音声言語情報処理 7−23
（1995. 7. 22）

認識単位の粒度自由・並列アーキテクチャとその実現のためのReference Interval-free 連続DP

岡 隆一、木山 次郎、伊藤 慶明
oka@trc.rwcp.or.jp
新情報処理開発機構 つくば研究センタ
〒305 つくば市竹園 1-6-1 つくば三井ビル 13F

ここでは、人間の高次の認識・思考過程をモデル化を行なうための Grain-free and Parallel architecture を提案した。これは、実時間応答性をもつ柔軟な Man-Machine インタフェースを設計するためのものである。このアーキテクチャを実現するに向いたアルゴリズムとして、既に提案している Reference Interval-free 連続 DP がある。さらに、「概念スポッティング」とよぶもの、またその実現のための「連続オートマトン」とよぶものをここで提案したアーキテクチャの中で位置づけた。これらは、いずれも音声認識の Non-categorical 方式を構成する要素であることを主張した。

Grain-free and Parallel Architecture and Reference Interval-free Continous Dynamic Programming

Ryuichi Oka, Jiro Kiyama, Yoshiaki Itoh
Tsukuba Research Center, Real World Computing Partnership

Grain-free and parallel architecture is proposed for realizing a real-time and flexible man-machine interface system. Under the architecture the system recognizes arbitrary unit of reconition and makes a real-time response at each time. The architecture is implemented by using an algorithm for spotting recognition called Reference Interval-free Continuous Dynamic Programming (RIFCDP) which is applicable to both spontaneous speech and motion image of human gesture. The RIFCDP has the ability of spotting for an endless stream of input by arbitrary parts of a single reference pattern. The real-time system under the architecture also integrates frame-wisely two kinds of spotting recognition result through the computation of an automaton called Continuous Automaton.

1. はじめに

Brooks によって提案された [1] ロボットの Subsumption architecture はロボットの設計に革新を与えたと評価されている。ロボットの実世界における実時間応答性という要求の評価基準で各時点ごとに subsumption 群を順序づけ、これによって各時点の行動を決定するということの有効性が評価されたためである。

さて、いまロボットではなく、知的な思考を行なう人間の脳の活動においても実時間応答性を要求される状況を考えてみよう。例えば、人間が音声やジェスチャで思考内容を計算機に実時間で伝え、また計算機からの実時間の応答によってまたその思考が進展や変化するという状況である。このとき、実時間応答性の要求は第一義的な要素となり、ロボットの実世界における実時間応答性の要求とほぼ同様の意味をもってくる。

この状況で、人間の思考過程に影響を及ぼしている要素群 は、Brooks のいうロボットの行動を定めている要素群ではなく、人間の脳での高次の認知レベルにある要素群を想定しなければならない。しかし、その役割を実現するアーキテクチャについては、Brooks のいう subsumption architecture によるものと類似してくるといえよう。すなわち、新しいアーキテクチャが考えられるとしたら、それは、“計算機にはとにかく実時間でそれなりの出力を要求するが、必ずしも「洗練され、深く認識した認識」に基づく出力は要求しないという状況” の扱いに好都合のものであるといえよう。これは、ロボットに「洗練され、よく計画するされた行動」を求めなかった Brooks の提案と通じるものがある。新しいアーキテクチャは、(ロボットの behaviour based ではなく) 人間の高次の思考過程を中心においた、音声や画像の認識に基づく新しい人間 - 機械系のありかたに関係するといえよう。

ここでは上記の立場にたつアーキテクチャを「粒度自由・並列アーキテクチャ」とよぶことにする。ここでいう粒度とはシステムの認識する単位の大きさを意味する。このアーキテクチャに基づけば、その利用を広げられる音声や画像の認識に基づくインターフェースの構築の可能性がある。

2. 音声とジェスチャの認識に基づくインタフェース

複数人のユーザが音声とジェスチャを交えて議論し、あるタスクを実行している状況を想定する。このとき、計算機には、ユーザの発する音声とジェスチャ動画像を理解し、ユーザのタスク実行の支援を行なうデータの提示をディスプレイや合成音声で行なうことが期待されるとしよう（図 1 参照）。 ここで計算機に要請されることとは、

- 音声とジェスチャ動画像からの実時間意図理解
- 理解に基づく実時間のデータ検索と実時間画像応答

である。複数人で議論がなされる状況では、議論の実時間的な進行に遅れない計算機からの応答があるとき、それはユーザの思考過程に実時間で影響を与え得る形の支援ができることになる。このようなマン・マシン・インタフェースを設計するには、第一に、音声とジェスチャ動画像の frame-wise の認識を実現しなければならないこと、第二には、それらの認識に基づく、データベースへの検索と結果の表現も実時間で行なえること、の 2 つを前提としなければならない。

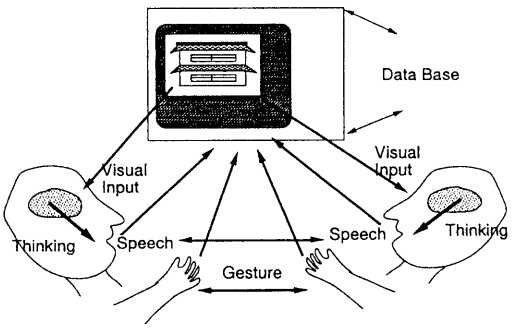

Figure 1: A man-machine communication system operated by speech and gesture expressions by users.

3. 実時間認識のための粒度自由・並列アーキテクチャ

図 1 の人間 - 機械応答系では、人間側により自由な対話を許すためには、音声の発声やジェスチャの動作についての制約を極力課さないようにしなければならない。特に人間側において、考えながら音声やジェスチャで意図を表現するとき特にこの要請は強い。また、そのような状態でこそ計算機の援助を必要とする場合が多い。そのとき、ユーザによる音声とジェスチャの表現について次のようなことが生じていると考えられる。

- 極めて冗長な音声やジェスチャ表現があり、その中で実際に意味のあると思われる部分はわずかであること、

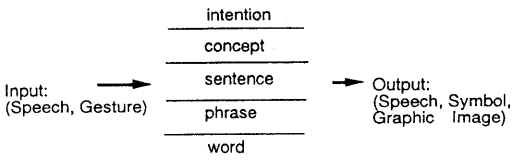

Figure 2: A Grain-free and Parallel architecture for modeling a real-time response system.

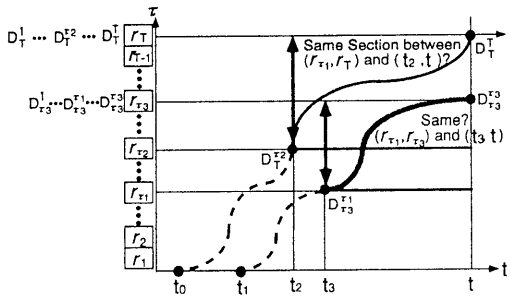

Figure 3: Schema of Reference Interval-free Continuous Dynamic Programming. A reference is often given by a feature sequence of a spoken story.

- その意味のある部分の認識単位は様々でありうること、
- 様々な認識単位の果たす役割はほぼ同等でありうること、

これらのことを考慮すると、われわれは図2に示されるアーキテクチャを採用するのが妥当と考える。

人間が音声やジェスチャで自らの意図を表す時、文章の音声やジェスチャ全体を使って表していても、実際に意味しているものは、その中の一部である、単語、句、文、概念、意図、である場合が多い。これらの単位は認知的な階層では異るところに位置づけられる。また、その分離は極めて困難である。一方、実時間応答を要求されるシステムの出力も、実際に人間が表現したい意味の部分によって作成されることを要求される。図2の意味するのはこのような状況で意味のある実時間応答を作成するために要求されるアーキテクチャである。すなわち、粒度の異る認識単位が並列に動作しうるためのアーキテクチャである。粒度の異る単位を入出力に対して並列に扱うということで、これを「粒度自由・並列アーキテクチャ」とよぶ。

4. Reference Interval-free 連続 DP

図2のようなアーキテクチャに好都合なアルゴリズムとはどのようなものであるかを考える。いま、音声の場合を考える。そこでは、単語、文節、文という単位性がある。また、それらを任意の個数繋ぎ合わせた単位というものも認識単位になりうる。このような任意性のある単位によって、入力音声から Frame - wise のスポッティングによって取り出すことができれば、前節で述べたシステムからの要請に応えるものとなる。我々はすでにこの目的に合致したアルゴリズムを「 Reference Interval- free 連続 DP (RIFCDP) 」[7] として提案している。この方式の概念図が図3に示されている。図3で横軸が入力音声のもつ時間軸、縦軸が標準パターンのもつ時間軸である。RIFCDP はよく知られている「連続 DP」という方式を発展させたものである。「連続 DP」というのは、標準パターンで表された認識単位をスポッティング的に入力音声中から認識する方式である。このとき、標準パターンの時間軸については最適な非線形の伸縮を行なう。スポッティングというのは、予め認識対象区間を切り出すことをしないものであり、認識とセグメンテーションが同時的に行なうものである。この「連続 DP」において、入力の各時刻で標準パターン側の時刻点に対応する点のすべてで、それまでの累積距離の履歴も保持しているとすると、各入力時刻で標準パターンの任意の区間に対する正規化マッチング距離は（保持している累積距離の履歴データの中で）その区間両端に対応しているものの累積距離の差分で計算することができる。これらを比較することによって、各時刻で、標準パターンの任意の区間によって入力音声中からスポッティングを行なうことができる。このように「連続 DP」を拡張したものが「RIFCDP」[7] である。

いま、1つの標準パターンとして、文、または続きの文、例えば物語を朗読したもの、等の音声を分析したものをとるとすれば、RIFCDP は単語、文節、文などの単位性に拘らず標準パターン内の任意の部分区間でもって入力音声中からスポッティング認識することができることになる。これはまさに、図2のアーキテクチャを実現するアルゴリズムとなっている。このときの RIFCDP の出力は（標準パターン内の）区間の時系列とすることができ、必ずしも単語、や文の記号表現とする必要はない。

RIFCDP はジェスチャの観測による動画像にも同じように適用できる。

5. 概念スポッティング

図2に基づくモデルにおいて、実際の実時間応答の出力を構成することを考える。そのために、著者らは「概念スポッティング」というものを提案している [3]。そこでは、システムの出力を「概念」と呼ん

でいる。また、システムへの入力は音声やジェスチャ動画像からのRIFCDPによる出力であるとする。

一般にスポッティングという手法はある単位概念をそれぞれ独立に実時間で抽出するのに向いている。特に、単位概念のもつ時間区間に部分的重なりや包含関係がある場合（これは階層の形成に関係する）にこの方法は問題なく適用できる。連続DPや先に示したRIFCDPはそれらを実現する手法であるが、「概念スポッティング」という考え方も手法は異るが同じ考え方に基づくものである。

5.1 出力単位としての画像ノードと概念スポッティング

図2では、人間の高度の認知レベルにある思考過程を環境とみており、ここからの計算機へ入力とその出力の対である“要素I/O”によって計算機内のモデルが構成されている。いま、以下の議論を分かり易くするために、計算機への入力は、人間の発声する音声と身振りや手振りのジェスチャとし、計算機からの出力はそれだけで一つの概念を表す画像とする。この概念は、人間の思考過程によくカップリングできるものであるとともに、それは計算機の中では一つの出力部分の状態表すものであるとしよう。

ここで、出力の候補群は計算機の中では、画像をノードとし、それらの依存関係を記号つきアークで表した、意味ネットワーク的な表現で表されているとしよう。このネットワークを対象にして、入力の音声からの”認識出力”列やジェスチャ動画像の”認識出力”列から画像表示された概念をスポッティングによってとり出すことを考える。スポッティングとは、入力の各時刻時刻において、（その信頼の程度つきの条件で）計算機から最終結果を出力するアーキテクチャであり、事前に対象とする入力の時間区間の区分することを行なわないものである。このとき、**すべての概念ノードは各時刻で直接に外部へ出力できるアーキテクチャとなっている**とする。また、このスポッティングも実時間計算方式であるが、これを実現するものとしてある並列オートマトン（連続オートマトンと呼ぶ）が提案されている[3]。

5.2 画像としての概念表現

ここでいう「概念」とはそれ自体で意味をもつ「画像」である。形式的には、1つの「概念」は充足度と画像特徴ベクトルの対で表される。充足度とは、その「概念」の完成度を示し、1つの画像特徴ベクトルは図4のように1枚の画像で表せる。

5.3 連続オートマトン

「連続オートマトン」と呼ぶものは、図5に示される意味ネットワークのような相互結合を許すオートマトンで概念スポッティングを実行するアルゴリズムである。オートマトンの特徴は、

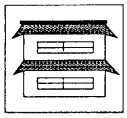

Figure 4: An Image Representing a Cencept. An image is frame-wisely constructed in a node for creating a visible output.

(a) 一つの概念を一つのノードに対応させる、

(b) ノードは充足度と画像特徴ベクトルをもつ、

(c) アークは1つの認識単位であるが、それは標準パターンの1つの区間特徴系列で表されている、

(d) 各時刻ごとに、その時刻のRIFCDP出力である区間特徴系列とすべてのアークの特徴系列との整合度によって、すべての画像ノードの状態を同期して一斉に更新する（図6）、

(e) 充足度はアークのもつ区間特徴系列とRIFCDP出力の区間特徴系列との類似度の累積によって更新される、

(f) ノードの連結性の有無に依存しない状態遷移による概念のスポッティング出力も可能（図6）、

である。

RIFCDPで出力される認識単位は、その粒度において様々なものがある。従って、その出力形式としては、記号化された表現でなく区間特徴系列であるとすることが自然である。RIFCDPの出力（区間特徴系列）を連続オートマトンの入力として、アークの区間特徴系列との距離によって、連続オートマトンの出力である画像を構成する。このようなアーキテクチャでは、形式的には音声の認識単位を記号で表すということが少なくとも明示的には現れない。

さて、Fahlmanは massively parallel architectures を要素の間を流れる信号の型によって3種類の分類をしている[4]。すなわち、(1) Message-passing system (2) Marker-passing system (NETL [5]) (3) Value-passing system (iterative relaxation algorithm [6]) である。「連続オートマトン」はFahlmanの示す上記の従来手法とは異っている。

明らかに「連続オートマトン」はMessage-passing systemではない。「連続オートマトン」は意味ネットワークを用いる点ではNETLと類似するが、NETLのようなMarker-passing systemでなくValue-passing systemの一種である。一方、Value-passing systemの典型である iterative relaxation algorithm が他の

ノードの状態値のみからの伝播によって状態値を更新するのに対し、「連続オートマトン」は各時刻の外部入力記号とすべてのアークに付随する記号との類似度によってドライブされて状態値が更新されるという特徴をもつ。Marker-passing と Value-passing を兼ね備えたものに CMU で開発された THISTLE があるが、これは「連続オートマトン」のようにすべてのノードの状態更新が各時刻の外部入力によってドライブされるものではない。外部入力からすべてのノードへの直接入力を想定できるものに Boltzmann machine があるが、これはシステムすべてのノードの状態を要素とするベクトルでシステムの状態を表す分散表現をとるのに対し、「連続オートマトン」は各ノード、各アークに意味をもたせた局所表現を採用している。

さて、「連続オートマトン」では、図6に示されるように、入力はすべてのアークにある記号表現に直接に接続し、またすべてのノードは出力に直接に接続することを意味している。このことによって、ノードからの直接出力が得られ、あらためて subroutine call を呼ばない。また、文脈依存を実現するためにネットワークにおける各ノードの依存関係とその状態の temporary initialization が行なわれる。

5.4 アルゴリズムの記述

5.4.1 Notations

概念に対応するノードを n 個考え、その i 番目を N_i, ($1 \leq i \leq n$) とする。N_i はその値域が $[0,1]$ である充足値 $S_i(t)$ と、特徴ベクトル $\mathbf{Q}_i(t)$ をもつとする。ノード N_j から ノード N_i へのアークに付随する区間特徴系列を「遷移単位」とよび、これを $w_{ji} \in W$ とする。遷移単位は、(部分) 画像を表すものと、画像の操作を表すものがある。これ以外に「ごみの単位」の $gb \in W$、「空単位」の $\phi \in W$ がある。入力の区間特徴 (標準パターンの任意の区間がなりうる) の列を、$\{u(t) : t = 0,1,2,..\}$ とし、$u(t) \in W$ とする。W は「遷移単位」、「入力の区間特徴」からなる集合である。また、$w_{ji} \neq \phi$ のとき、$D_{ji} = 1$ とし、そうでないとき、$D_{ji} = 0$ とする。さらに、一つの w_{ji} には一つの f_{ji} が対応し、f_{ji} は一つの部分画像を表す。また、f_{ji} を要素とする特徴ベクトルには整合と不整合を定める知識が事前に与えられているものとする。

5.4.2 充足度の更新式

ノード N_i に対応し、値域 $[0,1]$ をとる変数を $S_i(t)$, $(t = 0,1,2,3,...)$, $(1 \leq i \leq n)$, とし、その初期条件を $S_i(0) = 0$ とする。いま、t 時刻の入力 $u(t)$ とすべての W の要素の間には、その値域を $[0,1]$ とする類似度 $s(u(t), w_{ji})$ が定義されているとする。そのとき、

$$d_{ji}(t) = s(u(t), w_{ji}) \tag{1}$$

として、また、係数 $(0 < a < 1)$ として、$S_i(t)$ の更新式を、

$$S_i(t) = \begin{cases} \max_j\{D_{ji} \cdot ((1-a) \cdot S_j(t-1) + \\ a \cdot d_{ji}(t))\} \text{ if } u(t) \neq gb \\ S_i(t-1) \text{ if } u(t) = gb \end{cases} \tag{2}$$

とする。そのとき、出力としての記号番号 $i^*(t)$ は、$S_i(t) < h$ のとき、$i^*(t) = 0$ として、

$$i^*(t) = \arg\{\max_i(S_i(t) \geq h)\} \tag{3}$$

と定める。時刻 t の出力は $i^*(t) \neq 0$ の時、ノードにある特徴ベクトルを画像化したもので表現される。

5.4.3 特徴ベクトルの更新式

ノード N_i のもつ特徴ベクトル $\mathbf{Q}_i(t)$ の更新を考える。この特徴ベクトルは m 個の f_{ji} (部分画像) によって作られるとする。すなわち、$\mathbf{Q}_i(t) = (q_{i,1}(t), q_{i,2}(t), ..., q_{i,m}(t))$ とする。ここで、$q_{i,1}(t)$ は時刻 t に最近の過去の入力に依存する f_{ji} とし、$q_{i,2}(t)$ は $q_{i,1}(t)$ より次に以前の過去入力に依存する f_{ji} とする。以下同様とする。初期条件を、$Q_i(0) = (\phi, \phi, \phi, ..., \phi)$ とする。式 (3) に対応して、

$$j^* = \mathrm{Arg}\{\max_j\{D_{ji} \cdot ((1-a) \cdot S_j(t-1) + a \cdot d_{ji}(t))\}\}$$

として j^* が定まる。明らかに各時刻 t では1つの j^* が定まり、かつそれに対応する1つの i を i^* とする。$Q_i(t)$ の更新を $f_{j^*i^*}$ を用いて行なう。

(i) $i \neq i^*$ のとき、

$$q_{i,k}(t) = q_{i,k}(t-1), \ (1 \leq k \leq m). \tag{4}$$

(ii) $i = i^*$ については、

if $u(t) = gb$ のとき、

$$q_{i^*,k}(t) = q_{i^*,k}(t-1), \ (1 \leq k \leq m). \tag{5}$$

if $u(t) \neq gb$ かつ $f_{j^*i^*}$ が $Q_{j^*}(t-1)$ に整合のとき、

$$\begin{cases} q_{i^*,1}(t) = f_{j^*i^*} \\ q_{i^*,k}(t) = q_{j^*,k-1}(t-1), \ (2 \leq k \leq m). \end{cases} \tag{6}$$

式 (6) の更新データは画像に変換されるが、式 (6) の更新に利用されたノード番号 j^* の 時刻 t のデータのみを初期化、$S_{j^*}(t) = 0$, $Q_{j^*}(t) = (\phi, \phi, ..., \phi)$、する。

if $u(t) \neq gb$ かつ $f_{j^*i^*}$ が $Q_{j^*}(t-1)$ に不整合時、

$$q_{i^*,k}(t) = q_{i^*,k}(t-1), \ (1 \leq k \leq m). \tag{7}$$

上記のアルゴリズムにおいて、あるノードが状態が遷移して画像出力するとき、そのノードの状態を初期化することは、文脈の依存性 によって一時的にエージェントを取り除くことを意味する (図6)。この操作により、次にスポッティングされる概念候補を絞りこむことができるようになる。

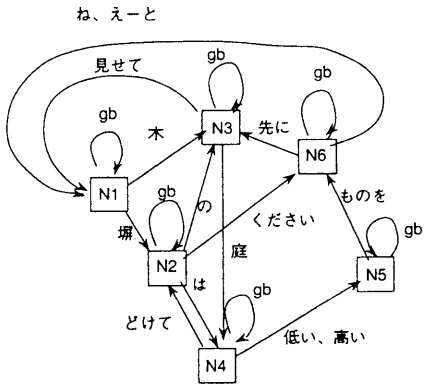

Figure 5: A Continuous Automaton for representing a data-base. Each square corresponds to a figure as a possible output of the system. Each arc symbol is represented by a segment interval of a standard pattern.

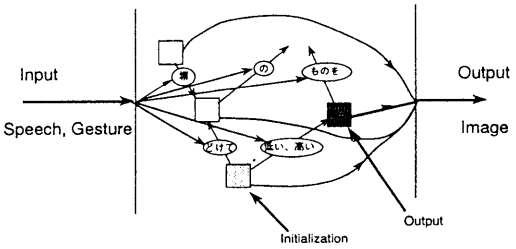

Figure 6: Each input is connected to all arcs of the Continuous Automaton. Each state of square node of the automaton represents an image. Each state of square node is directly connected to the output of the system. Temporary initialization of state implies deletion of a temporary garbage object as the effect of subsuming.

6. 音声や動画像の認識における Non-categorical Paradigm

上記に述べてきたことを整理すると、

- Reference pattern を Spontaneous speech やジェスチャ動画像の分析特徴系列からの作成、
- RIFCDP 適用による Reference pattern の区間出力、
- 画像自体をノードとし、区間特徴系列がアークに付加された意味ネットワークでモデルを作成、
- システムの出力を画像ノードの状態とする、

ということになる。ここにおいて、入力から出力に至る過程において、明示的には認識単位を記号化した表現はどこにも存在しないシステム構成となっている。このことを別表現すると、(音声やジェスチャの) 認識における non-categorical 方式であるということができる。

Non-categorical な方式で音声やジェスチャの認識システムを構築していくことで、マルチモーダルな情報の認識に基づくの Man-machine interface が今後一段と発展するのではないかということを主張したい。なぜなら、認識粒度からの自由、システムの出力作成における記号的知識との整合問題からの自由、モデル学習のし易さ、などの特徴をこの方式がもつからである。認識技術というものが100%の性能をもちえない前提でそれを用いた Man-machine interface を構築するとき、スポッティングによる情報の選択、リアルタイム応答による臨情感による計算機の認識エラーの回復、などの側面での技術の発展が必要である。

7. あとがき

本稿では人間の高次の認識・思考過程をモデル化を行なうための Grain-free and Parallel architecture を提案した。これは、実時間応答性をもつ柔軟な Man-Machine インタフェースを設計する手法にもなる。このアーキテクチャを実現するに向いたアルゴリズムとして、既に提案している Reference Interval-free 連続 DP がある。さらに、「概念スポッティング」とよぶもの、またその実現するための「連続オートマトン」とよぶものをここで提案したアーキテクチャの中で位置づけた。これらは、音声認識の Non - categorical 方式を構成する要素であることを主張した。

謝辞 本稿について様々なコメントを戴きました電総研の麻生 英樹、中島秀之の両氏に深謝致します。

参考文献

[1] R.Brooks:"A Robust Layered Control System for A Mobile Robot," IEEE J. of Robotics and Automaton, Vol.RA-2, No.1, pp.14-23, Mar. 1986.
[2] 木山、伊藤、岡 : "連続構造化法を用いた pan-frame-wise な文理解", 信学会時限専門委員会資料 SPREC-93-1(1993.7).
[3] 岡、伊藤、木山、張:"概念スポッティングのための画像オートマトン", 日本音響学会 H 7 春季講演発表会, 3-4-12 (1995-03).
[4] S.E.Fahlman:"Three Flavors of Parallelism", In Proceedings of the Fourth National Conference of the Canadian Society for Computational Studies of Intelligence, Saskatoon, Saskatchewan, May, 1982.
[5] S.E.Fahlman, G.E.Hinton and T.J.Sejnowski:"Massively Parallel Architectures for AI: NETL, Thistle, and Boltzmann Machines", In Proceedings of the National Conference on Artificial Intelligence AAAI-83, 109-113, Washington DC.
[6] L.S.Davis and A.Rosenfeld:"Cooperating Porocess for Low-level Vision: A survey", Artificial Intelligence, 1981, 3, 245-264.
[7] 伊藤、木山、小島、関、岡:"標準パターンの任意区間によるスポッティングにための Reference Interval-free 連続 DP (RIFCDP)", 信学会、音声研究会、(1995-06).