音声言語情報処理 29－35
(1999. 12. 21)

語彙制約なし音声認識へのアクセント句境界情報の利用

岩野 公司† 広瀬 啓吉‡
†東京大学大学院 工学系研究科
‡東京大学大学院 新領域創成科学研究科
〒113-8656 東京都文京区本郷 7-3-1
iwano@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp

あらまし 日本語連続音声認識の性能向上を目的として, 語彙制約なし音声認識にアクセント句境界情報を利用し, その精度を向上をさせることを考える. 具体的には, 語彙制約なし認識器と句境界検出とを融合したシステムを提案する. システム中には語彙制約なし音声認識部を 2 段階用意する. 前段認識部では句境界を考慮しないモーラ bigram を言語モデルとして用意し, 入力連続音声全体を認識する. 一方, 後段認識部では句境界をまたがるモーラ遷移を除いたモーラ bigram を用意し, 基本周波数パターンや前段の認識結果から抽出した句境界情報をもとに入力音声を句ごとにわけ再認識を行なう. 後段の言語モデルのパープレキシティが前段のものより小さくなることから認識率の改善につながる. 男性話者 2 名分, 各々学習用データ 450 文, 実験用データ 50 文を用い, 特定・不特定話者実験を行なったところ, 最高で約 2 % のモーラ認識性能の向上を確認した.

キーワード 連続音声認識, 語彙制約なし音声認識, 韻律情報, アクセント句境界検出

Use of Prosodic Word Boundary Information for Unlimited-Vocabulary Speech Recognition

Koji Iwano† Keikichi Hirose‡
†School of Engineering, University of Tokyo
‡School of Frontier Sciences, University of Tokyo
7-3-1 Hongo, Bunkyou-ku, 113-8656 Tokyo, JAPAN
iwano@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp

Abstract In order to improve the performance of Japanese continuous speech recognition, we are developing an unlimited-vocabulary speech recognition system using prosodic word boundary information. The system has two recognition stages with different language models: in the first stage it is the mora bigram obtained without taking prosodic word boundaries into account, and in the second stage it is the one obtained with taking them into account. Because of perplexity reduction from the first stage to the second stage, the better mora recognition rates are obtainable when input utterances are re-recognized after segmented into prosodic words, whose boundaries are detected using fundamental frequency contours and recognition results of the first stage. The system indicates maximum improvements in mora recognition rate by 2 % on both speaker-closed and -open experments using two male speakers' data (450 training data and 50 testing data per speaker).

key words Continuous Speech Recognition, Unlimited-Vocabulary Speech Recognition, Prosodic Information, Prosodic Word Boundary Detection

1 はじめに

現在の連続音声認識では高い認識性能を得るため，語彙や文法にある程度の制限を設けている．したがって，そういった語彙や文法の制約範囲外の入力への対処，すなわち，話し言葉の認識や未知語の処理などが解決すべき課題となっている [1]．このような課題に対し，語彙や文法の制約を持たない認識である「語彙制約なし音声認識」の重要性が見直されつつある．語彙制約なし音声認識は，入力音声を任意の音韻列として出力するもので，「音韻タイプライタ」とも呼ばれ [2]，単語を単位とした認識器と組み合わせることで未知語処理を実現する手法なども提案されている [3]．

一方，従来まではかえって認識を阻害するものとして音声認識プロセスに用いられることのなかった音声の韻律情報を，認識性能の向上に利用しようとする研究も進められている [4]．音声の韻律的特徴は，人間が音声を知覚・理解するうえで重要な役割を果しており，音声認識の水準を向上させるためにはこの特徴を利用することが不可欠である．

以上のような観点から，韻律的特徴から句境界情報を抽出し，語彙制約なし音声認識に利用することで，その精度を向上させることを考える．筆者らは，既に，基本周波数 (F_0) パターンからアクセント句境界を効果的に検出する手法として，「モーラ遷移確率モデル」というモーラ（拍）を単位とした統計的韻律モデルによるアクセント句境界検出法を提案している [5]．そこで，これを入力音声をモーラ系列として出力するような語彙制約なし認識器と融合し，その認識性能の向上をはかる．

本稿では，まず，アクセント句境界検出手法について説明する．その後，上記タイプライタ型認識と句境界検出を融合した認識システムについて説明し，評価実験を通して，句境界情報を利用したことによる認識率の向上について述べる．

2 アクセント句境界の検出

2.1 モーラ遷移確率モデル

アクセント句境界の検出を行なうために，アクセント句を Hidden Markov Model (HMM) でモデル化する．HMM は，局所的には定常であるが全体としては非定常であるような信号を表現するのに適している．韻律的特徴を反映する F_0 パターンも同様の特性を持っていることから，HMM による高精度なモデル化が期待されるが，音韻に比べ F_0 パターンの変化は時間的に緩やかで広範囲に渡るため，音韻と同様の 10 ms といった短い時間単位では非定常の性質が現れにくくなり，HMM が効果的に機能しない．このような問題を解決するため，より長い時間長を有するモーラを時間単位として導入した「モーラ遷移確率モデル」を提案している [5]．

2.2 アクセント句境界検出システム

モーラ遷移確率モデルを用いた，句境界検出システムを図 1 に示す．なお，文献 [5] の手法からは幾つかの改良を施してある．以下に，簡単に処理の流れを説明する．

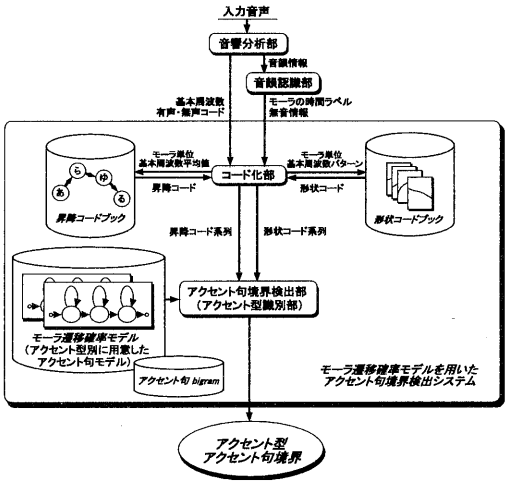

図 1: モーラ遷移確率モデルを用いたアクセント句境界検出システム

(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．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」「昇降コード」という 2 種類のコードを予め作成したコードブックを利用して割り当てる．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 7 種のアクセント句モデルと，言語モデルとしてアクセ

ント句の bigram を用意しておく．アクセント句モデルは離散型 HMM であり，その内部には，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 つの出力確率を乗じて得る．なお，モデルパラメータは Baum-Welch アルゴリズムによる学習によって求める．

(4) 入力の 2 コードの系列とアクセント句モデル・アクセント句 bigram との照合を Viterbi アルゴリズムによって行ない，入力音声を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する．同時に，それぞれのアクセント句の位置と長さについての情報も得られるので，アクセント句連鎖の結合部をアクセント句境界として検出する．

2.2.1 形状・昇降コード化

コード化に先だち，まず音韻境界情報を元に切り出されたモーラについて，F_0 の抽出誤り・モーラ境界のずれの影響を軽減するため，F_0 の孤立点集合の除去を行なっておく．孤立点集合を除去したあとのモーラ単位 F_0 パターンについて，有声部分がモーラ長の 20 % 以上を占めるものについては「有声モーラ」，それ以外のモーラを「無声モーラ」とする．また，音韻認識部で無音（ポーズ）と判定された区間については，おおよそモーラの平均時間長にあたる 100 ms で切り分け，それぞれを「無音モーラ」としておく．

有声モーラについては，この時点でモーラの F_0 の平均値を以下の式 (1) から求めておく．有声モーラ i について，$\overline{F_{0i}}$ は（対数をとった）F_0 の平均値，$F_{0i}(t)$ は時刻 t における（対数をとった）F_0 の値，$V_i(t)$ は時刻 t において有声であれば 1，無声であれば 0 となる関数，T_i はモーラ i の時間長である．

$$\overline{F_{0i}} = \frac{\sum_{t=0}^{T_i} F_{0i}(t) \cdot V_i(t)}{\sum_{t=0}^{T_i} V_i(t)} \tag{1}$$

有声モーラについてはさらに，モーラ長を一定とするように正規化処理を行なう．具体的には，モーラ中で最初に有声を観測した時刻から，最後に有声を確認した時刻までを改めて切り出し，各時刻での F_0 値から，式 (1) で求めた F_0 平均値を減じたあと，モーラ長を一定とするように時間軸・周波数軸の両方向に同じ比率で線形に伸縮を行なう．

形状コードの割り当ては，まず，「無音モーラ」に「無音コード」を，「無声モーラ」には「無声コード」を割り当て，「有声モーラ」については，LBG アルゴリズムによるクラスタリングで作成した 32 個のクラスタのうち，距離が最小となるものをコードとして割り当てる．このコード割り当てやクラスタリングに際して必要となる有声モーラ i, j 間の距離 D_{ij} は，以下の式 (2) で定義される．このとき，t' は正規化後の時刻，T_i' は正規化後のモーラ i の時間長，$\{F_{0i}(t')\}'$ は時刻 t' におけるモーラ i の正規化後の（対数をとった）F_0 である．

$$D_{ij} = \frac{\sum_{t'=0}^{T_i'} |\{F_{0i}(t')\}' - \{F_{0j}(t')\}'| \cdot V_i(t') \cdot V_j(t')}{\sum_{t'=0}^{T_i'} V_i(t') \cdot V_j(t')} \tag{2}$$

一方，昇降コード化では，まず「無音モーラ」と「無声モーラ」に関して，コード化対象となるモーラとその直前のモーラとの関係から 4 つの特別なコードを用意し割り当てる．4 つのコードとしては，1) 対象となるモーラ・直前のモーラともに「無音モーラ」，2) 対象となるモーラのみ「無音モーラ」，3) 直前のモーラのみ「無音モーラ」，4) 対象となるモーラ・直前のモーラのどちらかが「無声モーラ」，というそれぞれの場合について用意しておく．対象となるモーラ・直前のモーラともに「有声モーラ」となる場合については，対象となるモーラについて直前のモーラからの F_0 平均値の差を計算し，その値をコード化とクラスタリングに用いる．形状コードの場合と同様に，LBG アルゴリズムにより 32 個のクラスタを作成しておき，コード化に利用する．

従って，双方のコードのコードブックの大きさは合計で，形状コード 34，昇降コード 36 となる．

2.2.2 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．

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 上記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，それぞれの HMM のトポロジーは図 2 に示す通りである．

3 アクセント句境界検出と語彙制約なし音声認識の融合

図 3 にアクセント句境界検出とモーラタイプライタ型認識器を融合したシステムの構成を示す．認識部はモーラが定義された辞書とモーラ bigram を持ち，入力連続

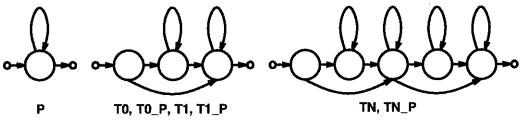

図 2: アクセント句モデルのトポロジー

音声をモーラ系列に変換するタイプライタ型認識器であり, このような認識部を 2 段用意する. まず前段の認識部において, 句境界情報を利用せずに入力音声に対するモーラ系列を導き出す. その結果から得られたモーラ境界の時刻情報と無音区間の情報を利用し, アクセント句境界の検出を行い, 後段の認識部では, その句境界検出結果をもとに入力音声を句ごとに分け再認識することで最終的な認識結果を得る.

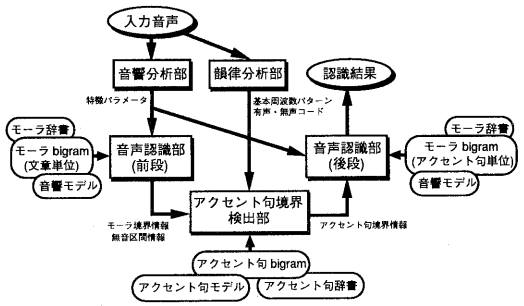

図 3: アクセント句境界検出とモーラタイプライタ型音声認識の融合システム

なお, 認識には HTK (Version 2.1)[6] を使用している. 音韻モデルには, 「日本語ディクテーション基本ソフトウェア 97 年度版 [7]」に含まれている, 混合数 16, 状態数 3000 の triphone モデルを利用した. モーラ辞書中には, 125 種類のモーラに加え「無音区間」を表す記号 (SP) を定義しておく. モーラ bigram は CMU-Cam Toolkit[8] を用いて構築する. その際, 学習データ中に出現しなかったモーラ bigram は, Linear Discounting に基づいた Back-off 平滑化を用いて unigram から値を推定する. また, カットオフは行なっていない. なお, 入力音声の音響分析条件は表 1 に示す通りである.

ここで, 前段の認識部のモーラ bigram は文章内全てのモーラ遷移を考慮して作成することになるが, 後段の認識部では, アクセント句内のみでのモーラ遷移を考慮したものを作成することになる. したがって, 後段の bigram はアクセント句境界をまたがって生起するモーラの遷移が除かれていることになる. 境界をまたがるモーラ遷移の種類は句内部のみでのモーラ遷移にくらべ変化に富んでいるため, これを除くことはパープレキシティの抑制につながる. これにより, 後段での認識性能が前段より向上することが期待される.

表 1: 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), ΔMFCC (12 次), 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

4 評価実験

4.1 音声試料・実験条件

音声試料として, 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のセット B の文音声データ, 男性話者 2 名 (MYI, MHT) 分 (同タスクでそれぞれ 500 文章) を用意した. 特定話者での実験と 2 話者間での実験の双方を行なうため, 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, 実験は全て open となる.

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

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

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

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

特定話者実験と 2 話者間での不特定話者実験を行なうため, 学習用・実験用データを以下のように組み合わせた 4 通りの実験条件を用意した. 各実験において, 学習用データは, アクセント句モデルの学習, 形状・昇降コードに対するクラスタリング, アクセント句 bigram の構築, モーラ bigram の構築に使用される.

(a) 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.

(b) 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.

(c) 学習用データに **T(MHT)**, 実験用データに **R(MYI)** を用いた不特定話者実験.

(d) 学習用データに **T(MYI)**, 実験用データに **R(MHT)** を用いた不特定話者実験.

4.2 アクセント句境界検出性能

2.2 節のアクセント句境界検出システムにおける, 検出性能を表 2 に示す. 評価値として, 境界検出率 R_d, 挿入誤り率 R_i を式 (3) のように定義しておく. 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. なお, ここでは, モーラの切り出しに利用する音韻境界の情報はタイプライタ型認識部で使用する triphone を用いた強制切り出しから得ている.

$$R_d = N_{cor}/N_{bou}, \quad R_i = N_{ins}/N_{bou} \qquad (3)$$

表 2: アクセント句境界検出性能

実験	R_d(%)	R_i(%)
(a)	72.70	12.27
(b)	75.38	12.31
(c)	70.25	11.66
(d)	73.85	14.77

4.3 音声認識性能の比較

タイプライタ型認識に用いたモーラ bigram の実験用データ 50 文に対するモーラあたりのテストセット・パープレキシティを表 3 に示す. 参考のため, 文節内部のモーラ遷移をもとに作成したモーラ bigram についても示しておく.

認識性能の評価に用いるモーラ認識率 C は式 (4) のように定義する.

$$C = (N_{all} - N_{del} - N_{sub} - N_{ins}) \;/\; N_{all} \qquad (4)$$

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

実験は言語モデルの重み S を 1.0 ~ 20.0 まで 1.0 刻みで変化させて行っており, 各実験で前・後段ともに同じ値を用いている. なお, この S は認識時に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. 前段におけるモーラ認識率を C_b, 前段で得られたモーラ境界情報を用いて行なった句境界検出結果をもとに後段で再認識したときのモーラ認識率を C_a として, アクセント句境界検出性能 R_d, R_i とあわせて図 4 に示しておく. 参考のため, モーラ境界情報を前段の認識結果からではなく, 強制切り出しによって求めた場合について, 1) 正しくアクセント句境界を与えたときの後段でのモーラ認識率 C_c と, 2) アクセント句境界を句境界検出部によって得たときのモーラ認識率 C_d (その時の句境界検出性能は表 2 の通り) も同図に示しておく.

表 3: モーラ bigram のテストセット・パープレキシティ (モーラあたり)

実験	文内部 (前段認識部)	アクセント句内部 (後段認識部)	文節内部
(a)	42.01	28.66	28.69
(b)	40.22	29.19	28.61
(c)	41.94	28.76	28.69
(d)	40.92	29.27	28.60

5 結論・考察

本論文では, F_0 パターンから検出したアクセント句境界を語彙制約なし認識であるモーラタイプライタ型の認識器に利用することよって, モーラ認識率を向上させる手法について説明し, その特定・不特定話者実験の結果と認識性能向上に関して報告した.

本手法ではアクセント句境界検出によって語彙制約なし音声認識の言語モデルを切替え, 性能向上を狙っているが, 表 3 にあるようにパープレキシティの面からみてもその妥当性が伺える. また, 図 4 の結果をみると, 正しいアクセント句境界を入力とした後段の認識率 (C_c) は, 全ての実験において, 文全体を入力とした前段での認識率 (C_b) より高く維持され, 最高で約 4 % の性能改善を示しており, この結果からもその効果が見てとれる.

前段の認識率 (C_b) と後段の認識率 (C_a) を比較すると, 性能の向上が $S = 7$ 付近で特によくみられ, 最高で約 2 % の性能改善を示した. しかし, 実験 **(b)**, **(c)** では性能の改善が顕著に現れていない. これは, 前・後段における言語モデルの重み S を独立に最適化することや句境界検出性能の改善などで改良が期待される. また, 全ての実験について, C_d が C_a より高い傾向を示しているが, これは, 句境界検出時に利用するモーラ境界情報の違いによるものであり, モーラ境界の精度の悪さや学

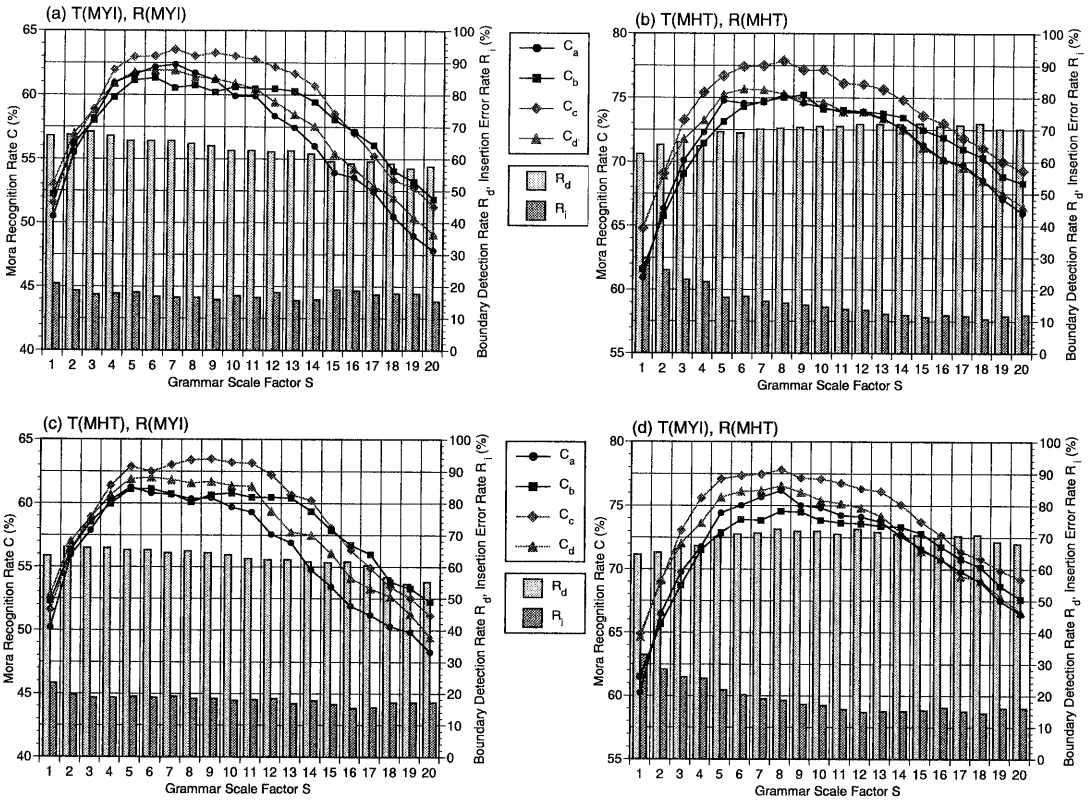

図 4: 前・後段での特定話者実験 (a)(b)・不特定話者実験 (c)(d) でのモーラ認識率と句境界検出性能

習用データとの不一致により C_a が劣化してしまうことに起因していると考えられる. したがって, この点についての改良も施す必要がある.

なお, 本手法では, 前段の認識結果ならびにアクセント句の境界検出結果として第一候補のみを用いているため, 複数の認識結果を使ったシステムの改良についても検討が望まれる.

参考文献

[1] 中川聖一, "小特集に寄せて –音声対話システム構築の課題–," 音響誌, Vol.54, No.11, pp.783-790 (1998-11).

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

[3] 北 研二, 江原暉将, 森元 逞, "連続音声認識における未知語処理," 音講論, Vol.I, pp.93-94 (1991-3).

[4] Y. Sagisaka, N. Campbell, and N. Higuchi, ed., Computing Prosody, Part IV, Springer-Verlag, New York (1997).

[5] 岩野公司, 広瀬啓吉, "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," 情処学論, Vol.40, No.4, pp.1357-1364 (1999-4).

[6] S. Young et. al., The HTK Book, v2.1, Cambridge University, Cambridge (1997).

[7] 河原達也 他, "日本語ディクテーション基本ソフトウェア (97 年度版)," 音響誌, Vol.55, No.3, pp.175-180 (1999-3).

[8] P. Clarkson and R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit," *Proc. EUROSPEECH'97*, Rhodes, pp.2707-2710 (1997-9).