モーラを単位とした基本周波数パターンの確率モデル化とそれによるアクセント句境界の検出

岩野公司† 広瀬啓吉†

アクセントやイントネーションといった音声の韻律的特徴から音声認識に有効な情報を高い精度で抽出するためには，1）特徴量の変動に対処するため韻律を確率モデルで表現することや，2）韻律的特徴のみならず音韻の情報をあわせて利用することが必要である．そこで本論文では，日本語連続音声の基本周波数（F_0）パターンに対し，音声認識プロセスで得られるモーラ境界情報を利用し，モーラを単位として F_0 パターンを確率モデル化する手法を提案する．2名の男性話者が，各々500文を発声したデータベースを用い，その50文を実験用データ，450文を学習用データとして，本モデルを用いたアクセント句境界検出とアクセント型認識の実験を行った．データベース付属の音素ラベルから作成した正解モーラ境界を用いたとき，特定話者で検出率約77%，挿入誤り率約15%，2話者間で検出率約76%，挿入誤り率約18%を得た．また，比較のため，モーラ境界を簡単な音声認識によって得た場合と，音韻境界情報を利用せずにフレーム単位でモデル化した場合の実験も行った．アクセント型の認識実験結果と併せ，得られた結果は，提案モデルの有効性を示すものであった．

A Statistical Modeling of Fundamental Frequency Contours in Moraic Unit and Its Use for the Detection of Prosodic Word Boundaries

KOJI IWANO† and KEIKICHI HIROSE†

In order to extract effective information for speech recognition with high accuracy from prosodic features of speech, such as accent and intonation, it is necessary to statistically model the prosody to cope with its feature variations, and to utilize segmental features to some extent. From this viewpoint, in this paper, we propose a statistical modeling of fundamental frequency contours of Japanese continuous speech in mora unit, where the mora boundary information is given during the speech recognition process. Experiments are conducted for the detection of prosodic word boundaries and recognition of accent types. Utterances of 500 sentences form each of two male speakers are used; 50 for the testing and the rest for the training. When mora boundaries obtained from the phone labels in the database are used, around 77% of correct boundary detection rates are obtained with insertion error rates around 15% for speaker dependent cases. For speaker independent cases, these rates are around 76% and around 18%, respectively. For comparison, two boundary detection experiments are further conducted; one using mora boundaries obtained from the speech recognition process instead of those obtained from the database, the other based on the frame-based modeling using no segmental boundary information. These results, together with similar results for accent type recognition indicate the validity of the proposed modeling.

1. はじめに

人間の音声は，大きく2つの特徴を有している．1つは母音や子音といった音響の違いを反映する「音韻的特徴」であり，もう1つは，単語のアクセントや文のイントネーションなどといった音韻的特徴以外の「韻律的特徴」である．現状の音声認識技術においては，もっぱら音韻的特徴のみが利用され，韻律的特徴はむしろ認識を阻害するものとして排除される傾向にあった．しかし，韻律的特徴が人間の音声知覚過程において重要な役割を果たしていることは明らかであり，より高い水準の音声認識を行うためにはこの特徴を利用することが不可欠である．そのような観点から，連続音声認識に韻律的特徴を利用するための研究が行われている．

日本語音声を対象とした研究においては，韻律的特徴として特に音声の高低に対応する基本周波数（F_0）

† 東京大学大学院工学系研究科
School of Engineering, University of Tokyo

を利用し，統語境界の検出[1)~5)]や文構造の解析[6),7)]，単語アクセント型の識別[8)]などを行うものが多い．これらは，認識に先だって韻律的特徴のみから結果を導き出し，それらを利用することで音韻面での認識の探索空間を抑えるといった効果を主に狙ったものである．しかし，韻律的特徴のみを利用する方法は，性能に限界があるだけでなく，得られた結果を音韻認識と融合することが難しいといった問題を抱えており，実際の認識システムへの導入はほとんど行われていない．そこで，韻律処理の処理にあたり，性能と融合性の両者の向上を狙い，韻律的特徴のみを利用するのではなく，音韻面から得られる情報をあわせて利用することを考える．

一方，韻律的特徴の処理にあたり，話者や発声ごとの特徴量のゆらぎに対処したモデル化が必要である．現在，音韻の認識に関しては確率モデルである隠れマルコフモデル（Hidden Markov Model: HMM）[9),10)]を利用する手法が最も優れているといわれている．これは HMM が特徴量の確率分布を用いてゆらぎを表現することが可能であり，かつ多量のデータに対してパラメータの学習アルゴリズムが確立しているなどといった利点によるためである．そこで，この HMM を用いて韻律をモデル化することを考える．HMM は局所的には定常であるが全体的には非定常な信号を表現するのに適しているモデルである．しかし，音韻的特徴に比べ音声の広範囲にわたって緩やかに現れる韻律的特徴を，音韻と同様に 10 ms 程度の短い時間単位（フレーム）でとらえてしまうと非定常な性質が現れにくくなるため，HMM が効果的に機能しなくなってしまう．そこで，韻律的特徴を扱ううえでの単位として，より長い時間長の「モーラ（拍）」を採用する．

以上のような観点から，本論文では，1）音韻境界であるモーラ境界の情報を韻律的特徴である F_0 パターンとともに利用し，2）F_0 パターンをモーラ単位でコード化し，そのコード系列を HMM への入力とする手法を提案する．さらに，このモデルを利用した日本語連続音声のアクセント句境界の検出手法を提案し，その性能評価を行う．

2. モーラと基本周波数（F_0）パターン

モーラは日本語の場合，おおよそ仮名 1 文字分の音声に相当する単位であり，100 ms 程度の時間長を有している．このモーラを単位とした F_0 の高低のパターンに基づいて単語アクセントの型分類が行われるといった事実もあり，モーラと F_0 パターンとの相性は非常に良い．**図 1** に日本語東京方言における 4 モーラ単語のアクセント型による分類を示す[11)]．アクセント型は，何番目のモーラで F_0 が降下するかを示している．なお，4 モーラ単語における 0 型と 4 型の違いは，これらの単語に助詞を付けて発声したときのアクセントの違いに現れる．

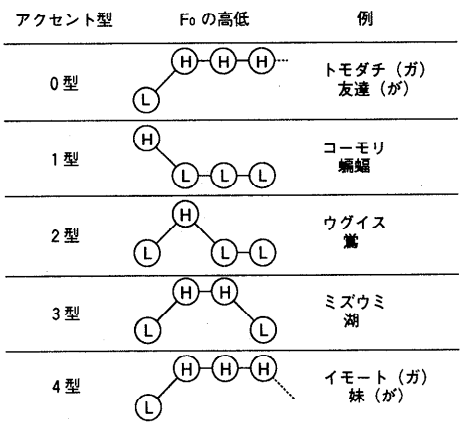


図 1 日本語東京方言における 4 モーラ単語のアクセント型
Fig. 1 Accent types for 4-mora words of Tokyo dialect in Japanese.

また，日本語連続音声は上記のアクセント型を有する句に分割されるが，これをアクセント句と呼ぶ．

3. モーラ遷移確率モデル

本論文では，アクセント句をアクセント型ごとに HMM でモデル化する．F_0 パターンをモーラ単位でとらえれば，アクセント句内の F_0 の上昇・平坦・下降というそれぞれの部分を 1 もしくは数モーラの定常状態と見なすことができ，また，アクセント句全体はそれぞれの定常状態をつなぎあわせた非定常なイベントと見なすことができる．したがって，HMM の特性を損なわずにモデル化が可能である．このようにして得られた確率モデルをモーラ単位で状態が遷移することから「モーラ遷移確率モデル」と呼ぶことにする．

このモーラ遷移確率モデルによる F_0 パターンのモデル化に関し利点をまとめると以下のようになる．

(1) HMM の特質を損なわないで統計的なモデル化が可能．
(2) モーラを基本として表現することの多い F_0 パターンのモデル化が容易．
(3) F_0 の観測されない無声部の扱いが容易．
(4) 学習データなどのデータサイズが小さい．
(5) 音素を基本とした認識システムとの融合が容易．

このうち (3) に関して簡単に説明しておく．音声中には，F_0 が観測されない無声区間が存在し，その扱いが問題となる．しかし，モーラを単位として F_0 を

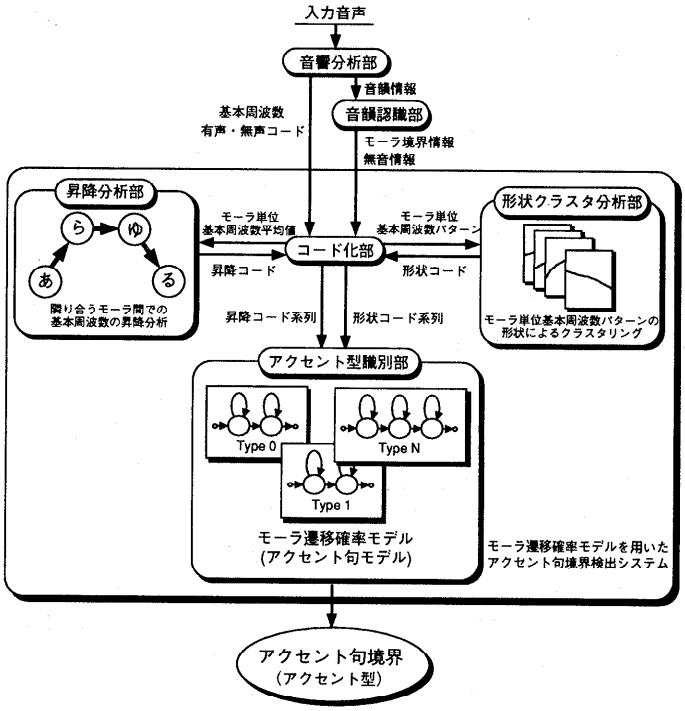


図 2 モーラ遷移確率モデルを用いたアクセント句境界検出システム
Fig. 2 Prosodic word boundary detection system based on the statistical modeling of F_0 contours in moraic unit.

見れば，大部分のモーラは F_0 の観測されやすい母音部を含んでいるため，子音部において F_0 が観測されなかったとしても F_0 を利用したパラメータ化を行いやすい．

4. モーラ遷移確率モデルを用いたアクセント句境界の検出

4.1 システム概要

モーラ遷移確率モデルを用いたアクセント句境界検出システムを提案する．システムの構成図を **図 2** に示す．まず，本システムにおける境界検出の流れを簡単に説明する．

(1) システムは，音響分析部から入力音声の F_0 パターンと有声・無声の識別コードの時系列データを受け取る．F_0 の抽出は文献 12) の手法を用いており，抽出された F_0 は対数をとっておく．また，音韻認識部からモーラ境界と無音（ポーズ）の情報を受け取る．

(2) コード化部では，入力の F_0 パターンをモーラ境界の情報を基にモーラ単位に切り分け，それぞれに「形状コード（Shape Code）」「昇降コード（ΔF_0 Code）」という 2 種類のコードを割り当てる．形状コードはモーラ単位の F_0 パターンそのものの形状を表すコードであり，昇降コードはあるモーラの F_0 の平均値がその直前のモーラから，どの程度上昇（下降）したかを数段階で示すコードである．このようにして，入力音声全体のコード系列を 2 系列作成し，アクセント型識別部に用意されているアクセント句モデルへの入力とする．

(3) アクセント型識別部には，アクセント型別にアクセント句モデルと，入力音声を任意個のアクセント句の並びとして表現した文法（言語モデル）を用意しておく．アクセント句モデルには離散型 HMM を用いており，その内部では，入力される 2 つのコード系列に対し別々に出力確率が与えられている．最終的な出力確率は，その 2 つの出力確率を適当に重み付けし，乗じることで得られる．

(4) 入力音声のアクセント型識別結果が，アクセント型で表記されたアクセント句の連鎖として出力される．同時に，それぞれのアクセント句の位置と長さについても結果が得られるので，アクセント句連鎖の結合部がアクセント句境界として検出されることになる．

4.2 コード化部

コード化の詳細について説明する．

モーラ境界情報を用いて F_0 パターンをモーラ単

位に切り分けるが，その際，長母音などの特殊拍はまとまった単位として扱われるため，各モーラに対する境界が完全な形では得られない．このような音節については，構成されるモーラ数で時間を等分することにより，便宜的に各モーラを切り分けている．得られたモーラ単位の F_0 パターンについて，有声・無声の識別コードを用い，有声部分が 10% 以下のモーラを「無声モーラ」，それ以外のモーラを「有声モーラ」として，区別して扱うこととする．また，無音（ポーズ）区間は，おおよそモーラの平均時間長にあたる 100 ms で切り分け，特別に「無音（ポーズ）モーラ」として扱うこととする．

それぞれのモーラ単位 F_0 パターンに対し，形状クラスタ分析部と昇降分析部での分析結果を利用して，「形状コード」と「昇降コード」を割り当て，2 つのコードの系列を作成する．

このように，本手法では離散コード化を導入しているが，これは形状コード化の際に，1）無声モーラや無音モーラに特別な意味を持たせることが容易であり，2）入力音声の F_0 抽出結果に多少誤りがあっても，抽出誤りを含まない形状クラスタパターンとのマッチングによってコード化を行うことで抽出誤りの影響を抑制できる，といった利点があるためである．

4.2.1 形状クラスタ分析

形状コード化を行うためには，あらかじめモーラ単位 F_0 パターンを形状別に分類しておく必要がある．そこで，形状クラスタ分析部では，モーラ単位 F_0 パターンを形状によってクラスタリングし，数個のクラスタを作成しておく．

クラスタリングに用いるデータには，F_0 の抽出誤りの影響を最小限にするため，すべて有声音で構成されている（無声部などによる F_0 の脱落のない）ものを選ぶ．クラスタリング手法としては，リーダー・アルゴリズム[13] を用いるが，このアルゴリズムはデータの順序によって結果が影響を受けやすいため，あらかじめ SL アルゴリズム[13] を利用して距離の近いものが集まるように並べかえを行っておく．双方のアルゴリズムとも，2 パターン間の距離が必要となるので距離の定義を行う必要がある．モーラ単位の F_0 パターンに対し，直流成分を除去した後，時間長を一定とするように時間軸方向で線形に伸縮を行い，それと同比率で周波数軸方向にも線形に伸縮を行う．このような正規化を行ったうえで，パターン i，j 間の形状に基づく距離 D_{ij} を以下のように定義する．

$$D_{ij} = \frac{\sum_{t'=0}^{T'} |\{F_{0i}(t')\}' - \{F_{0j}(t')\}'|}{T'} \quad (1)$$

ここで，t' は正規化後の時刻，T' は正規化後のモーラの時間長，$\{F_{0i}(t')\}'$ は時刻 t' におけるパターン i の正規化後の（対数表示された）F_0 である．この距離を用いて形状クラスタ分析を行い，9 個の形状クラスタを作成する．

4.2.2 形状コード化

形状コード化は以下のように行われる．

(1) 無音モーラに対し「無音コード」を割り当てる．

(2) 無声モーラに対し「無声コード」を割り当てる．

(3) 有声モーラに対しては，形状クラスタ分析部で得られた各クラスタの平均形状パターンとの比較を行い，距離が最小となるクラスタの番号をコードとして割り当てる．有声モーラにも無声部分が含まれている可能性があるため，そのモーラにおいて最初に有声を確認した時刻から最後に有声を確認した時刻までを切り出して改めてパターンと見なし，形状クラスタ分析部と同様な方法で正規化を行う．しかし，F_0 の抽出誤りや有声・無声の識別誤り，モーラ境界のずれなどの影響でまだ内部に無声部分が観測される可能性がある．そこで，こういった無声部分を距離計算から除外するため，関数 $V_p(t)$ を用意しておく．$V_p(t)$ はパターン p が時刻 t において無声であれば 0，有声であれば 1 となる関数である．正規化されたモーラ単位の F_0 パターン p と，クラスタの平均形状パターン c との間の距離 D_{pc} を以下のように定義する．

$$D_{pc} = \frac{\sum_{t'=0}^{T'} |\{F_{0c}(t')\}' - \{F_{0p}(t')\}'| \cdot V_p(t')}{\sum_{t'=0}^{T'} V_p(t')} \quad (2)$$

最終的に形状コードの数は，9 つの有声コードに無音コードと無声コードを加え 11 となる．

4.2.3 昇降分析

コード化部において「昇降コード」の割当てを行うためには，各モーラに対して F_0 の平均値を求め，さらに隣り合うモーラ間におけるその差の分布を分析しておく必要がある．

昇降分析に用いるデータとしては，隣り合うモーラが双方とも有声モーラとなるものを選ぶ．有声モーラ p の F_0 の平均値 $\overline{F_0}_p$ を以下のように定義したうえで，平均値の差のデータをとり，標準偏差 σ を求めておく．

$$\overline{F_0}_p = \frac{\sum_{t=0}^{T} F_{0p}(t) \cdot V_p(t)}{\sum_{t=0}^{T} V_p(t)} \quad (3)$$

4.2.4 昇降コード化

昇降コード化をするためには，無音モーラ，無声モーラについても F_0 の平均値を定義する必要がある．そこで，以下のように平均値を設定する．

(1) 無音モーラの F_0 の平均値を 0 とする．

(2) 無声モーラについては，その前後で最も近い有声モーラ，または無音モーラを探し出し，その 2 つの F_0 の平均値から直線補間して得られる値を平均値とする．

昇降コードの数を形状コード数と同じ 11 とするため，3σ 範囲の中央を 0 に平行移動させた後に，その範囲を 9 等分する．3σ 範囲外の領域と合わせた 11 の領域のうち，差の値がどこに属するかでコード化が行われる．

4.3 アクセント型識別部

アクセント型識別部には，アクセント型別にモーラ遷移確率モデルによって表現されたアクセント句（アクセント句モデル）とアクセント句の並びを記した文法（言語モデル）が用意されている．アクセント句モデルは離散型 HMM でありモデルパラメータ学習は Baum-Welch アルゴリズム[9),10)] によって行われ，確率計算には Viterbi アルゴリズム[9),10)] を用いる．アクセント句の並びに関する文法（言語モデル）には，制約を記述したネットワーク文法（制約文法）と連鎖確率の付いたアクセント句の 2 つ組を表現したもの（bi-gram）の 2 つを用意する．

4.3.1 アクセント句モデル

アクセント型の違いと無音区間の位置に注目し，アクセント句モデルとして以下の 7 種類を用意する．2 章で示したように，N モーラ単語については $N+1$ 個のアクセント型が存在する．しかし，0 型と N 型の区別はその単語（アクセント句）のみからでは判断できないため，本手法では両者を同じ型として扱っている．

T0, T0_P 0 型のアクセント句．または，モーラ数とアクセント型数が一致するアクセント句．

T1, T1_P 1 型のアクセント句．

TN, TN_P T0，T1 以外のアクセント句．

P 無音区間．

X_P（X = T0, T1, TN）は無音区間が後続するアクセント句を意味している．本来，無音区間はアクセント句ではないが，無音コードを吸収するため，便宜的にモデル P を用意している．また，状態数は，TN, TN_P は 3 状態，T0, T0_P, T1, T1_P は 2 状態，P は 1 状態としている．前述のとおり，このモデルへの入力は形状・昇降の 2 コードの系列となる．このとき，時刻 t で観測されたコードのベクトル $\boldsymbol{o_t}$ に対する状態 j での出力確率 $b_j(\boldsymbol{o_t})$ は以下のように定義される．

$$b_j(\boldsymbol{o_t}) = [P_{js}(o_{st})]^{\gamma_s}[P_{jr}(o_{rt})]^{\gamma_r} \quad (4)$$

ここで，o_{st}, o_{rt} は順に時刻 t で観測された形状コード，昇降コードを表し，$P_{js}(o_s)$, $P_{jr}(o_r)$ はそれぞれ状態 j において形状コード o_s，昇降コード o_r の生成される確率を示している．また，γ_s, γ_r は形状コード，昇降コードに対する重み付けの係数である．

4.3.2 文法（言語モデル）

アクセント句の並びについて記述した文法を，以下の 2 種類用意し，実験に用いる．

- 制約文法

「X_P というアクセント句モデルの後には必ず無音区間 P が出現し，文は X_P という句で終了する（X = T0, T1, TN）」という制約を人手により記した文法．

- アクセント句 bi-gram

モデル学習用データより作成される，連鎖確率付きのアクセント句の 2 つ組．

5. 評価実験

5.1 音声試料

音声試料は，ATR 自動翻訳電話研究所で作成された研究用日本語音声データベース[14)]（分析条件 16 bit, 10 kHz サンプリング）のセット B の文音声データ，男性話者 2 名（MYI, MHT）分を使用した．特定話者での実験と 2 話者間での実験の双方を行うため，モデル学習用データ（**T**）と実験用データ（**R**）を以下のようなデータセットに分けておく．各データセットは重なりがなく，実験はすべてオープンテストとなる．

T(MYI) 話者 MYI，450 文（B01～J50），
アクセント句数 3,023 個，無音区間 586 個．

R(MYI) 話者 MYI，50 文（A01～A50），
アクセント句数 326 個，無音区間 70 個．

T(MHT) 話者 MHT，450 文（B01～J50），
アクセント句数 3,167 個，無音区間 915 個．

R(MHT) 話者 MHT，50 文（A01～A50），
アクセント句数 325 個，無音区間 99 個．

形状クラスタリング・昇降分析・アクセント句 bi-gram 作成用データは，モデル学習時に使用するデータと同一のものを用いる．

ここで，モデル学習や bi-gram 作成時に必要となるアクセント句境界位置とアクセント型の情報について説明する．話者 MYI のデータに関しては，本データベースで提供されている言語・韻律情報データ中の

表 1 モデル学習用データ中の形状コードの種類と分布
Table 1 Features and distribution of shape codes in the training data for prosodic word HMMs.

コード	形状特徴	個数	
		T(MYI)	T(MHT)
1	無音	2,070	3,974
2	無声	1,814	1,430
3	平坦	2,238	3,694
4	緩上昇	1,057	1,338
5	中上昇	348	454
6	急上昇	301	404
7	緩下降	4,368	3,804
8	中下降	2,121	2,044
9	急降下	1,525	787
10	凸型 1	292	255
11	凸型 2	300	154
	合計	16,434	18,338

表 2 モデル学習用データ中の昇降コード分布
Table 2 Features and distribution of ΔF_0 codes in the training data for prosodic word HMMs.

コード	昇降（傾き）	個数	
		T(MYI)	T(MHT)
1	下降（負）	1,372	1,461
2	・	80	47
3	・	390	233
4	・	1,357	1,360
5	・	3,817	4,111
6	平坦（0）	5,578	7,203
7	・	1,583	1,596
8	・	764	581
9	・	235	154
10	・	90	57
11	上昇（正）	1,168	1,535
	合計	16,434	18,338

アクセント句境界とアクセント型をそのまま使用している．話者 MHT のデータに関しては，J-ToBI[15),16)] ラベルのデータを利用しており，BI 層・トーン層のラベル情報からアクセント句境界とアクセント型を定めた．したがって，両者のアクセント句境界の基準は若干異なっている．

5.2 コードの種類と分布

形状・昇降コードの種類とモデル学習用データ **T(MYI)**, **T(MHT)** 中でのコード分布を **表 1**, **表 2** に示す．

形状コードの「凸型 1」「凸型 2」は，それぞれ凸の頂上部分がモーラの中央より右寄りなもの（凸型 1）と左寄りなもの（凸型 2）を示している．昇降コードの番号は，小さい方がより急峻に下降し，大きい方がより急峻に上昇することを意味している．コード 1, 11 といった極端な勾配のコードの大部分は，差をとった隣接 2 モーラのうち，一方が（F_0 の平均値が 0 と見なされる）無音モーラとなるときに割り当てられている．この数が付近のコードに比べ極端に多くなっているのは，もともと隣接する 2 モーラともに有声モーラとなるものを昇降分析に用いたため，その差が無音モーラを含む場合に比べ小さかったことに起因している．

表 3 アクセント句境界検出実験の結果（特定話者・2 話者間）
Table 3 Results of prosodic word boundary detection on the experimets of speaker dependent and independent cases. Mora boundaries obtained from phone labels in the database are used.

実験	制約文法		bi-gram	
	R_d (%)	R_i (%)	R_d (%)	R_i (%)
(1a)	82.52	26.99	75.15	16.26
(1b)	84.92	25.23	80.31	14.46
(2a)	83.69	31.38	77.85	21.23
(2b)	79.75	26.69	74.23	15.64

5.3 アクセント句境界の検出

形状コード，昇降コードそれぞれに対する重み付け係数 γ_s, γ_r はともに 1.0 として実験を行った．

アクセント句境界の検出性能の評価値として，境界検出率 R_d，境界挿入誤り率 R_i を定義する．

$$R_d = \frac{N_{cor}}{N_{bou}} \times 100 \ (\%) \tag{5}$$

$$R_i = \frac{N_{ins}}{N_{bou}} \times 100 \ (\%) \tag{6}$$

このとき，アクセント句境界の数を N_{bou}，正しく検出出来た句境界数を N_{cor}，句境界の挿入誤り数を N_{ins} とする．その際，正解の境界位置から ±100 ms の範囲で検出されたものは正解としている．

5.3.1 特定話者・2 話者間実験

特定話者実験と 2 話者間での実験結果を **表 3** に示す．この実験では，モーラ境界・無音区間の情報はデータベース附属の音素ラベルから作成した．以下のように，モデル学習用データと実験用データの組合せを変えることで，合計 4 通りの実験を行った．

(1a) モデル学習用データに **T(MYI)**，実験用データに **R(MYI)** を用いた特定話者実験．

(1b) モデル学習用データに **T(MHT)**，実験用データに **R(MHT)** を用いた特定話者実験．

(2a) モデル学習用データに **T(MYI)**，実験用データに **R(MHT)** を用いた 2 話者間での実験．

(2b) モデル学習用データに **T(MHT)**，実験用データに **R(MYI)** を用いた 2 話者間での実験．

5.3.2 モーラ境界・無音情報の有効性の検証実験

音韻情報から得られるモーラ境界と無音区間の情報の有効性を検証するため，(1a) の特定話者実験に対し以下のように条件を変化させ実験を行った．

表 4 モーラ境界・無音情報の有効性検証実験の境界検出結果
Table 4 Results of prosodic word boundary detection on the experiments for speaker MYI: **(i)** utilizing phone labels and pause information in the database, **(ii)** utilizing phone labels and pause information obtained from a speech recognition process, and **(iii)** utilizing no segmental features.

実験	制約文法		bi-gram	
	R_d (%)	R_i (%)	R_d (%)	R_i (%)
(i)	82.52	26.99	75.15	16.26
(ii)	73.62	33.43	68.40	18.10
(iii)	58.28	111.35	52.45	99.08

(i) 正解のモーラ境界・無音情報を使用
モーラ境界と無音の情報をデータベース附属の音素ラベルを基に作成する.

(ii) 認識結果のモーラ境界・無音情報を使用
モーラ境界と無音の情報を，簡単な音韻認識によって得る．この音韻認識には，情報処理振興事業協会（IPA）の独創的情報技術育成事業の研究成果物「日本語ディクテーション基本ソフトウェア 97 年度版」に含まれている音韻モデルを利用し，言語モデルとしてはモーラの bi-gram をモデル学習用データから作成し使用する．モーラ認識率は約 40% であり，モーラ境界の検出性能をアクセント句境界検出性能と同様に計算したところ，±20 ms の誤差を許容して検出率 61.0%，挿入誤り率 32.7% であった.

(iii) モーラ境界・無音情報未使用（フレーム単位）
モーラ境界と無音の情報をまったく使わない．フレームシフトを 10 ms，基本周波数パターンを切り出す窓の長さを 100 ms として，各フレームの F_0 パターンに対し同様のクラスタ分析・昇降分析・コード化を行い，改めてモデル学習・検出実験を試みたもの（比較用）.

以上の条件による比較実験の結果を **表 4** に示す.

5.4 アクセント型の識別

アクセント句境界検出に際して，副産物として得られるアクセント型の識別結果についても触れておく.

実験データ中のアクセント句数を N，脱落誤り数を D，挿入誤り数を I，置換誤り数を S としたとき，アクセント型識別率として，

$$A = \frac{N-D-S-I}{N} \times 100 \quad (\%) \qquad (7)$$

を定義する.

アクセント型の区別は [0 型・1 型・それ以外] という 3 つのアクセント型の識別を行うものとする．特定話者実験と 2 話者間での実験で得られた結果を **表 5** に，特定話者実験に対しモーラ境界・無音情報の有効性を検証する実験で得られた結果を **表 6** に示す.

表 5 アクセント型識別実験の結果（特定話者・2 話者間）
Table 5 Results of accent type recognition on the experimets of speaker dependent and independent cases. Mora boundaries obtained from phone labels in the database are used.

実験	A (%)	
	制約文法	bi-gram
(1a)	54.60	59.20
(1b)	52.92	59.38
(2a)	48.00	57.85
(2b)	56.13	57.67

表 6 モーラ境界・無音情報の有効性検証実験のアクセント型識別結果
Table 6 Results of accent type recognition on the experiments for speaker MYI: **(i)** utilizing phone labels and pause information in the database, **(ii)** utilizing phone labels and pause information obtained from a speech recognition process, and **(iii)** utilizing no segmental features.

実験	A (%)	
	制約文法	bi-gram
(i)	54.60	59.20
(ii)	55.83	61.96
(iii)	7.06	19.63

5.5 実験結果の正解例

特定話者実験 (1a) において，アクセント句境界検出・アクセント型識別の双方とも正解した例を **図 3** にあげる．枠内に囲まれている部分がアクセント句モデル名で記されたアクセント型の識別結果であり，それぞれの結合部分をアクセント句境界として検出している.

6. 結論・考察

本論文では，韻律的特徴である F_0 パターンの統計的処理にあたり，モーラ境界（音韻境界）や無音の情報をあわせて利用する手法として，モーラ単位とした F_0 パターンの確率モデル化手法を提案した．また，このモデルを用いたアクセント句境界の検出実験を行った.

2 話者による特定話者実験では，bi-gram 利用時に両話者とも高い境界検出性能ならびにアクセント型識別性能を得ることを確認した．また，2 話者間での検出実験では，bi-gram 利用時に特定話者実験と比較して境界検出性能で約 5%，アクセント型識別で約 1.5% の性能劣化を確認した．特に，境界検出性能に関しては，2 話者間の実験のうち，1 つは境界検出率が低下しており，もう 1 つは挿入誤り率が上昇するといった傾向がみられているが，これは両話者のデータ間でア

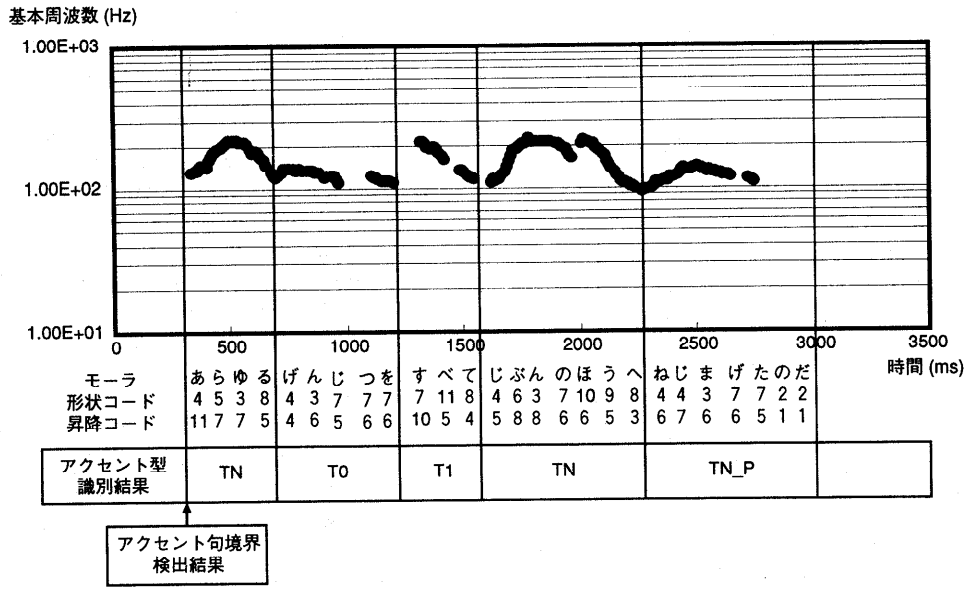


図 3 アクセント句境界検出・アクセント型識別の正解例

Fig. 3 An example of prosodic word boundary detection and accent type recognition. This figure shows a case of correct detection/recognition.

クセント句境界の基準が統一されていないことに起因していると考えている．

また，モーラ境界と無音の情報をデータベース中の音素ラベルから得た場合と，モーラの系列を出力する簡単な音韻認識から得た場合，さらに，まったく使わずに韻律的特徴のみをフレーム単位でモデル化した場合について実験を行い，比較を行った．その結果，精度の高いモーラ境界と無音の情報を多く使ったものほど境界検出性能が良好であることが分かり，このような情報が日本語音声における F_0 パターンの確率モデル化に際し，有効であることが確認できた．

アクセント句並びを表現する文法（言語モデル）については，すべての実験に対し 2 種類（制約文法と bi-gram）用意し，その数値の比較を行っている．境界検出性能に関しては，検出率・挿入誤り率の双方を考慮すると bi-gram を用いたときの性能の方が良好であった．アクセント型識別結果に関しても同様の結果であった．

句境界検出結果を従来手法と定量的に比較することは，1）対象とする境界の種類に違いがある，2）与える条件に違いがある，3）性能評価に用いる評価値の定義に統一性がない，といったことから一概には難しいが，文献 4) の手法は対象とする境界の種類が近く，また性能評価に用いる評価値の定義が同じであることからある程度の比較を行うことができる．この手法は，F_0 パターンの大局的特徴と局所的特徴のそれぞれをヒューリスティックな規則に基づいて分析し，相互の結果から最終的な句境界を定める手法であり，検出性能としては，検出率 が約 80% のとき挿入誤り率約 50%と報告されている．提案手法では，同等の検出率のとき挿入誤り率は特定話者で約 15%，別話者で約 25% となっており，さらに統計的な手法であるので学習データ数を増やすことによって検出性能の向上も見込める．このような点を考慮すると，確率モデルによる統計的手法が有効であったと考えられる．また，アクセント型識別結果については，文献 2) の結果と比較することができる．この手法は，本研究と同じく HMM を用いてアクセント句を確率モデル化しているが，F_0 の 1 次回帰係数，2 次回帰係数を特徴量として採用しており，フレーム単位で量子化を行って HMM の入力としている（その際，無音・無声区間に特別なコードを与えるところも類似している）．男性話者 1 名による特定話者・オープンの実験を行った結果，アクセント型識別率は 62.7% に達したと報告されているが，1 文中に含まれるアクセント句の数を既知として実験を行っているため，脱落誤りや挿入誤りが生じてない．本研究では，同様の実験で 59.4% を達成しているが，文中のアクセント句の数は未知という条件であり，脱落誤りや挿入誤りによって数値が劣化していることを考慮すると従来研究以上の性能を有していると考えられる．これは，音韻境界を使用し，フレームでなくモーラを単位として F_0 パターンを確率モデ

ル化したことの有効性を示唆している．

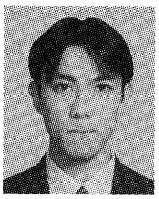
今後の課題としては，まず，連続型 HMM の導入による精度の向上があげられる．その際，モーラ単位の F_0 パターンに対して連続値によるパラメータ化を行う必要があるが，離散コード化の利点であった，1) 無声部や無音部に特別な意味を持たせることが容易，2) 抽出誤りを含まないパターンとのマッチングによる抽出誤りの影響の抑制，といった点を損なわないような工夫が必要である．また，不特定話者に対する検出性能向上のため，学習データの数を増やすことも検討しているが，そのためには基準の統一された韻律ラベル付き音声データベースを整備する必要がある．さらに，本システムの実際の認識システムへの応用も検討しており，具体的には語彙制約のない音声認識システムへの利用を考えている．

謝辞 本研究に用いた J-ToBI ラベルデータを提供してくださった ATR の関係各位に深謝いたします．

参考文献

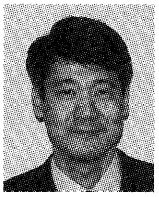
1) 鈴木良弥，関口芳廣，重永 実：日本語連続音声認識のための韻律情報を利用した句境界の抽出，電子情報通信学会論文誌，Vol.J72-D-II, No.10, pp.1609–1617 (1989).
2) 高橋 敏，松永昭一：統計的韻律モデルによる連続音声の句境界検出，電子情報通信学会技術研究報告，SP-90-71, pp.25–31 (1990).
3) Okawa, S., Endo, T., Kobayashi, T. and Shirai, K.: Phrase Recognition in Conversational Speech Using Prosodic and Phonemic Information, *IEICE Trans. Information and Systems*, Vol.E76-D, No.1, pp.44–50 (1993).
4) 今野博之，広瀬啓吉：韻律情報を利用した統語境界の抽出，電子情報通信学会技術研究報告，SP-93-112, pp.31–38 (1994).
5) 中井 満，シンガーハラルド，匂坂芳典，下平博：F_0 生成モデルを用いたテンプレートに基づく連続音声の句境界検出，電子情報通信学会論文誌，Vol.J80-D-II, No.10, pp.2605–2614 (1997).
6) 大平栄二，小松昭男，市川 熹：韻律情報を用いた音声会話文の文構造推定方式，電子情報通信学会論文誌，Vol.J72-A, No.1, pp.23–31 (1989).
7) 遠藤 隆，小林哲則，白井克彦：韻律情報を用いた構文推定とその音声認識への応用，電子情報通信学会技術研究報告，SP-91-103, pp.9–14 (1992).
8) 吉村 隆，速水 悟，田中和世：モーラ単位 HMM の接続による単語アクセントパターン識別，電子情報通信学会技術研究報告，SP-92-104, pp.9–14 (1992).
9) 中川聖一：確率モデルによる音声認識，chapter 3, pp.29–89, 電子情報通信学会 (1988).
10) Rabiner, L. and Juang, B.: *Fundamentals of Speech Recognition*, chapter 6, pp.321–386, Prentice Hall (1993).
11) 金田一春彦，秋永一枝：明解日本語アクセント辞典（第二版），三省堂 (1981).
12) Hirose, K., Fujisaki, H. and Seto, S.: A Scheme for Pitch Extraction of Speech Using Autocorrelation Function with Frame Length Proportional to the Time Lag, *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing*, ICASSP-92, Vol.1, pp.149–152 (1992).
13) Hartigan, J.A.: クラスター分析，マイクロソフトウェア (1983).
14) 阿部匡伸，匂坂芳典，梅田哲夫，桑原尚夫：研究用日本語データベース利用解説書（連続音声データ編），TR-I-0166, ATR 自動翻訳電話研究所 (1990).
15) Venditti, J.J.: Japanese ToBI Labelling Guidelines, Technical Report, Ohio-State University, Columbus, U.S.A. (1995).
16) ニックキャンベル：Tones and Break Indices（ToBI）システムと日本語への適用，日本音響学会誌，Vol.53, No.3, pp.223–229 (1997).

(平成 10 年 9 月 30 日受付)
(平成 11 年 2 月 8 日採録)



岩野 公司

1995 年東京大学工学部電子情報工学科卒業．1997 年同大学院工学系研究科情報工学専攻修士課程修了．現在，同大学院博士課程在学中．日本音響学会会員．



広瀬 啓吉（正会員）

1972 年東京大学工学部電気工学科卒業．1977 年同大学院博士課程修了．工学博士．同年同大学工学部電気工学科講師．1994 年同電子工学科教授．1996 年同大学院工学系研究科電子情報工学専攻教授．1999 年 4 月より新領域創成科学研究科基盤情報学専攻教授．1987 年米国 MIT 客員研究員．音声言語情報処理分野一般についての研究開発に従事，特に韻律に着目した研究．IEEE，米国音響学会，ESCA，電子情報通信学会，日本音響学会，人工知能学会，言語処理学会等会員．