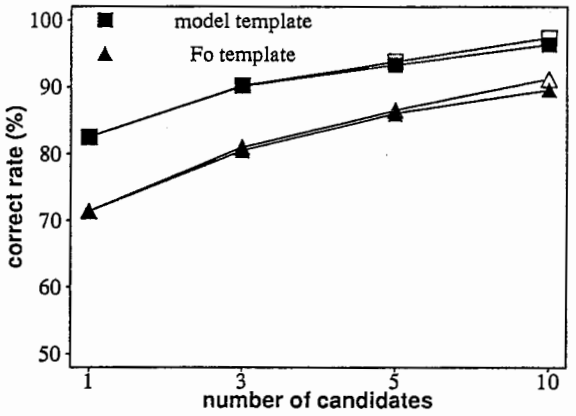


図 3.3: F_0 テンプレートとモデルテンプレートによる句境界検出率の比較 (話者 MYI)

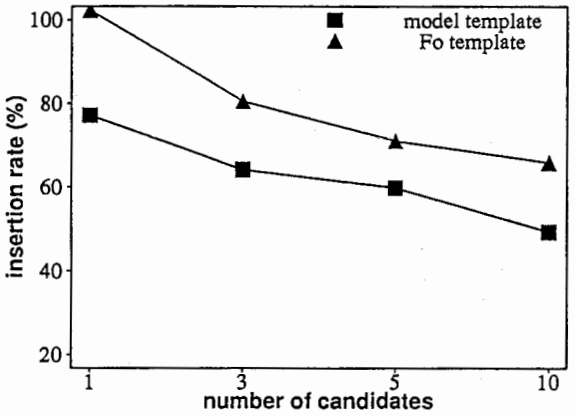


図 3.4: F_0 テンプレートとモデルテンプレートによる句境界挿入誤り率の比較 (話者 MYI)

4　考察

4.1　F_0 テンプレートとモデルテンプレートの比較

図 3.3、図 3.4はそれぞれ話者 MYI についての F_0 テンプレートとモデルテンプレートによる句境界検出率、句境界挿入誤り率の比較である。黒く塗り潰してプロットしたものは N 位候補を個別に評価したもののうち、最大の検出率を挙げた候補についてであり、白抜きでプロットしたものは N 位候補までの累積である。いずれの候補数においても F_0 テンプレートに比べて、モデルテンプレートによる句境界検出率は 7% 以上改善されている。また、どちらのテンプレートにおいても累積検出率に着目した場合、最大句境界検出率とあまり大きな差がないことがわかる。累積検出率とは平たく言えば 1 位候補で検出されなかった句境界が 2 位候補で検出されていれば、加算していくといったものであるが、単一候補による最大句境界検出率と複数候補からなる累積句境界検出率にあまり差がないというのは、挿入誤りが関与しているからであろう。イメージ的にはどの候補もほぼ同じくらいの数の自動検出境界を持っており、わずかなずれによって正解に入ったり入らなかったりしているような感じである。実際、正解率が上がるにしたがって挿入誤りが減っているわけだが、それでも句境界挿入誤り率の依然多いことが指摘される。

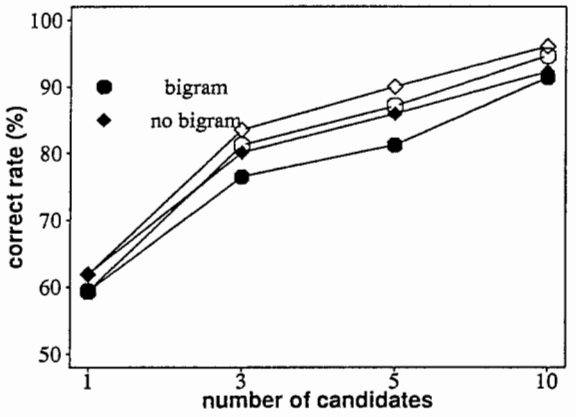


図 3.5: 等確率の遷移コスト、bigram による遷移コストを与えた場合の句境界検出率の比較 (話者 MYI)

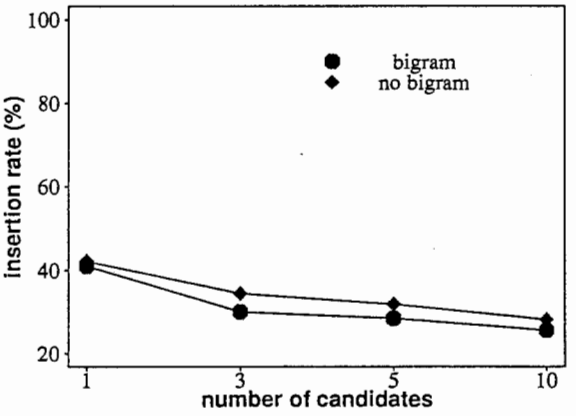


図 3.6: 等確率の遷移コスト、bigram による遷移コストを与えた場合の句境界挿入誤り率の比較 (話者 MYI)

4.2 bigram による接続コストの影響

図 3.5、図 3.6はそれぞれ話者 MYI についての等確率遷移コストを与えた場合と bigram による遷移コストを与えた場合の句境界検出率の比較である。当然のことながら、接続コストを与えたために可能な限り少ない接続回数で整合しようと作用し、結果的に検出数が減少する。このため、挿入誤りをかなり減少させることができるが、可能なテンプレートの組合わせが減少しているわけであるから、同時に句境界検出率も減少する。今回 bigram を接続コストに使用したのは、減少したテンプレートの組み合わせの中でもモデルに則した候補が上位に検出されることで句境界検出率が大きく減少することはないであろうと期待したからであるが、結果的には接続コストを等しくした場合の方が検出率は高くなった。ただし、その分、挿入誤り率は bigram を使った方が低くくなる。

句境界検出率と挿入誤り率のどちらを重視するかは常に問題となる。単一句境界候補の検出をしていたときには「挿入誤りは処理の過程で除去可能であるが、未検出の境界は後処理で検出されることは無い」という考えから、挿入誤りの増加を覚悟の上で句境界検出率を上げていた。しかし、N-best に拡張したことにより、検出されなかった句境界が 2 位以下の候補で補えるので、句境界検出率をできるだけ下げずに挿入誤り率を下げることを検討しなければならないであろう。そのような意味で bigram の接続コストによる挿入誤りの抑制は必要であると思える。ただし、今回のシステムは従来の F_0 テンプレートのシステムの枠組の中で行なったため、まだモデル化としては不十分である。

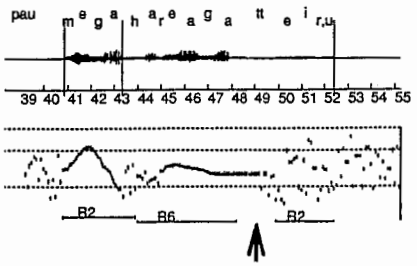


図 3.7: 単語内に誤ってポーズが検出された例

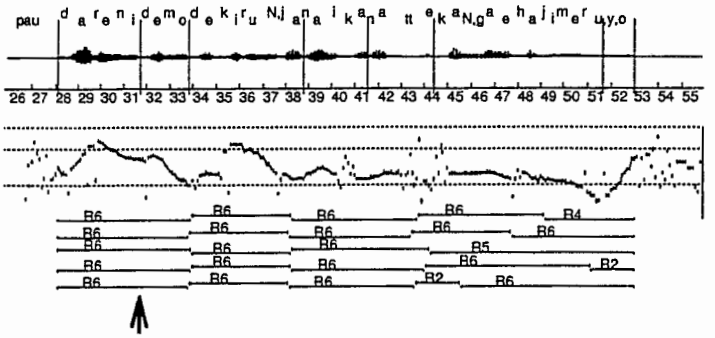


図 3.8: bigram による影響

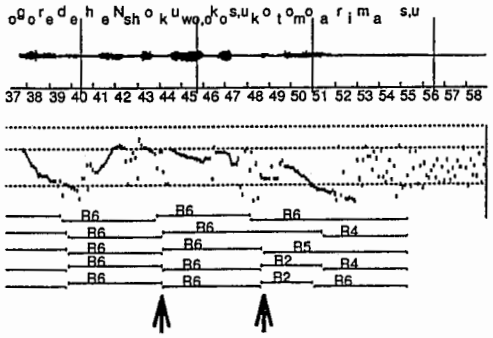


図 3.9: ピッチの抽出エラーによる影響

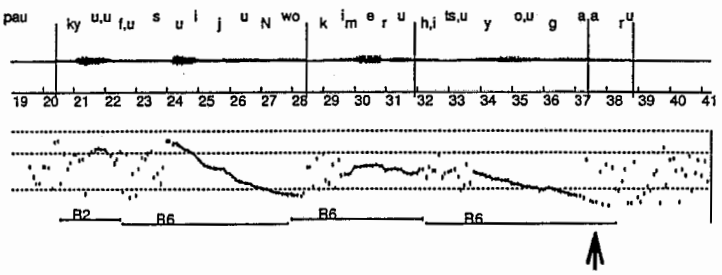


図 3.10: 抽出困難な例

4.3　句境界検出エラーについて

bigram を使用した実験においてどの候補においても検出されなかった句境界が 12 個 (/228 個) あった。そのうちわけは表 3.10の通りである。このうち、「(〜が, も) ある」のような抽出困難な句境界については一見したところアクセントが見られないので特に検出されなくても構わないと思われる。残りのものについては、それぞれポーズ検出精度、ピッチ抽出精度、bigram 制約等々、おおよそ予想通りのものが原因であった。なお、ポーズ検出率は 87.8% ポーズ挿入誤り率は 6.75% である。

表 3.10: 境界検出エラーのうちわけ

単語内のポーズエラー	3 個
bigram の制約の影響	2 個
ピッチ抽出精度の影響	2 個
抽出困難 ··· (〜が, も) ある	5 個

第 4 章

結論

1 まとめ

従来のパタン連続整合法ではアクセント句の F_0 パタンをクラスタリングすることによりアクセント F_0 テンプレートを作成し、句境界検出を行なっていた。しかし、アクセントモデルを仮定していないため単なるアクセントパタン認識の範疇に止まっていた。

そこで、本稿では藤崎らによって提案されているアクセントモデルを使用してアクセントモデルテンプレートによる句境界検出法を提案した。これによりテンプレートの整合規則がモデルによって図られ、処理速度などが改善できた。具体的には

- DP パスの傾斜 45° の整合
- テンプレートの遷移可能領域
- bigram

などの点において改良され、句境界検出率、句境界挿入誤り率ともに向上した。

	(従来)	bigram [累積]
句境界検出率	96.4% (89.6%)	91.4% [94.7%]
句境界挿入誤り率	49.3% (65.8%)	25.5%
処理速度	1/7 ~ 1/8	

2 今後の課題

各節において課題を提示してきたが、テンプレートの学習、整合における課題について、ここでもう一度まとめておく。

テンプレートの学習

今回、句境界検出結果を導くことを第一に研究を進めたため、テンプレートの学習方法は F_0 テンプレートのアルゴリズムを若干修正したものを使用したが、最適なクラスタリング方法について、さらに検討すべきである。例えば、

- パラメータと F_0 の変換を密にしたクラスタリング
- アクセント成分とフレーズ成分による separate VQ
- アクセント指令による振幅の差を吸収するような距離尺度

などである。これらの方法によってテンプレートの整合方法も変わるはずであり、よりピッチパタンのモデルの特徴を扱えるようになるであろう。

テンプレートの連続整合

現在のシステムにおいては

- テンプレートの接続コストの問題
- $F_{\min}$ の話者性に関する問題

などがある。前者については bigram などを誤りの抑制のために使用したが、今回の実験ではその効果があまりみられなかった。これはテンプレートの学習の問題とも関連しているが、現在のアクセントモデルテンプレートがモデルとして十分表現されていないからであろう。また後者の問題については従来法の R 型 F_0 テンプレートと同様に高さ方向へバイアスを加えることで対処したが、それはアクセント句毎に $F_{\min}$ の値が変化することを意味しており、ピッチパタンのモデルに矛盾している。

連続音声認識への応用

将来的にこれらの句境界情報を連続音声認識に統合していくことを考えなければならない。それには、どのようなシステムに対して、どのような形で情報を与えるのかを検討していかなければならないであろう。

謝辞

研究の機会を与えて戴いた 北陸先端科学技術大学院大学 下平博 助教授、 ATR 音声翻訳通信研究所 山崎泰弘 社長に深く感謝致します。研究を進めるにあたり藤崎モデルのパラメータに関して樋口宜男 第2研究室室長、平井俊男 研究員より貴重な御意見と研究データを戴いたことを心から感謝致します。第4研究室の Mark Seligman 客員研究員には ICASSP-95 のプロポーザルを英訳する上で御意見を戴いたことを感謝致します。また、音声データの句境界情報についてラベラの方々に御協力戴いたことを感謝致します。山形大学からの実習生 門前聖康 氏には研究をまとめる上でいろいろと検討して戴きました。ありがとうございました。最後に研究環境を整えて戴いた TSG の皆様、並びに ATR 音声翻訳通信研究所の皆様に感謝致します。

参考文献

[1] 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).

[2] H. Shimodaira and M. Kimura. “Accent Phrase Segmentation Using Pitch Pattern Clustering”. In *ICASSP-92*, pp. I–217–220, (1990).

[3] 中井満, 下平博, 嵯峨山茂樹. “ピッチパタンのクラスタリングに基づく不特定話者連続音声の句境界検出”. 電子情報通信学会論文誌, Vol. J77-A, pp. 206–214, (1994-02).

[4] H. Shimodaira and M. Nakai. “Prosodic Phrase Segmentation by Pitch Pattern Clustering”. In *ICASSP-94, 76.5*, (1994–04).

[5] H. Shimodaira and M. Nakai. “Robust Pitch Detection by Narrow Band Spectrum Analysis”. In *ICSLP-92*, pp. 1597–1600, (1992–10).

[6] 中井満, 下平博, 嵯峨山茂樹. “ピッチパタンのクラスタリングに基づく句境界検出法”. 日本音響学会 平成5年度秋季研究発表会 講演論文集 I, 1–8–23, pp. 45–46, (1993–10).

[7] H. Shimodaira and M. Nakai. “Accent Phrase Segmentation Using Transition Probabilities Between Pitch Pattern Templates”. In *EUROSPEECH-93*, pp. 1767–1770, (1993–09).

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

[9] 中井満, 下平博. “N-best 法を用いたアクセント句境界候補の検出”. 日本音響学会 平成6年度秋季研究発表会 講演論文集 I, (1994–11 予定).

[10] 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).

[11] Hiroya Fujisaki and Hisashi Kawai. “Realization of Linguistic Information in the Voice Fundamental Frequency Contour of the Spoken Japanese”. In *ICASSP-88*, pp. 663–666, (1988).

[12] 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).

[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] Y. Linde, A. Buzo, and R. M. Gray. “An Algorithm for Vector Quantizer Design”. *IEEE Trans. Commu.*, Vol. **COM-28**, 1, pp. 85–95, (1980–01).

[15] Hermann Ney. “The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition”. Vol. **ASSP-32**, 2, pp. 263–271, (1984–04).

[16] 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).

[17] 磯, 渡辺, 桑原. “音声データベース用文セットの設計”. 昭63 春音講論, 2–2–19, pp. 89–90, (1988–03).

付録 A

音声資料

連続音声データベース

MHT、MSH、MTK、の三名については平井らの研究 [12] により式 (2.5) で表されるパラメータが与えられている。ただし、それぞれの 503 文章中の 200 文程度であり、文章番号、文章グループについては特に定まってない。表 A.1 ～ A.3に本実験で使用した学習データの一覧を記す。

表 A.1: 話者 MHT の学習データ

話者	発話文章 (グループ・文章番号)														
MHT	B		02		04			07						13	14	15
				18			21	22		24	25	26		28	29	
			32						38	39						
		46		48												
	C						06			09		11	12			
			17	18					23			26	27			30
		31	32		34	35	36				40				44	
		46														
	D		02	03	04			07			10			13		
				18	19		21		23	24	25	26				
											40					
			47	48												
	E		02		04	05		07		09		11	12	13	14	15
		16		18	19		21		23	24			27	28	29	30
		31	32	33		35				39					44	
		46														
	F			03												
		16									25	26			29	
		31			34		36									
	G							07			10	11		13	14	
		16			19		21			24			27	28		
			47		49											
	H	01	02		04			07	08		10				14	
		16				20	21			24	25	26		28	29	
							36		38	39	40			43		
		46	47		49											
	I	01	02			05		07		09	10	11	12			15
		16		18	19	20		22		24	25	26	27	28	29	30
		31	32	33		35	36			39			42		44	
			47		49											
	J	01	02			05	06	07	08	09	10	11			14	15
			17		19	20	21	22		24	25	26		28	29	30
		31		33	34		36	37	38				42	43		45
		46		48	49	50	51	52	53							

Total: 193

表 A.2: 話者 MSH の学習データ

話者	発話文章 (グループ・文章番号)															
MSH	B		02					07					12	13	14	15
		16		18	19		21			24	25			28	29	
			32				36		38	39					44	
		46		48												
	C						06			09		11	12			15
		16	17	18		20		22	23	24		26	27		29	
		31	32		34	35	36				40			43	44	45
		46	47			50										
	D		02		04	05	06	07			10		12	13		
		16		18	19		21		23	24		26				
				33	34	35					40		42			
			47	48												
	E		02		04	05		07		09		11	12	13	14	15
		16		18	19		21		23	24			27	28	29	
		31	32	33		35				39					44	
		46														
	F	01		03	04											
		16	17					22			25	26	27			
		31			34		36									
	G		02					07				11		13	14	
		16		18	19		21		23	24			27	28		
														43		
			47		49											
	H	01	02		04			07	08		10				14	
		16				20	21			24	25	26		28	29	30
							36		38	39	40	41		43		
		46	47		49											
	I	01				05		07		09	10	11	12		14	15
		16	17		19			22		24	25	26	27	28	29	30
		31	32	33		35	36			39			42	43	44	45
			47		49											
	J	01	02		04	05	06	07	08	09	10	11	12		14	15
			17		19	20	21	22		24	25	26	27	28	29	30
		31	32	33	34	35			38		40				44	45
		46		48	49		51	52	53							

Total: 221

表 A.3: 話者 MTK の学習データ

話者	発話文章 (グループ・文章番号)															
MTK	B		02					07	08				12	13	14	15
		16		18			21	22			25	26		28	29	
			32				36		38	39						
		46		48												
	C									09		11	12			
		16	17	18					23	24		26				30
			32			35	36								44	45
		46														
	D		02	03	04			07					12			
					19		21			24	25					
						35					40					
			47													
	E		02		04	05		07		09		11	12	13	14	15
		16		18					23	24			27	28	29	
			32	33						39					44	
		46														
	F															
											25					
		31					36									
	G			03				07			10	11		13	14	
		16								24			27	28		
					49											
	H		02		04			07	08		10				14	
						20	21				25	26		28		
									38	39	40	41		43		
					49											
	I	01	02					07				11	12			15
		16						22	23	24						
				33		35									44	
			47		49											
	J	01	02			05	06		08		10	11		13	14	
		16	17		19	20	21	22		24	25	26	27	28	29	30
		31		33	34		36		38				42			45
		46		48	49		51		53							

Total: 151

付録 B

話者 MHO,FKN,FKS による実験結果

本文で挙げた 話者 MYI 以外に、 MHO(男性), FKN(女性), FKS(女性) についても句境界検出実験を行なった。

予め謝っておくが、女性話者の実験に対してピッチの抽出範囲を男性話者と同じ 70 ～ 300 Hz にしてしまうという実験ミスのため、半ピッチのエラーが続出し、ピッチの抽出エラーによる句境界挿入誤りが増加している。したがって、FKN、FKS については句境界検出システムに対して正当な評価はされていないことに注意して戴きたい。

表 B.1: F_0 テンプレートによる句境界検出精度 (話者 MHO)

MHO	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	75.2	(75.2)	[75.2]	97.2	(97.2)
3	85.3	(75.1)	[87.6]	81.8	(100.2)
5	88.1	(74.8)	[90.7]	77.5	(102.7)
10	93.0	(75.5)	[95.1]	69.4	(104.6)

表 B.2: モデルテンプレートによる句境界検出精度 (話者 MHO)

MHO	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	85.9	(85.9)	[85.9]	97.4	(97.4)
3	93.5	(85.2)	[94.1]	82.5	(100.4)
5	95.4	(85.5)	[96.4]	67.1	(102.2)
10	97.9	(83.7)	[98.4]	56.1	(108.0)

表 B.3: モデルテンプレート (等確率遷移) による句境界検出精度 (話者 MHO)

MHO	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	55.5	(55.5)	[55.5]	45.9	(45.9)
3	75.8	(60.5)	[81.5]	33.5	(51.2)
5	83.7	(60.6)	[88.9]	31.6	(56.8)
10	90.0	(62.6)	[95.2]	28.2	(62.5)

表 B.4: モデルテンプレート (Bigram) による句境界検出精度 (話者 MHO)

MHO	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	57.2	(57.2)	[57.2]	41.8	(41.8)
3	73.8	(56.1)	[79.7]	31.9	(47.3)
5	81.4	(59.0)	[87.9]	29.8	(52.2)
10	89.0	(61.7)	[95.2]	27.5	(60.8)

表 B.5: F_0 テンプレートによる句境界検出精度 (話者 FKN)

FKN	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	75.0	(75.0)	[75.0]	180.2	(180.2)
3	82.6	(73.9)	[83.1]	167.6	(183.2)
5	85.0	(73.2)	[85.2]	155.6	(182.0)
10	87.0	(72.4)	[87.6]	142.3	(181.6)

表 B.6: モデルテンプレートによる句境界検出精度 (話者 FKN)

FKN	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	69.4	(69.4)	[69.4]	161.2	(161.2)
3	83.6	(70.7)	[85.5]	140.6	(168.0)
5	88.5	(71.0)	[90.0]	129.2	(170.9)
10	93.3	(71.0)	[95.7]	123.0	(181.5)

表 B.7: モデルテンプレート (等確率遷移) による句境界検出精度 (話者 FKN)

FKN	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	50.7	(50.7)	[50.7]	116.6	(116.6)
3	73.2	(55.9)	[76.5]	108.7	(130.7)
5	82.9	(56.9)	[86.8]	110.7	(139.7)
10	90.8	(60.0)	[95.1]	100.0	(148.2)

表 B.8: モデルテンプレート (Bigram) による句境界検出精度 (話者 FKN)

FKN	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	49.3	(49.3)	[49.3]	110.2	(110.2)
3	71.2	(55.0)	[74.5]	108.8	(125.7)
5	82.0	(56.0)	[87.3]	101.5	(131.6)
10	90.2	(58.7)	[96.0]	91.7	(139.7)

表 B.9: F_0 テンプレートによる句境界検出精度 (話者 FKS)

FKS	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	67.8	(67.8)	[67.8]	122.3	(122.3)
3	79.0	(68.0)	[80.0]	107.3	(121.6)
5	82.1	(68.1)	[84.4]	105.1	(124.2)
10	86.0	(68.5)	[88.0]	125.0	(93.6)

表 B.10: モデルテンプレートによる句境界検出精度 (話者 FKS)

FKS	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	74.2	(74.2)	[74.2]	109.9	(109.9)
3	85.5	(73.2)	[87.2]	96.5	(115.0)
5	88.1	(72.1)	[90.1]	89.8	(120.0)
10	93.2	(72.3)	[95.9]	80.9	(126.6)

表 B.11: モデルテンプレート (等確率遷移) による句境界検出精度 (話者 FKS)

FKS	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	57.2	(57.2)	[57.2]	81.7	(81.7)
3	76.7	(60.9)	[81.7]	72.8	(89.4)
5	82.8	(61.6)	[88.5]	68.5	(95.0)
10	89.6	(62.6)	[95.5]	65.4	(101.9)

表 B.12: モデルテンプレート (Bigram) による句境界検出精度 (話者 FKS)

FKS	句境界検出率 (%)			句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	[累積]	最大検出率候補	(平均)
1	57.2	(57.2)	[57.2]	76.8	(76.8)
3	74.0	(60.1)	[78.4]	68.4	(85.1)
5	83.0	(60.5)	[88.6]	63.9	(89.2)
10	89.2	(61.5)	[95.1]	59.5	(95.8)

付録 C

通訳対話音声データベースによる実験結果

4名(A,B,C,D)による日本語・英語通訳対話文。このうち話者A(日本人の発呼者)のデータについて句境界検出した結果をまとめる。全体的に句境界検出率は低く、挿入誤り率が高い。これらの原因として、

- ポーズ検出率やピッチの抽出精度が悪い。
- 一文あたりが短くて句境界が無い。(「はい」など。)
- 冗長語を1つのアクセント句として認識できない。(「あの」や「えー」など。)
- 与えられている正解句境界が怪しい。

などが挙げられる。

トピック	旅行・ホテルの部屋の予約(非対面)	
テキスト	(男性 東京 声優)	TAS12001.A 23文
		TAS12002.A 20文
	(女性 大阪 プロナレーター)	TAS22001.A 13文
		TAS22002.A 16文
	(女性 大阪 テレビレポーター)	TAS32001.A 18文
		TAS32002.A 13文
		計103文章

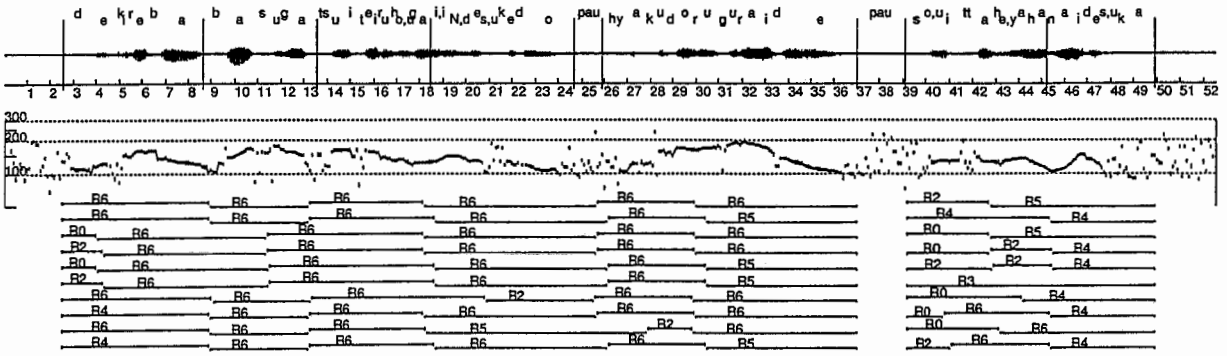


図 C.1: Spontaneous Speech の句境界検出 (良い例)

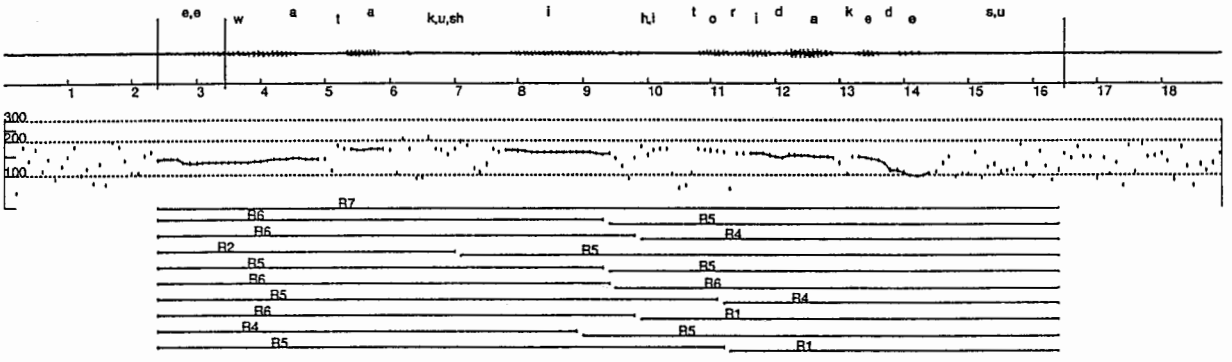


図 C.2: Spontaneous Speech の句境界検出 (悪い例)

表 C.1: モデルテンプレート (Bigram) による句境界検出精度 (話者 TAS1200[1-2].A)

	句境界検出率 (%)		句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	最大検出率候補	(平均)
1	51.1	(51.1)	97.7	(97.7)
3	65.9	(52.5)	103.4	(112.3)
5	77.5	(56.5)	102.2	(122.6)
10	87.0	(57.9)	103.3	(135.2)

表 C.2: モデルテンプレート (Bigram) による句境界検出精度 (話者 TAS2200[1-2].A)

	句境界検出率 (%)		句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	最大検出率候補	(平均)
1	34.7	(34.7)	131.9	(131.9)
3	58.3	(43.5)	120.8	(137.5)
5	77.8	(48.0)	112.5	(141.6)
10	79.2	(46.2)	113.9	(152.7)

表 C.3: モデルテンプレート (Bigram) による句境界検出精度 (話者 TAS3200[1-2].A)

	句境界検出率 (%)		句境界挿入誤り率 (%)	
候補数	最大検出率候補	(平均)	最大検出率候補	(平均)
1	42.2	(42.2)	168.9	(168.9)
3	68.9	(51.1)	162.2	(176.3)
5	75.6	(49.8)	151.1	(191.1)
10	91.3	(51.2)	147.8	(207.7)