韻律的特徴を用いた音声認識の高精度化
Use of prosody for speech recognition

東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo
峯松 信明
Nobuaki MINEMATSU

＜研究協力者＞
東京大学大学院情報理工学系研究科
広瀬啓吉　　村上隆夫

This paper reports some trials to use prosodic features to enhance speech recognition. The use is experimentally examined in different levels of interaction between the prosodic features and lexical information retrieval. The levels are parameter, phone, word, and phrase levels. In parameter level, quantitative analysis was done on correlation between F_0 and cepstrums. In phone level, biphones were used for cross-word triphones if a prosodic boundary or a pause was found between the two words. In word level, n-gram language models were divided into two types; models for crossing the prosodic boundary and those for not. Two trials were done in phrase level. Accent nucleus at the beginning of a phrase was focused and decoding strategy was modified if the nucleus was found there. In the other trial, beam width was dynamically controlled according to prosodic boundaries. Although not all of these trials were successful, every trial is described briefly in this paper.
Key words: speech recognition, prosody, accent nucleus, language model, acoustic model, decoding

1 F_0 変化に起因するスペクトル包絡変動の分析

1.1 背景と目的

従来の音声情報処理では音声生成のモデルとしてソース・フィルタモデルを仮定することが多く，その結果，音源の特性と声道の特性は独立に扱うことが多かった。音声認識の場合でも，ケプストラムリフタリングにより音源特性が分離できるとの仮定の下で処理系が構築されている。一方，波形接続型の音声合成では，音源特性と声道特性との独立性を仮定することで生じる品質劣化が避けられず，結局のところ，幾つかの F_0 レンジに対応する波形素片データベースをレンジの数だけ揃えることで解消している。

本研究ではこれらの動向を踏まえ，F_0 変化に起因するスペクトル変動をケプストラムをパラメータとして定量的に分析することとした。基本的には，同一音素の異発声（異 F_0）間において，ケプストラム距離と F_0 変動との依存性を見た。詳しい報告は参考文献 [1] を参照して戴きたい。

1.2 日本語母音音声を用いた分析

孤立発声された日本語五母音資料より，以下の条件を満たすフレーム対を抽出し分析対象とした。なお，孤立発声母音であるが，発声中，F_0 は故意に上昇・下降させる形で発声させている。

- 同一発声（同一話者）中の 2 フレーム
- 入り渡り，出渡り部を除く安定した 2 フレーム
- 2 フレーム間の時間差が 100[msec] 以上 300[msec] 以下，或は 100 [msec] 以下，かつ F_0 比が $2^{2/12}$ 以上。

条件を満たす全ての 2 フレーム対を抽出し，横軸に F_0 変化の大きさ，縦軸にスペクトル変動の大きさ（ケプストラム距離）をプロットしたものを図 1 に示す。F_0 変化は 12 が 1 オクターブの変化を意味する。図より F_0 変化の増加に伴いスペクトル変動も増加していることが分かる。図 2 はスペクトル変動を，母音別に，重回帰分析に基づいて予測した場合の予測誤差を示している。高精度の予測が可能である。なお図中の点線は 5 母音音響空間をコードブックサイズ 5 で VQ した場合の VQ 歪みであり，スペクトル変動の大きさを示す一つの尺度である。

1.3 日本語子音音声を用いた分析

孤立発声が困難であるため前後に母音を付与した発声（VCV）を用いた。この際，前後母音からの調音結合に基づくスペクトル変動が懸念されるため，母音からの影響が十分無視できる音声区間を実験的に検討し，該当区間のみを用いて分析を行なった。

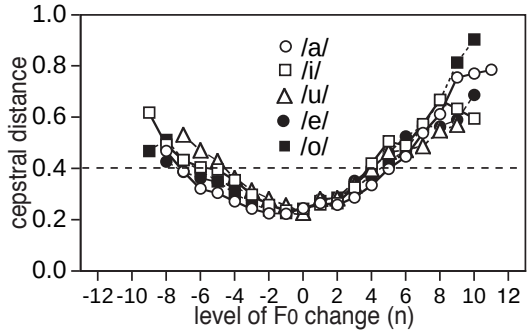

図 1: F_0 変化に起因するスペクトル包絡変動

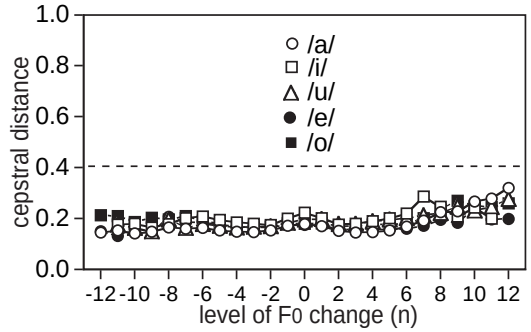

図 2: スペクトル包絡変動の予測誤差

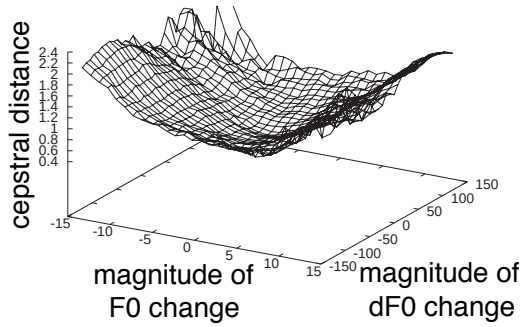

図 3: F_0，ΔF_0 に起因するスペクトル変動

また無声子音についても，前後の有声区間から補間された F_0 との依存関係について検証した。

フレーム対に関して，対間の F_0 差，ΔF_0 差とその時のスペクトル変動を計算した。話者 A，音韻/n/における結果を図 3 に示す。x 軸は F_0 変化の度合いを，y 軸は ΔF_0 差を示す。他話者の場合でもほぼ同様であった。図より F_0 及び ΔF_0 変化に依存してスペクトルが変動している様子が分かる。

ΔF_0 差がほぼ 0 のフレーム対のみを用いて分析を行なった。図 4 に種々の音韻における F_0 変化とスペクトル変動の関係を示す。スペクトルの F_0 への依存性は，分析対象音韻のうちで/n/が最も強く，/a/や/j/がそれに続き，/z/や/g/は有声子音の中でも低いことが分かった。また，無声子音では，/h/が/z/と同等のスペクトル変動を示したものの，/s/の F_0 依存性は極端に小さいことが分かった。

3 人の話者の F_0 変化とスペクトル変動の関係を

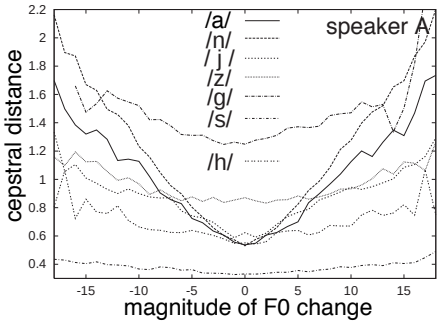

図 4: 音韻ごとの F_0 変化とスペクトル変動の関係

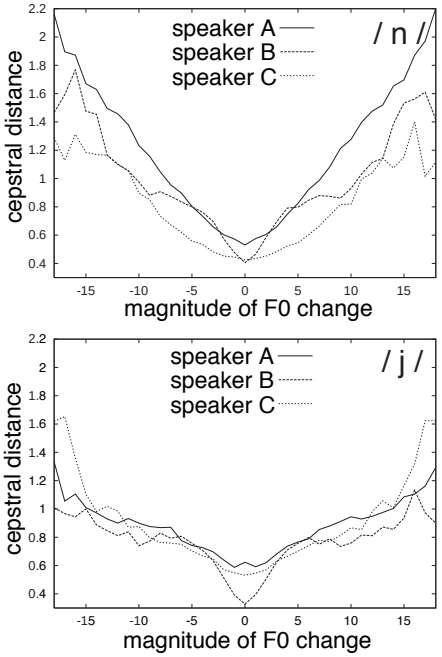

図 5: 話者ごとの F_0 変化とスペクトル変動の関係

分析した。結果を図 5 に示す（有声子音/n/と/j/）。どの話者でもスペクトルが F_0 に依存して変化している様子を確認することができた。これらの結果より，F_0 変化に基づく子音スペクトル変動を重回帰分析により予測した。予測精度は母音と同様に高かったが，母音，子音何れの場合においても，得られたモデルパラメータの話者依存性や音韻依存性は否めず，話者情報，音韻情報が不定の場合にはこれらの予測が極めて困難となる。即ち本分析によって得られた知見を，そのまま音声認識へと導入することは断念せざるを得なかった。

2 Cross-word 音響モデル選択における韻律利用

2.1 背景と目的

大語彙連続音声認識では，二段構成をとることが多いが，各段で使用される音響モデルの差異として，

単語境界（cross-word）における音素環境依存性がある。即ち，処理の高速化のために第一段では単語境界においては環境非依存のモデルを使用し，第二段では（高精度なモデルである）環境依存のモデルを使用する。しかし，多くの音韻（変形）規則が韻律句内の現象を説明していることから推察されるように，韻律句境界となっている単語境界では，調音結合の度合いが低くなることが容易に予想される。このような場合においても環境依存のモデルを使用することは認識率の低下に繋がる恐れがある。そこで，韻律句境界あるいはポーズ境界が仮説探索において単語境界と一致している場合には，音素環境依存モデルの使用を制限する方法について検討した。

2.2 韻律に基づく音響モデル選択の効果

cross-word 音響モデル利用の動的制御による効果を検討する。大語彙連続音声認識結果を表 1 に示す。SCR は文正解率であり，また，×RT は第二パスまで含めた RT ファクタである。韻律境界位置情報に基づいて CCD(cross-word context dependent) モデルを使い分ける提案手法によって，SCR が顕著に上昇している (約 14[%] の上昇率)。また，ビーム幅の動的制御及び CCD モデルの動的適用によって，ほぼ同一の WAR を保ったまま効果的に RT ファクタを低減させていることが分かる。本手法の詳細な報告は参考文献 [2] を参照して戴きたい。

3 韻律句境界を考慮した N-gram 言語モデル

3.1 背景と目的

韻律句境界やポーズ境界は，それ以外の音声区間と比較して音響的に特徴的な様態を呈するだけでなく，言語的にも特徴的な性質を示すことが推測される。即ち，言語的な情報ストリームの「切れ目」として存在していると考えれば，境界前後の言語的繋がりは希薄になると考えられる。例えば日本語の場合，文節境界の前後と，非文節境界の単語境界前後では後者の方が次単語の予測が容易である（即ち言語的繋がりが強固である）。大語彙連続音声認識の場合，この言語的繋がりは言語モデル（N-gram）として実装されるが，このモデルを境界時と非境界時とで区別して構築し，利用することを検討した。詳細な報告は参考文献 [3] を参照して戴きたい。

表 1: 韻律境界情報に依存した音響モデル選択の効果

strategies	WAR[%]	SCR[%]	×RT
within-word triphones	86.0	52.0	9.2
cross-word triphones	90.1	56.0	7.9
static beam width + CCDs with PBs	91.3	64.0	8.1
dynamic beam width + CCDs with PBs	89.5	62.0	6.5

3.2 二種類の言語モデルの実装

言語モデルの学習には，新聞数年分といった大規模テキストコーパスを要する。ある言語境界を跨ぐ遷移と跨がない遷移とを区別し，二種類の言語モデルを構築する場合，従来方法では，例えば境界を跨ぐ遷移を新聞数年間分と同様の量だけ集める必要が生じる。言語境界として韻律境界（即ち音声があって初めて定義できる境界）を用いた場合，韻律境界情報が付与された音声データが同規模だけ必要となり，これを用意することは現実的に不可能である。

この問題を品詞情報に着眼することで解決する。品詞に着眼すれば，その語彙（異なり品詞数）は異なり単語数より格段小さい為，小規模の音声データベースでも品詞に基づく bi-gram カウントは比較的安定した統計量として算出される。この品詞の bi-gram カウントを用いて，単語の bi-gram カウントを二種類（境界を跨ぐ遷移と跨がない遷移）推定し，それらを用いて二種類の単語 bi-gram を構築する。

w_i から w_j へ遷移する bi-gram カウントを N，w_i の品詞は $\mathrm{pos}(w_i)$，w_j の品詞は $\mathrm{pos}(w_j)$ であるとする。$\mathrm{pos}(w_i)$ から $\mathrm{pos}(w_j)$ へ遷移する品詞 bi-gram カウントにおいて，α%が言語境界と共に出現したとする。この場合，言語境界と共に出現した単語 bi-gram カウントを αN，言語境界とは別の個所で出現した単語 bi-gram カウントを $(1-\alpha)N$ として，二種類の単語 bi-gram を導出する。

上記アルゴリズムは，言語境界の有無の情報が品詞遷移の様子に明確に現れることを前提としている。表 2 に言語境界を跨ぐ場合の品詞遷移と跨がない場合の品詞遷移の様子を示す。言語境界の有無によって品詞遷移の様子に大きな差異があることが分かる。

3.3 評価実験

言語的境界（韻律句境界及びポーズ境界）に基づいて構築された二種類の言語モデルのパープレキシ

表 2: 言語境界を跨ぐ場合（上表）と跨がない場合（下表）の単語遷移における品詞分布 [%]

		遷移後			
		名詞	動詞	助詞	副詞
遷移前	名詞	71.1	13.4	1.4	2.8
	動詞	85.6	5.7	1.1	4.0
	助詞	51.1	34.3	0.2	6.1
	副詞	59.6	28.8	0.0	1.4

		遷移後			
		名詞	動詞	助詞	副詞
遷移前	名詞	8.9	5.2	67.5	0.1
	動詞	6.2	12.7	43.8	0.0
	助詞	6.8	47.9	36.9	0.3
	副詞	2.5	15.0	60.0	0.0

表 3: 評価実験結果 1（学習：話者 MYI による 453 文発声，評価：話者 MHT による 50 文発声）

	全遷移	非言語境界	言語境界
ベースライン	117.4	46.5	2082
提案モデル	110.4	44.4	1893
PP 削減率	6.0%	4.6%	9.1%

表 4: 評価実験結果 2（学習：話者 MHT による 453 文発声，評価：話者 MYI による 50 文発声）

	全遷移	非言語境界	言語境界
ベースライン	117.4	57.3	1436
提案モデル	112.8	55.3	1364
PP 削減率	3.9%	3.5%	5.0%

ティ（以下，PP と略す）に基づく評価を行なった。なお，評価文音声に対して韻律分析を行ない言語的境界を自動抽出し，また，音素アライメントをとり，自動抽出された言語境界がアライメントによって得られる単語境界位置と重なる場合に言語境界時の言語モデルを参照し，そうでない時に非言語境界時の言語モデルを参照することとした。

結果を表 3，表 4 に示す。話者 open，データ open の評価実験において，提案モデルがベースラインモデルに対して PP を減少させていることが分かる。これは，与えられたテキストをよりコンパクトに表現していることを意味し，同一テキストに対する音声認識のタスク難度を軽減させることが期待される。

しかし，実際に大語彙連続音声認識に対して本提案言語モデルを用いたところ，認識性能の改善は観測されたが，その度合いは非常に僅かなものであったことを付記しておく。

4 句頭のアクセント核に着眼した仮説探索制御

4.1 背景と目的

さて，音声からその語彙情報を確定する場合，人間は韻律の何を用いて，心的辞書検索を効率化しているのだろうか？韻律的特徴，特に F_0 は種々の要因により容易に変形する。このように不安定な物理現象を用いて処理系を構築する場合，いつ F_0 の挙動が安定するのか，そもそも常に F_0 をモニタリングすることが正しいのか，捨てるべき時もあるのではないだろうか，という視点も必要であろう。

ここでは，筆者が従来行なった，音声知覚・音声合成・音声認識・日本語音韻論研究などから示唆される「F_0 情報を参照するタイミング」を検討し，その検討を仮説探索において実装する。詳しい報告は参考文献 [4] を参照して戴きたい。

4.2 種々の先行研究に基づく韻律への着眼

4.2.1 音声知覚研究からの提言

日本語の単語アクセントは基本的に「F_0 の落ち(アクセント核)」の場所によって分類される。音声の入力は left-to-right であるため，語頭近くに核がある単語ほど，音声聴取のより早期の段階で，話者が意図した単語アクセント型が「既知」となり，心的辞書検索範囲を限定した語彙同定処理が可能となる。筆者の先行研究より以下が示されている。

- 単語の同定に必要な単語頭の音声長は，1 型の場合他型と比較して顕著に短くなる。
- 2 型の場合も語彙同定加速化は観測されるが，単語の親密度の影響を受ける。1 型の場合は低親密度の場合でも安定して高速化される。
- 単語音声を本来とは異なるアクセント型に変形しノイズ環境下で提示すると,「本来 1 型である単語を非 1 型とした場合」「本来非 1 型であるものを 1 型とした場合」に同定率が顕著に下がる。

これらの結果より「語頭アクセント核を key にした語彙検索過程が存在すること」「語頭から離れると（音声入力が継続されると），加速化効果は弱くなること」が示唆される。これは，分節的特徴が伝搬する音素（モーラ）の情報と，核が伝搬するアクセント型の情報とが，時間軸上でどのように統合されていくのか，という問いに対する一つの考えを示唆している（概念図を図 6 に示す）。しかしながら，「F_0 情報を参照するタイミング」として「語頭」と

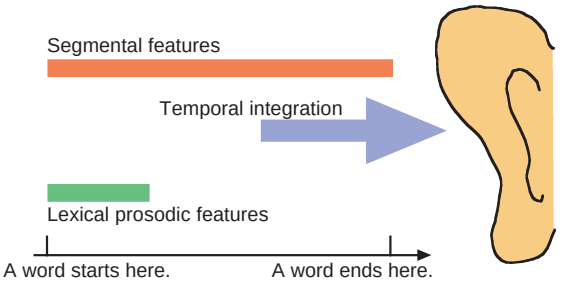

図 6: 音素情報と韻律情報の統合による孤立単語知覚

答えることは出来ない。日本語の場合，連続音声になると語レベルのアクセントは容易に変形してしまうからである（アクセント変形）。

4.2.2 音声合成研究からの提言

筆者は，日本語テキスト音声合成におけるアクセント変形規則の高精度化についても研究を行なっている [5]。これは匂坂らによって提案された変形規則の拡張に相当する。この規則を用いることで「上記した知覚実験結果が，連続音声中において利用可能か否か」が検討できる。この規則は「アクセント句境界が与えられた場合，各アクセント句が持つ高々 1 つのアクセント核はどこに位置すべきか」の情報を与える。着目する語の前後に単語が存在する場合，着目単語は直前・直後の単語によってアクセント型が変形する。さて，アクセント句頭に着目し，上記アクセント変形規則に対して「句頭にアクセント核が観測された場合，その句頭単語を孤立発声すると 1 型となるのか?」と問い合わせると「一部例外を除いて，必ず 1 型となる」となる。例外とは，変形後 2 型となる語において第 2 モーラが無声化すると核が前にずれ，1 型となる場合である。しかしこの場合，第 2 モーラは無声化母音となっており（その結果，母音直前の子音も無声子音のはず），第 2 モーラそのものに F_0 情報は存在しない。認識における韻律利用を「句頭の第 1,2 モーラが有声と判定された場合のみ」行なえば，この問題は回避できる。

4.2.3 音声認識研究からの提言

なぜ，句頭に着眼するのか？例えば，句境界時及び句内時の N-gram を個別に構築し，句境界及び句内における PP を求めると表 5 のようになる。両者の間には非常に大きな差がある。PP は情報論的に考えたブランチングファクタであるため，認識（同定処理）の困難さに直結する。即ち，句境界時では認識処理の困難さが 100 倍増加することを意味する。言語的複雑さが急増した状況下で，句頭に付随する新たな情報源に着眼することは至極自然である。

4.2.4 日本語音韻論研究からの提言

1 型単語はどの程度存在しているのだろうか？橋本らによれば，表 6 のような分布となる。その割合は非常に小さい。アクセント句が 3 単語で構成され，アクセント句頭におけるアクセント型分布が上記分布に沿うと考えた場合，1/3 x 0.143 = 約 5%，即ち，入力音声中の全単語の約 5%に焦点を充てた処理系の導入が「いつ韻律を使うべきか」の答えとなる。

先行研究 [5] では，アクセント属性推定アルゴリズムの開発のみならず，固有名詞に対するアクセント辞書の整備も行なった。新聞から得られる頻出 5000 語の固有名詞に対するアクセント型分布を求めると表 7 のようになった。1 型が多い。固有名詞の認識は言語属性が曖昧であるため，認識の難易度は増加する。アクセント核に着眼した処理系が，固有名詞認識精度を向上させることが期待される。

4.3 いつ使うのか，いつ捨てるのか？

ここまでの議論を図的にまとめると図 7 のようになる。音声認識では，フレーム同期で加算される音響スコアと，フレーム非同期に加算されていく言語スコアとの累積によりスコア計算が行なわれる。後者は，仮説上の単語境界において付与される N-gram スコアと新たな単語を仮説上に追加した場合に付与されるインサーションペナルティとに分かれる。本研究では，フレーム非同期に句頭にて加算される新スコアとして語レベルの韻律スコアを考える。韻律スコアの具体的な定義であるが，F_0 に関するパラメータ α に対して，その実現値 α_0 が得られた

表 5: 句内における PP と句境界における PP

	句内	句境界
ベースライン	25.7	2752
提案手法	23.9	2408

表 6: 単語アクセント型の分布 (%)

type-0	type-1	type-2	type-3	type-4	type-5
42.9	14.3	9.6	16.4	9.6	7.1

表 7: 固有名詞における単語アクセント型の分布 (%)

type-0	type-1	type-2	type-3	type-4	type-5
31.5	33.2	14.5	10.7	5.1	5.0

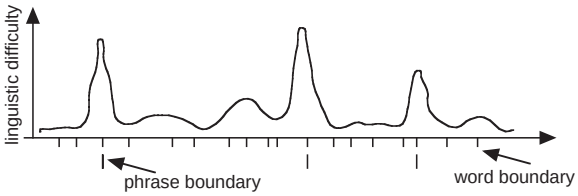

図 7: フレーム非同期に加算されるマッチングスコア

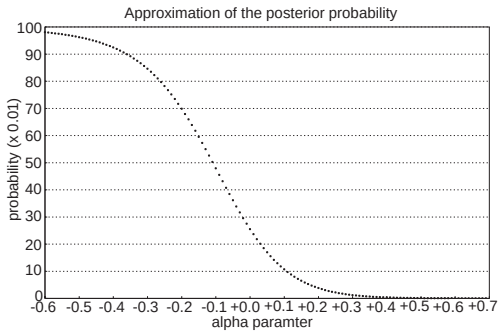

図 8: 事後確率の推定

時にその単語のアクセント型が 1 型となる事後確率 $P(\mathrm{at}=1|\alpha=\alpha_0)$（at = accent type）を考え，この事後確率の関数として韻律スコアを導入する。

韻律観測量 α として，対象となる単語の第 1 モーラ中の母音区間，第 2 モーラ中の母音区間における代表 F_0 値（F_{0rep}）を求め，その差分で定義した。

$$\alpha = F_{0rep}^{2st} - F_{0rep}^{1st}$$

任意母音の F_{0rep} は以下のようにして求める。

$$F_{0rep} = \frac{\sum_t w_t \log(F_{0t})}{\sum_t w_t}$$

ここで，F_{0t} は（母音区間内の）時刻 t の F_0，w_t は時刻 t のパワーである。事後確率の推定であるが，

$$\begin{aligned} P(\mathrm{at}=1|\alpha=\alpha_0) &= \frac{P(\mathrm{at}=1,\alpha=\alpha_0)}{P(\alpha=\alpha_0)} \\ &= \frac{P(\alpha=\alpha_0|\mathrm{at}=1)P(\mathrm{at}=1)}{\sum_{\mathrm{at}} P(\alpha=\alpha_0|\mathrm{at})P(\mathrm{at})} \end{aligned}$$

と変形し，事後確率を $P(\alpha=\alpha_0|\mathrm{at})$（多くの場合正規分布で近似）及び $P(\mathrm{at})$ とから推定した。ATR503 文 (MYI) に対して行なうと，図 8 のようになった。発音辞書中にその単語を孤立発声した際に 1 型となるか否かの情報を記載し，観測量 α から計算される事後確率値，及び認識対象の語彙のアクセント情報を元に韻律スコア（ボーナスあるいはペナルティー）を決定することで，仮説探索を修正する。

4.4 認識実験

仮説展開において現時刻が無音に相当する単語（句読点など）の最終フレームであった場合，以降

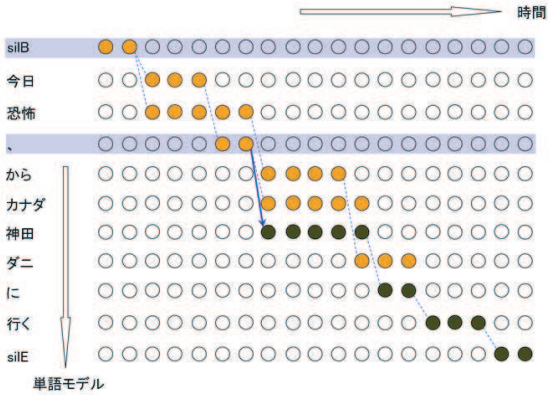

図 9: 第二パスにおける韻律スコアの導入

のフレーム系列を処理する際にアクセント処理系を駆動することで，提案手法は実装される。

本研究では，第二パス（リスコアリング）において提案手法を実装した。Julius の場合，第一パスでの認識結果がトレリスの形で第二パスに渡される。トレリスには認識仮説における単語群が時間情報と共に格納されているが，各単語候補は，その単語の直前の単語の情報も保持している。第二パスは後ろ向き探索が行なわれ，既に確定した単語系列を元に直前の単語を（逆向き trigram による言語スコアを参照して）決定する。当然単語候補は複数ある訳だが，各単語候補にはその単語候補の直前に位置すると考えられる単語の情報が参照可能である。この直前の単語が無音相当の単語であった場合，単語候補のスコアに韻律スコアを導入する（図 9 参照）。

韻律スコアとして 2 通りの方法を検討した。なお何れの場合も，1 型となる事後確率を p とした時 $p > 0.5$ の場合に限り，韻律スコア S を加えた。

ボーナス及びペナルティを導入

$$S = \begin{cases} 10(1+\log_2 p) & (\text{仮説中の単語が 1 型の場合}) \\ 10(1+\log_2(1-p)) & (\text{非 1 型の場合}) \end{cases}$$

ペナルティのみを導入

$$S = \begin{cases} 0 & (\text{仮説中の単語が 1 型の場合}) \\ 10(1+\log_2(1-p)) & (\text{非 1 型の場合}) \end{cases}$$

男性話者一名 (MHT) の ATR 音素バランス 503 文を用いて認識実験を行なった（音響モデルは不特定話者モデルを使用）。韻律スコアの導入前後で認識結果が変化した単語について表 8 にまとめる。なお，誤 → 誤とは，韻律スコア導入で認識結果は変わったが，両方とも誤りとなった場合を意味する。表より，ペナルティのみを導入した方が，誤認識の誘発を抑えながら認識性能を向上させていることが分か

表 8: 韻律スコア導入による効果

	ボーナス&ペナルティ	ペナルティのみ
誤 → 正	13 単語	10 単語
正 → 誤	12 単語	3 単語
誤 → 誤	25 単語	11 単語

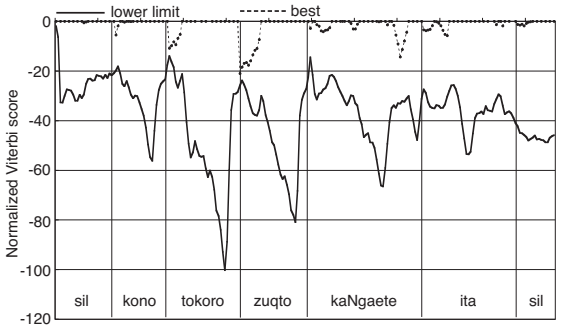

図 10: 静的なビーム幅を用いた探索処理における，アクティブな仮説に対する尤度変化

る。韻律スコア導入のために不正解となった単語について調査したところ，1) 語頭 2 モーラに対するアライメントにずれが生じ，不適切な F_0 パターンを参照していた，2)F_0 抽出の時点でグロスエラーが生じていた，の 2 つが主要因であった。これらはいずれも韻律処理を駆動する前処理のエラーであり，導入した韻律処理系そのものが誤認識を引き起こしている例は無かった。本研究では，韻律が「様々な要因によって容易に変化する現象である」という事実を念頭に置き，韻律情報と語彙情報とが，信頼性高く，関連する状況を種々の音声研究を参照する形で絞り込んでいるが，その効果であると考えている。

5 韻律句境界を考慮した仮説探索ビーム幅制御

5.1 背景と目的

木構造辞書を用いたデコーディングにおいて，ルートに近いノード探索時に正解単語がビームから除外されないためには，十分広いビーム幅が必要となる。図 10 はビーム幅固定の仮説探索において，アクティブな仮説に対する時間正規化ビタビスコアの最小値と，正解スコアを示している。単語尾に近づくにつれて，ビームに残る仮説中の最低スコアは減少する傾向が分かる。これは，不要なアクティブ仮説が増大していることを意味する。

ここでは，ビーム幅を単語頭では広く，単語尾に近づくにつれて徐々に狭くし，不要な仮説展開を抑える方式を提案し，実験的に検討する。なお，詳細な報告は参考文献 [6] を参照して戴きたい。

5.2 ビーム幅制御の実装

韻律句境界情報からは正解文における句境界位置が推定されるが，デコーディング処理中は，仮説展開において言語尤度が加算されるタイミングに同期したビーム幅制御も必要となる。次式に従ってビーム幅を動的に制御した。

$$
\begin{aligned}
P(t) &< P_{max}(t) - \hat{\lambda}(t) \\
\hat{\lambda}(t) &= \lambda(0) + \lambda_{\mathrm{var}}(t) \times \\
&\qquad \frac{\sum N_{\mathrm{word\ end}}(t) + \sum N_{\mathrm{phone\ branch}}(t)}{\sum N_{\mathrm{active}}(t)} \\
&\qquad \mathrm{where}, \lambda(0) \geq \lambda_{\mathrm{var}}(t) \geq 1
\end{aligned}
$$

$P_{max}(t)$ は時刻 t における最尤仮説の尤度，$\hat{\lambda}(t)$ が動的に制御されるビーム幅であり，対象とする仮説の尤度が第一式を満たした場合，その仮説は却下される。$\lambda_{\mathrm{var}}(t)$ は時刻 t を含む韻律句頭から韻律句尾にかけて定義される単調減少の関数であり，韻律句頭から韻律句尾にかけてビーム幅を狭める働きを持つ。また，$N_{\mathrm{word\ end}}(t)$，$N_{\mathrm{phone\ branch}}(t)$ は時刻 t が単語 (形態素) 尾である仮説数，及び，木構造辞書中の分岐ノードである仮説数である (即ち，言語尤度が加算されるノードである)。時刻 t においてこれらの条件を満たす仮説数の割合が高い場合，ビーム幅の削減を緩める必要がある。

5.3 評価実験

本研究で構成したマルチパス構成のデコーダを図 11 に示す。評価文音声としては，JNAS データベースの一部 (音響モデル，言語モデルの学習に使用されていない話者，新聞記事による 50 文) を利用した。なお，全て男声であり，一人 5 文ずつ合計 10 名による発声である。

まず，ビーム幅の動的制御による効果について検討する。静的なビーム幅制御に基づく単語正解率 (WAR[1]) を表 9 に示す。ビーム幅を動的に制御した場合の結果を表 10 に示す。いずれも第一パスにおける評価結果である。表より，WAR=86%の時は，ビーム幅を動的に制御することで active node 数を約 50%，RT ファクタを約 30%減少させることができ，ビーム幅の動的制御が大語彙連続音声認識において有効に寄与することが示された。

[1]Word Accuracy Rate。%Correct − %insertion により計算する。

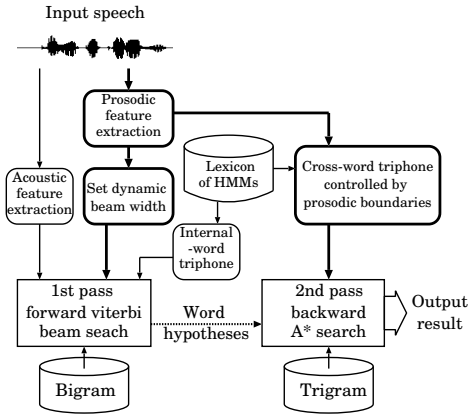

図 11: 提案する大語彙連続音声認識システム

表 9: 静的なビーム幅制御による単語正解率

beam width	WAR [%]	# average active nodes/frame	× RT
20	67.35	33.57	2.0
25	78.23	71.87	2.5
30	79.88	149.06	3.2
35	**86.04**	**282.81**	**4.4**
40	88.91	502.63	5.9
45	90.14	829.02	7.9
50	90.35	1273.89	10.5
55	91.58	1826.49	13.5

表 10: 動的なビーム幅制御による単語正解率

max. beam width	WAR [%]	# average active nodes/frame	× RT
40	78.64	59.44	2.2
50	**86.04**	**139.10**	**3.1**
60	87.06	302.04	4.5
70	89.53	594.12	6.6
80	90.76	1054.34	9.6
90	91.58	1661.17	12.7

6 まとめ

本研究では，韻律情報を大語彙連続音声認識に導入することを目的として，種々のレベルにおける韻律利用において多角的に検討を行なった。まず，パラメータレベルの検討として，F_0 とケプストラム係数の相関分析を行ない，また，単音レベルにおいては，句境界位置に基づいた，環境依存・非依存音響モデルの選択を行ない，その効果を見た。単語レベルにおいては句境界を跨ぐ場合と跨がない場合の単語遷移の言語的差異に基づき，二種類の N-gram 言語モデルを提案した。句レベルにおいては，句頭のアクセント核の有無に基づく仮説探索制御，更には，句を単位としたビーム幅制御についても実験的に検討し，その有効性を検証した。今後は更に上位のレベル，意味・理解のレベルにおける韻律利用についても検討を行ないたい。

参考文献

[1] N. Minematsu, K. Tsuda, and K. Hirose, “Quantitative analysis of F0-induced variations of cepstrum coefficients,” Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, pp.113–117 (2001)

[2] S. W. Lee, K. Hirose, and N. Minematsu, “Incorporation of prosodic modules for large vocabulary continuous speech recognition,” Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, pp.97–101 (2001)

[3] K. Hirose, N. Minematsu, and M. Terao, “Statistical language modeling with prosodic boundaries and its use for continuous speech recognition,” Proc. Int. Conf. Spoken Language Processing (ICSLP’2002), pp.937–940 (2002)

[4] 村上隆夫，峯松信明，広瀬啓吉，“音声認識における語レベルの韻律利用に関する実験的検討”，日本音響学会春季講演論文集，3-Q-24，pp.191–192 (2004-3)

[5] N. Minematsu, R. Kita, and K. Hirose, “Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese Text-to-speech Conversion,” Trans. IEICE, vol.E86-D, no.3, pp.550–557 (2003-3)

[6] S. W. Lee, K. Hirose, and N. Minematsu, “Efficient search strategy in large vocabulary continuous speech recognition using prosodic boundary information,” Proc. Int. Conf. Spoken Language Processing (ICSLP’2000), pp.274–277 (2000)