(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4485694号
(P4485694)
(45)発行日　平成22年6月23日(2010.6.23)　(24)登録日　平成22年4月2日(2010.4.2)

(51)Int.Cl.　F I
G10L 15/28 (2006.01)　G10L 15/28 210M

請求項の数 8 (全 12 頁)

(21)出願番号	特願2000-608365 (P2000-608365)
(86)(22)出願日	平成12年3月7日 (2000.3.7)
(65)公表番号	特表2002-540478 (P2002-540478A)
(43)公表日	平成14年11月26日 (2002.11.26)
(86)国際出願番号	PCT/EP2000/001965
(87)国際公開番号	WO2000/058945
(87)国際公開日	平成12年10月5日 (2000.10.5)
審査請求日	平成19年3月6日 (2007.3.6)
(31)優先権主張番号	99200949.8
(32)優先日	平成11年3月26日 (1999.3.26)
(33)優先権主張国	欧州特許庁 (EP)

(73)特許権者　590000248
コーニンクレッカ　フィリップス　エレクトロニクス　エヌ　ヴィ
オランダ国　５６２１　ベーアー　アインドーフェン　フルーネヴァウツウェッハ　1
(74)代理人　100087789
弁理士　津軽　進
(74)代理人　100114753
弁理士　宮崎　昭彦
(72)発明者　セレン　エリック
オランダ国　５６５６　アーアー　アインドーフェン　プロフ　ホルストラーン　６

最終頁に続く

(54)【発明の名称】並列する認識エンジン

(57)【特許請求の範囲】
【請求項１】
一連の音声単語を認識する大語彙音声認識システムであって、
前記一連の音声単語を表す時系列入力パターンを入力する入力手段、及び
多語彙音声認識装置と関連付けられた多語彙認識モデルを用いて、語彙から前記入力パターンを一連の単語として認識するように動作する当該多語彙音声認識装置、
を有する大語彙音声認識システムにおいて、Ｎ個の多語彙音声認識装置を有し、当該認識装置の各々はそれぞれ異なる多語彙認識モデルと関連付けられ、前記認識モデルの各々は前記大語彙の特定部分に目標を置き、複数の前記音声認識装置に前記入力パターンを送り、前記複数の音声認識装置により認識された単語列から認識された単語列を選択するように動作する制御器を有することを特徴とする大語彙音声認識システム。
【請求項２】
Ｍ＞Ｎである多語彙認識モデルをＭ個有し、少なくとも１つの前記音声認識装置に対し、認識コンテクストに依存して前記Ｍ個のモデルから前記関連する認識モデルを選択するように動作するモデル選択器を有する請求項１に記載のシステム。
【請求項３】
音声入力が関連する文書は、少なくとも１つの認識コンテクストを決定する請求項２に記載のシステム。
【請求項４】
前記文書は、ＨＴＭＬページのようなウェブページであり、前記文書のコンテクストは

、当該文書内で特定される又は当該文書に関連する請求項３に記載のシステム。
【請求項５】
前記モデル選択器は、前記文書における又は関連する単語に依存して前記認識モデルを選択するように動作する請求項３に記載のシステム。
【請求項６】
前記モデル選択器は、
前記認識装置の１つによりまだ使用されていないＮ－Ｍ認識モデルからテスト認識モデルを選択し、
前記テスト認識モデルで前記入力パターンの少なくとも一部を認識するようにテスト認識装置を制御し、及び
前記テスト認識装置の認識結果が前記認識装置の１つの認識結果よりも良好である場合、前記テスト認識モデルでの認識を可能にする
ように動作する請求項２に記載のシステム。
【請求項７】
前記認識モデルは、より一般的なコンテクストを持つモデルからより特定のコンテクストを持つモデルへ階層的に配され、階層において高位レベルでの階層的に関連するより一般的なモデルでの認識が他の認識モデルと関連付けられる少なくとも１つの認識装置の結果と比較される良好な認識結果を得る場合、前記モデル選択器は、より特定のモデルでの認識を可能にするように動作する請求項１に記載のシステム。
【請求項８】
前記システムは、インターネットのようなネットワークを介して接続されるサーバステーション及びユーザステーションを有し、前記ユーザステーションは、ユーザから入力パターンを入力し、当該入力パターンを表す信号を前記サーバステーションに転送し、前記サーバステーションは、前記認識装置及び制御器を有する請求項１に記載のシステム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、一連の音声単語を認識する大語彙音声認識システムに関する。このシステムは、この一連の音声単語を表す時系列入力パターンを入力する入力手段、及び前記音声認識装置に関連する多語彙認識モデルを用いて、語彙から入力パターンを一連の単語として認識するように動作する多語彙音声認識装置を有する。
【０００２】
【従来の技術】
インターネット環境において音声を認識するシステムは、米国特許公報ＵＳ５，８１９，２２０号から既知である。このシステムは、特に音声を用いてWWW上の情報源にアクセスすることを目的としている。ウェブとのインタフェースとして音声認識システムを構築することは、従来の音声認識の分野において生じる問題とは大いに異なる問題に直面する。ユーザは、如何なる項目に関するどんな文書にも事実上アクセスすることができるので、主要な問題は前記システムがサポートを必要とする大語彙である。これら大語彙に対する適切な認識モデル、例えば言語モデルを構築することは、不可能でないなら、非常に難しい。既知のシステムにおいて、統計上のn-gram言語モデル及び音響モデルを含む既定の認識モデルが使用される。この認識モデルは、ウェブトリガされた単語セット(web-triggered word set)を用いて動的に変更される。ＨＴＭＬ文書は、単語認識検索を高める確率に対する最終的な単語セットに含まれるべき単語セットを特定するのに使用されるリンク、例えばハイパーテキストリンクを含んでいる。このやり方で、音声認識スコアを計算するのに使用される単語セットは、前記ウェブトリガされた単語セット含むことでバイアス(biased)される。
【０００３】
【発明が解決しようとする課題】
既知のシステムは、適応後にバイアスされたモデルを得ることを可能にする開始モデルと

して適切な多語彙モデルを必要とする。実際に、このバイアスされたモデルは、現在の認識コンテクストに対し最適となる従来の多語彙モデルと見なされる。上述したように、それが開始モデルとしてのみ使用される場合も、適切な大語彙モデルを構築することは非常に難しい。更なる問題は、例えば検索エンジン上に存在するような特定のウェブサイト又はＨＴＭＬ文書若しくは書店のような大規模電子ショップへの入力を認識するようなある認識タスクに対して起こる。このような状況において、述べられる単語数は莫大である。従来の多語彙モデルは、一般的には、可能な単語の全範囲を効果的にカバーすることはできない。比較的少ない単語で開始モデルをバイアスさせることは、良好な認識モデルとなることはない。この開始モデルが既にかなり良好であると仮定する場合、適切なバイアスは、莫大な追加の単語セット及びかなりの量の処理を必要とする。

【０００４】

【課題を解決するための手段】

本発明の目的は、大語彙をよりよく扱うことが可能な認識システムを提供することである。

【０００５】

本目的を達成するために、前記システムは、Ｎ個の多語彙音声認識装置を有し、これら認識装置の各々はそれぞれ異なる多語彙認識モデルと関連し、これら認識モデルの各々は、前記大語彙の特定部分に目標を置き、前記システムは、入力パターンを複数の音声認識装置へ送り、これら複数の音声認識装置によって認識された単語列から認識された単語列を選択するように動作する制御器を有することを特徴とする。

【０００６】

各々が前記大語彙の一部に目標を置いた特定の認識モデルを具備する幾つかの認識装置を使用することで、大語彙に対する認識モデルを構築するタスクは、特定のコンテクストに対し大語彙モデルを構築する扱い易いタスクに分解される。上記コンテクストは、健康、エンターテイメント、コンピュータ、芸術、ビジネス、教育、政治、科学、ニュース、旅行等を含んでいる。これらコンテクストの各々は普通に、例えば言語の一般的な単語における語彙に重複すると理解されている。前記コンテクストは、これら共通する単語の統計及びこれらコンテクストに特有の専門用語において異なるであろう。入力を認識するためにこれらモデルの幾つかを使用することで、より広い範囲の発話が適切に学習されたモデルを使用して認識可能となる。幾つかのモデルを使用することの更なる利点は、認識中に良好な識別を可能にすることである。１つの大語彙が使用された場合、ある発話は、１つの特定の意味（及びスペル）で認識されるだけである。例として、ユーザが「カラー」のような音の単語を発音した場合、認識される単語列の大部分は、まさに共通の単語「カラー」を含むだろう。（ファッションのコンテクストの）カラー、カラードヘリング(collared herring)（ロール巻きにしたニシン）（食品のコンテクスト）又はカラーボーン(collar-bone)（鎖骨）（健康のコンテクスト）のカラーという単語が認識されることはほとんどない。これら特殊な単語は、一般的な単語の単語列が頻繁に発生することで必然的に優位となる大語彙において認識される機会はそんなにない。幾つかのモデルを使用することで、各モデルは、そこから選択が行われる１つ以上の候補単語列を特定する。この最終的な選択において、カラー(color)という単語列が選択されても、その中にカラー(collar)という代替の単語列がユーザに表示されてしまう。

【０００７】

好ましくは、前記認識装置は、ユーザが認識に関し著しく遅いと感じない感覚で並行して動作する。これは、各々が自己処理源を有する個々の認識エンジンを用いて達成される。代わりに、これは、従来の時分割技術を用いて「並行」して認識タスクを動作する十分パワフルな逐次処理器を用いて達成されてもよい。

【０００８】

並行する音声認識エンジンを用いることは公知であることに注意されたい。米国特許番号ＵＳ５，７５４，９７８は、認識エンジンを並行して用いていることが記載されている。これらエンジンの全ては、例えば９５％という比較的高い精度を有している。５％のエン

ジンの不正確さが重複しない場合、認識の精度は改善可能である。これら不正確さが完全に重複しないことを保証するために、これらエンジンは別々でもよい。代わりに、これらエンジンの１つへの入力信号が僅かに摂動するか、又はこれらエンジンの１つが僅かに摂動する場合に、前記エンジンは同じになる。比較器は、認識されたテキストを比較し、前記エンジンの出力間における一致の度合いに基づきテキストを容認又は拒絶する。このシステムは、大語彙には存在しない正確な認識エンジンを必要とするので、このシステムは、大語彙認識に対し何ら解決法を提供できない。どのシステムも大語彙の特定部分を目標にする別々のモデルを使用しない。

【０００９】

国際公開公報ＷＯ９８／１０４１３号は、並行に動作可能な任意の数の音声認識モジュールを備える対話システムを説明している。前記モジュールは、例えば孤立型数字認識、連続番号認識、少語彙単語認識、孤立型多語彙認識、連続単語認識、キーワード認識、単語列認識、アルファベット認識等の音声認識の特定型式を目標とする。この対話システムは、どの型式の入力をユーザが供給するか事前に分かり、それに応じて１つ以上の特定モジュールを活性化させる。例えば、ユーザが番号を話す必要がある場合、対話エンジンは、ユーザが数字又は連続番号として番号を話すことを可能にする孤立型数字認識及び連続番号認識を可能にする。このシステムは、大語彙を処理する解決法を供給しない。

【００１０】

本発明に係るシステムの認識モデルは、既定されていてもよい。好ましくは、従属請求項２に定められるように、モデル選択器は、認識するのにアクティブに使用されるモデルを少なくとも１つ動的に選択するのに使用される。この選択は、ユーザが入力するコンテクスト、例えばクエリー又は口述項目に依存する。好ましくは、このモデル選択器は、多くの認識モデルを選択する。実際には、少なくとも１つのモデルが一般項目に関する普通の日常語彙を示す。このようなモデルは、普通は、常に使用されている。

【００１１】

従属請求項３に定められる実施例において、文書は認識コンテクストを規定する。従属請求項５に定められるように、これは、前記文書に存在する単語を走査し、これら単語を認識するのに最適な認識モデル（例えば、文書と同様に大多数の単語又は単語列を持つこれらモデル）を決定することで行われる。

【００１２】

従属請求項４に定められる実施例において、コンテクストは、例えばコンテクストを特定する埋め込みタグを用いて、ウェブページに示される。このページは、例えばリンクを介してコンテクスト（又はコンテクスト識別子）を示してもよい。

【００１３】

従属請求項６に定められる実施例において、このシステムは、現在の認識タスクに適したこれらの認識モデルを特定することをアクティブに試みる。認識に対しアクティブに使用されるときの認識モデルに加えて、他のモデルは、これらの適性に対しテストされる。このテストは、使用されないモデルが、アクティブに使用されるモデルの１つより良好な結果を与えるかをチェックする追加の認識装置を１つ以上使用することで背景タスクとして実行されてもよい。代わりに、実際の認識装置は、この認識装置に十分な性能が残っている瞬間、例えばユーザが話していないときにテストモデルをテストするのに使用されてもよい。このテストは、ユーザの全ての入力を含んでいる。特に、ユーザが既に多くの音声入力を供給した場合、好ましくはこのテストが最新の入力に限定される。このやり方で、ユーザが項目を直ぐ変えるとき、より適したモデルが選択可能である。どのモデルが最適かを決める基準、すなわち最も正確な認識を提供する基準は、好ましくはスコア又は信頼手段のような認識の性能表示に基づいている。

【００１４】

従属請求項７に定められる実施例において、認識モデルは階層的に配される。これは、適切なモデルを選択することを容易にする。好ましくは、認識は多くの比較的に一般的なモデルで始まる。ある一般的なモデルが良好な認識結果を提供することが分かった場合、よ

り特定のモデルがこの認識を更に改善するようにテストされる。より特定のモデルの幾つかは、幾つかのより一般的なモデルによって共有されてもよい。ある瞬間に、特定のモデルの認識結果が悪くなった場合、この特定のモデルより階層的に上位にあるより一般的なモデルの幾つかが試されてもよい。これは、あるコンテクストから他のコンテクストへの円滑な遷移を可能にする。例えば、ユーザは、健康という一般的なコンテクストに関する入力を供給することで始めてもよい。ある瞬間において、ユーザが医療センター又は施設のようなより特定のコンテクストに最初に焦点を置き、更に健康ファームのような最も特定のコンテクストに下がって行くことが検出されてよい。特に、前記健康ファームは、興味のあるエリアに配され、これはユーザに休暇、旅行又は特に健康ファームのエリアにある旅行というより一般的なコンテクストに移動する気にさせる。

【００１５】

従属請求項８に定められるように、認識は、個々の認識サーバにより行われてもよい。インターネットのコンテクストにおいて、このようなサーバは、ネット上の個々のステーションであり、検索エンジンのような存在するステーション又は電子書店のようなサービスプロバイダで統合される。特に、多くのユーザに対し動作する認識サーバは、大多数のユーザに適する語彙をサポート可能にする必要がある。幾つかの、特定の多語彙モデルは、高い認識精度でこのタスクを良好に実行可能にする上記システムを利用する。

【００１６】

【発明の実施の形態】

本発明のこれら及び他の特徴は、図面に示される実施例から明白であり、これら図面を参照して説明する。

【００１７】

例えば多語彙連続音声認識システムのような音声認識システムは、入力パターンを認識するために、認識モデルの集合体を典型的に使用する。例えば、音響モデル及び語彙は、単語を認識するのに使用されてもよく、言語モデルは、基本的な認識結果を改善するのに使用されてよい。図１は、多語彙連続音声認識システム１００の典型的な構造を説明する（L. Rabiner, B-H. Juang著、"Fundamentals of speech recognition" Prentice Hall 1993, 頁434-454参照）。このシステム１００は、スペクトル分析サブシステム１１０及びユニット整合サブシステム１２０を有する。このスペクトル分析サブシステム１１０において、音声入力信号（ＳＩＳ）は、特徴である代表ベクトル（観測ベクトル：ＯＶ）を計算するために、スペクトル的及び／又は一時的に分析される。典型的に、この音声信号は、デジタル化（例えば６．６７ｋＨｚのレートでサンプル化）され、例えばプリエンファシス(pre-emphasis)を与えることで前処理される。連続するサンプルは、例えば３２ｍｓｅｃの音声信号に対応するフレームにグループ化（ブロック化）される。連続するフレームは部分的、例えば１６ｍｓｅｃ重複している。しばしば、線形予測分析（ＬＰＣ）のスペクトル分析法は、特徴である代表ベクトル（観測ベクトル）を各フレームに対し計算するのに使用される。この特徴ベクトルは、例えば２４，３２又は６３個の構成要素を有してもよい。多語彙連続音声認識への標準的アプローチは、音声生成の見込みモデルを仮定することであり、これによって、指定される単語列$W = w_1 w_2 w_3 \cdots w_q$は、一連の音響観測ベクトル$Y = y_1 y_2 y_3 \cdots y_T$（ｔ＝１，…，Ｔ）を生成する。認識誤りは、観測ベクトルの観測される列$y_1 y_2 y_3 \cdots y_T$の大半を発生させる前記単語列$W = w_1 w_2 w_3 \cdots w_q$を決定することで統計的に最小とすることができる。ここで観測ベクトルは、スペクトル分析サブシステム１１０の結果である。これは最大を決定することになり、帰納的確率は、全ての可能な単語列Ｗに対し、

ｍａｘ（Ｗ｜Ｙ）

となる。条件付き確率にベイズの定理(Bayes' theorem)を与えることで、Ｐ（Ｗ｜Ｙ）は、

Ｐ（Ｗ｜Ｙ）＝Ｐ（Ｙ｜Ｗ）・Ｐ（Ｗ）／Ｐ（Ｙ）

で与えられる。Ｐ（Ｙ）とＷとは独立しているので、最も起こりうる単語列は、全ての可能な単語列Ｗに対する以下の方程式

ａｒｇ　ｍａｘ　Ｐ（Ｙ｜Ｗ）・Ｐ（Ｗ）　　（１）
で与えられる。
【００１８】
ユニット整合サブシステム１２０において、音響モデルは、上記方程式（１）の第１項を供給する。この音響モデルは、与えられた単語列Ｗに対する一連の観測ベクトルＹの確率Ｐ（Ｙ｜Ｗ）を概算するのに使用される。多語彙システムに対し、これは音声認識ユニットの一覧と観測ベクトルとを整合させることで通常は実行される。音声認識ユニットは、一連の音響参照によって表される。音声認識ユニットの様々な形態が使用されてもよい。例えば、全体の単語又は単語の集合さえも１つの音声認識ユニットで表される。単語モデル（ＷＭ）は、与えられた語彙の単語各々に一連の音響参照の音声表記(transcription)を供給する。全体の単語が音声認識ユニットで表されるシステムに対し、単語モデルと音声認識ユニットとの間に直接的な関係が存在する。他のシステム、特に多語彙システムは、単音素、２音素、音節のようなサブ単語ユニットと、fenene及びfenoneのような派生ユニットとが言語的に基づく音声認識ユニットに使用してもよい。このようなシステムに対し、単語モデルは、語彙の単語に関連する一連のサブ単語ユニットを記載する辞書１３４と、複雑な音声認識ユニットの音響参照の列を記載するサブ単語モデル１３２とにより与えられる。単語モデル構成器１３６は、前記サブ単語モデル１３２及び辞書１３４に基づく単語モデルを有する。図２は、サブ単語ユニットに基づくシステムの単語モデル２２０を説明し、ここで、示される単語は、３つの一連のサブ単語モデル（２５０，２６０及び２７０）によってモデル化され、これらサブ単語モデルの各々は、４つの一連の音響参照（２５１，２５２，２５３，２５４；２６１から２６４；２７１から２７４）を具備する。図２に示される単語モデルは、隠れマルコフモデル(HMMs: Hidden Markov Models)に基づき、これは確率論的なモデル音声信号に広く使用されている。このモデルを使用する場合、各認識ユニット（単語モデル又はサブ単語モデル）は、これのパラメタがデータの学習セットから概算されるＨＭＭによって典型的に特徴付けられる。多語彙音声認識システムに対しては、多くの学習データがより多くのユニットに対しＨＭＭを適切に学習させる必要があるので、通常、サブ単語ユニットは例えば４０個の限定されたセットが使用される。ＨＭＭの状態は、音響参照に対応している。参照をモデル化し、離散又は連続確率密度を含む様々な技術が知られている。１つの特定の発音に関する音響参照の各列は、この発話の音響音声表記とも呼ばれる。ＨＭＭ以外の他の認識技術が使用される場合、音響音声表記の細部が異なることは明白である。
【００１９】
図１の単語レベル整合システム１３０は、音声認識ユニットの全列と観測ベクトルとを整合させ、前記ベクトルと列との整合の尤度(likelihood)を供給する。サブ単語ユニットが使用される場合、サブ単語ユニットの可能な列を辞書１３４の列に制限するために、辞書１３４を用いることで前記整合に制約が置かれる。これは結果を単語の可能な列に減少させる。
【００２０】
十分な認識のために、整合に更なる制約が置かれるので、調査される経路が言語モデル（ＬＭ）によって特定されるような適切な列である単語列に対応する経路となる、この言語モデルに基づく文章レベル整合システム１４０を使用することも好ましい。このように、この言語モデルは、前記方程式（１）の第２項Ｐ（Ｗ）を供給する。音響モデルの結果と言語モデルとの組合せは、認識された文章（ＲＳ）１５２であるユニット整合サブシステム１２０の結果となる。パターン認識に使用される言語モデルは、言語及び認識タスクの構文上及び／又は語義上の制約１４２を含んでもよい。構文上の制約に基づく言語モデルは、通常、文法１４４と呼ばれる。この言語モデルにより使用される文法１４４は、原則として、
$P(W) = P(W_1)\,P(W_2 \mid W_1) \cdot P(W_3 \mid W_1 W_2) \cdots P(W_q \mid W_1 W_2 W_3 \cdots W_q)$
で与えられる単語列$W = W_1 W_2 W_3 \cdots W_q$の確率を供給する。実際には、与えられる言

語における全単語と全列長とに関する条件単語確率を容易に概算するのは不可能なので、N-gram単語モデルが広く使用されている。N-gram単語モデルにおいて、項Ｐ（W_j｜W_1 W_2 w_3…W_{j-1}）は、Ｐ（W_j｜W_{j-N+1}…W_{j-1}）で近似される。実際には、bigram又はtrigramが使用される。trigramにおいて、項Ｐ（W_j｜W_1 W_2 W_3…W_{j-1}）は、Ｐ（W_j｜W_{j-2} W_{j-1}）で近似される。

【００２１】

図３は、本発明に係る音声認識システム３００のブロック図を示す。このシステムの作用の実施例は、特に、認識された音声がテキスト又は同様な表現に変換されるアプリケーションに関し記載される。このようなテキスト表現は、テキスト表現が、例えばワードプロセッサにおける文書、又はデータベースの領域を指定するテキスト領域に挿入される口述用途に使用されてもよい。口述に関し、現在の多語彙認識装置は、６０，０００語までのアクティブな語彙及び辞書をサポートしている。より多くの単語に対し十分に正確な認識を可能にするモデルを構築するために、十分に適切なデータを得ることは難しい。典型的に、ユーザは、制限された数の単語をアクティブな語彙／辞書に加える。これら単語は、（単語の音響音声表記も含む）３００，０００から５００，０００語の背景語彙から検索される。口述又は同様の用途に対し、例えば、大語彙は、少なくとも１００，０００のアクティブな単語又は３００，０００を越えるアクティブな単語からなる。特にリンク部分をクリックすることで全く異なるコンテクストが作られるインターネット環境に対して、背景語彙の単語の多くがアクティブに認識されることが好ましいことは明白である。それに添付された先行する名前の確率(prior name probability)のある形態でフラットリストとして通常はモデル化されるが、高品質な言語モデルは存在しない例えば名前を認識するような他の認識タスクに対して、５０，０００語以上の語彙が既に莫大に分類されている。

【００２２】

認識結果は、口述用途に使用する必要はないと理解される。会話システムのような他のシステムに対する入力として同じように使用される。ここで、認識された音声情報に依存することは、データベースから検索される、又は本を注文若しくは旅行を予約するような操作が達成される。

【００２３】

図３には孤立型システム３００が示されている。このシステムは、例えばＰＣのようなコンピュータ上で実行される。項目３１０は、ユーザから音声表示信号を入力する相互接続部を示す。例えば、マイクロホンがこの相互接続部３１０に接続されてもよい。音声表示信号は、事前に記録されてもよく、又は遠隔地から例えば電話若しくはネットワークを介して検索されることが分かる。このシステム３００は、ユーザからの入力を入力するためのインタフェース３２０を有する。これは、例えば従来の音響カードを使用して実行されてもよい。前記インタフェースがアナログ形態で音声を入力する入力部を持つ場合、このインタフェースは、このアナログ音声を音声認識システム３３０で更に処理するのに適した形態のデジタルサンプルに変換するＡ／Ｄ変換器を好ましくは有する。このインタフェースがデジタル形態で音声を入力する入力部を持つ場合、好ましくは、前記変換器は、前記デジタルデータを更なる処理をするのに適したデジタル形態に変換することが可能である。音声認識システム３３０は、図１のスペクトル分析サブシステム１１０に記載されるような入力信号を典型的には分析する。本発明に従い、音声認識システム３３０は、複数の多語彙音声認識装置を有し、これら各々は、それぞれが異なる多語彙認識モデルと関連している。図１に示されるような典型的な認識に対し、個々の認識装置は、図３の番号３３５より小さい番号で示されるような図１のモデル独立型スペクトル分析サブシステム１１０を割り当てることが可能である。図３は、３つの別々の認識装置３３１，３３２及び３３３を用いて説明している。これら認識装置は、同じアルゴリズムを使用してもよく、ここでは、語彙及び言語モデルのような使用するモデルに違いがある。音声認識は、好ましくはスピーカ独立であり、連続音声入力を可能にする。音声認識自体は公知であり、例えば米国シリアル番号０８／４２５，３０４（当方整理番号ＰＨＤ９１１３６）に対応す

るヨーロッパ番号ＥＰ９２２０２７８２．６、米国シリアル番号０８／７５１，３７７（当方整理番号ＰＨＤ９１１３８）に対応するＥＰ９２２０２７８３．４、米国特許番号ＵＳ５，６３４，０８３号（当方整理番号ＰＨＤ９３０３４）に対応するＥＰ９４２００４７５．５号のような様々な文書に開示され、これら全ては本出願の譲受人である。認識装置は、ほぼ同じ瞬間にこれら認識装置が同じ音声入力を別々に認識するような感覚、つまり“並行に”動作する。これは、例えばＶＬＩＷ処理器のような、“並行”動作処理器における別々の処理器又は処理ユニットのような、認識装置の各々に対し別々の情報源を用いることで達成される。同様の“並行”実行は、各認識装置が別々のタスクとして実施される十分高度な実行を持つ従来の逐次処理器でも得られる。好ましくは、前記認識は、単語が前記システムに入力された後、単語の認識時にあまり遅延が起こらないという感覚における“リアルタイム”である。

【００２４】

本発明に従って、多語彙音声認識装置の各々は、認識モデルの各々が大語彙の特定部分を目標としたそれぞれ異なる多語彙認識モデルと関連している。これらモデルは、好ましくは記憶装置３４０からロードされる。ここでの記載に関し、前記認識モデルは、１つの認識タスクに使用されるモデルのコヒーレントセットとするためのものである。例えば、図１を参照すると、認識モデルは、単語モデル（辞書１３４及びサブ単語モデル１３２）と、大語彙のある特定部分に対する言語モデル（文法１４４及び意味論上の制約１４２）とからなる。当然ながら、普通、様々な認識モデル間に重複が存在してもよいし、存在するであろう。このような重複は、通常は語彙の一部に起こる。言語モデルは、部分的又は完全に同じでもよい。簡単なシステムにおいて、認識モデルの数は、認識装置の数と一致する。つまり、各認識装置は、排他的な認識モデルと固定した１対１の関係で関連付けられる。好ましくは、このシステムは、以下に詳細に説明されるように、アクティブな認識装置よりも多くのモデルを有する。当該図は８個のモデル３４１から３４８を示す。

【００２５】

前記認識装置の出力は、認識された単語列の最終的な選択を行うための制御器３５０に送られる。個々の認識装置３３１から３３３は、認識された単語列を一つだけ生成する。代わりとして、（例えば単語グラフで表示される）多重列が生成されてもよい。好ましくは、個々の認識装置の結果は、制御器３５０がほとんどの単語列を選択することが可能である、例えば尤度のような情報又は信頼手段(confidence measures)を含んでいる。この制御器３５０は、音声入力を認識装置に送ることも担っている。アクティブな認識装置の数が一定である場合、この送信は不変である。この場合、制御器３５０は、送信に関する特別なタスクを持たない。

【００２６】

好ましい実施例において、前記システムはアクティブな認識装置（Ｎ）よりも多くの認識モデル（Ｍ）を有する。モデル選択器３６０は、認識コンテクストに依存して、Ｍ個のモデルから関連する認識モデルを少なくとも１つの音声認識装置に対し選択するのに使用される。このモデル選択器３６０は、アクティブな認識装置の各々に対するモデルを選択してよい。しかしながら、共通に使用される語彙をカバーする基本的な認識モデルは、常にアクティブであることが好ましい。このような状況において、モデル選択器３６０によって少なくとも１つのモデルを選択する必要はなく、認識装置に安定して割り当てられる。

【００２７】

他の実施例において、少なくとも１つの認識モデルは、音声入力が関係する文書により決定されるコンテクストに基づいて選択される。例えば、ユーザが健康を項目とする文書を口述する場合、１つの認識装置は、健康に関する音声を認識するのに最適な特定の認識モデルでロードされる。ユーザは、この文書に関するコンテクストを、例えばシステムのモデルに対応する可能なコンテクストのリストから選択することで明確に示される。この場合、システム３００は、上記リストを従来のやり方、例えばウインドウの選択ボックスを用いてユーザに示す。このシステムは、例えば、文書内に既に存在する又はこれまでに話されたテキストを走査し、どのモデルが上記テキスト（例えばモデルがこれまでのテキス

トと同様に多くの単語又は単語列と持つテキスト）を認識するのに最適かをチェックすることで、コンテクストを自動的に決定してもよい。コンテクスト識別子は、文書と関連付けられてもよく、最適なモデルを決定するためにシステム３００により得られてもよい。好ましくは、ＨＴＭＬページのようなウェブページに関連する音声に対し、前記文書のコンテクストがこの文書に指定される又はこの文章と関連していることが好ましい。これは、タグの形態で行われ、このタグは、音声が関連する本来のウェブページの制作者により封入されている。このタグは、スポーツ、健康、エンターテイメント等のテキスト項目の形態でコンテクストを明確に示す。仕様書は、コンテクスト番号のような識別子、又はコンテクストを指定する場所へのリンク（例えばハイパーリンク）の形態でのような間接的でもよい。後者の場合、システム３００は、（例えばコンテクスト番号を認識モデルの１つにマッピング、すなわちハイパーテキストリンクにアクセスし、コンテクスト情報を得ることで）内在するコンテクストの仕様書から実際のコンテクストを得ることが可能である。

【００２８】

好ましい実施例において、モデル選択器３６０は、手近で認識に最適な利用可能な認識モデルがどれかをチェックすることで認識をアクティブに改善するように試みている。このために、モデル選択器３６０は、認識装置３３４で示されるように、少なくとも１つのテスト認識装置を制御する。このテスト認識装置３３４は、アクティブな認識装置３３１から３３３によりまだ使用されていない認識モデルの１つに結合される。入力された音声の一部（又は全て）は、前記テスト認識装置にも与えられる。このテスト認識装置の結果は、制御器３５０による選択の結果又は個々のアクティブな認識装置３３１から３３３の結果と比較される。テスト認識装置３３４の認識結果がアクティブな認識装置３３１から３３３の１つの認識結果よりも良好となる場合、テスト認識モデル（すなわち、テスト認識装置３３４によって使用される瞬間のモデル）は、アクティブな認識装置の１つにより使用するためにロードされる。好ましくは、最悪の認識結果を与えた認識モデルは、（おそらく、常に使用される基本認識モデルを除いて）置き換えられる。

【００２９】

認識モデルは、より一般的なコンテクストを備えるモデルからより特定のコンテクストを備えるモデルへ階層的に配されることが好ましい。図４は、例えばエンターテイメント、健康、旅行及びコンピュータの個々の一般項目をカバーする４つの最も一般的なモデル４１０，４２０，４３０及び４４０を備える上記階層を示す。一般的なモデルは、項目内の全ての発行物に関する表示テキストを分析することで構築される。このモデル自体、モデルがどのように表示テキストから構築されるかは十分知られている。健康の一般的なモデルは、例えば医薬品、手術、食品／ダイエット、病院／医療センターに関する下位階層（すなわちより特定のモデル）と関連付けられてもよい。これらモデルの各々は、これらのより特定の項目に関するテキストを用いて作られる。当該図において、モデル４２２は、病院／医療センターに関連する。このコンテクスト内において、モデル４２４が健康ファームをカバーする更なる再分割が行われてもよい。健康ファームに関する文書が典型的に周辺区域を記載するので、健康ファームに関するテキストを分析することによって、自動的にある旅行項目に関する音声を認識するのにも適した認識モデルが作られる。これは、カテゴリ旅行モデル内のモデル４３２より下位階層にあるモデルとして使用するのに適した同じモデルを作る。あるモデルでの認識が良好な認識結果を得る場合、モデル選択器３６０は、より特定のモデルでの認識を可能とするように動作する。上記のより特定のモデル（すなわち階層的下層のモデル）は、より一般的なモデルと置き換えて使用されてよい。それは、より一般的なモデルに加えて使用されてもよい。より特定のモデルでの追加の認識は、より一般的なモデルのみを生じさせ、より一般的なモデルと同じ階層レベルにおいて、他の階層的に関連しないモデルと十分比較されて実行することが好ましい。例えば、スポーツ及び健康モデルは、階層的に関連せず（両方とも最高位のレベルである）、スポーツモデルの使用がより良好な認識結果を与え、より特定のスポーツモデルが使用されてよい。より特定の健康モデルを使用する必要が無くなる。実際には、健康モデルの認識

結果が非常に不十分な場合、このモデルでの認識は、より特定のスポーツモデルを持つ追加の認識に有利となるように終わる。例えばフットボール、野球、アスレチック、カーレース等のような幾つかのより特定のスポーツモデルが存在する場合、これらモデルの全ては検査される。この選択は、単に、既に認識された音声と特定のモデルの語彙との一致にも基づいている。ある瞬間における特定のモデルでの認識が不十分な結果を与える場合、認識は、好ましくは特定のモデルより階層的に上位の少なくとも１つのモデルで継続される。

【００３０】

図５に示されるような好ましい実施例において、認識システムが分散されている。この分散されたシステムは、サーバステーション５４０と、少なくとも１つのユーザステーションを有する。３つのユーザステーション５１０，５２０及び５３０が示され、ユーザステーション５２０にのみ、細部が示されている。これらステーションは、従来のコンピュータ技術を用いて実施される。例えば、ユーザステーション５２０は、デスクトップ型パーソナルコンピュータ又はワークステーションにより形成されるのに対し、サーバステーション５４０はＰＣサーバ又はワークステーションサーバにより形成される。これらコンピュータは、コンピュータの処理器にロードされた最適なプログラムの制御下で動作する。サーバステーション５４０及びユーザステーション５１０，５２０，５３０は、ネットワーク５５０を介して接続されている。このネットワーク５５０は、適切なネットワーク、例えばオフィス環境におけるローカルエリアネットワーク又は好ましくはインターネットであるワイドエリアネットワークでもよい。これらステーションは、ネットワーク５５０を介して通信するための通信手段５２２及び５４２をそれぞれ有する。ネットワーク５５０と組み合わせて使用する如何なる通信手段が使用されてもよい。典型的に、これら通信手段は、通信インタフェース又はモデムのようなハードウェアと、インターネットのＴＣＰ／ＩＰプロトコルのような特定の通信プロトコルをサポートするソフトウェアドライバのソフトウェアとの組合せにより形成される。ユーザステーション５２０は、例えばインタフェースを介してユーザから音声を入力する手段を有する。ユーザステーション５２０は、サーバステーション５４０に転送するのに適した音声信号を事前処理する手段を更に有する。例えば、ユーザステーションは、図１のスペクトル分析サブシステム１１０に類似のスペクトル分析サブシステム５２６を有する。サーバステーション５４０は、図３のシステム３００に記載される全ての他のタスクを実行する。例えば、サーバ５４０は、複数の（図３の認識システム３３５に類似の）認識装置を具備する認識システム５４３と、（図３の制御器３５０に類似の）制御器５４４、（図３の選択器３６０に類似の）モデル選択器５４５及び（図３の記憶装置３４０に類似の）モデルを記憶する記憶装置５４６を有する。

【図面の簡単な説明】

【図１】　図１は、多／大語彙認識装置の構造を示す。

【図２】　図２は、完全な単語モデル図を示す。

【図３】　図３は、本発明に係るシステムのブロック図を示す。

【図４】　図４は、認識モデルの階層図を示す。

【図５】　図５は、本発明に係る分配システムのブロック図を示す。

【図１】

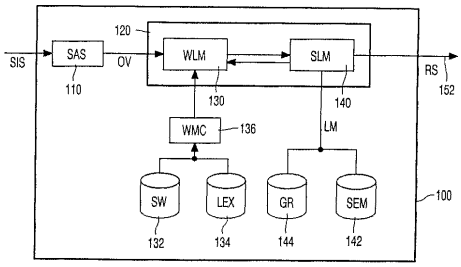

FIG. 1

【図２】

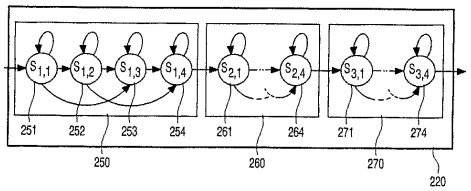

FIG. 2

【図３】

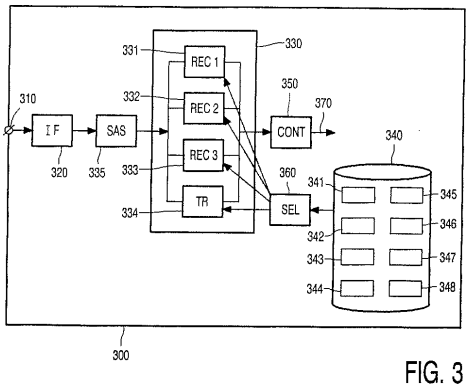

FIG. 3

【図４】

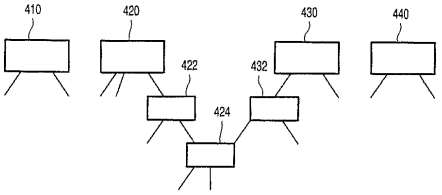

FIG. 4

【図５】

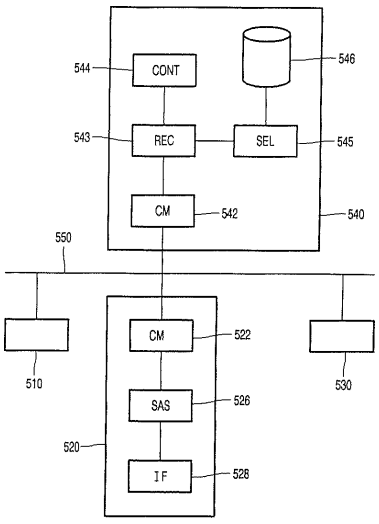

FIG. 5

フロントページの続き

(72)発明者　ベスリン　ステファン
オランダ国　５６５６　アーアー　アインドーフェン　プロフ　ホルストラーン　６
(72)発明者　ウルリッチ　メインハード
オランダ国　５６５６　アーアー　アインドーフェン　プロフ　ホルストラーン　６

審査官　井上　健一

(56)参考文献　米国特許第７２８６９８９（ＵＳ，Ｂ１）
特開平８－３１４４９６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G10L 15/28