卒業論文

連続音声認識における韻律情報に関する研究

～ＩＭＥＭ法を用いた句境界検出～

東北大学工学部　情報工学科　Ｂ４

三好 哲也

目 次

1 序論

1.1 研究の背景

近年, 電気機器や、ＯＡ機器などの機能の充実に伴ない操作法が複雑化してきており、それらの機械とコミュニケーションをするのが困難になってきている。
このため、人と機械が、自然にコミュニケーションをとる手段としては、音声によるものが最も望ましいと思われる。しかし、会話文の認識は、明確に発声された単語を認識する場合と比べて非常に困難であることは、容易に推測できる。このため、単語認識に用いられる局所的な音韻情報のほかに音声のもつ、イントネーション、アクセントといった韻律情報について検討する必要がある。

1.2 研究の目的

連続音声を認識する上で 韻律情報である、アクセントパタンなどにより句境界ごとに、分割することにより単語認識並みの認識率を得られることを考えると韻律情報を用いた句境界検出は、重要であるといえる。
また、韻律情報の最も基礎となるものは、ピッチ (基本周波数) であり、このピッチは、音声の生成過程では、声帯の振動数に相当する。
そこで、本研究では、励振波スペクトルを句境界情報を含む特徴パタンとして、瞬時化最大エントピー法（ＩMEM法）を用いて抽出し、これをパタンとして、Ｏｎｅ　ｓｔａｇｅ　ＤＰ法を用いて、句境界を検出することをその目的とする。

1.3 本論文の構成

１章では序論として、研究の背景及び目的、２章では、従来法である、ラグウインドウ法による、特徴パタンの抽出法、３章では、２段ＡＲ過程の非定常過程のスペクトル推定法による特徴パタンの抽出法、４章では、３章の方法によるパタンでの実験方法、５章では、実験の結果及び、考察、６章では、結論として、まとめ及び今後の課題を述べ、７、８章では謝辞及び参考文献を挙げる。

2　ラグウインドウ法を用いた特徴パタンの抽出法

2.1　原理

ラグウインドウ法は、フーリエ変換を用いて、パワースペクトルから、ピッチ構造を含んだスペクトルを分離する手法である。このラグウインドウ法によって、得られた、ピッチスペクトルを２０次元の特徴パタンとして、Ｏｎｅ　ｓｔａｇｅ　ＤＰ法により句境界検出を行なう。(以後、ＰＳＰパタンと呼ぶことにする。) ラグウインドウ法による分析系を図1に表わす。

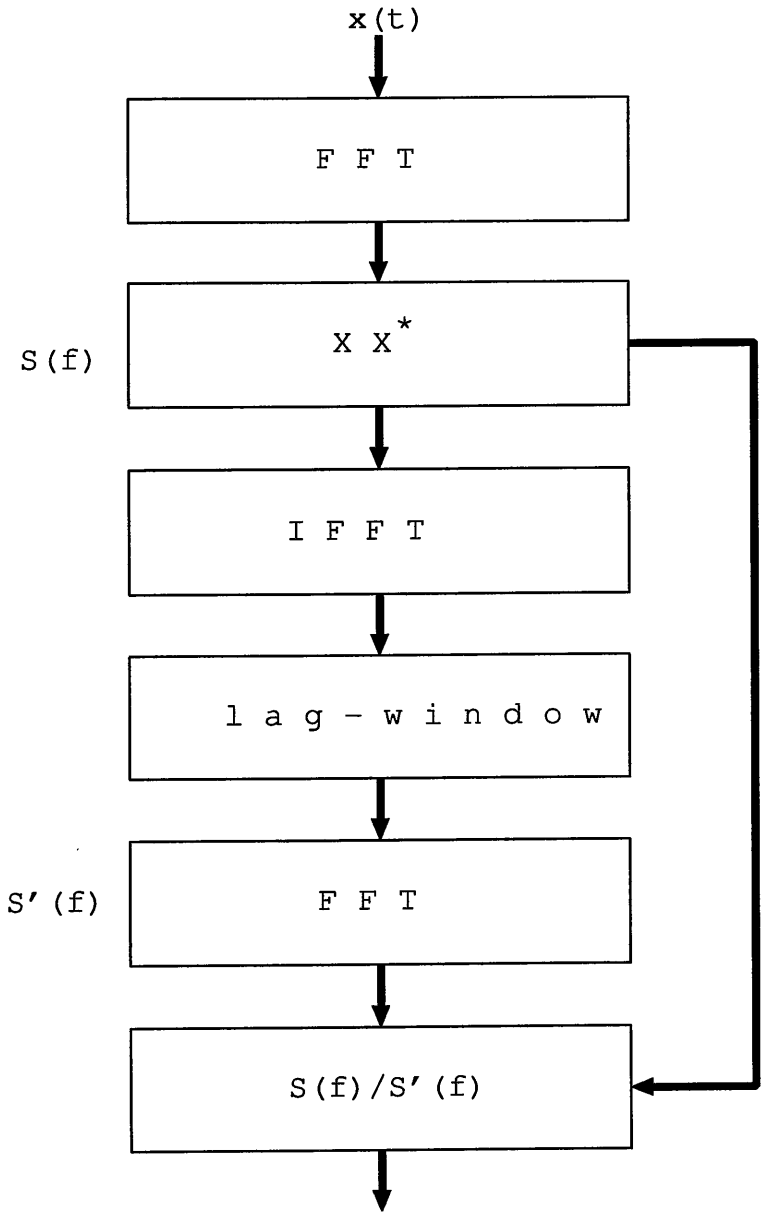

図 1: ラグウインドウ法による分析系

2.2 ラグウインドウ法の問題点

ラグウインドウ法はＦＦＴ法に基づく分析法である。ＦＦＴ法は、本来長いフレーム長が、必要とされるため、時間ごとに、変動する非定常過程のスペクトルを推定するには、フレーム間ごとに不規則な擾乱が生じ(時変性への追随能力が劣る。)、分析効率が、低い。このため、音声のような、非定常過程であるとみなせるような場合には、現象が時刻と共に顕著に変動し、時間依存性の情報が重要になる。そこで、スペクトルが、時刻に対して顕著に変動する非定常過程の正確なスペクトル推定法が必要となる。

3　２段ＡＲ過程の非定常スペクトル推定法を用いた特徴パタンの抽出法

3.1　音声生成のモデル

音声の生成の過程を２段階の過程で起こると仮定し、まず、１段階目で白色雑音から励振波が生成され、２段階目で音声波が生成されるする。この際の各段の過程をＡＲ過程とする。この過程を図2に表わす。

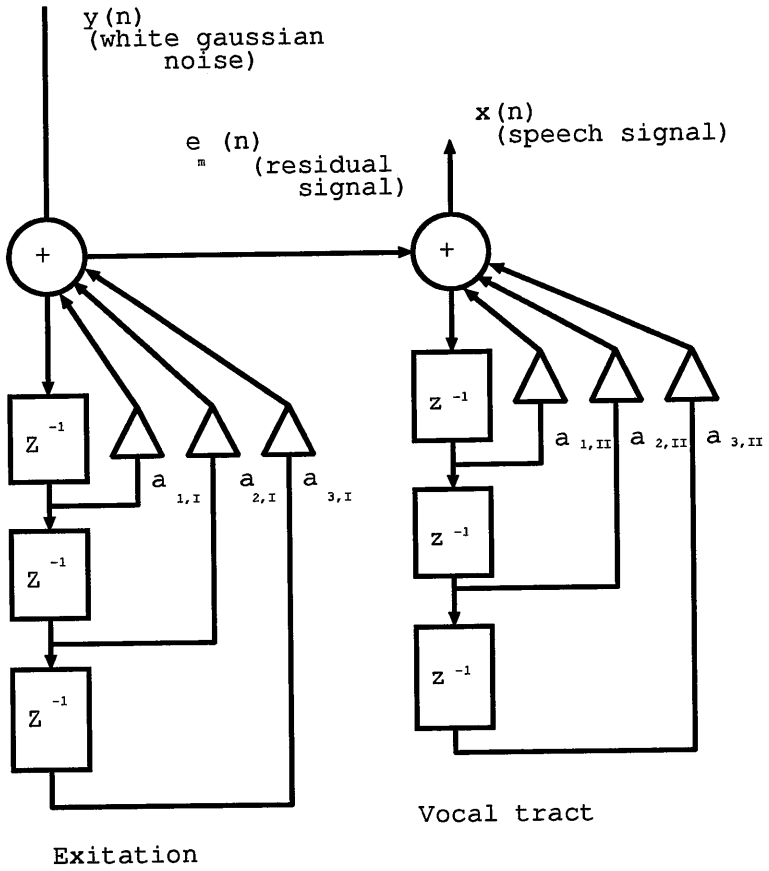

図 2: 音声生成モデル

3.2　原理

音声信号をまず、１段目の分析により、声道スペクトルを推定する。このとき、残差信号も同時に推定し得ることができる。この１段目で得られた、残差信号を低周波成分を効果的に推定するためにまず、ローパスフィルタを通過させ、次に間引きを行ない(本実験では、間引きを行なっていない。)、２段目の分析を行なうことにより、励振波スペクトルを推定することができる。この分析系の構成を図3に表わす。

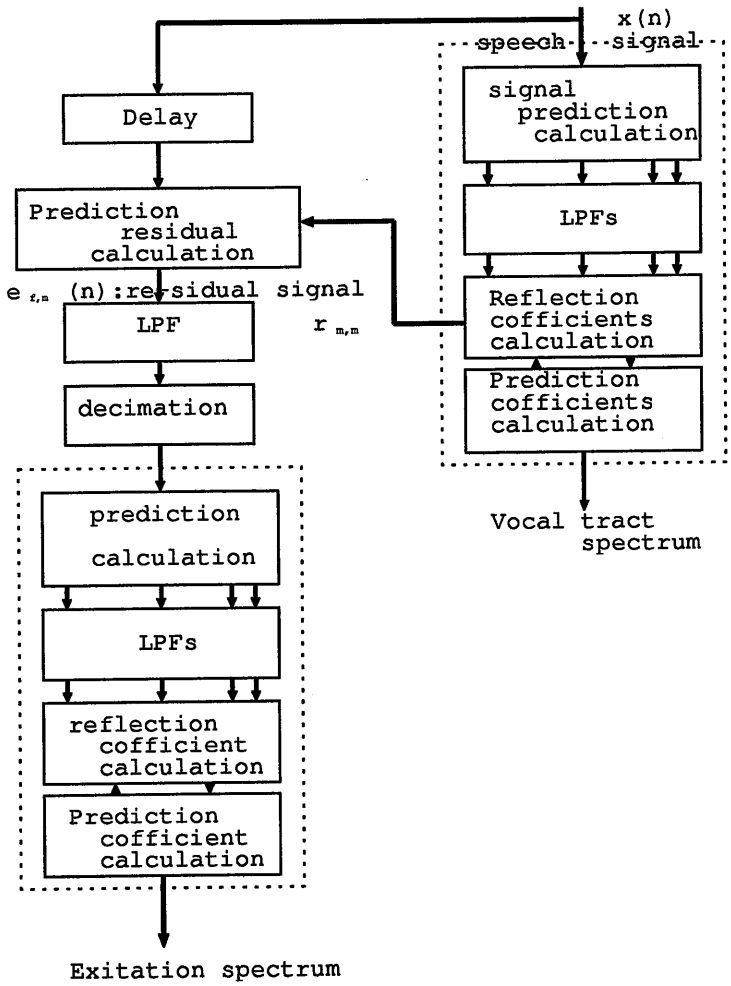

図 3: 励振波の分析系

この得られた励振波スペクトルをＰＳＰパタンと同様に２０次元のベクトルを抽出しこれを特徴パタンとする。この分析の際、非定常スペクトル推定法（瞬時化最大エントロピー法）を用いているため、フレームごとに不規則な変動が除かれた時間追随性のよい、スペクトルが得られる。そのために、これを特徴パタンとして使用することにより不規則な変動のない、句境界に関する情報が得られるのではないかと考えられる。

3.3　瞬時化最大エントロピー法の原理

瞬時化最大エントロピー法は、スペクトルが、時刻に対して顕著に変動する非定常過程のスペクトル推定の方法であり、時間依存性を損なわずに不規則な変動が除かれた、スペクトルが得られる特徴を持っている。
ここで、瞬時化最大エントロピー法の原理について述べる。
スペクトルのフーリエ変換が相関関数であるというＷｉｅｎｅｒ－Ｋｈｉｔｃｈｉｎｅの関係の制約のもとで、エントロピーを増加させないように未知部分を推定することであり、ＡＲ過程 ($x_n = a_1 x_{n-1} + a_2 x_{n-2} + \ldots + a_p x_{n-p} + r_n$) で表わされる数値フィルターの係数を推定することにより、スペクトルを推定する方法であり、すなわち、あるスペクトルをもつランダム波を発生させうる”白色雑音を入力する系”を探り出すということである。このとき、計算アルゴリズムは次のようになる。時刻ｎを終点とするフレームを考え、フレームの時刻をｎで表わすとすると最小にすべき評価関数は次式である

$$E_m(n) = \frac{1}{2}(e_{f,m}^2(n) + e_{b,m}^2(n)) \tag{1}$$

$$e_{f,m}(n) = x(n) + \sum_{i=1}^{m} \gamma_{m,i}(n) x(n-i) \tag{2}$$

$$e_{b,m}(n) = x(n-m) + \sum_{i=1}^{m} \gamma_{m,i}(n) x(n-m+i) \tag{3}$$

ここで、$e_{f,m}(n)$,$e_{b,m}(n)$ は、前方および後方予測残差、$\gamma_{m,i}(n)$ は、予測係数である。
予測係数$\gamma_{m,i}(n)$ について次式が成立する。

$$\gamma_{m,i}(n) = \gamma_{m-1,i}(n) + \gamma_{m,m}(n)\gamma_{m-1,m-i}(n) \tag{4}$$

但し$\gamma_{m,0} = 1$ である。式 (1) を最小とする反射係数$\gamma_{m,m}(n)$ は次式で表わせる。

$$\gamma_{m,m} = \frac{-2C_{m-1}(n)}{F_{m-1}(n) + B_{m-1}(n-1)} \tag{5}$$

$$C_m(n) = \sum_{i=0}^{m}\sum_{j=0}^{m} \gamma_{m,i}(n)\gamma_{m,j}(n)\phi(i, m+1-j; n)$$

$$F_m(n) = \sum_{i=0}^{m}\sum_{j=0}^{m} \gamma_{m,i}(n)\gamma_{m,j}(n)\phi(i, j; n)$$

$$B_m(n-1) = \sum_{i=0}^{m}\sum_{j=0}^{m} \gamma_{m,i}(n)\gamma_{m,j}(n)\phi(m+1-i, m+1-j; n) \tag{6}$$

$$\phi(i, j, ; n) = h(n) * (x(n-i)x(n-j)) \tag{7}$$

但しｈ（ｎ）は、ローパスフィルタのインパルス応答、＊は、畳み込みである。
式 (4)～(7) によって予測係数$\gamma_{m,i}(n)$ をｍ＝１からｍ＝Ｉ（分析次数）まで再帰的に計算し、$\gamma_{I,i}(n)$ を得る。

4　実験方法

4.1　学習

1. 入力された、音声信号を瞬時化最大エントロピー法で分析し、励振波スペクトルを推定する。これを、２０次元のベクトル（４６．９～４９２．９Ｈｚ、２３．４７Ｈｚごと）に抽出し、サンプリング周期１０ｍｓｅｃごとに求めることによりベクトル列を得ることができる。

2. 前段階で得られた、励振波スペクトルのベクトル列を視察によって得られた句境界ごとに切り出しそれを５００ｍｓｅｃに線形伸縮する。

3. この句境界ごとに切り出された励振波スペクトルのベクトル列をＬＢＧ法を用いて、クラスタリングし複数の代表パタンを作成し、これをテンプレートとする。

4.2　認識

1. 入力された、連続音声信号を同様に瞬時化最大エントロピー法で分析し、励振波スペクトルのベクトル列を抽出する。

2. この前段階で得たスペクトルのベクトル列を学習により得たテンプレートとＯｎｅ　ｓｔａｇｅ　ＤＰ法により区間全体で最小２乗誤差基準による最適なテンプレート列を求める。

3. その得られたテンプレート列の境目の時刻を調べることにより句境界が、検出することができる。

この実験では、テンプレート数は、１，２，４，８，１６，３２、また、各テンプレート数ごとにデルタピッチ寄与率を０．０，０．５，１．０で変化させて行ない計６＊３＝１８通りで認識を行なった。
また、今回は、 間引きを行なっていない。

ここで、デルタピッチとは、得られた振幅に対する重み付けされた最小２乗誤差直線の傾きであり次式で表わされる。

$$\min_{\Delta p_i, b} \sum_{k=-m/2}^{m/2} r_{i+k} w_k (p_{i+k} - (\Delta\ p_i k + b))^2$$

距離計算の際、パタンベクトル$\hat{P}_j \hat{P}_k$の距離は、次式で表わされ

$$D(\hat{P}_j, \hat{P}_k) = \sum_{i=1}^{L} (1-\alpha)|\hat{p_{ji}} - \hat{p_{ki}}|^2 + \alpha|\Delta\hat{p_{ji}} - \Delta\hat{p_{ki}}|^2$$

このとき、αをデルタピッチ寄与率と呼ぶ。また、デルタピッチ寄与率は、パタン間の距離におけるデルタピッチ距離のしめる割合で０～１の値をとる。
また、Ｏｎｅ　ｓｔａｇｅ　ＤＰ法のアルゴリズムについて次に述べる。
未知入力パタンのフレーム：$i = 1, \ldots, N$
テンプレート番号：$k = 1, \ldots, N$
テンプレートｋのフレーム：$j = 1, \ldots, j(k)$
(i, j, k)におけるフレーム間距離：$d(i, j, k)$
(i, j, l)における累積距離：$D(i, j, k)$
Ｓｔｅｐ１　intitialize　$D(1, j, k) = \sum_{n=1}^{j} d(1, n, k)$
Ｓｔｅｐ２
a)$for i := 2$　to　N　do　$step(b) - (e)$
b)$D(i, 1, k) := d(i, 1, k) + min(D(i-1, 1, k), D(i-1, J(k^*), k^*)$
$(k^* = 1, \ldots, K))$
c)$for j := 2$　to　$J(k)$　do　$step(e)$
e)$D(i, j, k) := d(i, j, k) + min(D(i-1, j, k)$
$D(i-1, j-1, k), D(i-1, j-1, k))$
Ｓｔｅｐ３　$Trace$　$back$　the　$best$　$path$

4.3 システム構成

本実験で用いたシステム構成を図4に表わす。ここで、図の点線枠の部分は、学習時に用いられ、実線枠部分は、認識時に用いられる。

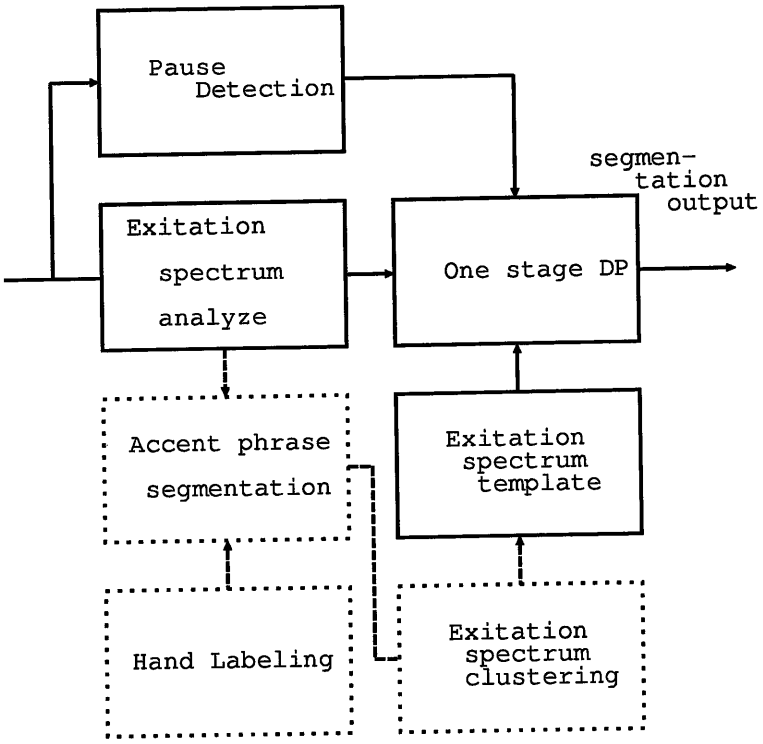

図 4: システム構成

4.4 使用データ

本実験で使用した、データ ベースを下の表に示す。

名称、分類	ＡＴＲ日本語音声データベース
テキスト	１００文章 内分け２グループ（Ａ，Ｂ）各５０文章
話者	男性 １名（ＭＹＩ）
Ａ／Ｄ変換	１２ｋＨｚサンプリング，１６ｂｉｔ

学習用にＡ、認識実験用にＢを用いた。

4.5 結果の評価

句境界の検出の正解の基準については、視察によって得られた、句境界の＋－１００ｍｓｅｃにあるとき、正解とした。また、評価基準として次式を用いた。

$$\text{句境界検出率} = \frac{\text{正解の句境界検出数の総数}}{\text{視察による句境界の総数}}$$

$$\text{誤認識率} = \frac{\text{誤検出した数の総数}}{\text{自動検出による句境界の総数}}$$

5　実験結果

オープン実験においては、テンプレート数８、デルタピッチ１．０のとき、認識率最高で、約７６％であった。

クローズ実験においては、テンプレート数３２、デルタピッチ０．０のとき、認識率最高で、約８０％であった。

オープン実験とクローズ実験の認識結果を比較すると、認識率が、極端なくい違いが認められないことから、ある程度、個人についての句境界の情報を含んでいると考えられる。

しかし、従来法のラグウインドウ法による、特徴パタンと比較すると、非常に認識率が悪かった。また、誤認識をしているところは、無声音の部分や有声音のパワーの弱いところであり、パタンには、きれいにスペクトルのピークが出ていなかった。これは、本来、励振波スペクトルを推定の際には、データを間引きするのだが、本実験では、間引きを行なっていないのが、原因と考えられる。間引きを行った場合と行なわない場合では、行った場合の方が、励振波スペクトルを推定する際行わない場合よりもピークが鋭敏に観察することができており、改善の余地がある。

もう１つ原因として考えられるのは、今回の実験では、パタンの抽出に、４６．９Ｈｚ～４９２．９Ｈｚの２０次元のベクトルであったが、本実験での分析法は、スペクトルのピークが、かなり狭い範囲で観測されるため、サンプリングの間隔の大きさによっては、ピークが抽出できずに直線的なパタンが抽出される可能性がある。しかし、サンプリングの間隔を狭めれば、当然、次元数が増加することになる。

認識率とデルタピッチ寄与率の関係のグラフを図5、図6、図7に表わす。また認識例を図8に示す。

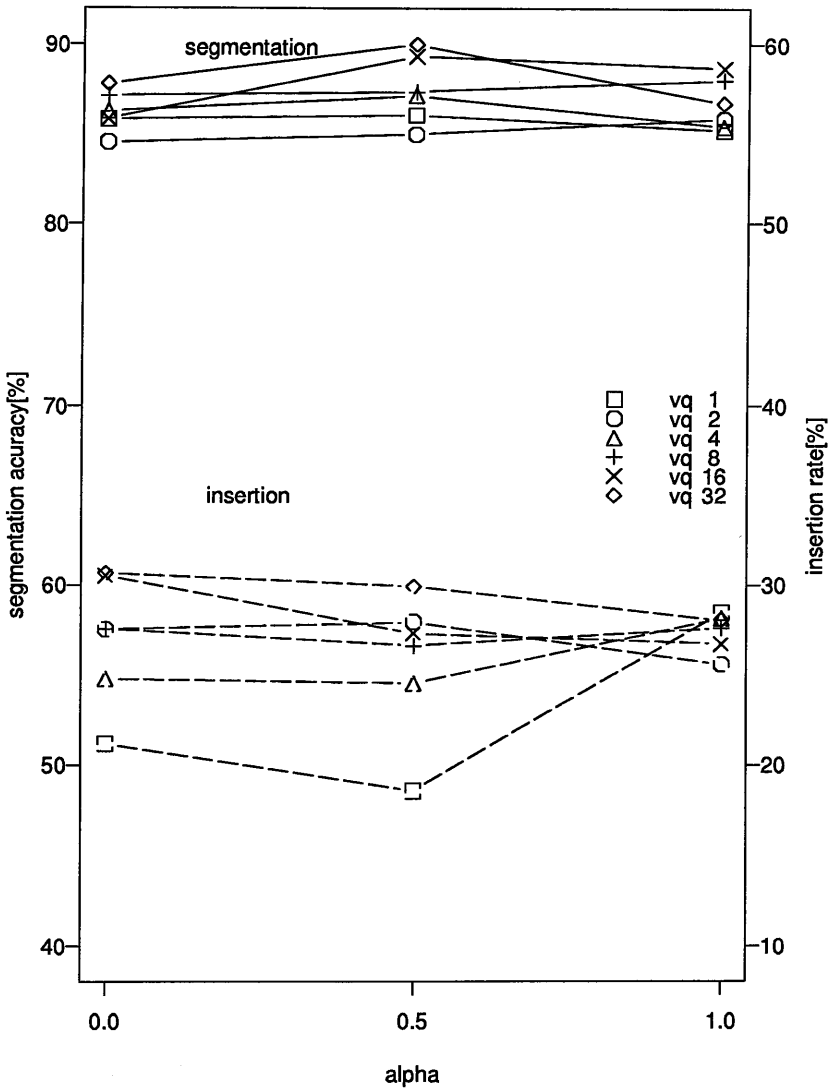

図 5: ＰＳＰパタンの認識率

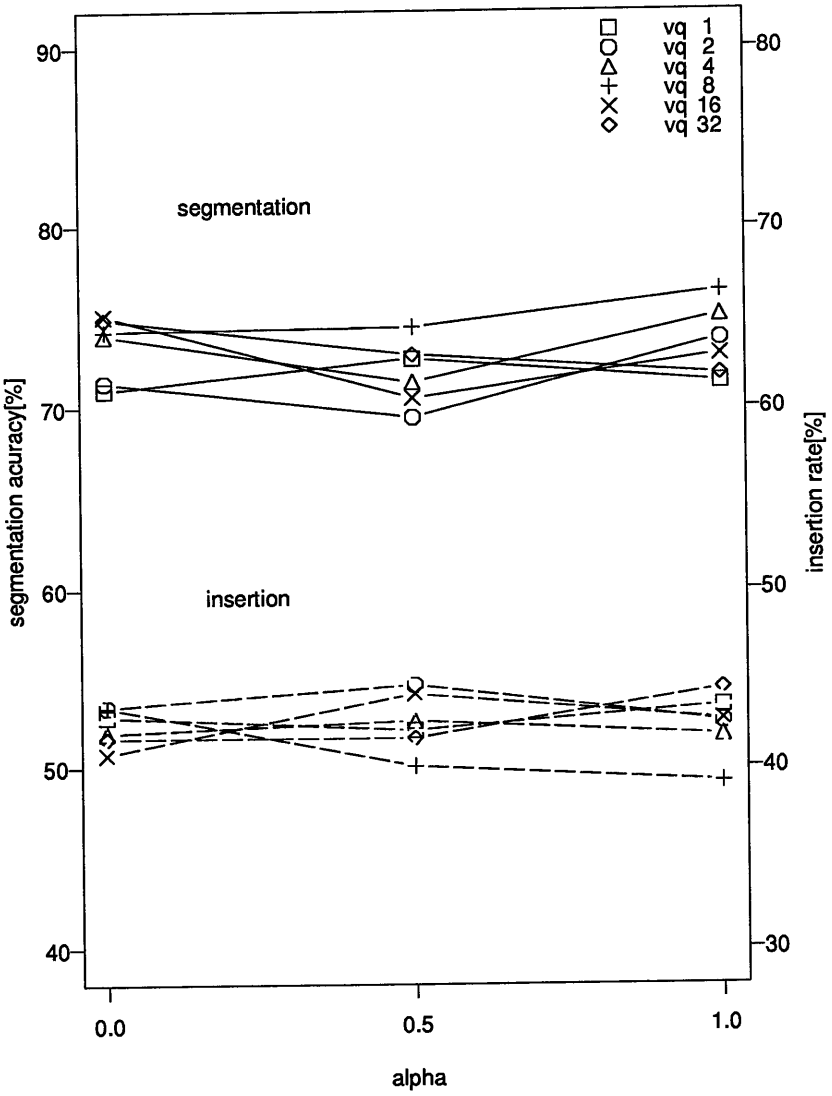

図 6: 励振波スペクトル (オープン実験)

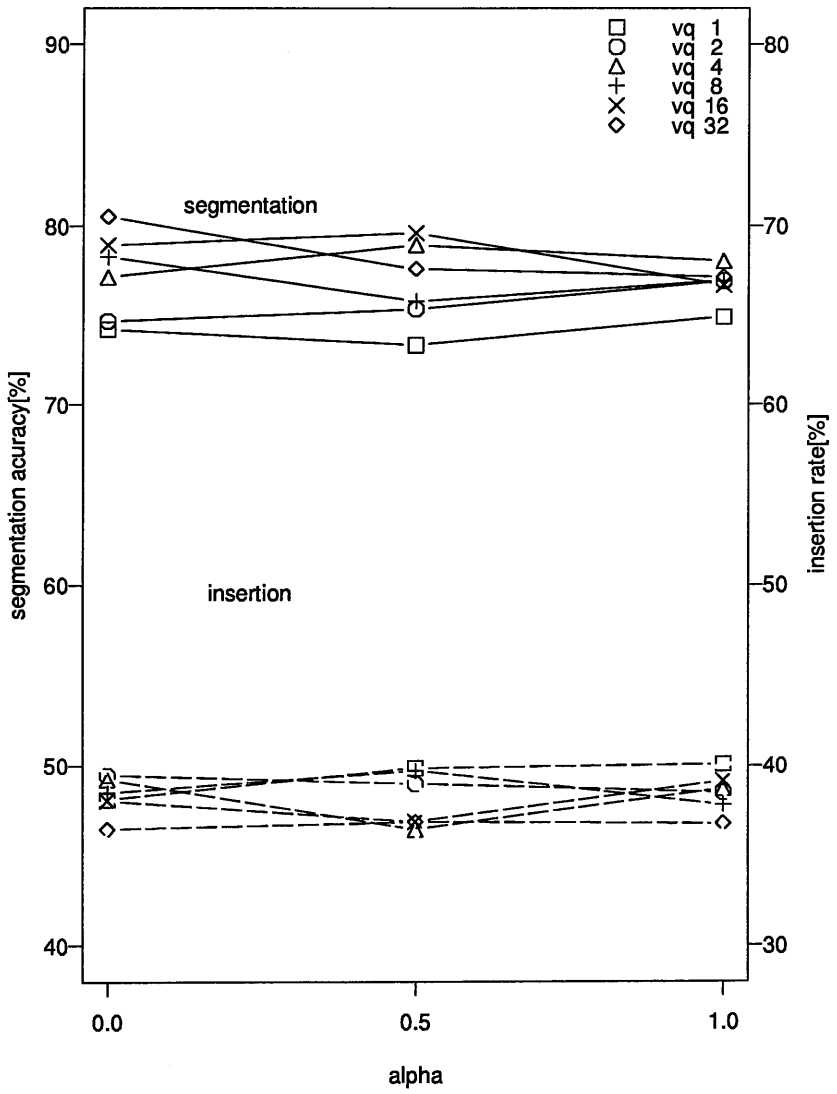

図 7: 励振波スペクトル (クローズ実験)

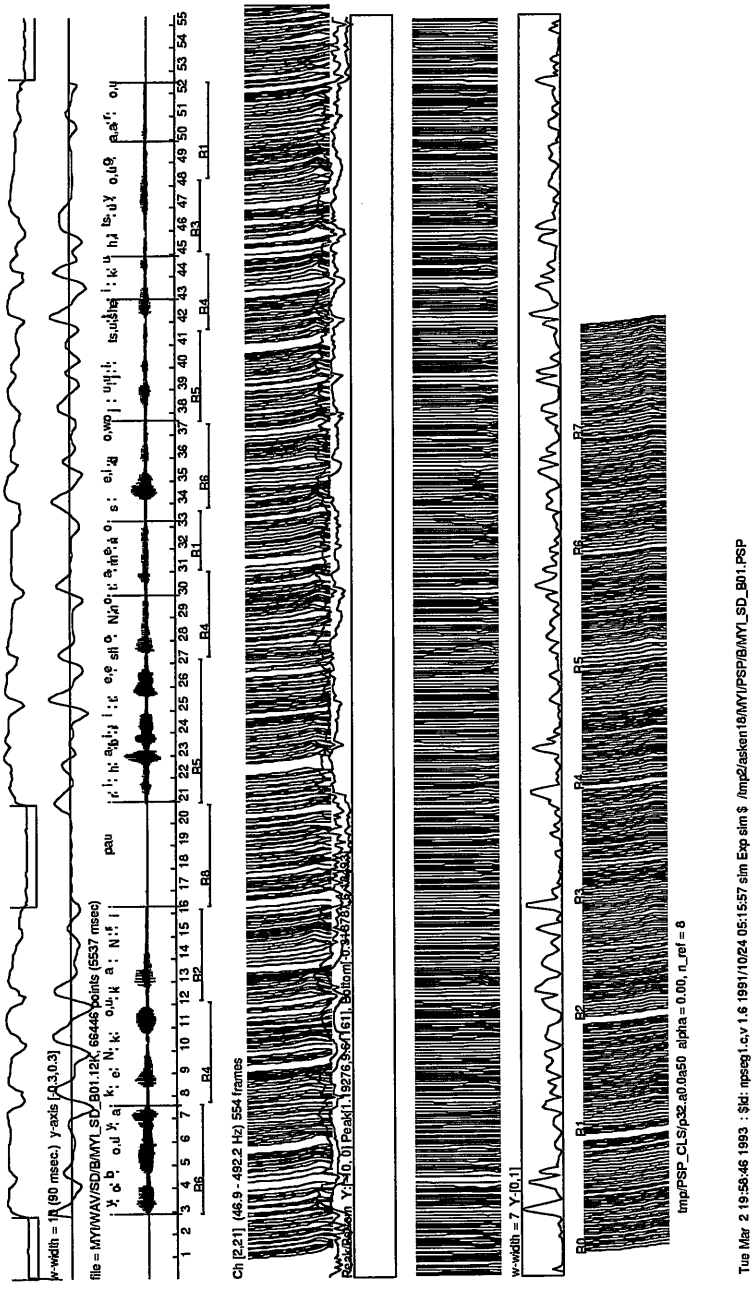
w-width = 10 (80 msec.) y-axis [-3,3.0,3]
file = MYI/WAV/SD/B/MYI_SD_B01.12K 68446 points (5537 msec)
pau
Ch [2,21] (46.9 - 492.2 Hz) 554 frames
w-width = 7 Y-[0,1]
tmp/PSP_CLS/p32.a0.0a50 alpha = 0.00, rl_ref = 8
Tue Mar 2 19:58:46 1993 : $Id: npseg1.c,v 1.6 1991/10/24 05:15:57 sim Exp sim $ /tmp2/asken18/MYI/PSP/B/MYI_SD_B01.PSP

図 8: 認識例

6 結論

6.1 まとめ

従来のラグウインドウ法と異なる非定常過程の分析法 (瞬時化最大エントピー法) で分析し特徴パタンの抽出を試みた。予備的な観察では、フレームごとで、不規則な擾乱の取り除かれた励振波スペクトルが、得られた。

認識実験では、オープン実験とクローズ実験から、ある程度の句境界情報を含む特徴パタンの抽出ができた。

本実験では、従来からある、ラグウインドウ法によって抽出された特徴パタンと比較すると、高精度なパタンの抽出はできなかった。

クラスタリングされた、代表パタンを調べてみると、視覚的に明らかにアクセントパタン (ピーク周波数の時間的変化) が表われているわけではなく、かなりスペクトルのピークが、平滑化されてしまっているため分かりずらいパタンとなってしまっている。これには、本分析法の特性上、スペクトルのピークの変化が、狭い範囲で生じるためであり、このためには、今回の実験で行なった、２０次元では、抽出しきれないことが分かった。

6.2 今後の課題

間引きを行った励振波スペクトル推定による特徴パタンの抽出することにより、スペクトルのピークが、きれいに観察されるため性能の改善が見込まれる。しかし、これは、その後の実験 (間引きを行なったもの) でも、ピークが、あまりでておらず平滑化されてしまっているのでどうも平滑化されてしまう原因は、実験結果で述べた、次元数が少ないことによるものであるようだ。このため、次元数をもっと大きくして実験する必要が、あるようだ。

また、本研究では、定常性に関するスペクトル推定による特徴パタンの作成であったが、この分析法 (瞬時化最大エントロピー法) は、非定常性に関するスペクトル推定も可能であるのでこ、定常性と非定常性の２つのスペクトルをそれぞれ特徴パタンとして抽出しこの２つのパタンを用いて句境界を検出する。

本実験で本方法の有効性を示すことが、出来ていないため、他の方法を考える必要性もあるであろう。

7 謝辞

本研究を行なうにあたり、御指導、御助言を賜わりました東北大学工学部阿曽弘具教授に心より感謝いたします。また、中井満氏、粟津辰功氏には、研究全般にわたり、親身な御指導を賜わりました。
最後に日頃より数多くの御指導、御討論、御協力を賜わりました、丸岡 研 (阿曽グループ) の皆様に感謝します。

8 参考文献

(1) 滝沢 由美、小田 啓介、渡辺 彰彦、深沢 敦司:"多段ＡＲ過程の非定常スペクトル推定", 信学技報ＤＳＰ９０－１, ｐｐ. １－８

(2) 滝沢 由美:"瞬時化最大エントロピー法に基づく非定常過程のスペクトル推定法", 電気情報通信学会論文誌 Ａ Ｖｏｌ. Ｊ７３－Ａ Ｎｏ. ６ ｐｐ. １０８３－１０９３ １９９０. ６

(3) Hiroshi Shimodaira and Masayuki kimura : "Accent Phrase Segmentation Using Pitch Pattern Clustering", IEEE 1992 ICASSP

(4) 下平 博、嵯峨山 茂樹、木村 正行:"ピッチパタン連続整合による連続音声のセグメンテーション", 信学技報ＳＰ９０－７２, ｐｐ. ３３－４０