(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4515054号
(P4515054)

(45)発行日　平成22年7月28日(2010.7.28)　(24)登録日　平成22年5月21日(2010.5.21)

(51)Int.Cl.		F I		
G10L 15/02	(2006.01)	G10L	15/02	300F
G10L 15/10	(2006.01)	G10L	15/10	300G
G10L 15/18	(2006.01)	G10L	15/18	200E
		G10L	15/18	300H

請求項の数 8　(全 26 頁)

(21)出願番号	特願2003-278640 (P2003-278640)
(22)出願日	平成15年7月23日 (2003.7.23)
(65)公開番号	特開2004-54298 (P2004-54298A)
(43)公開日	平成16年2月19日 (2004.2.19)
審査請求日	平成18年7月19日 (2006.7.19)
(31)優先権主張番号	60/398,166
(32)優先日	平成14年7月23日 (2002.7.23)
(33)優先権主張国	米国(US)
(31)優先権主張番号	60/405,971
(32)優先日	平成14年8月26日 (2002.8.26)
(33)優先権主張国	米国(US)
(31)優先権主張番号	10/267,522
(32)優先日	平成14年10月9日 (2002.10.9)
(33)優先権主張国	米国(US)

(73)特許権者　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100077481
弁理士　谷　義一
(74)代理人　100088915
弁理士　阿部　和夫
(72)発明者　デン　リ
アメリカ合衆国　９８０７４　ワシントン州　サマミッシュ　ノースイースト　３０　ストリート　２２３１０

最終頁に続く

(54)【発明の名称】音声認識の方法および音声信号を復号化する方法

(57)【特許請求の範囲】

【請求項１】

有限状態システムの現状態についてのスコアを生成することによって音声信号を復号化する方法であって、

先行する状態の終わりの最適な生成関係値に基づいて、前記現状態についての生成関係値を求めるステップであって、前記最適な生成関係値は、現状態についてのスコアを最大にする１組の連続的値内の生成関係値である、ステップと、

前記生成関係値を使用して音の尤度を判定するステップであって、該尤度は、前記先行する状態と前記現状態との間のパスに位置合せされる１組の観測ベクトルによって表されるステップと、

前記音の尤度を前記先行する状態からのスコアと組み合わせ、前記現状態についてのスコアを判定するステップであって、前記先行する状態からの前記スコアは軌跡の離散的クラスに関連し、該軌跡の離散的クラスにおいて可能な軌跡の連続値はクラスタ化され、該クラスは最適な生成関係値のクラスと合致する、ステップと

を備えたことを特徴とする音声信号を復号化する方法。

【請求項２】

複数のクラス中の生成関係値の各クラスについて別々のスコアを生成するのに使用されることを特徴とする請求項１に記載の音声信号を復号化する方法。

【請求項３】

前記生成関係値は、前記現状態での現在時刻と前記先行する状態の終わりとの間の時間

の長さにさらに基づくことを特徴とする請求項１又は２に記載の音声信号を復号化する方法。
【請求項４】
前記生成関係値は、時間依存性補間重みを使用して計算されることを特徴とする請求項１から３のいずれかに記載の音声信号を復号化する方法。
【請求項５】
前記時間依存性補間重みは、音に関連する時定数にさらに依存することを特徴とする請求項４に記載の音声信号を復号化する方法。
【請求項６】
前記生成関係値を使用して音の尤度を判定するステップは、前記生成関係値に、音に依存しない値を掛けるステップを含むことを特徴とする請求項５に記載の音声信号を復号化する方法。
【請求項７】
前記生成関係値を使用して音の尤度を判定するステップは、音がノイズであるとき、前記生成関係値に０を掛けるステップを含むことを特徴とする請求項５に記載の音声信号を復号化する方法。
【請求項８】
音声信号(u_j)の一部を表す観測値を受け取るステップと、
先行する音韻単位の終わりに対応する時間の軌跡を初期値(g_0^j)とし、仮説音韻単位の生成関係ターゲット(T^{uj})との間の線形補間である生成関係ダイナミックス値を使用して、前記仮説音韻単位についての予測値
【数１】

$$g_k^j = \beta^{u_j}(k) \cdot g_0^j + (1 - \beta^{u_j}(k)) \cdot T^{u_j}$$

を識別するステップであって、前記線形補間は時間依存性補間重み
【数２】

$$\beta^{u_j}(k) = (1 + \gamma^{u_j} k) \cdot e^{-\gamma^{u_j} k}$$

を使用し、ここで、γ^{uj}は前記音声信号(u_j)に関連する時定数である、ステップと、
前記予測値(g(t))を特徴ベクトル(o(t))にマッピンクする音依存性マッピング関数(h^{uj})を用いて、
【数３】

$$z(t) = g(t) + w(t)$$

$$o(t) = h^{u_j}(z(t)) + v(t)$$

と表し、ここで、w(t)，v(t)は前記観測値の期待値からの変位を表し、前記仮説音韻単位についての尤度
【数４】

$$p(o(t)|z(t), u)$$

を判定するステップと

を備えたことを特徴とする音声認識の方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明はパターン認識に関連した音声認識の方法および音声信号を復号化する方法に関する。

【背景技術】

【０００２】

音声認識システムなどのパターン認識システムは、入力信号を取得し、信号を復号化して、信号によって表されるパターンを見つけようと試みる。例えば音声認識システムでは、音声信号（しばしばテスト信号と呼ばれる）が認識システムによって受領され、復号化されて、音声信号によって表される単語の文字列が識別される。

【０００３】

多くの音声認識システムでは、接続された状態の単一の列によって音声単位が表される隠れマルコフモデル（hidden Markov model）が使用される。トレーニング信号を使用して、状態を占有する確率分布と、状態間を遷移する確率分布が音声単位ごとに求められる。音声信号を復号化するために、信号がフレーム単位に分割され、各フレームは特徴ベクトルに変換される。次いで特徴ベクトルは各状態についての分布と比較され、フレームによって表すことのできる最も可能性の高い一続きのＨＭＭ（hidden Markov model）状態が識別される。次いでそのシーケンスに対応する音声単位が選択される。

【０００４】

ＨＭＭベースの認識システムは、比較的単純な音声認識作業では良好に動作するが、音声のいくつかの重要な動的性質を直接にはモデル化しない（かつ、会話的な音声などの難しい作業についての動作は不十分であることが知られている）。その結果、ＨＭＭベースの認識システムは、トレーニングに使用される音声信号と、復号化する音声信号との動的な分節（articulation）の差を吸収することができない。例えば、くだけた発話設定では、話者は、自分の音声をハイポアーティキュレート（ｈｙｐｏ－ａｒｔｉｃｕｌａｔｅ）またはアンダーアーティキュレート（ｕｎｄｅｒ　ａｒｔｉｃｕｌａｔｅ）する傾向がある。これは、ユーザの音声分節の軌跡が、次のターゲットに再び向けられる前に、意図されるターゲットに達しない可能性があることを意味する。トレーニング信号は一般に、ハイポアーティキュレートされた音声よりも完全に分節された音声素材を話者が提供する、音声の「リーディング」スタイルを用いて形成されるので、ハイポアーティキュレートされた音声は、トレーニングされたＨＭＭ状態とは合致しない。その結果、レコグナイザ（recognizer）が提供する認識結果は、くだけた音声に対しては理想的な認識結果からほど遠い。

【０００５】

同様の問題がハイパーアーティキュレートされた音声でも生じる。ハイパーアーティキュレートされた音声では、話者は、自分の音声の相異なる音を区別することができるように余分な労力を払う。この余分な労力には、似た音の音声単位とより良好に区別できるようにある音声単位の音を変更すること、ある音声単位の音をより長く保つこと、または各音が近隣の音と区別して知覚されるように、各音の間を急速に遷移させることが含まれる可能性がある。これらの各機構は、ＨＭＭシステムを使用して音声を認識することを難しくする。各技法により得られる、音声信号に対する１組の特徴ベクトルは、トレーニングデータ中に存在する特徴ベクトルとは十分に合致しないからである。

【０００６】

ＨＭＭシステムはまた、発話する速度の変化に対処する上でも難点がある。したがって、誰かがトレーニング信号より遅く、または速く発話した場合、ＨＭＭシステムは、音声信号の復号化でより多くの誤りをする傾向がある。

【０００７】

ＨＭＭシステムの代替方法が提案されている。具体的には、音声信号の生成関係パラメ

ータ（ｐｒｏｄｕｃｔｉｏｎ－ｒｅｌａｔｅｄ　ｐａｒａｍｅｔｅｒ）の軌跡（trajectory）または挙動を直接モデル化することが提案されている。しかし、諸提案のどれも、音声の動的性質を完全にはモデル化していない。具体的には、各モデルは、話者が音声単位に対する所望のターゲットに近づくときに生じる、時間に依存する軌跡の変化に対処していない。加えて、軌跡についての連続的値に基づく確率決定を可能にし、同時に管理可能な軌跡状態の数に検索スペースを制限する復号化手段を各モデルは提供していない。

【０００８】

このことに照らして、音声軌跡の動的性質がより良好にモデル化され、復号化を管理可能とするように、他のモデル変数によって音声の生成関係挙動を明示的にモデル化する音声認識フレームワークが必要である。

【０００９】

いくつかの文献に上述のような従来の技術に関連した技術内容が開示されている（例えば、非特許文献１参照）。

【００１０】

【非特許文献１】L.Deng著「A Dynamic, Feature-Based Approach to the Interface Between Phonology and Phonetics for Speech Modeling and Recognition」 Speech Communication、 Vol. 24, No. 4, １９９８年、pp.299-323

【発明の開示】

【発明が解決しようとする課題】

【００１１】

従来のシステムには上述したような種々の問題があり、さらなる改善が望まれている。

【００１２】

本発明は、このような状況に鑑みてなされたもので、その目的とするところは、容易に理想的な認識結果を得ることができる音声認識の方法および音声信号を復号化する方法を提供することにある。

【課題を解決するための手段】

【００１３】

時間依存性補間重みを使用して、以前の時刻での生成関係ダイナミックス値とターゲットとの間の線形補間を実施することにより分節ダイナミックス値を識別する音声認識の方法が提供される。次いで生成関係ダイナミックス値を使用して予測音響特徴値を形成し、その予測音響特徴値を、観測した音響特徴値と比較して、観測した音響特徴値が所与の音韻単位によって生成された尤度を求める。

【００１４】

ある実施形態では、以前の時刻での生成関係ダイナミックス値は、１組の連続的値から選択される。加えて、音韻単位の尤度が、以前の時刻での生成関係ダイナミックス値の離散的クラスに関連するスコアと組み合わされ、現状態についてのスコアが求められる。

【発明の効果】

【００１５】

以上説明したように本発明によれば、容易に理想的な認識結果を得ることができる。

【発明を実施するための最良の形態】

【００１６】

以下、図面を参照して本発明の実施形態を詳細に説明する。

【００１７】

図１に、本発明を実施することのできる適切なコンピューティングシステム環境１００の一例を示す。コンピューティングシステム環境１００は、適切なコンピューティング環境の一例に過ぎず、本発明の使用法または機能の範囲に関して何らかの制限を示唆するものではない。例示的動作環境１００に図示する構成要素のうちのいずれか１つ、あるいはそれらの組合せに関係する何らかの依存関係または要件をコンピューティング環境１００が有するものと解釈すべきでもない。

【００１８】

本発明は、他の多数の汎用／専用コンピューティングシステム環境／構成で動作可能である。本発明と共に使用するのに適した周知のコンピューティングシステム、環境、および／または構成の例には、限定はしないが、パーソナルコンピュータ、サーバコンピュータ、ハンドヘルド／ラップトップ装置、マルチプロセッサシステム、マイクロプロセッサベースのシステム、セットトップボックス、プログラマブル家庭用電化製品、ネットワークＰＣ、ミニコンピュータ、メインフレームコンピュータ、電話システム、ならびに上記のシステムまたは装置のいずれかを含む分散コンピューティング環境などが含まれる。

【００１９】

本発明は、コンピュータが実行中の、プログラムモジュールなどのコンピュータ実行可能命令の一般的状況で説明することができる。一般に、プログラムモジュールは、特定のタスクを実行し、または特定の抽象データ型を実装するルーチン、プログラム、オブジェクト、コンポーネント、データ構造などを含む。本発明はまた、通信ネットワークを介してリンクされるリモートプロセッサによってタスクが実行される分散コンピューティング環境でも実施することができる。分散コンピューティング環境では、プログラムモジュールは、メモリ記憶装置を含む、ローカルコンピュータ記憶媒体とリモートコンピュータ記憶媒体のどちらにも位置することができる。

【００２０】

図１を参照すると、本発明を実施する例示的システムは、コンピュータ１１０の形態の汎用コンピューティング装置を含む。コンピュータ１１０の構成要素は、限定はしないが、プロセッサ１２０と、システムメモリ１３０と、システムメモリを含む様々なシステム構成要素をプロセッサ１２０に結合するシステムバス１２１とを含むことができる。システムバス１２１は、メモリバスまたはメモリコントローラと、周辺バスと、様々なバスアーキテクチャのうちのいずれかを用いるローカルバスとを含むいくつかのタイプのバス構造のうちのいずれでもよい。例えば、限定はしないが、このようなアーキテクチャには、ＩＳＡ（Ｉｎｄｕｓｔｒｙ Ｓｔａｎｄａｒｄ Ａｒｃｈｉｔｅｃｔｕｒｅ）バス、ＭＣＡ（Ｍｉｃｒｏ Ｃｈａｎｎｅｌ Ａｒｃｈｉｔｅｃｔｕｒｅ）バス、ＥＩＳＡ（Ｅｎｈａｎｃｅｄ ＩＳＡ）バス、ＶＥＳＡ（Ｖｉｄｅｏ Ｅｌｅｃｔｒｏｎｉｃｓ Ｓｔａｎｄａｒｄｓ Ａｓｓｏｃｉａｔｉｏｎ）ローカルバス、およびメザニンバスとも呼ばれるＰＣＩ（Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ）バスが含まれる。

【００２１】

コンピュータ１１０は、一般に様々なコンピュータ可読媒体を含む。コンピュータ可読媒体は、コンピュータ１１０がアクセス可能である入手可能などんな媒体でもよく、それには揮発性媒体と不揮発性媒体、リムーバブル媒体とノンリムーバブル媒体のどちらも含まれる。例えば、限定はしないが、コンピュータ可読媒体は、コンピュータ記憶媒体および通信媒体を含むことができる。コンピュータ記憶媒体には、コンピュータ可読命令、データ構造、プログラムモジュール、または他のデータなどの情報を格納するための何らかの方法または技術で実装される、揮発性媒体と不揮発性媒体、リムーバブル媒体とノンリムーバブル媒体のどちらも含まれる。コンピュータ記憶媒体には、限定はしないが、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ（electrically erasable programmable read-only memory）、フラッシュメモリ、または他のメモリ技術、ＣＤ（compact disk）－ＲＯＭ、デジタルバーサタイルディスク（ＤＶＤ）、または他の光記憶装置、磁気カセット、磁気テープ、磁気ディスク記憶装置、または他の磁気記憶装置、あるいは、所望の情報を格納するのに使用することができ、コンピュータ１１０でアクセスすることができる他のどんな媒体も含まれる。通信媒体は一般に、コンピュータ可読命令、データ構造、プログラムモジュール、または他のデータを、搬送波または他の移送機構などの被変調データ信号でに実装し、その通信媒体にはどんな情報送達媒体も含まれる。「被変調データ信号」という用語は、信号の特性集合のうちの１つまたは複数を有する信号、または情報を符号化するように変化する信号を意味する。例えば、限定はしないが、通信媒体には、有線ネットワークまたは直接有線接続などの有線媒体、ならびに音響媒体、ＲＦ（radio frequencies）媒体

、赤外線媒体、および他の無線媒体などの無線媒体が含まれる。上記のいずれの組合せも、コンピュータ可読媒体の範囲内に含まれるべきである。

【００２２】

システムメモリ１３０は、読取り専用メモリ（ＲＯＭ）１３１およびランダムアクセスメモリ（ＲＡＭ）１３２などの揮発性メモリおよび／または不揮発性メモリの形態のコンピュータ記憶媒体を含む。起動中などにコンピュータ１１０内の要素間で情報を転送する助けになる基本ルーチンを含むＢＩＯＳ（Basic Input/Output System）１３３が、一般にＲＯＭ　１３１内に格納される。一般に、プロセッサ１２０にとって即座にアクセス可能であり、かつ／またはプロセッサ１２０が現在操作しているデータおよび／またはプログラムモジュールをＲＡＭ　１３２は含む。例えば、限定はしないが、図１にオペレーティングシステム１３４、アプリケーションプログラム１３５、他のプログラムモジュール１３６、およびプログラムデータ１３７を示す。

【００２３】

コンピュータ１１０はまた、他の取外し可能／取外し不能な、揮発性／不揮発性コンピュータ記憶媒体も含むことができる。単なる一例であるが、図１に、ノンリムーバブル不揮発性磁気媒体を読み書きするハードディスクドライブ１４１と、リムーバブル不揮発性磁気ディスク１５２を読み書きする磁気ディスクドライブ１５１と、ＣＤ　ＲＯＭまたは他の光媒体などのリムーバブル不揮発性光ディスク１５６を読み書きする光ディスクドライブ１５５とを示す。例示的動作環境で使用することのできる他の取外し可能／取外し不能な揮発性／不揮発性コンピュータ記憶媒体には、限定はしないが、磁気テープカセット、フラッシュメモリカード、デジタルバーサタイルディスク、デジタルビデオテープ、ソリッドステートＲＡＭ、およびソリッドステートＲＯＭなどが含まれる。ハードディスクドライブ１４１は一般に、インタフェース１４０などのノンリムーバブルメモリインタフェースを介してシステムバス１２１に接続され、磁気ディスクドライブ１５１および光ディスクドライブ１５５は一般に、インタフェース１５０などのリムーバブルメモリインタフェースによってシステムバス１２１に接続される。

【００２４】

上記で解説し、図１に図示するドライブとその関連するコンピュータ記憶媒体は、コンピュータ１１０に対してコンピュータ可読命令、データ構造、プログラムモジュール、および他のデータの記憶を実現する。例えば図１では、ハードディスクドライブ１４１がオペレーティングシステム１４４、アプリケーションプログラム１４５、他のプログラムモジュール１４６、およびプログラムデータ１４７を格納するものとして図示している。これらの構成要素は、オペレーティングシステム１３４、アプリケーションプログラム１３５、他のプログラムモジュール１３６、およびプログラムデータ１３７と同じであっても、異なっていてもよいことに留意されたい。オペレーティングシステム１４４、アプリケーションプログラム１４５、他のプログラムモジュール１４６、およびプログラムデータ１４７には、少なくともこれらが相異なるコピーであることを示すために異なる符号を付けてある。

【００２５】

ユーザは、キーボード１６２と、マイクロフォン１６３と、マウス、トラックボール、またはタッチパッドなどのポインティングデバイス１６１などの入力装置を介して、コマンドおよび情報をコンピュータ１１０に入力することができる。他の入力装置（図示せず）には、ジョイスティック、ゲームパッド、サテライトディッシュ、スキャナなどを含めることができる。これらの入力装置や他の入力装置はしばしば、システムバスに結合されるユーザ入力インタフェース１６０を介してプロセッサ１２０に接続されるが、パラレルポート、ゲームポート、またはユニバーサルシリアルバス（ＵＳＢ）などの他のインタフェースおよびバス構造によって接続することもできる。モニタ１９１または他のタイプのディスプレイ装置もまた、ビデオインタフェース１９０などのインタフェースを介してシステムバス１２１に接続される。モニタに加えて、コンピュータは、スピーカ１９７やプリンタ１９６などの他の周辺出力装置も含むことができ、その周辺出力装置は、出力周辺

インタフェース１９０を介して接続することができる。

【００２６】

コンピュータ１１０は、リモートコンピュータ１８０などの１つまたは複数のリモートコンピュータへの論理接続を使用して、ネットワーク環境で動作することができる。リモートコンピュータ１８０は、パーソナルコンピュータ、ハンドヘルド装置、サーバ、ルータ、ネットワークＰＣ、ピア装置、または他の共通ネットワークノードでよく、一般に、コンピュータ１１０に関して上記で述べた要素のうちの多数またはすべてを含む。図１に示す論理接続は、ローカルエリアネットワーク（ＬＡＮ）１７１および広域ネットワーク（ＷＡＮ）１７３を含むが、他のネットワークも含むことができる。このようなネットワーキング環境は、オフィス、企業全体のコンピュータネットワーク、イントラネット、およびインターネットで一般的なものである。

【００２７】

ＬＡＮネットワーキング環境で使用する際、コンピュータ１１０は、ネットワークインタフェース／アダプタ１７０を介してＬＡＮ　１７１に接続される。ＷＡＮネットワーキング環境で使用する際、コンピュータ１１０は一般に、インターネットなどのＷＡＮ　１７３を介して通信を確立するためのモデム１７２または他の手段を含む。モデム１７２は内蔵でも外付けでもよく、ユーザ入力インタフェース１６０、または他の適切な機構を介してシステムバス１２１に接続することができる。ネットワーク環境では、コンピュータ１００に関して示したプログラムモジュールまたはその一部を、リモートメモリ記憶装置内に格納することができる。例えば、限定はしないが、図１に、リモートアプリケーションプログラム１８５がリモートコンピュータ１８０上に常駐するものとして示す。図示するネットワーク接続は例示的なものであって、コンピュータ間の通信リンクを確立する他の手段も使用できることを理解されたい。

【００２８】

図２は、例示的コンピューティング環境であるモバイル装置２００のブロック図である。モバイル装置２００は、マイクロプロセッサ２０２、メモリ２０４、入出力（Ｉ／Ｏ）構成要素２０６、ならびにリモートコンピュータまたは他のモバイル装置と通信する通信インタフェース２０８を含む。一実施形態では、上記の構成要素は、適切なバス２１０を介して互いに通信するように結合される。

【００２９】

メモリ２０４は、モバイル装置２００への総電力が遮断されたときにメモリ２０４中に格納された情報が失われないようにバッテリバックアップモジュール（図示せず）を備えるランダムアクセスメモリ（ＲＡＭ）などの不揮発性電子メモリとして実装される。メモリ２０４の一部は、プログラム実行用のアドレス指定可能メモリとして割り振ることが好ましく、メモリ２０４の別の部分は、ディスクドライブ上の記憶をシミュレートするためなどの記憶用に使用することが好ましい。

【００３０】

メモリ２０４は、オペレーティングシステム２１２、アプリケーションプログラム２１４、およびオブジェクトストア２１６を含む。動作中、オペレーティングシステム２１２をプロセッサ２０２によってメモリ２０４から実行することが好ましい。ある好適実施形態でのオペレーティングシステム２１２は、Ｍｉｃｒｏｓｏｆｔ　Ｃｏｒｐｏｒａｔｉｏｎから市販されているＷＩＮＤＯＷＳ（登録商標）ＣＥブランドのオペレーティングシステムである。好ましくは、オペレーティングシステム２１２は、モバイル装置用に設計され、１組の公開されたアプリケーションプログラミングインタフェース／メソッドを介してアプリケーション２１４が使用することのできるデータベース機能を実装する。オブジェクトストア２１６内のオブジェクトは、アプリケーション２１４およびオペレーティングシステム２１２によって維持され、公開されたアプリケーションプログラミングインタフェース／メソッドに対する呼出しに少なくとも部分的に応答する。

【００３１】

通信インタフェース２０８は、モバイル装置２００が情報を送受信することを可能にす

る多数の装置および技術を表す。この装置には、ほんの少数の例を挙げれば、有線／無線モデム、サテライト受信機、および放送チューナが含まれる。モバイル装置２００はまた、コンピュータと直接接続して、それらの間でデータを交換することもできる。そのような場合、通信インタフェース２０８は、赤外線トランシーバまたはシリアル／パラレル通信接続でよい。これらのすべては、ストリーミング情報を伝送することができる。

【００３２】

入出力構成要素２０６は、タッチセンシティブ画面、ボタン、ローラ、およびマイクロフォンなどの様々な入力装置と、オーディオジェネレータ、振動装置、およびディスプレイを含む様々な出力装置とを含む。上記で列挙した装置は例であり、かつモバイル装置２００上にすべて存在する必要はない。加えて、他の入出力装置を、本発明の範囲内のモバイル装置２００に取り付けることができ、または本発明の範囲内のモバイル装置２００に関して見つけることができる。

【００３３】

本発明は、音声の生成モデルを提供する。このモデルの下では、音声は、一続きの音韻単位の言語的定義を音声的に実施する話者による試行として表される。この試行の間、話者は、前の音韻単位の終わりに関連する値から、現在の音韻単位に関連するターゲットまでの軌跡をたどる生成関係値を生成する。この軌跡は、前の音韻単位の終わりの値と、ターゲット値の間の補間としてモデル化され、この補間は、２つの補間点間の重み付けが音韻単位の長さにわたって変化する時間依存性重みを含む。

【００３４】

本発明のモデルは隠れ軌跡モデル（Ｈｉｄｄｅｎ　Ｔｒａｊｅｃｔｏｒｙ　Ｍｏｄｅｌ）と呼ばれる。隠れ軌跡モデルは、ダイナミックスがリカーシブ（recursive）形式によってではなく明示的形式の時間依存性によって規定されるという点で、隠れ動的モデル（Ｈｉｄｄｅｎ－Ｄｙｎａｍｉｃ　Ｍｏｄｅｌ）の特別な形である。この隠れ軌跡モデルは、隠れ生成関係パラメータ（ボーカルトラック共鳴など）を説明する動的／軌跡モデル構成要素と、生成関係パラメータを、メル周波数ケプストラル係数などの観測可能な音響特徴に変換するマッピングモデル構成要素の２つの層を含む。動的モデルは、対応する境界（$\tau_1 \ldots \tau_{N+1}$）と共に音シーケンス（$u_1, \ldots, u_j, \ldots, u_N$）が与えられた場合、生成関係パラメータについての一続きの軌跡値（ｚ（１），．．．，ｚ（ｔ），．．．，ｚ（Ｔ））を予測する。マッピングモデルは、その一続きの軌跡値、音シーケンス、および音境界が与えられた場合、一続きの音響観測ベクトルを予測する。

【００３５】

どちらの層でも、統計モデルが与えられる。

【００３６】

【数１】

$$z(t) = g(t) + w(t) \qquad \text{式１}$$

$$o(t) = h^{u_j}(z(t)) + v(t) \qquad \text{式２}$$

【００３７】

上式で、ｇ（ｔ）は予測される軌跡であり、

【００３８】

【数２】

h^{u_j}

【００３９】

は生成関係パラメータを特徴空間にマッピングする音依存性マッピング関数である。

【００４０】

加数ｗ（ｔ）およびｖ（ｔ）は、平均値０を有し、実際の観測値の、期待値からの変位をモデル化する共分散行列$Q = C_{ww}$および$R = C_{vv}$をそれぞれ有するｉ．ｉ．ｄ．ガウスノイズを表す。すなわち、

【００４１】
【数３】

$$p(z(t)|g(t)) = N(z(t); g(t), Q) \qquad \text{式３}$$

$$p(o(t)|z(t), u) = N(o(t); h^{u_j}(z(t)), R) \qquad \text{式４}$$

【００４２】
本発明の下では、言語中の各音について、別々の１組の軌跡パラメータがトレーニングされる。加えて、音内の軌跡が、音の始まりでの軌跡の値と、音の開始から経過した時間とに基づいて決定される。したがって、ローカル時間ｋ＝ｔ－τ$_{j}$に対して、音ｕ$_{j}$内の軌跡を説明することができる。ただし、τ$_{j}$は音ｕ$_{j}$が始まる時間であり、各音は持続時間Ｋ^{j}＝τ$_{j+1}$－τ$_{j}$を有する。
【００４３】
本発明の下では、任意のローカル時間インデックスｋでの、音ｕ$_{j}$についての軌跡は以下のように定義される。
【００４４】
【数４】

$$g_k^j = \beta^{u_j}(k) \cdot g_0^j + (1 - \beta^{u_j}(k)) \cdot T^{u_j} \qquad \text{式５}$$

$$\beta^{u_j}(k) = (1 + \gamma^{u_j} k) \cdot e^{-\gamma^{u_j} k} \qquad \text{式６}$$

【００４５】
上式で、
【００４６】
【数５】

$$T^{u_j}$$

【００４７】
は、音ｕ$_{j}$の軌跡に対するターゲットであり、
【００４８】
【数６】

$$\gamma^{u_j}$$

【００４９】
は、音ｕ$_{j}$に関連する時定数であり、
【００５０】
【数７】

$$g_0^j$$

【００５１】
は、その音に進入したときの軌跡の初期値である。式６の右辺の括弧で閉じられた項は、臨界ダンピング関数であることに留意されたい。本発明が、
【００５２】
【数８】

$$g_0^j = g_{K^{j-1}}^{j-1} \qquad \text{式７}$$

【００５３】
を保証することによって音の間の連続性を実現することにも留意されたい。上式で、
【００５４】
【数９】

$$g_0^j$$

【００５５】
は音ｕ$_{j}$の始めでの軌跡の値であり、
【００５６】
【数１０】
$g_{K^{j-1}}^{j-1}$

【００５７】
は音ｕ$_{j-1}$の終わりでの軌跡の値である。
【００５８】
式５および７は２つの重要な性質を有する。第１に、時刻ｋ＝０では、軌跡は、先行する音についての軌跡の終わりの値と等しい。第２に、音が長期間にわたって持続する場合、軌跡はターゲット
【００５９】
【数１１】
T^{u_j}

【００６０】
に達する。したがって、式５で計算される軌跡は、音
【００６１】
【数１２】
g_0^j

【００６２】
の始まりでの生成関係ダイナミックス値と、生成関係ターゲット
【００６３】
【数１３】
T^{u_j}

【００６４】
との間の、時間依存性補間重み
【００６５】
【数１４】
$\beta^{u_j}(k)$

【００６６】
を用いた線形補間である。
【００６７】
本発明の下では、生成関係軌跡を音響特徴にマッピングするのに使用する音依存性マッピング関数
【００６８】
【数１５】
h^{u_j}

【００６９】
は、以下のように定義される区分的線形関数である。
【００７０】
【数１６】
$$h^{u_j}(z(t)) = H_m Z(t) + h_m \qquad \text{式８}$$

【００７１】

上式でｍは、フレームまたは音全体に対して一定である混合インデックスである。一実施形態では、H_mおよびh_mは音に依存しない。別の実施形態では、H_mおよびh_mは、音、音クラス、または左から右に配置されたＨＭＭ状態に対応するサブフォン（sub-phone）ユニットに依存するようにされる。

【００７２】

本発明の一態様の下では、静寂およびノイズに対して予測されるベクトルが、$H_m = 0$と仮定することによって形成される。その結果、静寂およびノイズに対して予測される特徴ベクトルは、生成関係値の軌跡に依存しない。これは、静寂およびノイズが音声生成での中断を表す生成モデルに適合する。

【００７３】

このマッピング関数を使用すると、式２および４は以下のようになる。

【００７４】

【数１７】

$$o(t) = H_m z(t) + h_m + v(t) \qquad \text{式９}$$

$$p(o(t)|z(t),u) = N(o(t); H_m z(t) + h_m, R_m) \qquad \text{式１０}$$

【００７５】

モデルパラメータ

【００７６】

【数１８】

T^{u_j}、γ^{u_j}

【００７７】

、H_m、h_m、Ｑ、R_mは、予測最大化トレーニングアルゴリズムを使用してトレーニングされる。このアルゴリズムはＥステップを含む。Ｅステップでは、１組のトレーニング観測ベクトルがモデルパラメータの初期推定と共に使用され、十分な統計が発生し、混合重み、軌跡、および軌跡の２乗を含むある隠れ変数の値が予測される。具体的には、Ｅステップは以下の計算を含む。

【００７８】

【数１９】

$$\omega_m^n = P(m|o^n) = \frac{P(o^n|m)P(m)}{\sum_{m'=1}^{M} P(o^n|m')P(m')} \qquad \text{式１１}$$

$$p(o^n|m) = \prod_{k=1}^{K^n} p(o_k^n|m) = \prod_{k=1}^{K^n} N(o_k^n; H_m g_k^n + h_m, S_m) \qquad \text{式１２}$$

$$E\left[\left(z_k^n|o^n, m\right)\right] = \left[H_m^{TRANS} R_m^{-1} H_m + Q^{-1}\right]^{-1}\left[H_m^{TRANS} R_m^{-1}(o_k^n - h_m) + Q^{-1} g_k^n\right] \qquad \text{式１３}$$

$$E\left[z_k^n z_k^{n\,trans}|o^n, m\right] = E\left[z_k^n z_k^{n\,trans}|o_k^n, m\right]$$

$$= \left[H_m^{TRANS} R_m^{-1} H_m + Q^{-1}\right]^{-1} + E\left[\left(z_k^n|o^n, m\right)\right] E\left[\left(z_k^n|o^n, m\right)\right]^{Trans} \qquad \text{式１４}$$

【００７９】

上式で、

【００８０】

【数２０】

$$S_m = H_m Q H_m^{trans} + R_m \qquad \text{式１５}$$

【００８１】

【数２１】

ω_m^n

【００８２】
は、音の境界によって定義される、トークンｎについての混合重みであり、
【００８３】
【数２２】

o_k^n

【００８４】
は、トークンｎについてのｋ番目に観測されるトレーニングベクトルであり、
【００８５】
【数２３】

g_k^n

【００８６】
は、ローカル時間点ｋでのトークンｎ中の予測される軌跡の値であり、
【００８７】
【数２４】

z_k^n

【００８８】
は、ローカル時間点ｋでのトークンｎ中の実際の軌跡の値であり、
ｍはトークンｎと共に使用する混合成分であり、
Ｍは、トークンｎに関連する混合成分の数であり、
各混合成分の確率Ｐ（ｍ）は一様で、１／Ｍに等しく、
「ｔｒａｎｓ」は行列の転置を表し、
【００８９】
【数２５】

$E[x|y]$

【００９０】
は、ｙが与えられた場合のｘの期待値を表す。
【００９１】
上記はＥＭアルゴリズム中のＥステップを完了する。Ｍステップの最初の反復を実施するために、モデルパラメータの初期推定を与えなければならない。一実施形態の下では、
【００９２】
【数２６】

T^{u_j}

【００９３】
および
【００９４】
【数２７】

γ^{u_j}

【００９５】
についての初期推定は、Ｋｌａｔｔ音声合成器と、あるスペクトル分析結果とを組み合わせた情報を用いて選択される。次いで、周知の位置合せ用の既存のＨＭＭシステム／技法を用いて、１組のトレーニング観測ベクトルについて位置合せ境界（$\tau_1 \ldots \tau_{N+1}$

）が決定される。ターゲット

【００９６】

【数２８】

$$T^{u_j}$$

【００９７】

、時定数

【００９８】

【数２９】

$$\gamma^{u_j}$$

【００９９】

、および位置合せ境界（$\tau_1 \ldots \tau_{N+1}$）を使用して、１組の軌跡ｇ（ｔ）が上記の式５を使用して推定される。当初、各軌跡を求める際にノイズが０であり、かつ観測ベクトルを混合に対してランダムに割り当てると仮定すると、H_mおよびh_mは、観測される特徴ベクトルｏ（ｔ）と計算される特徴ベクトルとの間の平方誤差の和を最小にするように推定される。ただし誤差は以下のように計算される。

【０１００】

$$v(t) = o(t) - (H_m g(t) + h_m) \qquad \text{式１６}$$

各混合についてｇ（ｔ）、H_m、およびh_mを決定した後、各混合について共分散行列R_mを以下のように推定することができる。

【０１０１】

【数３０】

$$\begin{aligned} R_m &= E\{vv^{trans}\} \\ &= \frac{1}{T}\sum_t v(t)\cdot v(t)^{trans} \\ &= \frac{1}{T}\sum_t (o(t)-(H_m g(t)+h_m))\cdot(o(t)-(H_m g(t)+h_m))^{trans} \end{aligned} \qquad \text{式１７}$$

【０１０２】

Ｑの推定は、観測ノイズｖ（ｔ）が最小となるように、軌跡ｇ（ｔ）の決定の際にノイズｗ（ｔ）をまず推定することによって決定される。これにより、以下の式が得られる。

【０１０３】

【数３１】

$$w(t) = (H_m^{trans} H_m)^{-1} H_m^{trans} \cdot (o(t) - (H_m g(t) + h_m)) \qquad \text{式１８}$$

【０１０４】

次いでＱを以下のように初期化する。

【０１０５】

【数３２】

$$\begin{aligned} Q &= E\{ww^{trans}\} \\ &= \frac{1}{T}\sum_t w(t)\cdot w(t)^{trans} \end{aligned} \qquad \text{式１９}$$

【０１０６】

一実施形態の下では、R_mおよびＱは対角行列であると仮定され、したがって行列の対角成分だけが計算される。

【０１０７】

初期モデルパラメータを上記のように得た後、それらをＥステップの結果と共に使用して、Ｍステップでのモデルパラメータを推定する。具体的には、モデルパラメータは以下

のように計算される。

【０１０８】

【数３３】

$$\hat{T} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K^n}\sum_{m=1}^{M^n}\omega_m^n\left(E\left[z_k^n \middle| o^n, m\right]\right) - g_o^n(k)}{\sum_{n=1}^{N} K^n \cdot \sum_{m=1}^{M}\omega_m^n} \qquad \text{式２０}$$

【０１０９】

上式で、

【０１１０】

【数３４】

$$\hat{T}$$

【０１１１】

はＴ、

【０１１２】

【数３５】

$$g_0^n(k) = (g_{K^{j-1}} - T)(1+\gamma k)\exp(-\gamma k)$$

【０１１３】

に対する更新であり、

およびＴは、以前に推定したターゲットである。

【０１１４】

【数３６】

$$\hat{H}_m = (A - CB)\cdot(I - DB)^{-1} \qquad \text{式２１}$$

【０１１５】

ただし、

【０１１６】

【数３７】

$$A = \left\{\sum_{n=1}^{N}\omega_m^n\sum_{k=1}^{K^n}\left(o_k^n \cdot E\left\{z_k^n\right\}^{trans}\right)\right\}\cdot X \qquad \text{式２２}$$

$$B = \left\{\sum_{n=1}^{N}\omega_m^n\sum_{k=1}^{K^n}E\left\{z_k^n\right\}^{trans}\right\}\cdot X \qquad \text{式２３}$$

$$C = \left\{\sum_{n=1}^{N}\omega_m^n\sum_{k=1}^{K^n}\left(o_k^n\right)\right\}\cdot Y \qquad \text{式２４}$$

$$D = \left\{\sum_{n=1}^{N}\omega_m^n\sum_{k=1}^{K^n}E\left\{z_k^n\right\}\right\}\cdot Y \qquad \text{式２５}$$

$$X = \left\{\sum_{n=1}^{N}\omega_m^n\sum_{k=1}^{K^n}E\left\{z_k^n z_k^{n\,trans}\right\}\right\}^{-1} \qquad \text{式２６}$$

$$Y = \frac{1}{\sum_{n=1}^{N}K^n\omega_m^n} \qquad \text{式２７}$$

【０１１７】

であり、Ｉは恒等行列である。

【０１１８】
【数３８】

$$\hat{h}_m = \frac{\sum_{n=1}^{N}\left[\omega_m^n \sum_{k=1}^{K^n}\left(o_k^n - \hat{H}_m E\left[z_k^n \middle| o^n, m\right]\right)\right]}{\sum_{n=1}^{N} K^n \omega_m^n} \qquad \text{式２８}$$

$$\hat{R}_m = \frac{\sum_{n=1}^{N}\left[\omega_m^n \sum_{k=1}^{K^n}\left(o_k^n - \hat{H}_m E\left[z_k^n \middle| o^n, m\right] - \hat{h}_m\right)\left(o_k^n - \hat{H}_m E\left[z_k^n \middle| o^n, m\right] - \hat{h}_m\right)^{trans}\right]}{\sum_{n=1}^{N} K^n \omega_m^n} \qquad \text{式２９}$$

$$\hat{Q} = \frac{\sum_{n=1}^{N}\left[\sum_{k=1}^{K^n}\left(E\left[z_k^n \middle| o^n\right] - g_k^n\right)\left(E\left[z_k^n \middle| o^n\right] - g_k^n\right)^{trans}\right]}{\sum_{n=1}^{N} K^n} \qquad \text{式３０}$$

【０１１９】
かつ
【０１２０】
【数３９】

$\hat{\gamma}$

【０１２１】
は勾配降下アルゴリズムを使用して求められる。ただし
【０１２２】
【数４０】

$\hat{\gamma}$

【０１２３】
は、その変化が反復の間のしきい値よりも小さくなるまで漸進的に更新される。具体的には、
【０１２４】
【数４１】

$\hat{\gamma}$

【０１２５】
は
【０１２６】
【数４２】

$$\hat{\gamma}^{r+1} = \hat{\gamma}^r + \varepsilon(-2)Q^{-1}\sum_{n=1}^{N}\sum_{k=1}^{K^n}\left(z_k^n - g^n\right)\cdot\left(\hat{\gamma}^r \cdot k^2 \cdot \left(g_0^n - \hat{T}\right)\cdot e^{-ik}\right) \qquad \text{式３１}$$

【０１２７】
を使用して更新される。
【０１２８】
ＥステップおよびＭステップは、最終の１組のモデルパラメータに達するように何回か反復することができる。その最終の１組のパラメータを求めた後、それを使用して、観測した１組の音響ベクトルを復号化することができる。
【０１２９】
復号化作業は、一続きの音響観測を生成した可能性の最も高い音のシーケンスを見つけるものである。本発明の一実施形態の下では、２つの状態を接続するエッジによって各音が表される有限状態変換器（transducer）を使用して復号化が実施される。ただし、時間内の状態の位置は、復号化中に決定される。したがって、復号化プロセスは、エッジのシ

ーケンス（Ｅ＝（ｅ$_1$，．．．，ｅ$_j$，．．．，ｅ$_n$））と、音響観測値のシーケンス（Ｏ＝（ｏ$_1$，．．．，ｏ$_j$，．．．，ｏ$_L$））を生成した可能性が最も高い状態位置とを見つけるものである。

【０１３０】

【数４３】

$$\hat{E} = \arg\max_{E} P(E|O) \qquad \text{式３２}$$

【０１３１】

上式は、ベイズの公式を使用し、分母を無視すると、以下のようになる。

【０１３２】

【数４４】

$$\hat{E} = \arg\max_{E} P(O|E)P(E) \qquad \text{式３３}$$

【０１３３】

式３３の確率は以下のように求められる。

【０１３４】

【数４５】

$$\arg\max_{E} P(O|E)P(E) = \max_{n \in \{n_{term}\}} \max_{g} \{H_{c(g)}(L,n)\} \qquad \text{式３４}$$

【０１３５】

上式で、

ｎ$_{term}$は、すべての終端状態の集合であり、

Ｌは、最後の観測の時間インデックスであり、

Ｈ$_{c(g)}$（Ｌ，ｎ）は、軌跡ｃ（ｇ）のクラス内の軌跡ｇを有する、時刻Ｌでのノードｎへの最高のスコアリングパスである。ただし、最高のスコアリングパスは以下のように求められる。

【０１３６】

【数４６】

$$H_{c(g)}(T,n) = \max_{e:n_{e2}=n} \max_{\Delta t} \{Q_{c(g),\Delta t}(t,e)\} \qquad \text{式３５}$$

【０１３７】

上式で、

ｅ：ｎ$_{e2}$＝ｎは、状態ｎで終了するエッジの集合であり、

Δｔはｔ－ｔ’に等しい。ただしｔ’は、現在の音またはエッジが始まった時間インデックスであり、ｔは状態ｎについての時間インデックスであり、したがってΔｔは、現在の音についての可能な持続時間であり、

【０１３８】

【数４７】

$$Q_{c(g),\Delta t}(t,e) = \max_{g'} \{\delta(c(g) = c(G_{g'}(\Delta t,e))) \cdot R_{c(g'),\Delta t}(t,e)\} \qquad \text{式３６}$$

【０１３９】

上式で、

Ｑ$_{c(g),\Delta t}$（ｔ，ｅ）は、時刻ｔ’＝ｔ－Δｔにエッジに進入し、エッジｅに沿うクラスｃ（ｇ）の軌跡に沿って時刻ｔにノードｎに進入する最良のパスの確率と理解することができる。

【０１４０】

ｇ’は、前のエッジの終わりまでの軌跡の履歴であり、

ｃ（・）は、可能な軌跡の連続的な値がクラスタ化された、軌跡の離散的クラスを示し、

【０１４１】

【数４８】

$$\delta(\underset{E}{A=B}) = \begin{cases} 1 & A=Bの場合 \\ 0 & 上記以外の場合 \end{cases} \qquad 式３７$$

$$G_{g'}(\Delta t, e) = \beta(\Delta t)\cdot g' + (1-\beta(\Delta t))\cdot T \qquad 式３８$$

【０１４２】

【数４９】

$$R_{c(g'),\Delta t}(t,e) = \left\{ p(o(t)\middle|\Delta t, e, g'_{opt}) \frac{P(\Delta t|e)}{P(\Delta t-1|e)} \cdot R_{c(g'),\Delta t-1}(t-1,e) \right\} \qquad 式３９$$

【０１４３】

ただし

【０１４４】

【数５０】

$$R_{c(g'),0}(t,e) = P(\Delta t=0|e)\cdot P(e)\cdot H_{c(g')}(t, n_{el}) \qquad 式４０$$

$$p(o(t)\middle|\Delta t, e, g'_{opt}) = \sum_m P(m|u,g)\cdot N(o(t); H_m g^e_{\Delta t} + h_m; S_m) \qquad 式４１$$

【０１４５】

これは、エッジｅに関連する仮説音についての尤度であり、仮説持続期間は以下のように近似される。

【０１４６】

【数５１】

$$p(o(t)\middle|\Delta t, e, g'_{opt}) \approx C \max_m \left\{ N(o(t); H_m g^e_{\Delta t} + h_m; S_m) \right\} \qquad 式４２$$

【０１４７】

上式で、

S_mは上記の式１５で定義される。

【０１４８】

【数５２】

g'_{opt}

【０１４９】

は、

【０１５０】

【数５３】

$g^e_{\Delta t}$

【０１５１】

を計算するのに使用するときに、式３６の右辺を最大にするｇ’であり、現在の音についての開始生成関係値を表す。

【０１５２】

【数５４】

$g^e_{\Delta t}$

【０１５３】

は、Δｔ、

【０１５４】

【数５５】

$$g'_{opt}$$

【０１５５】

、およびエッジｅに関連する軌跡パラメータが与えられたときの、生成関係ダイナミックス値の期待値であり、

Ｃは定数であり、

【０１５６】

【数５６】

$$H_m g^e_{\Delta t} + h_m$$

【０１５７】

は、観測ベクトルについての予測値を表す。ただしH_mおよびh_mはエッジｅに関連する音に依存しない。

【０１５８】

式４２の右辺の計算は、観測ベクトルと

【０１５９】

【数５７】

$$H_m g^e_{\Delta t} + h_m$$

【０１６０】

の差を取ることを含む。

【０１６１】

式３９は、エッジ（ｔ’＝ｔ－Δｔ）に割り当てられた最初の時間インデックスから現在の時間インデックスｔまでの各時間インデックスについて、単一の音を表すエッジｅ内で反復的に実施される。式４０に示すように、最初の時間インデックスに達したとき、現在のエッジの最初の状態と、現在のエッジの最初の時間インデックスとを使用して式３５を評価することによって別の反復が実施される。この反復は、式４０の最後の項によって示される。

【０１６２】

式３９の持続時間モデル、

【０１６３】

【数５８】

$$P(\Delta t|e)$$

【０１６４】

、および

【０１６５】

【数５９】

$$P(\Delta t - 1|e)$$

【０１６６】

は、トレーニング観測シーケンスをセグメント化する従来技術のＨＭＭモデルを使用してトレーニングすることができる。セグメント化したテキストを使用して、各音についての持続期間が識別され、その持続期間は、各音についての持続期間確率分布を構築するのに使用される。

【０１６７】

上記のリカーシブなフレームワークで示したが、図３のフローチャートを参照しながら以下で説明するように、式３２～４２の復号化は、本発明の一実施形態の下では反復的に実施される。図３に示す方法では、観測される信号の各フレームで、１組の最良のパスス

コアが求められる。具体的には、特定のフレームで利用可能な各状態について、上記の式３５で求められる別々の$H_{c(g)}(t,n)$値が、軌跡の各クラスについて求められる。例えば４つの状態４００、４０２、４０４、および４０６を含む図４の単純な状態図の場合、軌跡のクラスが３つだけ存在することを条件として、図５の$H_{c(g)}(t,n)$値が生成される。したがって、値５００、５０２、および５０４は、第１フレーム（時間インデックスt_1）の間に、それぞれ状態n_1のクラスｃ（ｇ）１、ｃ（ｇ）２、およびｃ（ｇ）３について生成される。値５０６、５０８、および５１０は、第２フレーム（時間インデックスt_2）の間に、それぞれ状態n_1のクラスｃ（ｇ）１、ｃ（ｇ）２、およびｃ（ｇ）３について生成される。図６に、観測されるベクトルの第４フレームについてのスコアを計算中の図５のダイアグラムを示す。図６は、以下の図３の説明で参照する。

【０１６８】

図３のステップ３００では、最初に観測したベクトルを選択し、時間インデックスを１に設定する。ステップ３０２では、時間インデックスの識別中に利用可能な状態を識別する。状態は、時間インデックスがその状態と有限状態システムの開始状態n_0の間のエッジ数に等しい場合、その時間インデックスで利用可能である。例えば、図４の単純化した状態システムでは、４つの状態４００、４０２、４０４、および４０６が存在する。時間インデックス１の間、利用可能な状態は４０２および４０４だけである。状態４００と状態４０４の間の各エッジについて少なくとも１つの特徴ベクトルを有するための十分な観測された特徴ベクトルが存在しないからである。時間インデックス２の間、状態４０２、４０４、および４０６はすべて利用可能となる。

【０１６９】

他の実施形態では、時間インデックスが、状態と開始状態n_0の間のエッジ数よりも大きいある量である場合、その状態は利用可能とみなされない。これにより、生じる可能性の低い状態についての一時的な位置をなくすることによって実行される計算の数が削減される。

【０１７０】

利用可能な状態を識別した後、ステップ３０４でその状態のうちの１つを選択する。次いでステップ３０６で、軌跡のクラスｃ（ｇ）を選択する。したがって図６では、ステップ３０４および３０６により、状態n_2およびクラスｃ（ｇ）１が選択される。これは、$H_{c(g)}(t,n)$値を生成する対象の状態およびクラスである。

【０１７１】

ステップ３０８では、選択した状態に向かうエッジを選択する。例えば、状態４０６に対して図４のエッジe_3を選択する。エッジを選択した後、Δｔをステップ３１０で選択する。この選択により、特定の時間インデックスでの以前の状態が識別される。したがって例えば、ステップ３０８および３１０でエッジe_3およびΔｔ＝３を選択することにより、図６の前状態のような、時間インデックスｔ１での状態ｎ１の発生が選択される。

【０１７２】

ステップ３１２では、現在選択している軌跡のクラス、選択したエッジ、および選択したΔｔについて、式３６に示す計算を最大にするｇ’を識別する。式３９および４０に示すように、式３６の計算は、Δｔが交差する時間インデックスのすべてに沿ってリカーシブに実行され、選択したエッジの始めの状態に関連する$H_{c(g')}(t-\Delta t, n_{e1})$の値に依存する。値$H_{c(g')}(t-\Delta t, n_{e1})$は、以前の時間インデックスで図３の方法を使用して先に計算されていること、および値は、式３６にとって最適なｇ’のクラスに基づいて選択されることに留意されたい。したがって、ｇ’がクラスｃ（ｇ）３内にある場合、図６の直線６００で示すようにエッジe_3およびΔｔ＝３を使用して、$H_{c(g)3}(t_1, n_1)$が、時刻t_4での状態N_2についての式３９および４０の計算で使用される。

【０１７３】

ステップ３１４では、この方法は、考慮すべき追加のΔｔが存在するかどうかを判定する。考慮すべき追加のΔｔが存在する場合、プロセスはステップ３１０に戻り、異なるΔ

ｔを選択する。次いで、新しいΔｔを使用してステップ３１２を反復し、式３６を使用して新しいスコアを識別する。

【０１７４】

Δｔがもはや存在しないとき、式３６を使用して計算したスコアを互いに比較し、ステップ３１６で、最高のスコアを現在のエッジについてのスコアとして選択する。例えば、図６の直線６００、６０２、および６０４で表されているように、エッジe_3に関して、Δｔ＝１、Δｔ＝２、Δｔ＝３のそれぞれについて別々のスコアを計算する。次いで最高のスコアをステップ３１６で選択し、それによって状態と、観測した音声フレームの間の特定の位置合せを選択する。

【０１７５】

ステップ３１８では、この方法は、現状態で終了するさらに別のエッジが存在するかどうかを判定する。さらに別のエッジが存在する場合、プロセスはステップ３０８に戻り、ステップ３１０、３１２、３１４、および３１６を反復して、新しく選択したエッジについてのエッジスコアを生成する。例えば、図６では、エッジe_4についてエッジスコアが計算される。

【０１７６】

ステップ３１８で、現状態で終了するエッジがもはや存在しない場合、最高のエッジスコアを、選択したクラスおよび状態に対する$H_{c(g)}$（ｔ，ｎ）パススコアとして選択する。

【０１７７】

ステップ３２２では、プロセスは、$H_{c(g)}$（ｔ，ｎ）パススコアを必要とする追加のクラスが存在するかどうかを判定する。追加のクラスが存在する場合、プロセスはステップ３０６に戻り、次の軌跡のクラスを選択する。次いでステップ３０８～３２０を反復し、新しく選択したクラスについて$H_{c(g)}$（ｔ，ｎ）パススコアを生成する。ステップ３２２で別のクラスが存在しない場合、プロセスは、スコアリングする必要のある別の利用可能な状態が存在するかどうかをステップ３２４で判定する。別の利用可能な状態が存在する場合、プロセスはステップ３０４に戻って次の利用可能な状態を選択し、ステップ３０６～３２２を反復して、その状態についての$H_{c(g)}$（ｔ，ｎ）パススコアの集合を生成する。

【０１７８】

現在の時間インデックスで利用可能な状態のすべてについてスコアを生成したとき、プロセスは、別の観測ベクトルが存在するかどうかをステップ３２６で判定する。別の観測ベクトルが存在する場合、プロセスはステップ３００に戻り、ステップ３００で次の観測ベクトルを選択し、時間インデックスを１つ増分する。次いで、新しいベクトルおよび時間インデックスについてステップ３０２～３２４を反復する。最後の観測ベクトルをステップ３２６で処理したとき、クラスの如何に関わらず、最高のスコアを有する終端状態を、復号化したパス中の最終状態として選択する。次いで、各状態で最高の$H_{c(g)}$（ｔ，ｎ）スコアを生成するのに使用したエッジおよびΔｔに沿ってトレースバックすることにより、復号化パスを判定する。

【０１７９】

本発明の下で軌跡のクラスを使用することにより、このようにして復号化を実行することが可能となる。このようなクラスがない場合、状態数が無限となるので、有限状態変換器を使用することができない。その理由は、軌跡（ｇ）が連続的であり、したがって任意の値を取れるからである。軌跡の代わりに軌跡のクラスについてのスコアを生成することにより、有限の状態数を有することが可能となり、図３で述べた復号化が可能となる。各状態に対して計算されるスコアの数を制限するためにクラスを使用するが、最適な軌跡ｇ’は、ｇ’についての連続的な値の集合から選択され、かつｇ’は、スコアを生成するのに使用される（上記の式４１を参照）ことにも留意されたい。これにより、上記の有限状態復号化が依然として可能となると共に、スコアの判定が正確となる。

【０１８０】

図７に、本発明を使用することができる音声認識システムのブロック図を与える。図７では、トレーナまたはユーザのいずれかである話者７００が、マイクロフォン７０４に向かって発話する。マイクロフォン７０４はまた、１つまたは複数のノイズ源７０２から加算ノイズも受け取る。マイクロフォン７０４によって検出された可聴信号は電気信号に変換され、アナログ－デジタル（Ａ－Ｄ）変換モジュール７０６に供給される。

【０１８１】

Ａ－Ｄ変換モジュール７０６は、マイクロフォン７０４からのアナログ信号を一連のデジタル値に変換する。いくつかの実施形態では、Ａ－Ｄ変換モジュール７０６は、１６ｋＨｚおよびサンプル当り１６ビットでアナログ信号をサンプリングし、それによって毎秒３２キロバイトの音声データを生成する。これらのデジタル値はフレームコンストラクタ７０７に供給され、一実施形態ではフレームコンストラクタ７０７は、１０ミリ秒ずれて開始する２５ミリ秒フレームに値をグループ化する。

【０１８２】

フレームコンストラクタ７０７によって生成されるデータのフレームは特徴抽出モジュール７０８に供給され、特徴抽出モジュール７０８は各フレームから特徴を抽出する。特徴抽出モジュールの例には、線形予測符号化（ＬＰＣ）を実施するモジュール、ＬＰＣ導出ケプストラム、知覚線形予測（ＰＬＰ）、聴覚モデル特徴抽出、およびメル周波数ケプストラム係数（ＭＦＣＣ）特徴抽出が含まれる。本発明はこれらの特徴抽出モジュールに限定されず、本発明の状況内で他のモジュールを使用できることに留意されたい。

【０１８３】

入力信号がトレーニング信号である場合、この一連の特徴ベクトルはトレーナ７２４に供給され、トレーナ７２４は特徴ベクトルおよびトレーニングテキスト７２６を使用して、本発明の生成モデル７２８をトレーニングする。例えば、前述のＥＭトレーニングアルゴリズムを使用して、生成モデルをトレーニングすることができる。

【０１８４】

入力信号がテスト信号である場合、特徴ベクトルがデコーダ７１２に供給され、デコーダ７１２は、特徴ベクトルのストリーム、レキシコン７１４、言語モデル７１６、および生成モデル７２８に基づいて、最も可能性の高い一続きの単語を識別する。一実施形態の下では、レキシコン（lexicon）７１４は、一続きの特徴ベクトルから単語を識別するために、デコーダ７１２によって走査される有限状態ネットワークを定義する。

【０１８５】

最も可能性の高い一続きの仮説単語が、信頼度測定モジュール７２０に供給される。信頼度測定モジュール７２０は、２次音響モデル（図示せず）に部分的に基づいて、音声レコグナイザによってどの単語が不適切に識別された可能性が最も高いかを識別する。次いで信頼度測定モジュール７２０は、どの単語が不適切に識別された可能性があるかを示す識別子と共に、一続きの仮説単語を出力モジュール７２２に供給する。本発明を実施するのに信頼度測定モジュール７２０は不要であることを当業者は理解されよう。

【０１８６】

特定の実施形態を参照しながら本発明を説明したが、本発明の主旨および範囲から逸脱することなく、形態および細部に変更を加えることができることを当業者は理解されよう。

【図面の簡単な説明】

【０１８７】

【図１】本発明の実施形態の一コンピューティング環境のブロック図である。

【図２】本発明の実施形態の代替コンピューティング環境のブロック図である。

【図３】本発明の実施形態の復号化の方法のフローチャートの図である。

【図４】本発明の実施形態の単純な有限状態図である。

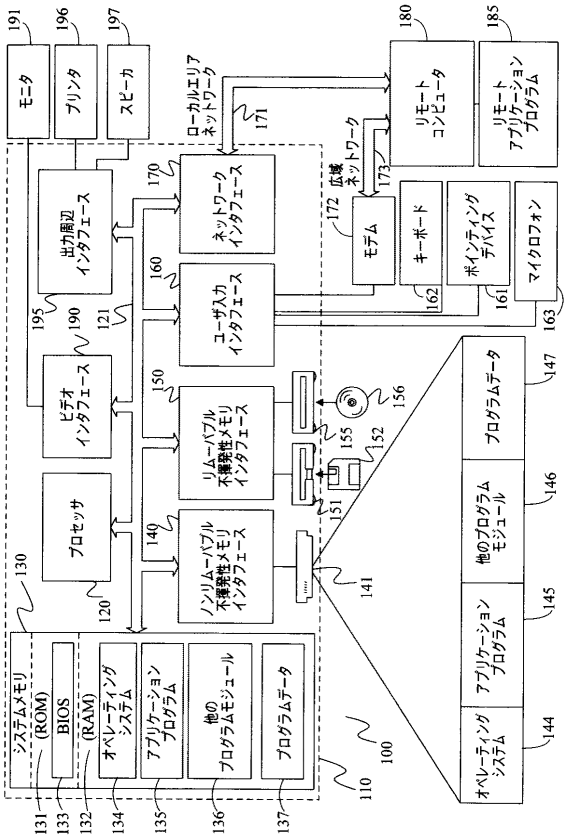
【図５】本発明の実施形態の様々な状態および時間でのクラスパススコアを示す状態図である。

【図６】本発明の実施形態の新しい時間インデックスでの状態についてクラスパススコア

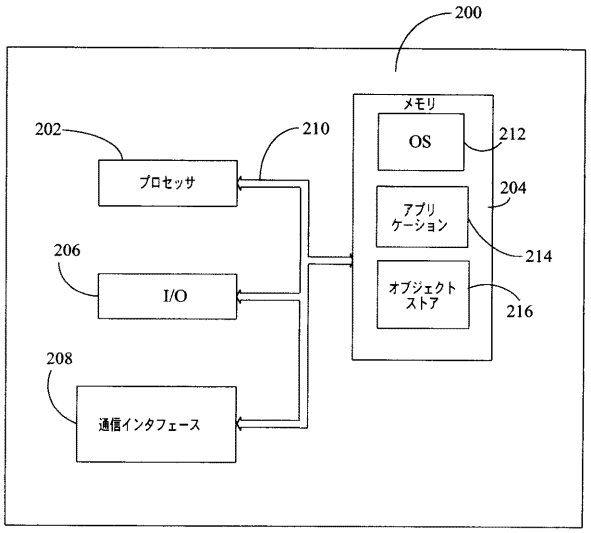
を計算するのに使用するパス接続を示す状態図である。
【図７】本発明の実施形態の本発明の一実施形態の下での音声認識システムのブロック図である。
【符号の説明】
【０１８８】
１００　コンピューティングシステム環境
１１０　コンピュータ
１２０　プロセッサ
１２１　システムバス
１３０　システムメモリ
１３１　読取り専用メモリ（ＲＯＭ）
１３２　ランダムアクセスメモリ（ＲＡＭ）
１３３　ＢＩＯＳ
１３４　オペレーティングシステム
１３５　アプリケーションプログラム
１３６　他のプログラムモジュール
１３７　プログラムデータ
１４０　インタフェース
１４１　ハードディスクドライブ
１４４　オペレーティングシステム
１４５　アプリケーションプログラム
１４６　他のプログラムモジュール
１４７　プログラムデータ
１５０　インタフェース
１５１　磁気ディスクドライブ
１５２　リムーバブル不揮発性磁気ディスク
１５５　光ディスクドライブ
１５６　リムーバブル不揮発性光ディスク
１６０　ユーザ入力インタフェース
１７０　ネットワークインタフェース／アダプタ
１７１　ローカルエリアネットワーク（ＬＡＮ）
１７２　モデム
１７３　広域ネットワーク（ＷＡＮ）
１８０　リモートコンピュータ
１８５　リモートアプリケーションプログラム
１９０　ビデオインタフェース
１９１　モニタ
１９５　出力周辺インタフェース
１９６　プリンタ
１９７　スピーカ
２００　モバイル装置
２０２　マイクロプロセッサ
２０４　メモリ
２０６　入出力（Ｉ／Ｏ）構成要素
２０８　通信インタフェース
２１０　バス
２１２　オペレーティングシステム
２１４　アプリケーションプログラム
２１６　オブジェクトストア
４００、４０２、４０４、４０６　状態

７００　話者
７０２　ノイズ源
７０４　マイクロフォン
７０６　アナログ－デジタル変換モジュール
７０７　フレームコンストラクタ
７０８　特徴抽出モジュール
７１２　デコーダ
７１４　レキシコン
７１６　言語モデル
７２０　信頼度測定モジュール
７２２　出力モジュール
７２４　トレーナ
７２６　トレーニングテキスト
７２８　生成モデル

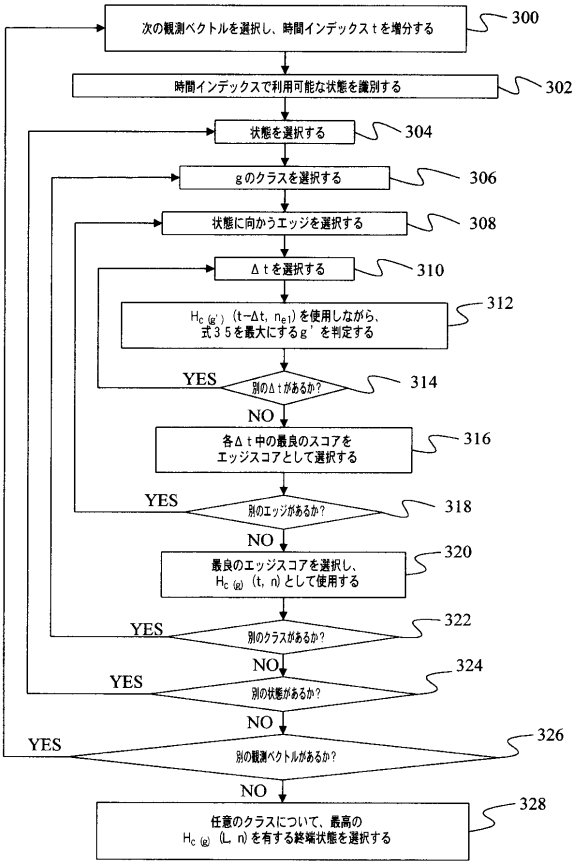
【図１】



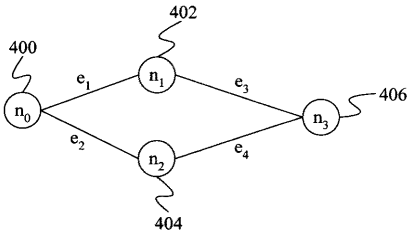
【図２】



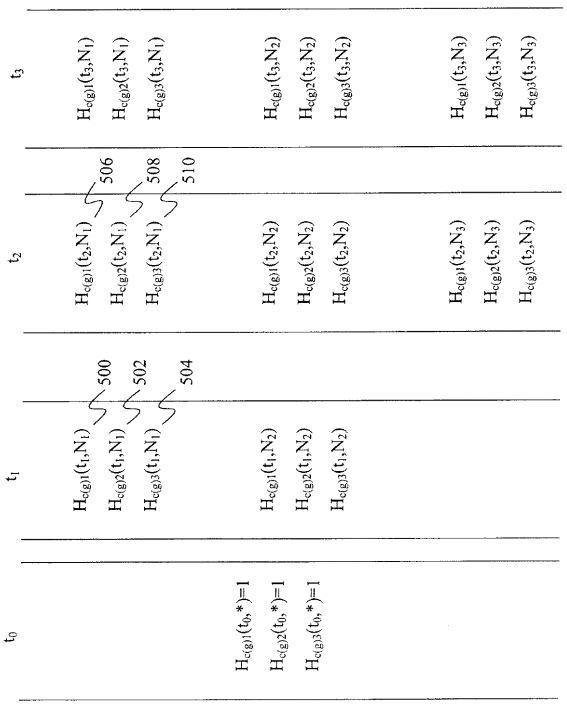
【図３】


次の観測ベクトルを選択し、時間インデックスｔを増分する
300
時間インデックスで利用可能な状態を識別する
302
状態を選択する
304
ｇのクラスを選択する
306
状態に向かうエッジを選択する
308
Δｔを選択する
310
$H_{c(g')}(t-\Delta t, n_{s1})$を使用しながら、
式３５を最大にするｇ'を判定する
312
YES
別のΔｔがあるか?
314
NO
各Δｔ中の最良のスコアを
エッジスコアとして選択する
316
YES
別のエッジがあるか?
318
NO
最良のエッジスコアを選択し、
$H_{c(g)}(t, n)$として使用する
320
YES
別のクラスがあるか?
322
NO
YES
別の状態があるか?
324
NO
YES
別の観測ベクトルがあるか?
326
NO
任意のクラスについて、最高の
$H_{c(g)}(L, n)$を有する終端状態を選択する
328

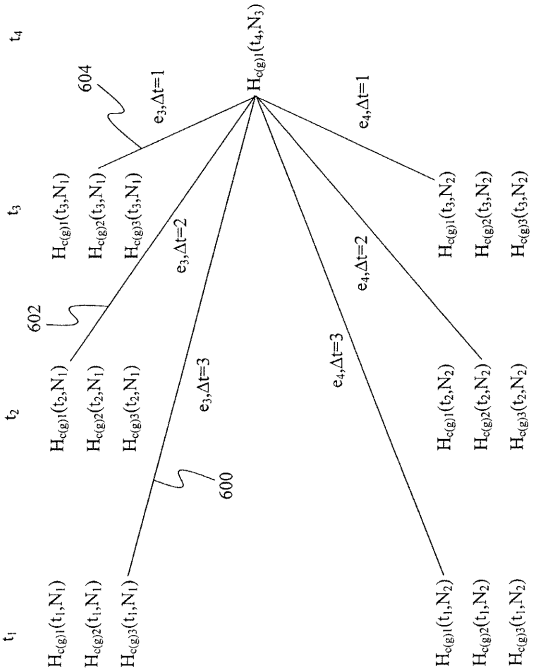
【図４】


400
402
404
406
n_0
n_1
n_2
n_3
e_1
e_2
e_3
e_4

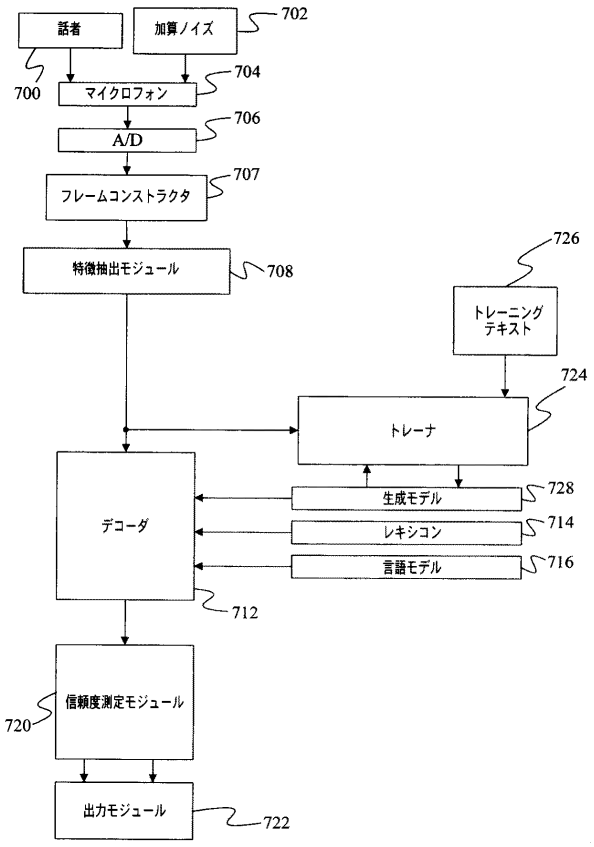
【図５】


t_0
t_1
t_2
t_3
$H_{c(g)1}(t_0,*)=1$
$H_{c(g)2}(t_0,*)=1$
$H_{c(g)3}(t_0,*)=1$
500
502
504
506
508
510

【図６】


t_1
t_2
t_3
t_4
600
602
604
$e_3,\Delta t=1$
$e_3,\Delta t=2$
$e_3,\Delta t=3$
$e_4,\Delta t=1$
$e_4,\Delta t=2$
$e_4,\Delta t=3$

【図７】


話者
加算ノイズ
702
700
マイクロフォン
704
A/D
706
フレームコンストラクタ
707
特徴抽出モジュール
708
726
トレーニング
テキスト
724
トレーナ
デコーダ
生成モデル
728
レキシコン
714
言語モデル
716
712
信頼度測定モジュール
720
出力モジュール
722

フロントページの続き

(72)発明者 チョウ ジャン－ライ
中華人民共和国 ペキン チャオヤン ディストリクト シャオヤオジュ ベイリ １１３ ビルディング １５０３
(72)発明者 フランク トルステン ベルント ザイデ
中華人民共和国 ペキン チャオヤン ディストリクト シャオユン ロード ナンバー３２ コンコルディア プラザ アパートメント ビー－１３０２
(72)発明者 アセラ ジェイ．アール．グナワルデナ
アメリカ合衆国 ９８１０１ ワシントン州 シアトル アラスカン ウェイ １９５０ ナンバー３２３
(72)発明者 ハガイ アティアス
アメリカ合衆国 ９８１２１ ワシントン州 シアトル ウォール ストリート ５００ アパートメント ９０１
(72)発明者 アレハンドロ アセロ
アメリカ合衆国 ９８００６ ワシントン州 ベルビュー １６３ プレイス サウスイースト ６５２５
(72)発明者 ホアン シュエドン
アメリカ合衆国 ９８００６－５３４５ ワシントン州 ベルビュー １５５ アベニュー サウスイースト ６１７９

審査官 山下 剛史

(56)参考文献 特開平０７－１９９３７９（ＪＰ，Ａ）
特開平０８－０９５５９２（ＪＰ，Ａ）
特開平１０－０６３２９１（ＪＰ，Ａ）
特開平０８－０２２２９６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ １５／００－１５／２８