(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4732030号
(P4732030)
(45)発行日　平成23年7月27日(2011.7.27)　(24)登録日　平成23年4月28日(2011.4.28)

(51)Int.Cl.　F I
G10L 15/08 (2006.01)　G10L 15/08 200Z

請求項の数 3　(全 9 頁)

(21)出願番号　特願2005-192199(P2005-192199)
(22)出願日　平成17年6月30日(2005.6.30)
(65)公開番号　特開2007-11027(P2007-11027A)
(43)公開日　平成19年1月18日(2007.1.18)
審査請求日　平成20年6月24日(2008.6.24)

(73)特許権者　000001007
キヤノン株式会社
東京都大田区下丸子３丁目３０番２号
(74)代理人　100126240
弁理士　阿部　琢磨
(74)代理人　100124442
弁理士　黒岩　創吾
(72)発明者　久保山　英生
東京都大田区下丸子３丁目３０番２号キヤノン株式会社内
(72)発明者　山本　寛樹
東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

審査官　間宮　嘉誉

最終頁に続く

(54)【発明の名称】情報処理装置およびその制御方法

(57)【特許請求の範囲】
【請求項１】
音素モデルの尤度を計算する情報処理装置の制御方法であって、
隣接する音素に依存して決まる複数の音素モデルの全てについて尤度を計算して得られた各音素モデルの尤度のうちの最大値を仮説の音響モデルの尤度とし、更に前記複数の音素モデルのうち尤度が最大値となる音素モデルを記憶手段に記憶したフレームから、現フレームが所定数のフレームを経過しているか否かを判定する判定工程と、
前記判定工程で、現フレームが前記所定数のフレームを経過していると判定された場合、前記複数の音素モデルの全てについて尤度を計算して得られた各音素モデルの尤度のうちの最大値を仮説の音響モデルの尤度とし、更に前記複数の音素モデルのうち尤度が最大値となる音素モデルを前記記憶手段に記憶し、前記判定工程で、現フレームが前記所定数のフレームを経過していないと判定された場合、前記記憶手段に記憶された音素モデルのみについて尤度を計算して得られた尤度を仮説の音響モデルの尤度とする計算工程とを有する制御方法。
【請求項２】
請求項１に記載の制御方法をコンピュータに実行させるためのプログラム。
【請求項３】
音素モデルの尤度を計算する情報処理装置であって、
隣接する音素に依存して決まる複数の音素モデルの全てについて尤度を計算して得られた各音素モデルの尤度のうちの最大値を仮説の音響モデルの尤度とし、更に前記複数の音

素モデルのうち尤度が最大値となる音素モデルを記憶手段に記憶したフレームから、現フレームが所定数のフレームを経過しているか否かを判定する判定手段と、

前記判定手段で、現フレームが前記所定数のフレームを経過していると判定された場合、前記複数の音素モデルの全てについて尤度を計算して得られた各音素モデルの尤度のうちの最大値を仮説の音響モデルの尤度とし、更に前記複数の音素モデルのうち尤度が最大値となる音素モデルを前記記憶手段に記憶し、前記判定手段で、現フレームが前記所定数のフレームを経過していないと判定された場合、前記記憶手段に記憶された音素モデルのみについて尤度を計算して得られた尤度を仮説の音響モデルの尤度とする計算手段とを有する情報処理装置。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、音声を認識する音声認識方法に関する。

【背景技術】

【０００２】

音声認識を行う際に、音素やトライフォンなど、単語より小さいサブワードを用いてモデル化する手法がある。特に、トライフォンのような隣接環境に依存してモデルを分けることで、モデルを詳細に分けるような方法が広く用いられている。例えば、トライフォン「ＳＩＬ－ａ＋ｋ」は、「ａ」という音の中でも直前の音が「ＳＩＬ（無音）」、直後の音が「ｋ」であることを表し、音素「ａ」でモデル化するよりも詳細にモデル化できるため、高い認識率を得ることができる。

【０００３】

しかしながら、トライフォンのような隣接環境に依存するモデルを用いる場合、隣接環境が複数表れる場合（例えば連続単語認識における単語境界）では、その隣接環境の数に応じて仮説を展開しなければならない。図５は、「白」、「黒」、「栗」、「赤」の繰り返し発声を認識することができる認識文法におけるサブワード系列及び仮説における尤度計算を表す図である。同図（ａ）において、５０１はサブワードであり、同図では中心音素と前後の隣接環境の音素に応じて決まるトライフォンを用いている。サブワード５０１は、一般的に同図（ｂ）に示すような１つ以上の状態を持つＨＭＭでモデル化される。５０２はサブワード５０１の一状態に対応する仮説であり、尤度計算では各仮説において尤度Ｓ（ａ，ｂ）を求める。５０３は、仮説を結ぶリンクである。尤度計算には各仮説のＨＭＭ状態における音声入力信号の出力確率や状態間をリンクに従って遷移する遷移確率などによって計算する。ここで上記のような文法では、各単語の単語境界において、サブワード５０１が複数の隣接環境に依存するため、隣接環境の数に応じて仮説を用意しなければならない。すなわち、単語先頭のサブワード（図５において、「＊－ｓｈ＋ｉ」、「＊－ｋ＋ｕ」、「＊－ｋ＋ｕ」、「＊－ａ＋ｋ」）の前環境には、「ＳＩＬ」および単語末尾音素の「ｏ」、「ｉ」、「ａ」を、単語終端のサブワード（同図において、「ｒ－ｏ＋＊」、「ｒ－ｏ＋＊」、「ｒ－ｉ＋＊」、「ｋ－ａ＋＊」）の後環境には、「ＳＩＬ」および単語先頭音素の「ｓｈ」、「ｋ」、「ａ」を考慮してそれぞれサブワード及び仮説を展開する必要がある。これを記述すると図６のように単語境界でサブワード及び仮説が拡がり、このように膨大に増えた仮説に対する尤度計算時間がかかってしまう。

【０００４】

この問題に対して、特許文献１では、単語内の隣接環境にのみ依存させることにより、単語境界の仮説展開を抑制している。図７（ａ）に、単語境界において音素モデルを利用したサブワード系列を、図７（ｂ）に、単語境界において、片方の隣接環境のみ依存するモデルを利用したサブワード系列を示す。このようなモデルを単語境界に利用することで、図６のような仮説展開を抑制することは可能であるが、一方で単語境界においてはその他の場所に比べて詳細ではないモデルを使うことになるので、認識率の低下を招く。そこで特許文献２では単語境界を単語間単語として単語と分けて仮説を生成して接続した方法が提案されているが、単語間単語において仮説が拡がることには変わりなく、また単語間

単語が多くの単語で共有できなければ効果は薄い。また特許文献３では隣接環境依存モデルの内部状態を共有化して木構造で表現した方法が提案されているが、状態でやはり隣接サブワードに依存して拡がりを持ち、十分に抑制するに至ってはいない。

【特許文献１】特開平０５－２２４６９２号公報
【特許文献２】特開平１１－０４５０９７号公報
【特許文献３】特開２００３－２０８１９５号公報

【発明の開示】
【発明が解決しようとする課題】
【０００５】

本発明の目的は、トライフォンなどの隣接環境に依存するサブワードのモデルを用いて音声認識を行う際に、複数の隣接環境に応じて仮説が展開されることを抑制し、音声認識の処理を高速化することである。

【課題を解決するための手段】
【０００６】

上記課題を解決するために、本発明の情報処理方法は、音素モデルの尤度を計算する情報処理装置の制御方法であって、隣接する音素に依存して決まる複数の音素モデルの全てについて尤度を計算して得られた各音素モデルの尤度のうちの最大値を仮説の音響モデルの尤度とし、更に前記複数の音素モデルのうち尤度が最大値となる音素モデルを記憶手段に記憶したフレームから、現フレームが所定数のフレームを経過しているか否かを判定する判定工程と、前記判定工程で、現フレームが前記所定数のフレームを経過していると判定された場合、前記複数の音素モデルの全てについて尤度を計算して得られた各音素モデルの尤度のうちの最大値を仮説の音響モデルの尤度とし、更に前記複数の音素モデルのうち尤度が最大値となる音素モデルを前記記憶手段に記憶し、前記判定工程で、現フレームが前記所定数のフレームを経過していないと判定された場合、前記記憶手段に記憶された音素モデルのみについて尤度を計算して得られた尤度を仮説の音響モデルの尤度とする計算工程とを有する。

【発明の効果】
【０００７】

本発明によれば、隣接環境に依存するサブワードのモデルを用いて音声認識を行う際に、複数の隣接環境に対して仮説を展開せずに各仮説で複数の隣接環境に対応するサブワードの中で最大尤度を求めることにより、仮説数の増大を抑制し、音声認識の処理を高速化することができる。

【発明を実施するための最良の形態】
【０００８】

以下、図面を参照しながら本発明の好適な実施例について説明していく。

【実施例１】
【０００９】

図１に、本実施例における音声認識装置の機能構成を表すブロック図を示す。同図において、１０１は、入力音声信号を分析して音声特徴量を得る音響処理部である。１０２は、サブワードの音響的特徴をＨＭＭなどによってモデル化したサブワードモデルを格納する音響モデルである。１０３は、認識可能な語彙、および文法あるいは接続確率を有する言語モデルである。１０４は、音響処理部１０１が求めた音声特徴量を入力とし、音響モデル、言語モデルを基に、仮説を生成して尤度計算を行う尤度計算部である。１０５は、尤度計算部１０４が行う尤度計算の際に、各仮説において隣接するサブワードに依存して決まる一つ以上のサブワードモデルを参照するサブワードモデル参照部である。

【００１０】

図２に、本発明の尤度計算部１０４における尤度計算の様子を示す。同図において、（ａ）は、隣接環境に依存するサブワードとしてトライフォンを用い、「白」、「黒」、「栗」、「赤」の繰り返し発声を認識することができる認識文法におけるサブワード系列を表す図である。２０１はサブワードであり、同図では中心音素と前後の隣接環境の音素に

応じて決まるトライフォンを用いている。（ｂ）は、単語「赤」の終端サブワード「ｋ－ａ＋＊」を詳細に表した図であり、２０２は、サブワード２０１のモデルの一状態に対応する仮説である。２０３は仮説を結ぶリンクである。

【００１１】

図１、図２を用いて本実施例における尤度計算について説明する。本実施例においては、尤度計算部１０４は、隣接環境の数に関わらず各中心音素について一つのサブワードを持つ。すなわち、図２（ａ）に示すとおり、単語「赤」の終端では、後環境音素「ＳＩＬ」、「ｓｈ」、「ｋ」、「ａ」に応じてそれぞれサブワードおよび仮説を生成するのではなく、「ｋ－ａ＋＊」一つに対応する仮説の系列を生成する。仮説における尤度計算では、後環境音素「ＳＩＬ」、「ｓｈ」、「ｋ」、「ａ」に応じたトライフォン「ｋ－ａ＋ＳＩＬ」、「ｋ－ａ＋ｓｈ」、「ｋ－ａ＋ｋ」、「ｋ－ａ＋ａ」及び仮説の状態番号を基にサブワードモデル参照部１０５がサブワードモデルを参照する。ここでサブワードモデルのリストを仮説ごとに保持していても良いし、ある一つのテーブルまたはハッシュに仮説とサブワードモデルのリストを対応付けて保持しておき、仮説のＩＤをキーとして参照しても良い。このようにして参照したそれぞれのサブワードモデルに対して、尤度計算部１０４が尤度Ｓ（ａ，ｂ）を求め（ａはトライフォン、ｂは状態番号）、図２（ｂ）に示すとおり、その最大尤度をサブワード「ｋ－ａ＋＊」の仮説における音響モデルの尤度とする。これにそれまでに計算された仮説の累積尤度を加えることで、仮説の累積尤度が計算される。

【００１２】

また、本発明は、図２のように認識語彙数だけサブワード系列を並べた尤度計算方法に限らず、図８に示すように先頭からのサブワード系列を共有化して木構造にした場合（同図では、「黒」、「栗」においてサブワード「＊－ｋ＋ｕ」、「ｋ－ｕ＋ｒ」を共有している）にも、単語境界において全く同様の方法で仮説展開を抑制することができる。

【００１３】

このような構成とすることで、隣接環境に依存するサブワードのモデルを用いて音声認識を行う際に、複数の隣接環境に対して仮説を展開せずに各仮説で複数の隣接環境に対応するサブワードの中で最大尤度を求めることにより、仮説数の増大を抑制し、音声認識の処理を高速化することができる。

【実施例２】

【００１４】

上記実施例では、複数の隣接音素環境に対応するサブワードの仮説の拡がりを抑えることができる。本実施例ではさらに、各仮説においてサブワードモデル参照部１０５が参照したサブワードモデルに対して、尤度Ｓ（ａ，ｂ）を求める計算回数を削減する。図３に本実施例における尤度計算のフローチャートを示す。それぞれの仮説において以下の処理を行う。まずステップＳ３０１において、所定の条件であるか否かを判定する（この条件については後述する）。条件を満たさない場合、ステップＳ３０２において、上記実施例と同様にサブワードモデル参照部１０５が参照する全てのサブワードモデルについて尤度を計算し、その最大値を仮説の音響モデルの尤度とする。次にステップＳ３０３において、ステップＳ３０２で得た尤度最大値を与えたサブワードモデルを記憶する。

【００１５】

そしてステップＳ３０５で最終フレームでない場合、ステップＳ３０６により次フレームへループする。一方、ステップＳ３０１において条件を満たす場合、ステップＳ３０４に進み、現フレーム以前にステップＳ３０３で計算したサブワードモデルのみについて尤度を計算し、状態２０２の尤度とする。

【００１６】

ここで所定の条件としては、「ステップＳ３０２、Ｓ３０３を実行したフレームから所定フレームを経過していないこと」あるいは「前フレームと現フレームの入力音声信号（あるいはその音声特徴量）の距離が所定値未満であること」あるいはその両方などが用いられるが、本発明で定める条件はこれらに限るものではない。すなわち、ある仮説に対し

て最大値を与えるサブワードモデルが同じになる可能性が高い、という仮定ができる条件であれば良い。

【００１７】

これにより、ステップＳ３０２の最大値計算を、ステップＳ３０４の記憶してあるサブワードモデルのみの計算に近似し、尤度計算回数を削減することができる。

【実施例３】

【００１８】

上記実施例では連続音声認識の単語境界において単語間接続によるサブワードの仮説展開を抑える例として説明したが、本発明はこれに限るものではない。単語内部の仮説であっても隣接環境が複数存在する仮説において適用可能である。図４は、「白」、「黒」、「栗」、「赤」を孤立単語認識することができる認識文法において、サブワードを共有して木構造を形成した木構造サブワード系列である。同図（ａ）では、従来の木構造生成方法によって先頭から共通するサブワードを共有し、「黒」、「栗」においてサブワード「ＳＩＬ－ｋ＋ｕ」、「ｋ－ｕ＋ｒ」を共有している。ここで本発明を適用すると、同図（ｂ）に示すとおり、サブワード「ｋ－ｕ＋＊」を用意して共有し、このサブワードに対する仮説においては上記実施例のように尤度計算を行うことにより、仮説数を削減することが可能となる。

【実施例４】

【００１９】

上記実施例では隣接環境に依存するサブワードとしてトライフォンを用いた例で説明したが、本発明はこれに限るものではなく、前後いずれかの環境にのみ依存するダイフォンや、その他様々な隣接環境に依存するサブワードについても適用可能である。また上記実施例では無音モデル「ＳＩＬ」については隣接環境に依存しないモデルを用いた例の図になっているが、本発明はこれに限るものではない。「ＳＩＬ」モデルについても同様に隣接環境依存モデルを用いることができ、その際には「ＳＩＬ」モデルについても本発明によって仮説の展開を抑制することができる。

【００２０】

なお、本発明の目的は、前述した実施例の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記憶媒体に格納されたプログラムコードを読み出し実行することによっても達成されることは言うまでもない。

【００２１】

この場合、記憶媒体から読み出されたプログラムコード自体が前述した実施例の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。

【００２２】

プログラムコードを供給するための記憶媒体としては、例えば、フレキシブルディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、磁気テープ、不揮発性のメモリカード、ＲＯＭなどを用いることができる。

【００２３】

また、コンピュータが読出したプログラムコードを実行することにより、前述した実施例の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているＯＳ（オペレーティングシステム）などが実際の処理の一部または全部を行い、その処理によって前述した実施例の機能が実現される場合も含まれることは言うまでもない。

【００２４】

さらに、記憶媒体から読出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行い、その処理によって前述した実施例の

機能が実現される場合も含まれることは言うまでもない。

【図面の簡単な説明】

【００２５】

【図１】実施例係る音声認識装置の機能構成を表すブロック図である。

【図２】実施例において生成されるサブワード系列及びその仮説における尤度計算を表す図である。

【図３】実施例２における尤度計算のフローチャートである。

【図４】実施例３における木構造サブワード系列を表す図である。

【図５】従来方法において生成される仮説展開前のサブワード系列、及び仮説における尤度計算を表す図である。

【図６】従来方法において単語境界を仮説展開して生成されるサブワード系列を表す図である。

【図７】従来方法において単語内の隣接環境にのみ依存するサブワードを使ったサブワード系列を表す図である。

【図８】実施例において生成される木構造サブワード系列及びその仮説における尤度計算を表す図である。

【符号の説明】

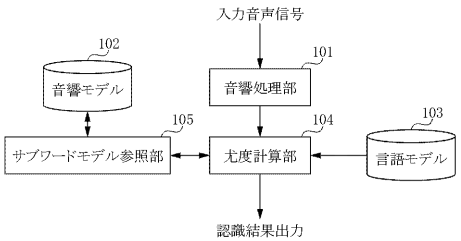
【００２６】

２０１　サブワード

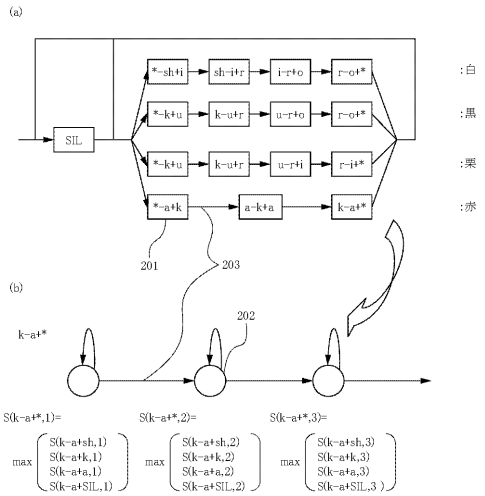
２０２　仮説

２０３　リンク

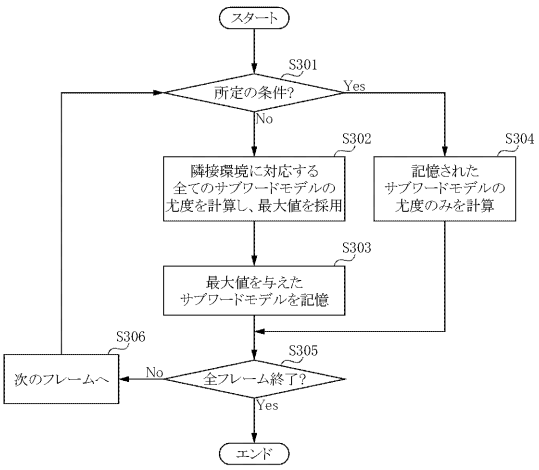
【図１】



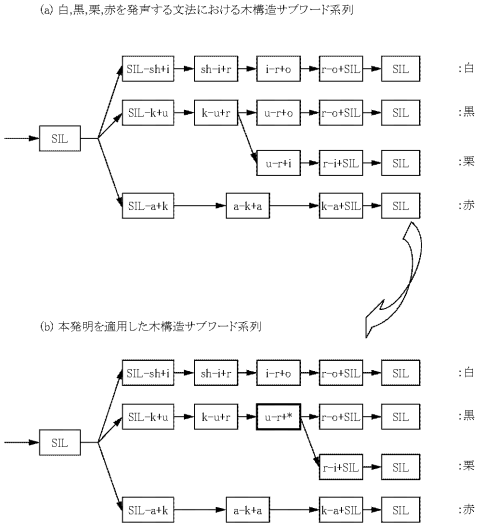
【図２】



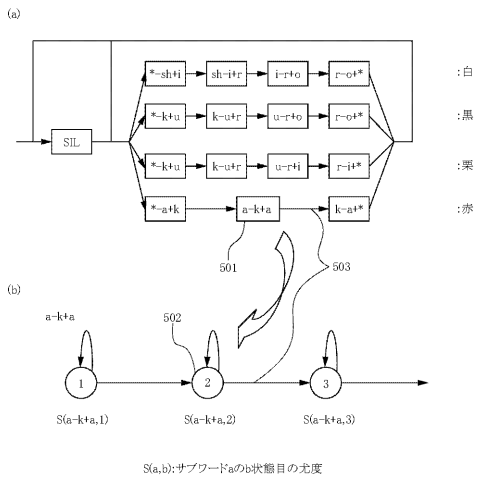
【図３】



【図４】

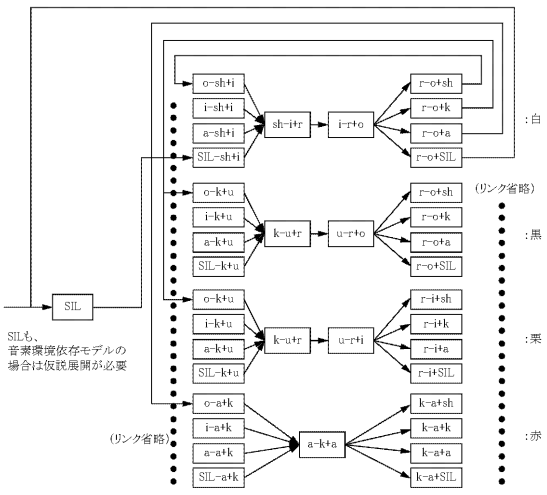


【図５】

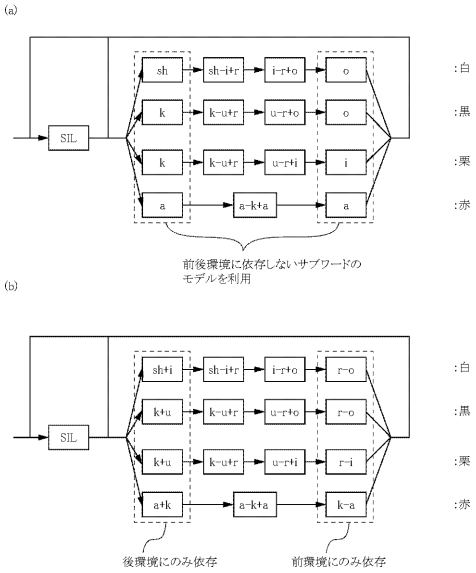


S(a,b):サブワードaのb状態目の尤度

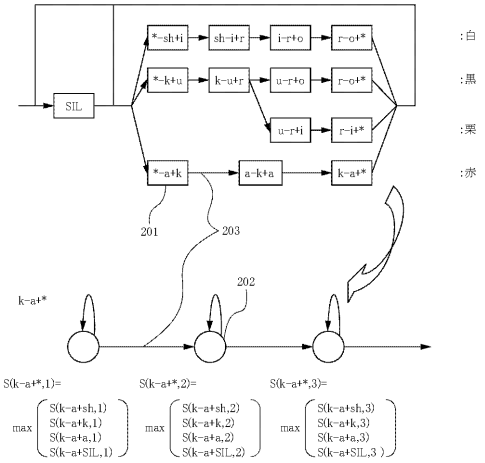
【図６】



【図７】


(a)
SIL
sh
k
k
a
sh-i+r
k-u+r
k-u+r
a-k+a
i-r+o
u-r+o
u-r+i
o
o
i
a
:白
:黒
:栗
:赤
前後環境に依存しないサブワードの
モデルを利用
(b)
SIL
sh+i
k+u
k+u
a+k
sh-i+r
k-u+r
k-u+r
a-k+a
i-r+o
u-r+o
u-r+i
r-o
r-o
r-i
k-a
:白
:黒
:栗
:赤
後環境にのみ依存
前環境にのみ依存

【図８】


SIL
*-sh+i
*-k+u
*-a+k
sh-i+r
k-u+r
a-k+a
i-r+o
u-r+o
u-r+i
r-o+*
r-o+*
r-i+*
k-a+*
:白
:黒
:栗
:赤
201
203
202
k-a+*
S(k-a+*,1)=
max
S(k-a+sh,1)
S(k-a+k,1)
S(k-a+a,1)
S(k-a+SIL,1)
S(k-a+*,2)=
max
S(k-a+sh,2)
S(k-a+k,2)
S(k-a+a,2)
S(k-a+SIL,2)
S(k-a+*,3)=
max
S(k-a+sh,3)
S(k-a+k,3)
S(k-a+a,3)
S(k-a+SIL,3)

フロントページの続き

(56)参考文献 特開平９－１２７９７７（ＪＰ，Ａ）
特開２０００－２５０５８０（ＪＰ，Ａ）
特開２００３－５７８７（ＪＰ，Ａ）
特開２００６－２９３０３３（ＪＰ，Ａ）
特開平９－６８９９６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１７／００
ＩＥＥＥ Ｘｐｌｏｒｅ
ＣｉＮｉｉ
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩ）
ＪＳＴ７５８０（ＪＤｒｅａｍＩＩ）