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

特許第4528535号
(P4528535)

(45)発行日　平成22年8月18日(2010.8.18)　(24)登録日　平成22年6月11日(2010.6.11)

(51)Int.Cl.　F I
G10L 15/06 (2006.01)　G10L 15/06 300C
G10L 15/18 (2006.01)　G10L 15/18 300G

請求項の数 10　外国語出願　(全 17 頁)

(21)出願番号　特願2004-37147(P2004-37147)
(22)出願日　平成16年2月13日(2004.2.13)
(65)公開番号　特開2004-246368(P2004-246368A)
(43)公開日　平成16年9月2日(2004.9.2)
審査請求日　平成19年1月12日(2007.1.12)
(31)優先権主張番号　10/365,850
(32)優先日　平成15年2月13日(2003.2.13)
(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　ミリンド　マハジャン
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ノースイースト　９７　ウェイ　１７４３０

最終頁に続く

(54)【発明の名称】テキストから単語誤り率を予測するための方法および装置

(57)【特許請求の範囲】
【請求項１】
実際の音声単位シーケンスを含む学習テキストから音声信号を生成するステップと、
生成された音声信号をデコードするステップと、
デコードされた音声信号を使用して、予測される音声単位シーケンスを生成するステップと、
前記学習テキストに含まれる実際の音声単位シーケンスおよび前記生成された予測される音声単位シーケンスを使用して、音声認識システムのパフォーマンスを識別するためのコンフュージョンモデルを構築するステップと、
前記コンフュージョンモデルを使用して、あるテキストが与えられた場合の音声認識システムの誤り率を識別するステップと
を含むことを特徴とする音声認識システムをモデル化する方法。
【請求項２】
請求項１に記載の方法であって、
前記コンフュージョンモデルおよび言語モデルを使用してテストテキストをデコードするステップと、
前記コンフュージョンモデルを使用して、前記予測される音声単位シーケンスを表す少なくとも１組のパスを表すテストテキスト用のネットワークを構築するステップと、
構築されたネットワークをスキャンしてパスを探索することで、予測される単語シーケンスを生成するステップと、をさらに備え、

前記誤り率を識別するステップは、以下の式１に従って予測される単語シーケンスに確率を割り当てることを特徴とする方法。

【数１】

$$p(W_p|W_c)=\frac{P(W_p)}{\sum_{W_{p'}:\varphi(W_{p'})=\varphi(W_p)}P(W_{p'})}\cdot\sum_{\bar{t}:\varphi(\bar{t})=\varphi(W_p)}p(\bar{t}|\varphi_c) \qquad \text{式1}$$

ただし、式中ｐ（Ｗｐ｜Ｗｃ）は、実際の単語シーケンスＷｃが与えられた場合の予測される単語シーケンスＷｐの確率、Ｐ（Ｗｐ）は予測される単語シーケンスの言語モデル確率、分母の総和は、予測される単語シーケンスＷｐと同じ音声単位シーケンスを有する予測されるすべての単語シーケンスのすべての言語モデル確率の合計、

【数２】

$p(\bar{t}|\varphi_c)$

はネットワークを通るパス

【数３】

$\bar{t}$

に沿った全確率であり、

【数４】

$\varphi(W_p)$

は、単語シーケンスＷｐに対応する音声単位シーケンスを表し、

【数５】

$\varphi(\bar{t})$

は、パス

【数６】

$\bar{t}$

に沿った予測される音声単位シーケンスを表す。

【請求項３】

請求項２に記載の方法であって、

前記誤り率を識別するステップは、予測される単語シーケンス内の単語と、実際の単語シーケンス内の単語との間の差を識別することで、予測される単語シーケンスと実際の単語シーケンスとの間の誤差を決定することを特徴とする方法。

【請求項４】

請求項３に記載の方法であって、

前記誤り率を識別するステップは、前記決定した誤差を以下の式２に適用して単語誤り率の期待値を算出することを特徴とする方法。

【数７】

$$E[WER]=\frac{\sum_{i=1}^{I}E[e_i]}{\sum_{i=1}^{I}N_i} \qquad \text{式2}$$

ただし、式中Ｅ［ＷＥＲ］は、テストテキストの単語誤り率の期待値、Ｅ［ｅ$_i$］はテストテキスト内のシーケンスｉでの誤差数の期待値、Ｎ$_i$は実際のシーケンスｉ内の単語数、およびＩはテストテキスト内の文の総数である。

【請求項５】

前記テストテキストをデコードするステップは、

前記第１の言語モデルおよび前記コンフュージョンモデルを使用して前記テストテキストをデコードし、予測される第１の組の音声単位を生成するステップと、
前記第１の言語モデルと異なる第２の言語モデルおよび前記コンフュージョンモデルを使用して前記テストテキストをデコードし、予測される第２の組の音声単位を生成するステップと、を備え、
前記誤り率を識別するステップは、生成された予測される第１および第２の組の音声単位を使用して前記誤り率をそれぞれ識別し、識別された誤り率を比較することで、前記第１の言語モデルのパフォーマンスを前記第２の言語モデルのパフォーマンスと比較するステップを備える
ことを特徴とする請求項２に記載の方法。

【請求項６】
前記音声信号をデコードするステップは、学習言語モデルを使用して前記音声信号をデコードすることを含み、前記学習言語モデルは、前記テストテキストのデコードに使用する前記言語モデルと異なる機能を備えることを特徴とする請求項２に記載の方法。

【請求項７】
前記誤り率を使用して前記言語モデルの学習を行うステップをさらに備えることを特徴とする請求項１に記載の方法。

【請求項８】
前記コンフュージョンモデルは、実際の音声単位から少なくとも１つの予測される音声単位への変換、およびこうした変換を行う確率をそれぞれ提供する１組のルールを含むことを特徴とする請求項１に記載の方法。

【請求項９】
前記ルールは、実際のシーケンス内の１の音声単位の右および／または左の文脈に依存するルールを含むことを特徴とする請求項８に記載の方法。

【請求項１０】
前記コンフュージョンモデルを構築するステップは、
学習テキストから生成された音声信号をデコードし、予測される音声単位シーケンスを生成すること、
前記学習テキスト内の学習音声単位のシーケンスを識別すること、
前記予測される音声単位シーケンスを前記学習音声単位のシーケンスに合わせて配列すること、および
前記配列されたシーケンスを使用して前記コンフュージョンモデルの学習を行うこと
を含むことを特徴とする請求項１に記載の方法。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は音声認識に関する。より詳細には、本発明は音声認識システムのパフォーマンスのモデル化に関する。

【背景技術】
【０００２】
音声認識では、音響モデルおよび言語モデルを使用して音響信号を単語シーケンスに変換する。音響モデルは、音響信号の特徴を、確率を備える単音などの可能なサブワード音声単位（ｓｕｂ－ｗｏｒｄ　ｓｐｅｅｃｈ　ｕｎｉｔ）シーケンスに変換する。言語モデルは、音響モデルによって識別された音のシーケンスから形成することができる様々な単語シーケンスの確率分布を提供する。

【０００３】
音響モデルの学習は、一般に、話者に既知のテキストを読ませ、次いで学習音声から学習テキストを予測するように音響モデルを作り上げることによって行われる。言語モデルの学習は、一般に、テキストの大きいコーパスから、コーパス内の様々な単語シーケンスの確率を単に識別することによって行われる。

【発明の開示】
【発明が解決しようとする課題】
【０００４】
得られる音声認識システムのパフォーマンスは、音響モデルおよび言語モデルの学習に使用される学習テキストにある程度依存する。その結果、音声認識システムは、一部のタスクドメインにおいて、他のタスクドメインにおけるより良好に機能する。音声認識システムが特定のタスクドメインでどのように動作するかを決定するために、タスクの実行時にユーザが使用すると思われる単語を誰かが話して、システムによってデコード可能な音響データを生成する必要がある。音声認識システムのパフォーマンスを決定するのに十分な量の音響データを生成するために人を雇うことは、費用がかかり、音声で実行可能になるコンピュータアプリケーションの開発の障壁となっている。
【０００５】
さらに、音響データを生成することは費用がかかるため、こうしたデータは、言語モデルの学習に使用するコーパス全体については生成されていなかった。その結果、言語モデルの学習は一般に、音響モデルが言語モデルコーパスに対してどのように機能するかを検査することなく行われていた。したがって、音響データを必要とすることなく音響モデルと言語モデルの組合せのパフォーマンスの測定にテキストのコーパスを使用できるようにするシステムを得ることが有益となる。これによって、音響モデルとの組合せで言語モデルの識別学習（ｄｅｓｃｒｉｍｉｎａｔｉｖｅ　ｔｒａｉｎｉｎｇ）が可能となる。
【課題を解決するための手段】
【０００６】
音声認識システムをモデル化する方法は、学習テキストから生成された音声信号をデコードして予測される音声単位シーケンスを生成することを含む。学習テキストは実際の音声単位シーケンスを含んでおり、これを予測される音声単位シーケンスとともに使用してコンフュージョンモデル（ｃｏｎｆｕｓｉｏｎ　ｍｏｄｅｌ）が形成される。別の実施形態では、コンフュージョンモデルを使用してテキストをデコードして、音声認識システムがテキストに基づいて音声をデコードした場合に予想される誤り率を識別する。
【発明を実施するための最良の形態】
【０００７】
図１は、本発明を実施するのに適したコンピューティングシステム環境１００の例を示している。コンピューティングシステム環境１００は、適したコンピューティング環境の一例にすぎず、本発明の使用または機能の範囲に関する限定を示唆するものではない。また、コンピューティング環境１００を、動作環境１００の例に示した構成要素のいずれか１つ、またはその組合せに関連する依存性または必要条件を有しているものと解釈すべきではない。
【０００８】
本発明は、他の多くの汎用または専用コンピューティングシステム環境または構成で動作可能である。本発明との使用に適したよく知られている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、それだけには限定されないが、パーソナルコンピュータ、サーバーコンピュータ、ハンドヘルドまたはラップトップ装置、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、プログラム可能な家庭用電化製品、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、テレフォニーシステム、上記の任意のシステムまたは装置を含む分散コンピューティング環境などがある。
【０００９】
本発明は、コンピュータによって実行される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な文脈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する、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、構成要素、データ構造などを含む。本発明は、タスクが通信ネットワークによってリンクされているリモート処理装置によって実行される分散コンピューティング

環境で実施するように設計されている。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を、メモリ記憶装置を含むローカルおよびリモートのコンピュータ記憶媒体に置くことができる。

【００１０】

図１を参照すると、本発明を実施するシステムの例は、汎用コンピューティング装置をコンピュータ１１０の形で含んでいる。コンピュータ１１０の構成要素は、それだけには限定されないが、処理装置１２０、システムメモリ１３０、およびシステムメモリを含む様々なシステム構成要素を処理装置１２０に結合するシステムバス１２１を含む。システムバス１２１は、様々なバスアーキテクチャのうちの任意のものを使用するメモリバスまたはメモリコントローラ、周辺機器バス、およびローカルバスを含むいくつかのタイプのバス構造のうちどんなものでもよい。こうしたアーキテクチャには、それだけには限定されないが一例として、業界標準アーキテクチャ（ＩＳＡ）バス、マイクロチャネルアーキテクチャ（ＭＣＡ）バス、拡張ＩＳＡ（ＥＩＳＡ）バス、ビデオ電子装置規格化協会（ＶＥＳＡ）ローカルバス、およびメザニンバスとしても知られている周辺機器相互接続（ＰＣＩ）バスなどがある。

【００１１】

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からアクセスできる使用可能な任意の媒体とすることができ、揮発性および不揮発性媒体、リムーバブルおよび非リムーバブル媒体を含む。コンピュータ可読媒体は、それだけには限定されないが一例として、コンピュータ記憶媒体および通信媒体を含み得る。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、他のデータなど、情報を記憶するための任意の方法または技術で実施される揮発性および不揮発性のリムーバブルおよび非リムーバブル媒体がある。コンピュータ記憶媒体には、それだけには限定されないが、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリまたは他のメモリ技術、ＣＤ－ＲＯＭ、デジタル多用途ディスク（ＤＶＤ）または他の光ディスク記憶装置、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、または所望の情報の格納に使用でき、コンピュータ１１０からアクセスできる他の任意の媒体などがある。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変調されたデータ信号に組み込む。これには任意の情報配送媒体がある。「変調されたデータ信号」という用語は、１つまたは複数のその特徴が、信号内の情報を符号化するように設定または変更された信号を意味する。通信媒体には、それだけには限定されないが一例として、有線ネットワーク、直接配線された接続などの有線媒体、および音響、ＲＦ、赤外線、その他の無線媒体などの無線媒体がある。また、上記のどんな組合せでもコンピュータ可読媒体の範囲内に含まれるものとする。

【００１２】

システムメモリ１３０は、読取り専用メモリ（ＲＯＭ）１３１やランダムアクセスメモリ（ＲＡＭ）１３２など、揮発性および／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。基本入出力システム１３３（ＢＩＯＳ）は、例えば起動中など、コンピュータ１１０内の要素間での情報の転送を助ける基本ルーチンを含み、一般にＲＯＭ１３１に格納されている。ＲＡＭ１３２は一般に、処理装置１２０から直接アクセス可能な、かつ／または処理装置１２０が現在処理しているデータおよび／またはプログラムモジュールを含む。図１は、それだけには限定されないが一例として、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を示している。

【００１３】

コンピュータ１１０は、他のリムーバブル／非リムーバブル、揮発性／不揮発性コンピュータ記憶媒体を含むこともできる。一例にすぎないが、図１は、非リムーバブル不揮発性磁気媒体から読み取り、あるいはそこに書き込むハードディスクドライブ１４１、リムーバブル不揮発性磁気ディスク１５２から読み取り、あるいはそこに書き込む磁気ディス

クドライブ１５１、およびＣＤ－ＲＯＭや他の光媒体など、リムーバブル不揮発性光ディスク１５６から読み取り、あるいはそこに書き込む光ディスクドライブ１５５を示している。動作環境の例で使用できる他のリムーバブル／非リムーバブル、揮発性／不揮発性コンピュータ記憶媒体には、それだけには限定されないが、磁気テープカセット、フラッシュメモリカード、デジタル多用途ディスク、デジタルビデオテープ、半導体ＲＡＭ、半導体ＲＯＭなどがある。ハードディスクドライブ１４１は一般に、インターフェイス１４０などの非リムーバブルメモリインターフェイ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、インターフェイス１５０などのリムーバブルメモリインターフェイスによってシステムバス１２１に接続される。

【００１４】

上述し、図１に示したドライブおよびその関連のコンピュータ記憶媒体は、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、およびコンピュータ１１０の他のデータの記憶を提供する。図１では例えば、ハードディスクドライブ１４１は、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を記憶するものとして示されている。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注意されたい。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は少なくとも異なるコピーであることを示すために、ここではそれらに異なる番号を付している。

【００１５】

ユーザは、キーボード１６２、マイクロフォン１６３、および一般にマウス、トラックボール、タッチパッドなどのポインティング装置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、衛星放送受信アンテナ、スキャナなどがある。これらおよび他の入力装置は、しばしばシステムバスに結合されているユーザ入力インターフェイス１６０を介して処理装置１２０に接続されるが、パラレルポート、ゲームポート、ユニバーサルシリアルバス（ＵＳＢ）など他のインターフェイスおよびバス構造で接続してもよい。モニタ１９１または他のタイプの表示装置もまた、ビデオインターフェイス１９０などのインターフェイ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、出力周辺インターフェイス１９５などを介して接続できるスピーカー１９７、プリンタ１９６などの他の周辺出力装置を含むこともできる。

【００１６】

コンピュータ１１０は、リモートコンピュータ１８０など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式環境で操作される。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、サーバー、ルーター、ネットワークＰＣ、ピア装置、または他の一般のネットワークノードでよく、一般にコンピュータ１１０に関連して上述した多くまたはすべての要素を含む。図１に示した論理接続は、ローカルエリアネットワーク（ＬＡＮ）１７１および広域エリアネットワーク（ＷＡＮ）１７３を含むが、他のネットワークを含んでいてもよい。こうしたネットワーキング環境は、オフィス、全社規模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はごく一般的である。

【００１７】

ＬＡＮネットワーキング環境で使用する場合、コンピュータ１１０は、ネットワークインターフェイスまたはアダプタ１７０を介してＬＡＮ１７１に接続される。ＷＡＮネットワーキング環境で使用する場合、コンピュータ１１０は一般に、モデム１７２、またはインターネットなどＷＡＮ１７３にわたって通信を確立する他の手段を含む。モデム１７２は、内蔵のものでも外付けのものでもよく、ユーザ入力インターフェイス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。ネットワーク式環境では、コンピュータ１１０に関連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に格納することができる。図１は、それだけには限定されないが一例として、リモートアプリケーションプログラム１８５をメモリコンピュータ１８０上に存在するものとして示している。図示したネットワーク接続は例であり、コンピュータ間の通信リンクを確立する他の手段を使用してもよいことは理解されよう。

【００１８】

本発明では、音声認識システムのパフォーマンスのモデルを構築し、それを使用して、あるテキストが与えられた場合の音声認識システムの見込み誤り率（ｌｉｋｅｌｙ ｅｒｒｏｒ ｒａｔｅ）を識別する。以下の説明で、このモデルをコンフュージョンモデルと呼ぶ。コンフュージョンモデルを使用することによって、特定のテキストから生成された音声をデコードする際に音声認識システムがどのように機能するかを推定するのに音響データは必要なくなる。

【００１９】

図２は、本発明によるコンフュージョンモデルを構築し、使用する方法を示すフロー図である。図２のステップ２００で、音響モデル３００の学習が行われる。これには、話者３０８に学習テキスト３０４の一部を朗読させて音響信号を生成することが必要となる。この音響信号は受信機３０９によって検出され、電子信号に変換される。電子信号は、特徴抽出機に提供され、そこで電子信号から１つまたは複数の特徴が抽出される。こうした特徴ベクトルを識別する方法は、当分野ではよく知られており、これには、特徴ベクトルごとに１３のケプストラム値を生成する１３次元メル周波数ケプストラム係数（ＭＦＣＣ：ＭｅｌＦｒｅｑｕｅｎｃｙ Ｃｅｐｓｔｒｕｍ Ｃｏｅｆｆｉｃｉｅｎｔｓ）抽出などがある。ケプストラル特徴ベクトルは、電子信号の対応するフレーム内にある音声信号のスペクトルの内容を表す。抽出された特徴ベクトルは、トレーナ３０２に提供され、トレーナはその特徴を使用して音響モデル３００の学習を行う。音響モデルを構築する方法は、当分野ではよく知られている。音響モデルは任意の所望の形態でよく、これには、それだけには限定されないが、隠れマルコフモデルなどがある。

【００２０】

ステップ２０２で、話者３０８が学習テキスト３０４の一部を話して、学習が行われた音響モデルを使用してデコードされるテスト信号が生成される。音響信号は、受信機３０９および特徴抽出機３１０によって特徴ベクトルに変換され、特徴ベクトルは、デコーダ３１２に提供される。

【００２１】

デコーダ３１２は、音響モデル３００、辞書３１５、および学習言語モデル３１４を使用して、特徴を、予測される音声単位シーケンスにデコードする。多くの実施形態では、音声単位とは、単音、２音、３音など音の単位（ｐｈｏｎｅｔｉｃ ｕｎｉｔ）である。

【００２２】

ステップ２００および２０２は、１個抜き（ｌｅａｖｅ－ｏｎｅ－ｏｕｔ）手法を使用して行うことができ、学習テキスト３０４が複数の組の学習データに分割されることに注意されたい。１組を除く全部を使用して音響モデルが構築され、モデルの構築に使用されない組は、音響モデルおよび言語モデルを使用してデコードされる。次いでステップ２００および２０２が繰り返され、デコードすべき組として、学習データから別の１組が選択され、残りの組から音響モデルが構築される。これは学習テキスト３０４内のデータの組ごとに繰り返され、その結果学習テキスト３０４内のデータの組ごとに個々の組の予測される音声単位が提供される。

【００２３】

学習データのデコードに使用する言語モデルの学習が学習データを使用して行われていれば、同様の１個抜き手順を使用して言語モデルの学習を行う必要がある。１個抜き手法は、予測される音声単位における偏りを避けるのに有益である。

【００２４】

ステップ２０４で、予測される音声単位シーケンスが、学習テキスト３０４からの実際の音声単位シーケンスに合わせて配列される。これは、図３の配列モジュール３１６によって行われる。一実施形態では、この配列は動的プログラミングを使用して行われ、可能な配列が互いに比較され、ある目的関数の最適化に基づいて可能な配列の最適なものが選択される。一実施形態では、この目的関数は、配列が終了した後で予測される音声単位シーケンスが実際の音声単位シーケンスと異なる度合いを示す誤差関数である。一部の実施形態では、誤差関数は簡単な二値関数であり、並べられた２つの音声単位が一致しない場合は誤差値１が生成され、２つの音声単位が一致する場合は誤差値０が生成される。別の実施形態では、異なるタイプの誤差に異なる重みが付加される。例えば、似た音を提供する音声単位は、大幅に異なる音を提供する音声単位より誤差値が小さい可能性がある。

【００２５】

予測されるシーケンスを実際の音声単位シーケンスに合わせて配列するステップは、コンフュージョンモデルが、学習が可能となる前に、並べられた音声単位を必要とする場合にのみ、ステップ２０４で行われる。隠れマルコフモデルなど、いくつかのタイプのコンフュージョンモデルでは、こうした配列は不要である。というのは、モデルの学習には、音声単位を配列するステップが本来含まれているからである。

【００２６】

ステップ２０６で、コンフュージョンモデルトレーナ３１８によってコンフュージョンモデル３２０が構築される。一実施形態では、コンフュージョンモデル３２０として隠れマルコフモデルが構築される。

【００２７】

図４は、本発明のコンフュージョンモデルの一実施形態による４状態隠れマルコフモデルを示す。本発明のこの実施形態では、個々の隠れマルコフモデルが実際の音声単位ごとに構築される。各隠れマルコフモデルは、その特定の実際の音声単位から生成できる予測される音声単位を出力として生成する。

【００２８】

４状態モデルには、状態４００、４０２、４０４、および４０６がある。コンフュージョンモデルでは、状態４００および４０６は、非生成状態（ｎｏｎ－ｅｍｉｔｔｉｎｇ ｓｔａｔｅ）であり、こうした状態を離れる遷移からは、予測される音声単位は生成されない。一方、状態４０２および４０４を離れる遷移からは、こうした遷移が使用されるたびに予測される音声単位が１つ生成される。状態４０２および４０４では、可能な予測される各音声単位を生成する個々の出力確率が存在する。

【００２９】

これらの状態は、関連する確率を有する遷移によって連結される。ある状態を離れる遷移の確率の合計は１に等しい。したがって、遷移４０８および４１０の遷移確率の合計は１に等しく、また遷移４１２および４１４の確率の合計は１に等しい。

【００３０】

図４のモデルを通る様々なパスは、様々な置換および挿入の可能性を表す。例えば、遷移４１０に沿った状態４００から状態４０６へのパスは、実際の音声単位の代わりに置き換えられる音声単位が予測される音声単位シーケンス内にないことを表す。遷移４０８および４１４に沿った、状態４００から４０２を通る４０６までのパスは、実際の音声単位が予測される音声単位１つに置き換えられることを表す。このパスは、実際の音声単位が予測される音声単位と同じである状況を含む。遷移４０８、４１２、および４１８に沿った、状態４００から状態４０２および４０４を通る４０６までのパスは、実際の音声単位が予測される音声単位２つに置き換えられることを表す。予測される音声単位の一方は状態４０２からの遷移で生成され、第２の音声単位は状態４０４からの遷移で生成される。

【００３１】

状態４０４は、実際の音声単位の代わりに任意の数の予測される音声単位を得る機構を提供する自己遷移４１６を含む。

【００３２】

図４のモデルなどの各隠れマルコフモデルの学習は、まず、状態４０２および４０４における予測される音声単位ごとに等しい確率を想定し、遷移ごとに等しい遷移確率を想定することによって行われる。この簡単なモデルを使用して、実際の音声単位と予測される音声単位の間の最適な配列を識別し、それによってＨＭＭごとに状態確率および遷移確率の学習に使用できる配列が提供される。更新されたモデルで、配列が調整され、新しい配列を使用してモデルの再学習が行われる。これはモデルが安定するまで続けられる。この学習手順は、ビタビ学習としても知られている。この学習プロセスのわずかに異なるバリエーションでは、実際の音声単位と予測される音声単位の間の複数の配列を考えることができる。これは、Ｆｏｒｗａｒｄ－Ｂａｃｋｗａｒｄ学習またはＢａｕｍ－Ｗｅｌｃｈ学習として知られている。

【００３３】

他の実施形態では、コンフュージョンモデルは、各ルールが実際の音声単位シーケンス内の単一の音声単位から予測される音声単位シーケンス内の０個、１個、または複数個の音声単位への変換の可能性を提供するルールベースのモデルとして構築される。本発明の実施形態では、各ルールは、文脈とは無関係とすることも文脈に依存することもできる。文脈依存型ルールでは、ルールは、実際のシーケンス内の単一の音声単位の左、右、または左右の文脈に依存することができる。さらに、本発明では、単一の音声単位の左または右の任意の数の音声単位を使用することができ、異なる文脈長さの組合せをいっしょに使用して、一方のルールが一方の文脈長さを使用し、もう一方のルールがより長い文脈長さを使用する状態で、異なる２つのルールが実際の音声単位を予測される音声単位シーケンスに変換する異なる２つの確率を提供できるようにすることができる。内挿法により確率を組み合わせる、または文脈に基づいて単一のルール確率を選択することができる。

【００３４】

ルール確率の学習を行うには、実際の音声単位シーケンスと予測される音声単位シーケンスの間の配列が検査されて、実際の音声単位を予測される音声単位シーケンスに変換するために各ルールを使用できる回数を決定する。この数を、実際の音声単位シーケンス内の単一の音声単位がルールに関連する特定の文脈内に見つかる回数で割る。したがって、各確率は、実際の音声単位、および必要に応じて実際の音声単位シーケンス内の特定の文脈が与えられた場合に予測される音声単位シーケンスを生成する尤度を示す。

【００３５】

本発明の一実施形態では、実際の音声単位の置換を必要とすることなく音声単位を予測されるシーケンスに挿入することができる。これは、配列前の実際の音声単位シーケンス内の各音声単位の間に空の音声単位εを挿入することによって達成される。配列中、予測されるシーケンス内には、こうした空の音声単位の相手として配列されるものがない。しかし、ε音声単位は、時々、予測されるシーケンス内の１つまたは複数の音声単位に合わせて配列されることがある。その結果、空の音声単位から配列内で見つかる予測される音声単位に変換するルールが生成される。

【００３６】

本発明の一実施形態では、ルールごとの確率が生成された後、一部のルールが削除されてルールセット内のルールの数を低減させ、それによってデコードが簡素化される。この削除は、文脈および単一の音声単位が実際の音声単位シーケンス内で見つかる回数に基づいて、または文脈および実際の音声単位が予測される音声単位の特定のシーケンスを生成する回数に基づいて行うことができる。後者の場合、削除されたが、別の予測される音声単位シーケンスを提供したルールと同じ文脈を共有するルールの確率を、同じ文脈および実際の音声単位での確率の合計が１に等しくなるように再計算する必要がある。

【００３７】

ステップ２０６でコンフュージョンモデルが構築された後、これを使用して、音響信号の生成の必要なく音響モデルのパフォーマンスをモデル化することができる。ステップ２０８で、デコーダ５０２は、コンフュージョンモデル５０４、辞書５０８、および言語モデル５０６を使用して、図５のテストテキスト５００をデコードする。デコーダ５０２は

、１組の予測される単語シーケンスを、コンフュージョンモデル５０４、辞書５０８、および言語モデル５０６を使用して計算されたシーケンスごとの確率とともに生成する。また、デコーダ５０２は、テストテキスト５００内の実際の単語シーケンスを渡す。

【００３８】

デコーダ５０２が使用する言語モデル５０６は、ほとんどの実施形態では、学習言語モデル３１４とは異なる。一般に、学習言語モデル３１４は脆弱な言語モデルとなるように選択されており、したがってコンフュージョンモデルは、音響モデル３００の弱点を示す。一方、言語モデル５０６は、音声認識システムで使用される言語モデルとより類似した強力な言語モデルである。一部の実施形態では、異なる言語モデルを同じコンフュージョンモデル５０４およびテストテキスト５００とともに使用して、算出された単語誤り率に基づいて２つの言語モデルの相対的パフォーマンスを決定する。これについて以下でさらに詳しく論じる。

【００３９】

ルールベースのコンフュージョンモデルでは、デコーダ５０２は、ルールをコンフュージョンモデル５０４に適用して、図６のパスなど、可能な予測される音声単位シーケンスを表す１組のパスを表すネットワークを構築する。図６には、状態６００、６０２、６０４、６０６、６０８、および６１０の間に遷移を示している。任意の２つの状態の間の遷移は、テストテキスト５００で表される実際の音声単位シーケンス内の特定の音声単位の代わりに使用されるある確率を有する予測される音声単位シーケンスを表す。例えば、状態６０２と６０４の間の遷移は、音声単位Ｐ１の代わりに使用できる予測されるすべての音声単位シーケンスを表す。同様に、状態６０４と６０６の間の遷移は、実際の音声単位シーケンスの音声単位Ｐ１と音声単位Ｐ２の間に挿入することができるすべての遷移を表す。遷移は、予測される音声単位シーケンス内に音声単位がない場合も提供できることに注意されたい。したがって、状態６０２と６０４の間の遷移の１つは、音声単位がないことを表すεを含み得る。これは、予測される音声単位シーケンス内から実際の音声単位Ｐ１が単に除去されることを示す。ネットワークは、空の音声単位εに対応する状態および遷移も含む。これらは、図６に示すように、状態６００と６０２、状態６０４と６０６の間に、実際の音声単位とともに交互の位置に挿入される。これらの遷移は、空の音声単位εに対応する挿入ルールに対応する。

【００４０】

図６では、各遷移は特定の確率を有している。このテストテキスト用に生成されたネットワークから予測される単語シーケンスを生成するために、デコーダは、ネットワークをスキャンすることによってネットワーク内のパスを探索する。ここでは、深さ優先探索や幅優先探索などの標準グラフスキャン戦略（ｇｒａｐｈ　ｔｒａｖｅｒｓａｌ　ｓｔｒａｔｅｇｉｅｓ）を使用することができる。デコーダが探索した各パスは、予測される音声単位シーケンスに対応する。

【００４１】

効率を高め、デコーダが探索するパスの数を制限するために、デコーダは、スキャン中に生成された部分パスを削除する（すなわちそれ以上の考察から除外する）ことを選択することができる。例えば、部分パスが、任意の単語シーケンスに対応する任意の音声単位シーケンスの先頭部分に一致していない予測される音声単位シーケンスに対応している場合、こうした部分パスを削除することができる。他の部分パスまたは完全パスに比べて低い確率を有する部分パスも削除することができる。

【００４２】

削除についてさらに説明するために、デコーダは、まず実際の音声単位シーケンスと同じ予測される音声単位シーケンスを生成してもよい。こうしたシーケンスは一般に、高い確率を有しており、低い確率の他のパスの識別を助ける。

【００４３】

デコーダ５０２は、特定のパスに沿って終了状態６１２に到達するとき、探索されたパスに対応する予測される音声単位シーケンスを、その関連する確率とともに使用すること

ができる。次いでデコーダは、単語単位シーケンスに対応する音声単位シーケンスを予測される音声単位シーケンスと照合することによって、予測される音声単位シーケンスを生成することができるすべての単語シーケンスを識別する。

【００４４】

デコーダは、ネットワーク内のすべてのパスを探索した後、下記の式に従って予測される単語シーケンスに確率を割り当てる。説明を簡潔にするために、式の導出の際に単純化するための仮説をいくつか立てたことに注意されたい。例えば、単語ごとに単一の発音（すなわち１つの音声単位シーケンス）があると仮定した。

【００４５】

【数１】

$$p(W_p|W_c)=\frac{P(W_p)}{\sum_{W_{p'}:\varphi(W_{p'})=\varphi(W_p)}P(W_{p'})}\cdot\sum_{\bar{t}:\varphi(\bar{t})=\varphi(W_p)}p(\bar{t}|\varphi_c)$$ 式1

【００４６】

式中ｐ（Ｗ$_{p}$｜Ｗ$_{c}$）は、実際の単語シーケンスＷ$_{c}$が与えられた場合の予測される単語シーケンスＷ$_{p}$の確率、Ｐ（Ｗ$_{p}$）は予測される単語シーケンスの言語モデル確率、分母の総和は、予測される単語シーケンスＷ$_{p}$と同じ音声単位シーケンスを有する予測されるすべての単語シーケンスのすべての言語モデル確率の合計、および

【００４７】

【数２】

$p(\bar{t}|\varphi_c)$

【００４８】

はネットワークを通るパス

【００４９】

【数３】

$\bar{t}$

【００５０】

に沿った全確率であり、Ｗ$_{p}$のシーケンスと同じ音声単位シーケンスを生成するすべてのパスについて和を取る。

【００５１】

【数４】

$\varphi(W_p)$

【００５２】

は、単語シーケンスＷ$_{p}$に対応する音声単位シーケンスを表し、

【００５３】

【数５】

$\varphi(\bar{t})$

【００５４】

は、パス

【００５５】

【数６】

$\bar{t}$

【００５６】

に沿った予測される音声単位シーケンスを表す。パスごとの確率

【００５７】
【数７】

$$p(\bar{t}|\varphi_c)$$

【００５８】
は、そのパスに沿った各遷移に関連する個々の確率の積として求められる。デコーダのトレリスを通る異なるパスは、ε遷移のために、同じ予測される音声単位シーケンスを有し得ることに注意されたい。ε遷移は、パスの類似性を決定するために音声単位として見なされない。例えば、シーケンスｔ－ε－ｉｙは、シーケンスｔ－ｉｙ－εと同じであると見なされることになる。というのは、シーケンスの類似性を決定する際にε音声単位は無視されるからである。
【００５９】
予測される単語シーケンスごとの確率が決定された後、ステップ２１０で、予測される単語シーケンスと実際の単語シーケンスの間の誤差が単語誤り率計算機５１０によって識別される。誤差は、予測される単語シーケンス内の単語と、デコーダ５０２によって提供される実際の単語シーケンス内の単語の間の差を識別することによって決定される。
【００６０】
ステップ２１２で、誤差を使用してテストテキスト５００の単語誤り率を生成する。一実施形態では、単語誤り率の期待値は、次のように決定される。
【００６１】
【数８】

$$E[WER]=\frac{\sum_{i=1}^{I}E[e_i]}{\sum_{i=1}^{I}N_i} \qquad \text{式2}$$

【００６２】
式中Ｅ［ＷＥＲ］は、テストテキスト５００の単語誤り率の期待値、Ｅ［ｅ$_i$］はテストテキスト内のシーケンスｉでの誤差数の期待値、Ｎ$_i$は実際のシーケンスｉ内の単語数、およびＩはテストテキスト５００内の文の総数である。
【００６３】
一実施形態では、Ｅ［ｅ$_i$］は文ｉについて予測される単語シーケンスごとの誤差数と、その予測される単語シーケンスの確率の積の合計に等しい。これは、次のような式で書き表すこともできる。
【００６４】
【数９】

$$E[e_i]=\sum_{W_p}\mathrm{ErrCount}(W_p,W_c)*p(W_p|W_c) \qquad \text{式3}$$

【００６５】
ここで、Ｗ$_c$はシーケンスｉ内の実際の単語シーケンス、ＥｒｒＣｏｕｎｔ（Ｗ$_p$，Ｗ$_c$）は予測される単語シーケンスＷ$_p$を実際の単語シーケンスＷ$_c$と照合することによって識別される誤差数、およびｐ（Ｗ$_p$｜Ｗ$_c$）は実際の単語シーケンスＷ$_c$が与えられた場合の予測される単語シーケンスＷ$_p$の確率である。予測されるすべての単語シーケンスについて和を取る。
【００６６】
他の実施形態では、ある文の予想される誤差数は、生じた誤差のタイプに基づいて誤差に重み付けすることによって計算される。言い換えれば、ＥｒｒＣｏｕｎｔ（Ｗ$_p$，Ｗ$_c$）の計算は、誤差のタイプに基づいて誤差に重み付けすることによって行われる。
【００６７】

したがって、式２を使用してテストテキストの単語誤り率を生成することが可能となる。テストテキスト内の各文は別々に検査されるため、音声認識システムがデコードするときに多数の誤差を生成しそうなテストテキスト内の単語シーケンスを識別することも可能である。上に示したシステムでは、音声認識システムがそのテキストに対してどのように機能するかを決定するのに音響データは必要ないことに注意されたい。これによって、様々なタスクに関係する音声認識システムの評価コストが大幅に低減される。また、これによって開発者は、ユーザに入力を要求し、その結果音声認識システムがより容易にデコードできる単語シーケンスを使用するようユーザを導くやり方を変えることができる。

【００６８】

さらに、本発明によって複数の言語モデルを互いに比較することができるようになる。これは、１つの言語モデルを使用して単語誤り率を決定し、次いで第２の言語モデルを使用して誤り率を決定することによって行うことができる。次いで単語誤り率が互いに比較されて言語モデルの相対的なパフォーマンスが決定される。

【００６９】

さらに本発明によって、言語モデルの学習を、ステップ２１２で計算された単語誤り率に基づく識別学習を使用して音響モデルで行うことができる。こうした学習で、言語モデルは、単語誤り率を改善するように変更される。この学習は一部音響モデルのパフォーマンスに基づいているため、結果として得られる言語モデルは、音響モデルのパフォーマンスを参照することなしに学習が行われる言語モデルより良好に機能すると考えられる。

【００７０】

本発明は、特定の実施形態に関連して説明してきたが、本発明の意図および範囲から逸脱することなく、形態および詳細に変更を加えることができることを当分野の技術者であれば理解されよう。

【図面の簡単な説明】

【００７１】

【図１】本発明の実施形態を実施できる一般のコンピューティング環境を示すブロック図である。

【図２】本発明の実施形態によるコンフュージョンモデルの構築および使用の方法を示すフロー図である。

【図３】本発明の実施形態によるコンフュージョンモデルの学習に使用する構成要素を示すブロック図である。

【図４】本発明の一実施形態によるＨＭＭコンフュージョンモデルを示す状態図である。

【図５】テキストおよびコンフュージョンモデルを使用して単語誤り率を決定するデコーダを示すブロック図である。

【図６】本発明の一実施形態によるデコード中に形成されたパスを示すトレリス図である。

【符号の説明】

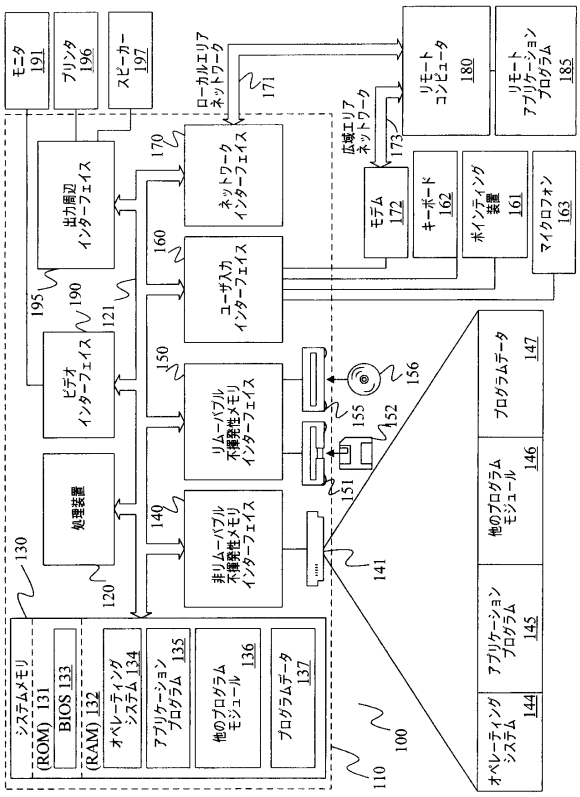
【００７２】

１００　コンピューティングシステム環境
１１０　コンピュータ
１２０　処理装置
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　基本入出力システム（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

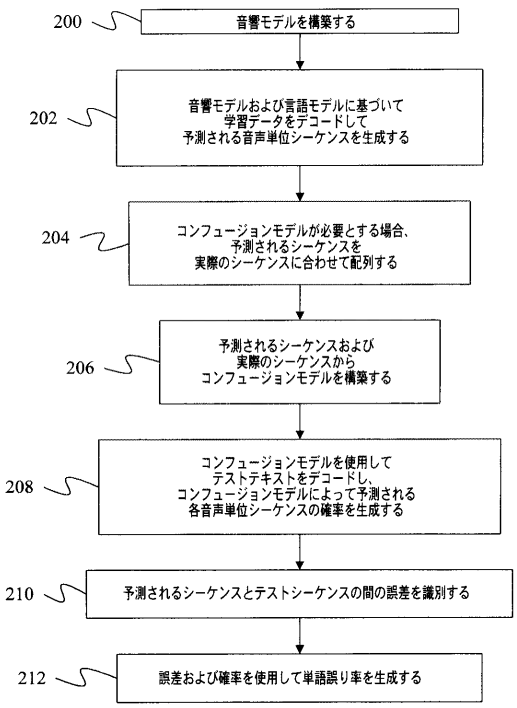
１４０　インターフェイス
１４１　ハードディスクドライブ
１４４　オペレーティングシステム
１４５　アプリケーションプログラム
１４６　他のプログラムモジュール
１４７　プログラムデータ
１５０　インターフェイス
１５１　磁気ディスクドライブ
１５２　リムーバブル不揮発性磁気ディスク
１５５　光ディスクドライブ
１５６　リムーバブル不揮発性光ディスク
１６０　ユーザ入力インターフェイス
１６１　ポインティング装置
１６２　キーボード
１６３　マイクロフォン
１７０　ネットワークインターフェイスまたはアダプタ
１７１　ローカルエリアネットワーク（ＬＡＮ）
１７２　モデム
１７３　広域エリアネットワーク（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　ビデオインターフェイス
１９１　モニタ
１９５　出力周辺インターフェイス
１９６　プリンタ
１９７　スピーカー
３００　音響モデル
３０２　トレーナ
３０４　学習テキスト
３０８　話者
３０９　受信機
３１０　特徴抽出機
３１２　デコーダ
３１４　学習言語モデル
３１５　辞書
３１６　配列モジュール
３１８　コンフュージョンモデルトレーナ
３２０　コンフュージョンモデル
４００　状態
４０２　状態
４０４　状態
４０６　状態
４０８　遷移
４１０　遷移
４１２　遷移
４１４　遷移
５００　テストテキスト
５０２　デコーダ
５０４　コンフュージョンモデル
５０６　言語モデル

５０８　辞書
５１０　単語誤り率計算機
６００　状態
６０２　状態
６０４　状態
６０６　状態
６０８　状態
６１０　状態
６１２　最終状態

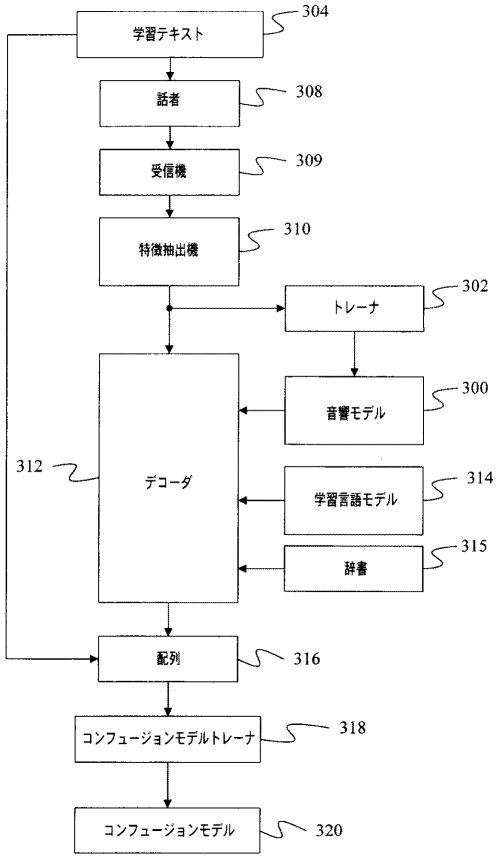
【図１】



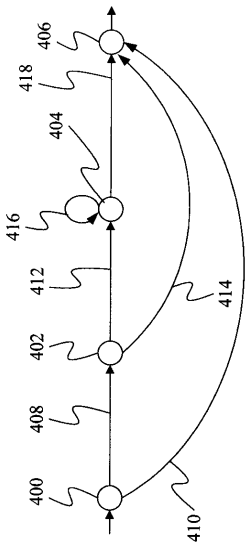
【図２】



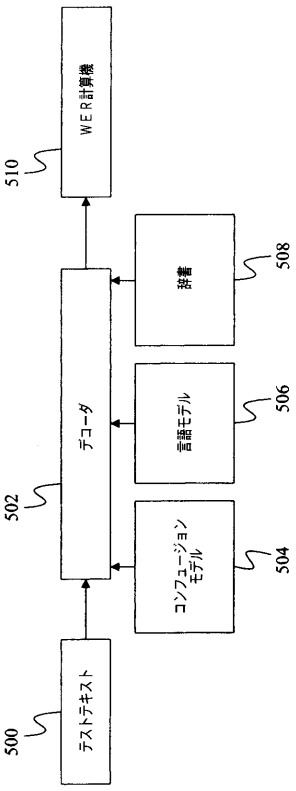
【図３】


学習テキスト
304
話者
308
受信機
309
特徴抽出機
310
トレーナ
302
音響モデル
300
デコーダ
312
学習言語モデル
314
辞書
315
配列
316
コンフュージョンモデルトレーナ
318
コンフュージョンモデル
320

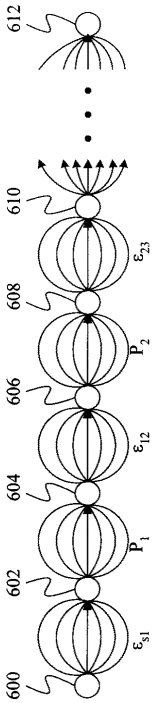
【図４】


406
418
404
416
412
414
402
408
400
410

【図５】


WER計算機
510
辞書
508
デコーダ
言語モデル
506
502
コンフュージョンモデル
504
テストテキスト
500

【図６】


612
610
608
606
604
602
600
ε_{23}
P_2
ε_{12}
P_1
ε_{s1}

フロントページの続き

(72)発明者　デン　ヤンガン
アメリカ合衆国　２１２０４　メリーランド州　トーソン　スティーブンソン　レーン　３４０　アパートメント　ナンバーシー４
(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　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　チプリアン　チェルバ
アメリカ合衆国　９８１１２　ワシントン州　シアトル　４３　アベニュー　イースト　１９２８　アパートメント　４

審査官　井上　健一

(56)参考文献　特表２００３－５３５４１０（ＪＰ，Ａ）
特開平０６－３１８０９６（ＪＰ，Ａ）
特開平１０－０９７２７１（ＪＰ，Ａ）
特開２００４－０６９８５８（ＪＰ，Ａ）

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