音声言語情報処理 25−15
（1999. 2. 6）

# モーラ遷移確率モデルによるアクセント句境界検出と連続音声認識への応用

岩野 公司　　広瀬 啓吉
東京大学 工学部

筆者らは，モーラ遷移確率モデルを用いた連続音声のアクセント句境界検出手法を提案している．この手法は，モーラ遷移確率モデルによってアクセント句の $F_0$ パターンをアクセント型別にモデル化し，これらのモデルとのマッチングによって入力音声のアクセント句の並びを求め，その連結部をアクセント句境界として検出するものである．モーラ境界を既知とし，ATR の連続音声データベースを用いたアクセント句境界検出実験を行なった結果，特定話者実験で境界検出率 77.0 %，挿入誤り率 14.7 % を得た．本稿では，この句境界検出法を，語彙制約のない連続音声認識と融合させたシステムを提案し，融合による認識性能の改善について報告する．句境界検出実験と同じデータを用いて調べたところ，数パーセントの認識率の向上が確認された．

# Detecting Prosodic Word Boundaries Using Statistical Models of Moraic Transition and Its Use for Continuous Speech Recognition

Koji Iwano　　Keikichi Hirose
Faculty of Engineering, University of Tokyo

A method was proposed formerly for prosodic word boundary detection in continuous speech based on the statistical modeling of moraic transitions of fundamental frequency ($F_0$) contours. In the method, $F_0$ contours of prosodic words were modeled separately according to their accent types. An input utterance was matched against the models and was divided into constituent prosodic words. By doing so, prosodic word boundaries can be obtained. The method was applied to the boundary detection experiments of ATR continuous speech corpus. With mora boundary locations given in the corpus, detection rate was 77.0 % and insertion error rate was 14.7 % for speaker dependent experiment. Then the method was integrated into a continuous speech recognition scheme with unlimited vocabulary. A few percentage improvement was observed in mora recognition rate for the above corpus.

## 1 はじめに

音声の韻律的特徴は，人間が音声を認識し理解する上で重要な役割を果たしていることが知られている．しかし，現状の音声認識システムでは，この特徴は殆ど用いられていない．そこで，より高精度かつ高度な音声認識の実現にむけ，認識システム内で韻律的特徴を積極的に利用しようとする研究がすすめられている．

すでに，韻律的特徴を利用した統語境界の検出や構文の推定に関する研究は数多く報告されているが，これらは音声認識プロセスと独立して処理を行なうことを想定しており，韻律的特徴のみを用いて結果を導くものが主である．しかしながら，このような手法は，1) 認識に有効な情報を十分な精度で得ることができない，あるいは，2) 韻律のみから得た結果を音韻や統語の情報に反映させることが難しい，といった問題を抱えており，実システムへの融合がすすんでいない．

そこで，韻律の処理にあたり，韻律的特徴のみでなく，音声認識の結果から得られる音韻の

情報を併せて利用することでこれらの問題点を改善することを考える．具体的な手法として，筆者らはモーラ遷移確率モデルによる基本周波数 ($F_0$) パターンのモデル化を提案し [1]，本モデルを用いた日本語 4 モーラ単語のアクセント型識別 [2]，日本語連続音声の文節境界検出 [1]，アクセント句境界検出とアクセント型識別 [3] について報告を行なっている．

モーラ遷移確率モデルは，音韻認識から得られたモーラの境界情報を利用し，入力音声の基本周波数パターンをモーラ単位で切り分け，コード化し，そのコード系列を入力とする確率モデルであり，具体的には隠れマルコフモデル (Hidden Markov Model : HMM) を使用している．したがって，1) 韻律的特徴の揺らぎに対処することが可能で，2) 超分節的特徴である韻律パターンを，フレームといった短い時間単位でなく，モーラという比較的長い，かつ韻律面での現象を表現するのに適した単位で HMM に入力することができるため，高い処理性能が期待できる．さらに，韻律処理の結果がモーラを基準にして得られるため，音素を基本とした音韻認識との融合性にも富んでいる．

特に，文献 [3] で提案したアクセント句境界検出手法は，本モデルを用いてアクセント句をアクセント型別にモデル化し，アクセント句並びに関する文法 (言語モデル) を用いることで，連続音声中のアクセント型とアクセント句境界を同時に導出するもので，それ以前の統語境界検出方式 [1] より良好な検出性能を得ることがわかっている．そこで，このアクセント句境界検出手法を，実際の音声認識システムと融合することで，認識性能の向上を試みる．音声認識システムとしては，モーラ bi-gram を言語モデルとしてもつ語彙制約なし連続音声認識システムを用いることとする．

本論文では，前半でアクセント句境界検出システム [3] についての概略とその性能に関して述べ，後半で音声認識システムへの融合と作成した融合システムの認識性能に関して論じる．

## 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

### 2.1 句境界検出システムの概要

句境界検出システムの構成図を図 1 に示す．

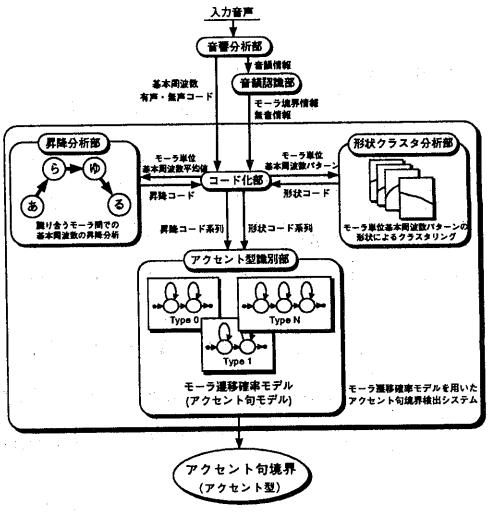


図 1: モーラ遷移確率モデルを用いたアクセント句境界検出システム構成図

システムは，まず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 [4] 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．

コード化部では，音韻認識部から得られるモーラ境界と無音 (ポーズ) の情報を元に，入力の $F_0$ パターンを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 ($\Delta F_0$ 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 $F_0$ パターン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たアクセント句モデル (モーラ遷移確率モデルによってモデル

化されたアクセント句) への入力とする.

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. このようにすることで, 入力音声のアクセント型識別結果は, 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, それぞれのアクセント句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.

### 2.2 形状コード

まず, モーラ単位の $F_0$ パターンを正規化する. 具体的には, モーラ単位 $F_0$ パターンの平均値が 0 になるように直流成分を除去したのち, モーラ時間長を決められた値とするように, 時間軸・周波数軸とも同じ割合で線形に伸縮を行なう.

形状コード化を行なうにあたり, モーラ単位 $F_0$ パターンの形状を分類しておく必要があるので, 無音や無声部分を含まないモーラ単位 $F_0$ パターンを用いてクラスタリングを行なう.

クラスタリング用のデータには, ATR の研究用日本語音声データベースのセット B, 話者 MYI の文音声データ (503 文) 中の 160 文から, 無音や無声部分を含まない 983 モーラを用いた. クラスタリングの結果, 9 個の形状クラスタが得られるので, これを有声部分を多く含むモーラに対するコード割り当てに用いる. さらに, 大部分が無声となるモーラや無音区間に対しそれぞれコードを設け, コード数を合計 11 とした.

コードの割り当ては, 以下のように行なう.

(1) 無音区間は 100 ms に切り分け (無音モーラ) , それぞれに「無音コード」を割り当てる.

(2) モーラ中の有声部分が 10 % 以下のモーラ (無声モーラ) には「無声コード」を割り当てる.

(3) それ以外のモーラ (有声モーラ) については, その形状が一番近いクラスタを求め, そのクラスタに定めたコード番号を割り当てる. その際, モーラ中に含まれる無声部分を無視するように距離計算を行なう.

### 2.3 昇降コード

2 つの隣接するモーラの $F_0$ 平均値の差に関して, コード数が形状コードと同じ 11 となるように分類を行なう. 形状クラスタリングに用いたものと同じデータベース (503 文) 中から, 2 つの隣接するモーラが双方とも有声モーラとなる 11,779 組を用い, その $F_0$ 平均値の差の標準偏差 $\sigma$ を求める. 有声モーラ中には, 無声部分も含まれていることがあるが, $F_0$ 平均値の計算には, 有声部分のみを使うこととする. 得られた $\sigma$ を元に, 3 $\sigma$ 範囲を 0 が中心になるように平行移動した上で 9 等分する. 3 $\sigma$ 範囲を越える領域を含めると, 全体は 11 空間に分割されるので, どの領域に差が当てはまるかを調べることによって, 11 コードの割り当てを行なうことができる.

コード化を行なう際には, 無音モーラ・無声モーラについても $F_0$ 平均値を定めておく必要があるので, 以下のように定義する.

(1) 無音モーラの $F_0$ 平均値は 0 とする.

(2) 無声モーラに関して, 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 つの基本周波数の平均値から直線補間して得られる値を平均値とする.

## 3 アクセント句モデル

入力が離散コードであるため, 離散型 HMM を用いてアクセント句をモデル化する. 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.

**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核位置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.

ここで, 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.

モデルへの入力が形状・昇降の 2 コードの系列となるため, 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 $b_j(\boldsymbol{o}_t)$ を以下のように定義する.

$$b_j(\boldsymbol{o}_t) = [P_{js}(o_{st})]^{\gamma_s}[P_{jr}(o_{rt})]^{\gamma_r} \quad (1)$$

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, $\gamma_s$, $\gamma_r$ 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.

アクセント句モデルの学習には, クラスタリングに用いたデータと同じ ATR の音声データベースのセット B, 話者 MYI の文音声データ 503 文を用いた. 学習用データベース中のアクセント句数は 3,365, 無音区間の数は 658 であった. なお, 学習は Baum-Welch アルゴリズムによって行なっている.

### 3.1 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 (言語モデル) を 2 種類用意する. 1 つは「制約文法」であり,「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である. もう 1 つは,「アクセント句 bi-gram」で, 連鎖確率付きのアクセント句の 2 つ組を示す言語モデルである. bi-gram は, アクセント句モデル学習用データと同じデータを用いて作成している.

## 4 句境界検出性能の評価

評価用データには, モデル学習用データ 503 文中の 50 文を用いた. この中のアクセント句数は 326, 無音区間の数は 70 である. 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より得ている. 形状・昇降コードに対する重み付け係数は共に 1.0 とした. また, 最適パスの計算には Viterbi アルゴリズムを用いている.

アクセント句境界の検出率を $R_d$, 挿入誤り率を $R_i$ として, 以下のように定義する.

$$R_d = \frac{N_{cor}}{N_{bou}} \times 100 \ (\%) \quad (2)$$

$$R_i = \frac{N_{ins}}{N_{bou}} \times 100 \ (\%) \quad (3)$$

このとき, アクセント句境界の数を $N_{bou}$, 正しく検出できた句境界数を $N_{cor}$, 句境界の挿入誤り数を $N_{ins}$ とする. なお, 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. 表 1 に 2 つの文法による検出性能実験の結果を示す.

表 1: アクセント句境界検出性能 (特定話者・closed test)

| 文法 | $R_d$(%) | $R_i$(%) |
|---|---|---|
| 制約文法 | 83.74 | 25.46 |
| bi-gram | 76.99 | 14.72 |

この結果から, 言語モデルとしては bi-gram を用いた方が性能が良好であることがわかる. そこで, 次節の融合システムにおけるアクセント句境界検出部では, アクセント句 bi-gram を用いている.

## 5 融合システムの作成

音声認識システムとアクセント句境界検出を融合したシステムを作成した. 図 2 に融合システムの構成図を示す. 音声認識システムとしては, 単語辞書を用いない語彙制約なし音声

認識システムを用いる．語彙制約なし音声認識としては，言語モデルとして音韻や音節の連鎖確率を用いる方法 [5] などが考えられるが，ここでは，モーラの連鎖確率を言語モデルとして持つ認識を持つ認識システムを利用することとした．したがって，認識結果としてモーラ系列が出力されることになる．

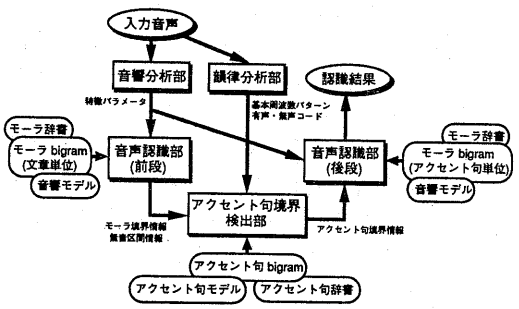


図 2: 融合システムの構成図

融合システム中では，図 2 のように 2 段階で語彙制約なし音声認識を行なっている．前段の音声認識では，韻律情報を使わずに処理を行ない，得られたモーラ境界の情報と無音区間の情報をアクセント句境界検出部に引き渡す．後段の音声認識は，アクセント句境界の結果を基に入力音声を切り分け，それぞれの句毎に再認識を行ない，それらをまとめて最終的な認識結果とする．なお，前段・後段の認識は共に HTK (Version 2.0)[6] を用いており，入力音声の音響分析条件は表 2 に示す通りである．

また，前段・後段の認識部で使用されているモーラ辞書・音韻モデル・モーラ bi-gram について説明しておく．

(1) モーラ辞書中には，152 種類のモーラが定義されている．さらに「無音区間」も便宜的にモーラ辞書に載せ，モーラと同等に扱う．

(2) 音韻モデルとしては，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である「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，混合数 16，状態数 3000 の tri-phone モデルを利用している．

表 2: 音声認識のための音響分析条件

| | |
|---|---|
| 標本化周波数 | 20 kHz |
| 分析窓 | Hamming 窓 |
| 分析窓長 | 25 ms |
| フレーム周期 | 10 ms |
| 高域強調 | $1 - 0.97z^{-1}$ |
| 特徴ベクトル | MFCC (12 次),<br>ΔMFCC (12 次),<br>Δ 対数パワー (計 25 次) |
| フィルタバンク | 24 チャネル |
| ケプストラム平均除去 | 発声単位で実行 |

(3) モーラ bi-gram は，ATR 音声データベースのセット B，話者 MYI の文音声データ (503 文) から，HTK で提供されているバックオフスムージング (back-off smoothing) 手法を利用して作成している．前段の認識部では文章内でのモーラ連鎖確率を表記した bi-gram を，後段の認識部ではアクセント句内でのモーラ連鎖確率を表記した bi-gram を利用している．したがって後者は，2 つのアクセント句をまたいでおこるモーラ連鎖は言語モデルの構築に利用されないことになる．それぞれのテストセットパープレキシティは，前段で約 38，後段で約 17 であった．

## 6 融合システムの性能評価

認識実験には，節 4 のアクセント句境界検出実験で用いたものと同じ評価用データ (50 文) を用いる．含まれているモーラ数は 1,541 である．

モーラの認識性能として，前段の認識で得られたモーラ認識率 $C_{bm}$ と後段の認識で得られたモーラ認識率 $C_{am}$ を以下のように定義する．

$$C_{bm}, C_{am} = \frac{N_{mora} - N_{del} - N_{subst} - N_{ins}}{N_{mora}} \times 100 \quad (\%) \quad (4)$$

このとき, $N_{mora}$ は総モーラ数, $N_{del}$ は脱落誤り数, $N_{subst}$ は置換誤り数, $N_{ins}$ は挿入誤り数である.

このモーラ認識率を, アクセント句境界の検出率 $R_d$, 挿入誤り率 $R_i$ と併せて図 3 に示す.

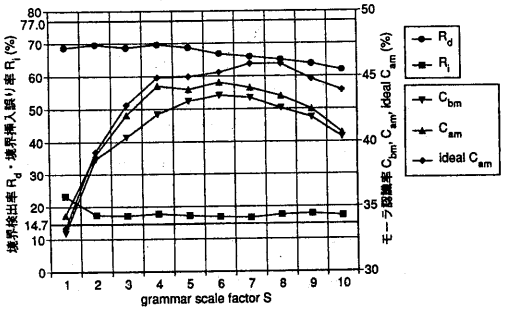


図 3: モーラ認識率とアクセント句境界検出率

図 3 中の横軸 $S$ は前段・後段の認識時に設定する *grammar scale factor* であり言語モデルから得られる (対数) 確率値を $S$ 倍して作用させることを意味している. *ideal* $C_{am}$ はアクセント句境界を正しく入力した場合の後段のモーラ認識率をしてしている. また, 横線で示されている句境界検出率 (77.0 %) と挿入誤り率 (14.7%) は, 4 節で得られた結果であり, 正しいモーラ境界を入力した場合の句境界検出性能を示している.

アクセント句境界の検出は, $S$ を 2 ～ 4 とした (前段の) 認識結果を利用したときに良好な性能を得ており, 検出率が約 70 %, 挿入誤り率が約 17 % となった. これは, 前段のモーラ認識に関し, $S$ が小さい時には挿入誤りが多く, 大きい時には脱落誤りが多くなる傾向があるため, 両者のバランスのとれた箇所でモーラ境界情報の精度が高かったことに起因していると考えられる. なお, $S = 3$ の時の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. モーラ認識率の改善度も句境界検出性能が良好なところで大きく, 最高で約 2 % のモーラ認識率の向上が確認された.

## 7 結論

本稿では, モーラ遷移確率モデルを用いたアクセント句境界検出手法とその性能について述べ, さらに, この句境界検出システムを語彙制約なし音声認識と融合したシステムを提案し, 融合よる認識性能の向上についての実験結果を示した. その結果, 最高で約 2 % の認識率の向上が確認された.

現段階では, 実験を特定話者に限定して行なっているが, 今後は, 不特定話者への拡張を考えている. 特に, アクセント句モデルを不特定話者に対応させるためには, 韻律ラベルの付いた複数話者の音声データを多量に準備する必要があるので, 自動ラベリングも視野に入れたデータ作成を検討している.

## 参考文献


[1] 岩野 公司, 広瀬 啓吉, "モーラ遷移確率モデルを用いた統語境界の検出," 信学技報, SP97-14, pp.33-40 (1997-6).

[2] K.Hirose and K.Iwano, "Accent Type Recognition and Syntactic Boundary Detection of Japanese Using Statistical Modeling of Moraic Transitions of Fundamental Frequency Contours," *Proc. IEEE ICASSP'98*, Seattle, Vol.1, pp.25-28 (1998-5).

[3] 岩野 公司, 広瀬 啓吉, "モーラ遷移確率モデルを用いたアクセント型の識別とアクセント句境界の検出," 信学技報, SP98-23, pp.1-8 (1998-6).

[4] K.Hirose, H.Fujisaki and S.Seto, "A Scheme for Pitch Extraction of Speech Using Autocorrelation Function with Frame Length Proportional to The Time Lag," *Proc. IEEE ICASSP'92*, San Francisco, Vol.1, pp.149-152 (1992-3).

[5] T.Kawabata, T.Hanazawa, K.Itoh, K.Shikano, "Japanese Phonetic Typewriter Using HMM Phone Recognition and Stochastic Phone-Sequence Modeling," *IEICE Trans.*, Vol.E74, No.7, pp.1783-1787 (1991-7).

[6] S.Young, J.Jansen, J.Odell, D.Ollason, P.Woodland, *The HTK Book v2.0*, Cambridge University (1995).