論文の内容の要旨

論文題目：Incorporation of Prosodic modules for Large Vocabulary Continuous Speech Recognition
（大語彙連続音声認識における韻律モジュールの導入）

氏名：李 時旭

音声認識技術は，近年の計算機技術の向上，及び，音声/言語現象を数理統計的にモデル化する方法論の確立によって飛躍的に進歩し，パソコン上で動作する実用アプリケーションも数多く市販されるようになった。本研究では，韻律的特徴を有効に音声認識技術への導入することを検討した。具体的には，大語彙連続音声認識における仮説探索処理における韻律句境界情報の利用を行なった。

韻律句境界情報を利用した大語彙連続音声認識の高精度化

木構造辞書を用いた大語彙連続音声認識における仮説探索過程では，factoringによって照合対象語彙が決定する以前から(単語レベルでの)言語尤度を推定し，音響尤度との統合が図られる。一般的な factoring では，木構造辞書中のある分岐ノードに接続されている単語群に対する言語尤度として，その単語群中の最大言語尤度を採用することが多い。その結果，ルートノードに近い分岐ノードでは，正解が含まれない単語群の言語尤度がより高く評価され，正解を含む単語群がビーム外に追いやられる可能性がある。このような事態は，ビーム幅そのものを十分に広くすることで避けることができるが，逆に認識時間を増大させる結果となる。一方，デコーディング処理が木構造辞書の単語尾ノードに近づくにつれて，言語的にも音響的にも，入力音声が照合対象としている単語であるか否かの信頼度(即ち，言語尤度，音響尤度の信頼性)はより高くなる。その結果，ビーム幅を削減したとしても認識精度には影響を与えないことが予想される。以上の考察より本研究では，韻律情報より推定される韻律句境界情報を利用して，ビーム幅を動的に制御する方式について検討した。また，多くの仮説探索器(デコーダ)は，二段構成をとることが多い。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ことが多い。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，韻律句境界における音素環境依存モデルの使用を制限したデコーディング手法についても検討した。

韻律句境界の検出

まず、F0 パターンとパワーパターンを用いて韻律句境界の推定を行なった。なお，ここでは文節を単位とした統語境界も，韻律句境界として考えている。まず，パワーパターン中に観測される谷を用いて句境界候補を算出する。次に，観測された F0 パターンと線分近似を行なった F0 の大局近似パターン中の谷を参照することで更に候補を追加する。これらに対して予め規定された規則を用いて句境界を推定する。この規則では閾値処理が行なわれ，主に句境界検出における挿入誤りを制御する働きを持つ。先行研究における評価実験では，30%の挿入誤り時に 80%の句境界検出を実現している。

韻律句境界を利用したビーム幅の動的制御

上述したように，ルートに近いノードにおける factored 言語尤度によって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１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，正解文に対するスコアを示したものである。図より明らかに，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観測される。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。

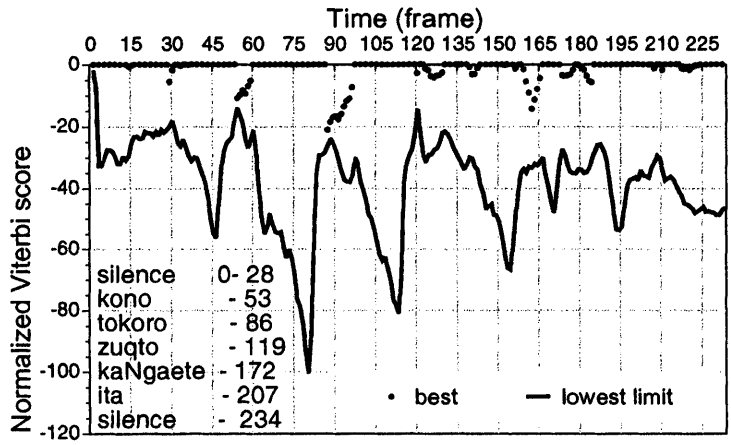

図１．静的なビーム幅を用いた探索処理における，アクティブな仮説に対する尤度変化

以上の考察より，ビーム幅を単語頭においては広く，単語尾に近づくにつれて徐々に狭く制御することで，不要な仮説展開を抑える方式を提案する。なお，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。従ってビーム幅を動的に制御した。

韻律境界を考慮した cross-word 音響モデル

仮説中の単語境界における右側音素環境は(未決定の)次単語に依存するため，単語境界時の環境依存モデルは第二パス(即ちリスコアリング)時に導入されることが多い。この場合前節で述べたように，無条件に cross-word モデルを導入することは認識率の劣化を招く恐れがある。即ち，韻律境界として出現した単語境界では，調音結合の度合いが弱くなることが予測さ

れる。そこで本研究では，韻律境界としての単語境界に対しては，非 cross-word モデルを利用する方式を検討した。

大語彙連続音声認識実験による評価

本研究で構成したマルチパス構成のデコーダを図２に示す。また，音響モデルは状態数 3,000 の triphone を，言語モデルには毎日新聞記事より構築された bigram，trigram を利用した。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。なお，音響分析条件は表１に示す通りである。

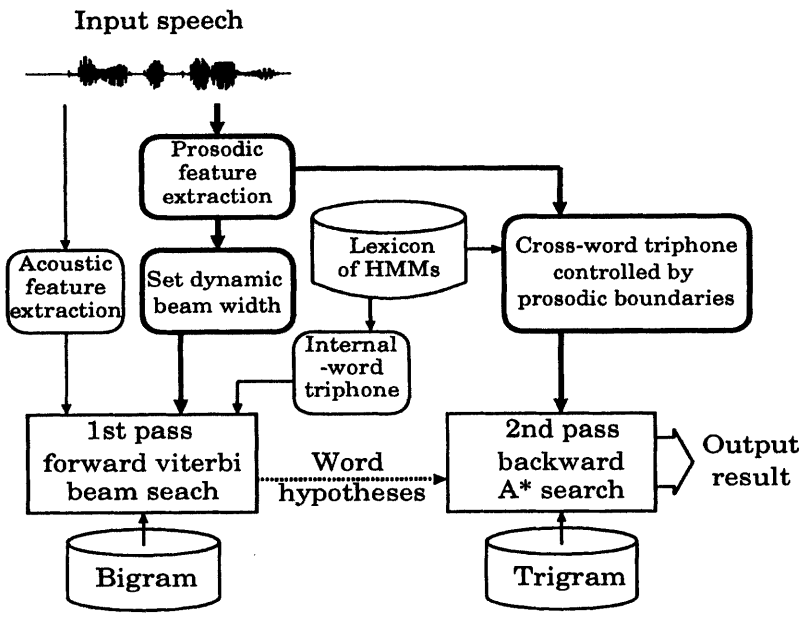

図２．韻律句境界情報を利用した大語彙連続音声認識システム

表１．音響分析条件

サンプリング	16kHz/16bit
高域強調	$1-0.97z^{-1}$
フレーム窓	ハミング窓(25 [msec])
フレームシフト	10 [msec]
音響的特徴	MFCC(12)+ △MFCC(12)+△パワー(1)
分析チャネル数	24

表２．静的ビーム幅制御による単語正解率

Beam width	WAR(%)	# ave. active nodes/frame	Times(xRT)
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３．動的なビーム幅制御による単語正解率

Max. beam width	WAR(%)	# ave. active nodes/frame	Times(xRT)
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

表４．韻律境界情報に依存したcross-wordモデル利用の効果

strategies	WAR(%)	SCR(%)	Times(xRT)
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

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:%Correct - %insertion)を表２に示す。ビーム幅を動的に制御した場合の結果を表３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%，RealTime(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。

次に，cross-word音響モデル利用の動的制御による効果を検討する。結果を表４に示す。SCRは文正解率であり，また，Times(xRT)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。

まとめ

本研究では，音声認識における韻律的特徴利用を念頭に置き，「大語彙連続音声認識の仮説探索におけるビーム幅，cross-word音響モデル利用を韻律句境界情報に基づいて制御する方式」を提案し，その有効性を実験的に示すことができた。