

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第4510953号
(P4510953)

(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14日 (2010. 5. 14)

(51) Int. Cl.	F I
G 1 0 L 15/06 (2006.01)	G 1 0 L 15/06 3 0 0 D
	G 1 0 L 15/06 3 1 0 S
	G 1 0 L 15/06 5 0 0 L

請求項の数 20 外国語出願 (全 25 頁)

(21) 出願番号	特願平11-168768	(73) 特許権者	599083248
(22) 出願日	平成11年6月15日 (1999. 6. 15)		ドラゴン・システムズ・インコーポレーテッド
(65) 公開番号	特開2000-35795 (P2000-35795A)		アメリカ合衆国マサチューセッツ州02160, ニュートン, ネバダ・ストリート
(43) 公開日	平成12年2月2日 (2000. 2. 2)		320
審査請求日	平成18年6月13日 (2006. 6. 13)	(74) 代理人	100089705
(31) 優先権主張番号	09/094609		弁理士 社本 一夫
(32) 優先日	平成10年6月15日 (1998. 6. 15)	(74) 代理人	100071124
(33) 優先権主張国	米国 (US)		弁理士 今井 庄亮
		(74) 代理人	100076691
			弁理士 増井 忠武
		(74) 代理人	100075236
			弁理士 栗田 忠彦

最終頁に続く

(54) 【発明の名称】 音声認識におけるノンインタラクティブ方式のエンロールメント

(57) 【特許請求の範囲】

【請求項 1】

音声認識システムのユーザをエンロールするための、コンピュータが実行するエンロールメント方法において、

複数のユーザ・アタランズを含みエンロール用テキストに略々対応しているユーザ音声を表すデータを収集するデータ収集処理を実行し、

1つのユーザ・アタランズに対応したデータの音響コンテンツを解析する解析処理を実行し、

前記解析の結果に基づいて、当該ユーザ・アタランズが前記エンロール用テキストのいずれかの部分とマッチングしているか否かを判定し、

当該ユーザ・アタランズが前記エンロール用テキストのいずれかの部分とマッチングしていたならば、当該ユーザ・アタランズの音響コンテンツに基づいて、前記エンロール用テキストの当該部分に対応した音響モデルを更新し、

当該ユーザ・アタランズが前記エンロール用テキストのいずれかの部分とマッチングしているか否かを判定する際に、ユーザが前記エンロール用テキストの単語を読み飛ばしても当該ユーザ・アタランズのマッチングの有無を判定し得るようにした、

ことを特徴とする方法。

【請求項 2】

前記データ収集処理において、前記方法を実行するコンピュータとは物理的に分離した録音装置を用いて録音されたデータを収集することを特徴とする請求項 1 記載の方法。

【請求項 3】

前記録音装置がデジタル録音装置であり、
前記データ収集処理において、前記デジタル録音装置からファイルを受取るようにした
、
ことを特徴とする請求項 2 記載の方法。

【請求項 4】

前記データ収集処理において、ユーザの音声を録音装置で再生することによって生成された信号を受取ることを特徴とする請求項 3 記載の方法。

【請求項 5】

前記録音装置がアナログ録音装置であることを特徴とする請求項 4 記載の方法。

10

【請求項 6】

前記エンロール用テキストの一部分をアクティブ部分に指定するアクティブ部分指定処理を実行し、1つのユーザ・アタランスに対応したデータの音響コンテンツを解析する前記解析処理において、前記エンロール用テキストのアクティブ部分に対して特定の相対位置にあるデータを解析することを特徴とする請求項 1 記載の方法。

【請求項 7】

直前に解析したアタランスの前記エンロール用テキスト上における位置を識別し、前記エンロール用テキストの一部分をアクティブ部分に指定する前記アクティブ部分指定処理において、前記識別位置に基づいてアクティブ部分の指定を行うことを特徴とする請求項 6 記載の方法。

20

【請求項 8】

前記アクティブ部分指定処理において、前記識別位置の前後に存在するテキストを含む部分をアクティブ部分に指定することを特徴とする請求項 7 記載の方法。

【請求項 9】

前記アクティブ部分指定処理において、前記識別位置を含むパラグラフ、前記識別位置に先行するパラグラフ、及び前記識別位置に後続するパラグラフを含む部分をアクティブ部分に指定することを特徴とする請求項 8 記載の方法。

【請求項 10】

前記エンロール用テキストのアクティブ部分に対して特定の相対位置にあるデータを解析する際に、当該データと前記エンロール用テキストの当該アクティブ部分に含まれている単語のモデルとの間のマッチングを探索することを特徴とする請求項 6 記載の方法。

30

【請求項 11】

前記エンロール用テキストのアクティブ部分に対して特定の相対位置にあるデータを解析する際に、前記エンロール用テキストの当該アクティブ部分に対応したエンロール用文法を使用することを特徴とする請求項 6 記載の方法。

【請求項 12】

当該ユーザ・アタランスが前記エンロール用テキストのいずれかの部分とマッチングしているか否かを判定する際に、前記エンロール用テキストに対応したエンロール用文法を使用することを特徴とする請求項 1 記載の方法。

【請求項 13】

当該ユーザ・アタランスが前記エンロール用テキストのいずれかの部分とマッチングしているか否かを判定する際に、更にリジェクション・文法を使用することを特徴とする請求項 12 記載の方法。

40

【請求項 14】

前記リジェクション・文法が、音素文法から成ることを特徴とする請求項 13 記載の方法。

【請求項 15】

前記リジェクション・文法が、前記エンロール用文法が使用する音素集合より小さな音素集合を使用してアタランスをモデル化する文法であることを特徴とする請求項 14 記載の方法。

50

【請求項 16】

当該ユーザ・アタランスが前記エンロール用テキストのいかなる部分ともマッチングしない場合に、当該ユーザ・アタランスを無視することを特徴とする請求項 1 記載の方法。

【請求項 17】

音声認識システムのユーザをエンロールするための、コンピュータが実行するエンロールメント方法において、

ポータブル録音装置を用いてエンロール用テキストに略々対応したユーザの音声を録音し、

録音した音声をコンピュータへ転送し、

前記コンピュータを使用して、前記録音音声の音響コンテンツを解析し、

前記コンピュータを使用して、前記解析の結果に基づいて、前記エンロール用テキストの複数の部分の夫々にマッチングする前記音声の複数の部分を識別し、

前記コンピュータを使用して、前記音声のマッチング部分の音響コンテンツに基づいて、前記エンロール用テキストのマッチング部分に対応した夫々の音響モデルを更新する、

ことを特徴とする方法。

【請求項 18】

録音した前記音声、前記エンロール用テキストのある部分を読み飛ばし、または、前記エンロール用テキストのある部分を反復していることを特徴とする請求項 17 記載の方法。

【請求項 19】

プログラムを記録したコンピュータが読取り可能な格納媒体であって、該プログラムはコンピュータに、

複数のユーザ・アタランスを含みエンロール用テキストに略々対応しているユーザ音声を表すデータを収集するデータ収集処理を実行させ、

1つのユーザ・アタランスに対応したデータの音響コンテンツを解析する解析処理を実行させ、

前記解析の結果に基づいて、当該ユーザ・アタランスが前記エンロール用テキストのいずれかの部分とマッチングしているか否かを判定させ、

当該ユーザ・アタランスの音響コンテンツに基づいて、前記エンロール用テキストの当該ユーザ・アタランスにマッチングしている部分に対応した音響モデルを更新させるための命令を含み、

以上の命令がコンピュータのコンフィギュレーションを設定することにより、ユーザが前記エンロール用テキストの単語を読み飛ばしても当該ユーザ・アタランスのマッチングの有無を判定し得るようにした、ことを特徴とするプログラムを記録したコンピュータ読取り可能な記録媒体。

【請求項 20】

ユーザをエンロールする音声認識システムにおいて、

音声信号を受取るための入力デバイスを備え、

プロセッサを備え、

複数のユーザ・アタランスを含みエンロール用テキストに略々対応しているユーザ音声を表すデータを収集するデータ収集処理を実行するように前記プロセッサが設定されており、

1つのユーザ・アタランスに対応したデータの音響コンテンツを解析する解析処理を実行するように前記プロセッサが設定されており、

前記解析の結果に基づいて、当該ユーザ・アタランスが前記エンロール用テキストのいずれかの部分とマッチングしているか否かを判定するように前記プロセッサが設定されており、

当該ユーザ・アタランスの音響コンテンツに基づいて、前記エンロール用テキストの当該ユーザ・アタランスにマッチングしている部分に対応した音響モデルを更新するように前記プロセッサが設定されており、

ユーザが前記エンロール用テキストの単語を読み飛ばしても当該ユーザ・アタランスのマッチングの有無を判定し得るように前記プロセッサが設定されている、ことを特徴とする音声認識システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、音声認識におけるエンロールメントに関する。

【0002】

【従来の技術】

音声認識システムは、ユーザの音声を解析して、ユーザが何を話しているかを判定するシステムである。音声認識システムの殆どは、フレーム方式のシステムである。フレーム方式のシステムでは、プロセッサが、認識対象音声を表している信号を非常に短い時間部分に分割して、一連のデジタル・フレームを生成する。

【0003】

連続音声認識システムは、ユーザが個々の単語ごと、ないしは語句ごとに区切って発声する必要がなく、連続して発声した一連の単語ないし語句を認識する能力を備えた音声認識システムである。これに対して、分離音声認識システムは、個々に区切って発声した単語ないし語句を認識するようにした音声認識システムであり、ユーザは、個々の単語ないし語句を発声するたびに、その後に短い間を置かなければならない。一般的に、連続音声認識システムは、分離音声認識システムと比べて誤認識の発生頻度が高いが、それは、連続した音声を認識するための認識作業がそれだけ複雑だからである。

【0004】

連続音声認識システムのプロセッサは、一般的に、音声の「アタランス (utterance: 前後を沈黙で区切られたひとまとまりの発声)」を解析するように設定されている。1個のアタランスに含まれるフレームの数は一定しておらず、所定長さ以上のポーズ (沈黙) が発生した場合に、そのポーズまでの音声部分をもって、1個のアタランスと見なすようにしている。

【0005】

ユーザが何を話しているかを判定するために、プロセッサはまず、1個のアタランスに含まれている複数のデジタル・フレームの各々に対して最良のマッチングを示す夫々の音響モデルを判定し、続いて、それら音響モデルに対応するテキストを判定する。様々な単語、語句、それにコマンドの集合をボキャブラリと呼び、1個の音響モデルは、ボキャブラリのうちの、1個の単語、1個の語句、または1個のコマンドに対応していることもある。1個の単語が複数の音素によって構成されているとき、それら音素の集合により、その単語の表音綴り字 (phonetic spelling) が表されている。音響モデルのうちには、沈黙を表す音響モデルや、種々の環境ノイズを表す音響モデルも含まれている。

【0006】

最良のマッチングを示す音響モデルの集合に対応する単語ないし語句は、認識候補と呼ばれる。プロセッサは、1個のアタランスに対して、ただ1つの認識候補しか生成しない (即ち、ただ1つの単語列ないし語句列しか生成しない) こともある。複数の認識候補を生成することもあり、後者の場合には、それら認識候補を列挙したリストを生成することになる。

【0007】

一般的に、音響モデルとは、その音響モデルに対応した音声単位 (例えば、音素) が、様々な発声者によってどのように発声されているかを記述したデータのことをいう。音響モデルがある特定のユーザの音声を表すときの精度を向上させて誤認識の発生頻度を低下させるために、音声認識システムは、音響モデルに対して、特定のユーザの音声に適応させるための修正を加えるようにしている。この音響モデルに対する修正は、音声認識システムの使用開始に際して実行する初期エンロールメント・セッションにおいて実行され、ま

10

20

30

40

50

た、その後の音声認識システム使用中にも実行することができ、いずれの場合も、ユーザの音声のサンプルを収集し、その収集したサンプルに基づいて音響モデルに修正を加える。

【0008】

従来の音声認識システムに組込まれているエンロールメント・セッションは、一般的に、ユーザに所定の単語リストの単語を読み上げさせたり、ユーザにプロンプトを発して特別の単語を読み上げさせるようにしたものであった。例えば、米国、マサチューセッツ州、ニュートン市に所在のドラゴン・システムズ社（Dragon Systems, Inc.）が販売している「Dragon Dictate for Windows」という製品に組込まれている短時間で実行できるエンロールメント・セッションでは、新ユーザにプロンプトを発して、少数の単語から成る単語リスト中の各単語を発声させ、そのユーザの発声に基づいて音響モデルの適応化を実行するようにしている。

10

【0009】

更にその他の様々なエンロールメント方式がこれまでに採用されている。例えば、米国、マサチューセッツ州、ニュートン市に所在のドラゴン・システムズ社（Dragon Systems, Inc.）が販売している「Naturally Speaking」という製品に組込まれているインタラクティブ方式のエンロールメント・セッションは、適当なエンロール用テキストを選択して新ユーザに読み上げさせるようにしており、その際に、そのエンロール用テキスト中のユーザが読み上げるべき位置を、エンロールメント用のディスプレイ（例えば矢印）によって指示するようにしている。

20

【0010】

【発明の概要】

本発明は、音声認識システムのユーザをエンロールするための、ノンインタラクティブ方式のエンロールメント方法を提供するものである。この方法はノンインタラクティブ方式であるため、ユーザは、例えばポータブル録音装置を用いてエンロール用音声を録音し、その録音した音声を後刻ダウンロードして処理させることによって、音声認識システムの音響モデルをより精緻なものに更新することができる。この方法によれば、音声はエンロール用テキストに略々対応しているだけでよく、ユーザが、単語、語句、センテンス、ないしはパラグラフを読み飛ばしたり、反復して読み上げたりしても全く問題を生じない。この方法は、ユーザ音声を、エンロール用テキストに対比させて解析するようにしており、それによって、その音声のうちの、エンロール用テキストのいずれかの部分とマッチングする部分を識別し、そして、そのユーザ音声のマッチング部分に基づいて、エンロール用テキストのマッチング部分に対応した夫々の音響モデルを更新するようにしている。この方法を用いれば、エンロールメント・プロセスの融通性を増大させることができ、それによって、エンロールメント作業をより容易に実行できるようにすることができる。

30

【0011】

大局的にとらえた特徴の1つは次の通りである。まず、音声認識システムのユーザをコンピュータがエンロールする際に、エンロール用テキストに略々対応しているユーザの音声を表すデータを収集する。コンピュータは、1つのユーザ・アタランスの音響コンテンツを解析し、その解析の結果に基づいて、当該ユーザ・アタランスが前記エンロール用テキストのいずれかの部分とマッチングしているか否かを判定する。当該ユーザ・アタランスが前記エンロール用テキストのいずれかの部分とマッチングしていたならば、当該ユーザ・アタランスの音響コンテンツに基づいて、前記エンロール用テキストの当該部分に対応した音響モデルを更新する。ユーザが前記エンロール用テキストの単語、センテンス、ないしはパラグラフを読み飛ばし、或いは、反復して読み上げた場合でも、当該ユーザ・アタランスが前記エンロール用テキストのいずれかの部分とマッチングしているか否かを判定することができる。

40

【0012】

具体的な実施の形態は、以下に列挙する特徴の1つまたは幾つかを備えたものとすることができる。本発明によるエンロールメント作業は、インタラクティブ方式で行われるもの

50

ではない。そのため、ユーザ音声を表す前記データは、コンピュータとは物理的に分離した録音装置を用いて録音されたデータとすることができる。例えば、その録音装置をデジタル録音装置として、そのデジタル録音装置からファイルを受取ること、前記データを収集するようにすることができる。また、ユーザの音声を録音装置で再生することによって生成された信号を受取ること、前記データを収集するようにすることもでき、この場合の録音装置は、例えばアナログ録音装置であってもよい。

【0013】

ユーザ・アタランスの解析を実行する前に、コンピュータが、前記データを複数のグループに区分するようにし、区分した各グループが、ユーザが発声したアタランスの1個ずつを表しているようにしてもよい。

10

【0014】

コンピュータが、前記エンロール用テキストの一部分をアクティブ部分に指定して、1個のユーザ・アタランスに対応したデータの音響コンテンツを解析する際に、そのエンロール用テキストのアクティブ部分に対して特定の相対位置にあるデータを解析するようにしてもよい。また、コンピュータが、直前に解析したアタランスの前記エンロール用テキスト上における位置を識別し、更に、前記エンロール用テキストの一部分をアクティブ部分に指定する際に、前記識別位置に基づいてアクティブ部分の指定を行うようにしてもよい。その場合、前記識別位置の前後に存在するテキストを含む部分をもってアクティブ部分としてもよい。例えば、前記識別位置を含むパラグラフとそのパラグラフの前後に位置するパラグラフとを含む部分をもってアクティブ部分とすることができる。

20

【0015】

コンピュータが、前記データと、前記エンロール用テキストのアクティブ部分に含まれている単語のモデルとの間のマッチングを探索するようにしてもよい。また、その際には、コンピュータが、前記エンロール用テキストのアクティブ部分に対応したエンロール用グラマーを使用するようにしてもよい。

【0016】

また、リジェクション・グラマーを使用して、当該ユーザ・アタランスが前記エンロール用テキストのいずれかの部分とマッチングしているか否かを判定するようにしてもよい。前記リジェクション・グラマーを、音素グラマーから成るものとしてもよく、また、前記リジェクション・グラマーを、前記エンロール用グラマーが使用する音素集合より小さな音素集合を使用してアタランスをモデル化するグラマーとしてもよい。

30

【0017】

各々が対応したエンロール用グラマーを有する複数のエンロール用テキストのうちから、適当なエンロール用テキストを選択して使用するようにしてもよい。また、ユーザからエンロール用テキストを受取って使用するようにしてもよい。更に、その場合には、ユーザ・アタランスがそのエンロール用テキストのいずれか部分にマッチングしているか否かを判定するために使用する、その受取ったエンロール用テキストに対応したエンロール用グラマーは、新たに生成するようにすればよい。

【0018】

当該ユーザ・アタランスが前記エンロール用テキストのいかなる部分ともマッチングしない場合には、当該ユーザ・アタランスを無視するようにしてもよい。大局的にとらえたもう1つの特徴は、次のようなものである。音声認識システムのユーザをエンロールするために、ポータブル録音装置を用いてユーザの音声を録音し、その録音した音声をコンピュータへ転送する。続いて、コンピュータを使用して、その録音した音声の音響コンテンツを解析し、その解析の結果に基づいて、前記エンロール用テキストの複数の部分の夫々にマッチングする前記音声の複数の部分を識別し、そして、前記音声のマッチング部分の音響コンテンツに基づいて、前記エンロール用テキストのマッチング部分に対応した夫々の音響モデルを更新する。録音した音声は、前記エンロール用テキストのある部分を読み飛ばしていても、また、前記エンロール用テキストのある部分を反復して読み上げていてもかまわない。

40

50

【0019】

大局的にとらえた更に別の特徴は、次のようなものであり、即ち、物理的に独立した録音装置を用いてエンロール用テキストに対応したデータを収集すると共に、そのエンロール用テキストの一部分をアクティブ部分に指定し、そのエンロール用テキストのアクティブ部分に対して特定の相対位置にあるアタランスの音響コンテンツを解析するというものである。

【0020】

その他の特徴及び利点は、以下の詳細な説明、図面、及び請求項から明らかな通りである。

【0021】

10

【詳細な説明】

図1において、音声認識システム100は、複数の入出力（I/O）デバイス（図示例では、マイクロホン105、マウス110、キーボード115、及びディスプレイ120）と、汎用コンピュータ125とを備えている。汎用コンピュータ125は、プロセッサ（CPU）130、I/Oユニット135、及びサウンド・カード140を備えている。メモリ145には、データ及びプログラムが格納されており、格納されているプログラムには、オペレーティング・システム150、アプリケーション・プログラム155（例えば、ワード・プロセッサのプログラム等）、それに音声認識ソフトウェア160などが含まれている。

【0022】

20

マイクロホン105は、ユーザの音声をひろって、その音声をアナログ信号の形でサウンド・カード140へ伝達する。サウンド・カード140では、その信号がA/Dコンバータで処理され、それによって、そのアナログ信号がデジタル・サンプル集合に変換される。プロセッサ130は、オペレーティング・システム150及び音声認識ソフトウェア160の制御下において、ユーザの連続音声に含まれている複数のアタランスの各々を識別する。アタランスとは、十分に長い所定時間長さ（例えば、160～250ミリ秒）の休止（ポーズ）によって前後が区切られた、ひとまとまりの発声部分をいう。アタランスには、ユーザの音声中の1個の単語だけから成るアタランスもあれば、2個以上の単語から成るアタランスもある。

【0023】

30

この音声認識システムには更に、アナログ・レコーダ・ポート165、及び／または、デジタル・レコーダ・ポート170を装備することができる。アナログ・レコーダ・ポート165は、サウンド・カード140に接続しており、ハンドヘルド型レコーダ（手持型録音装置）を用いて録音した音声をサウンド・カード140へ入力する際にこのポート165を使用する。このアナログ・レコーダ・ポート165は、ハンドヘルド型レコーダをこのポート165に挿入したときに、そのレコーダのスピーカと対向する位置にマイクロホンを配設した構成のものとしてもよく、また、マイクロホン105をアナログ・レコーダ・ポートとして利用するようにしてもよい。また更に、アナログ・レコーダ・ポート165をテープ・プレーヤで構成し、ハンドヘルド型レコーダで録音したテープを、そのテープ・プレーヤに挿入することによって、そのテープに録音されている情報がサウンド・

40

【0024】

デジタル・レコーダ・ポート170は、例えばハンドヘルド型デジタル録音装置を使用して生成したデジタル・ファイルを転送するためのポートとして構成することができる。またその場合には、そのデジタル・ファイルを直接、メモリ145へ転送する構成としてもよい。更に、デジタル・レコーダ・ポート170は、コンピュータ125の記憶装置（例えばフロッピー・ドライブ）で構成することもでき、また、リモート・ロケーションとの間で電気通信を行うためのモデム・カードやネットワーク・カードで構成してもよい。

【0025】

50

図2は、音声認識ソフトウェア160の構成要素を示した図である。理解を容易にするために、以下の説明では、それら構成要素が、ある特定の結果を得るための動作を実行するというように説明する。ただし実際には、それら構成要素の各々は、プロセッサ130が特定の処理手順に従って動作することによって実現されるものである。

【0026】

まず、フロントエンド処理モジュール200は、サウンド・カード140から（またはデジタル・レコーダ・ポート170から）転送されてくる一連のデジタル・サンプル205を、1個のアタランスの周波数成分を表す一連のパラメータ・フレーム210へ変換する。各々のフレームは、24個のパラメータを含んでおり、1個のアタランスの非常に短い時間部分（例えば10ミリ秒に相当する部分）を表している。

10

【0027】

認識部215は、ある1個のアタランスに対応した一連のフレームを受取ったならば、それらフレームに処理を施して、当該アタランスに対応したテキストを識別する。そのために、認識部215は、そのテキストに関する幾つかの仮定を考察して、それら仮定の夫々のスコアを付与する。各々の仮定に付与するスコアは、その仮定がユーザの音声に対応している確率を反映した値を有する。処理を容易にするために、このスコアの値は、負の対数値で表すようにしている。この場合、スコアの値が小さいほど、マッチングは良好であり（即ち、仮定が真実である確率が高い）、一方、スコアの値が大きいほど、マッチングの見込みは小さい（即ち、仮定が真実である確率が低い）。従って、スコアの値が大きければ、それだけマッチングの尤度は低下する。認識部215は、1個のアタランスについて以上の処理を完了したならば、スコアが良好な幾つかの仮定を、複数の認識候補を列挙したリスト形で、制御／インターフェース・モジュール220へ供給する。この場合、そのリストに列挙されている認識候補は、その各々が、1つずつの仮定に対応しており、また、その各々にスコアが付与されている。それら認識候補のうちには、テキストに対応した認識候補が含まれていることもあり、また、コマンドに対応した認識候補が含まれていることもある。また、コマンドには、単語から成るコマンド、語句から成るコマンド、それに、センテンスから成るコマンドなどがある。

20

【0028】

認識部215は、1個のアタランスを構成している複数のフレーム210を処理する際に、1個または2個以上の制約グラマー225に従ってその処理を実行する。制約グラマーは、テンプレートと呼ばれることもあり、また、規制ルールと呼ばれることもある。制約グラマーのうちには、あるアタランスに対応することのできる単語の条件をを規定しているもの、単語どうしの間の並び順の条件を規定しているもの、単語どうしの間の文法形態の条件を規定しているもの、それに、単語どうしの間の並び順の条件と文法形態の条件との両方を規定しているものなどがある。例えば、メニュー操作コマンドに関係した制約グラマーのうちには、メニューに示される項目（例えば「ファイル」や「編集」等）だけを集めた制約グラマーや、メニュー画面内での移動を行わせるコマンド単語（例えば「アップ（up）」、「ダウン（down）」、「トップ（top）」、「ボトム（bottom）」等）だけを集めた制約グラマーがある。制約グラマーは、アクティブ状態にされたり、非アクティブ状態にされたりするため、その時々で、アクティブ状態にある制約グラマーは異なっている。例えば、制約グラマーのうちには、特定のアプリケーション・プログラム155に関連付けられているものがあり、そのような制約グラマーは、ユーザがそのアプリケーション・プログラムを開いたならばアクティブ状態にされ、ユーザがそのアプリケーション・プログラムを閉じたならば非アクティブ状態にされる。認識部215は、アクティブ状態にある制約グラマーに違反する仮定を廃棄する。更に、認識部215は、ある種の制約グラマーに関連付けられている仮定のスコアに対しては、その制約グラマーの特性に基づいて調整を加えることがある。

30

40

【0029】

音声認識ソフトウェア160が使用する制約グラマー225のうちの1つに、ラージ・ボキャブラリ・ディクテーション・グラマーがある。ラージ・ボキャブラリ・ディクテーシ

50

ジョン・グラマーは、アクティブ・ボキャブラリ 230 に含まれている単語を識別するためのグラマーであり、アクティブ・ボキャブラリ 230 は、ソフトウェアにとって既知の単語から成るボキャブラリである。ラージ・ボキャブラリ・ディクテーション・グラマーは、更に、夫々の単語の出現頻度も示している。ラージ・ボキャブラリ・ディクテーション・グラマーに用いる言語モデルは、各々の単語の出現頻度を、その単語の前後にどのような単語があるかを考慮せずに示したユニグラム・モデルとしてもよく、或いは、各々の単語の出現頻度を、その単語に先行する単語がどのようなものであるかを考慮して示したバイグラム・モデルとしてもよい。例えばバイグラム・モデルを使用する場合には、名詞や形容詞は、動詞や前置詞と比べて、「ザ (the)」の後にくる確率が高いということを規定することができる。

10

【0030】

アクティブ・ボキャブラリ 230 に含まれている各単語は、発音モデルで表されており、より詳しくは、各単語は、その単語の表音綴り字を成す音素列によって表されている。また、各音素は、3 個のノードから成るトライフォンで表すことができる。トライフォンは前後関係を考慮した音素であり、例えば「abc」というトライフォンは、音素「b」の前に音素「a」が存在し、音素「b」の後に音素「c」が存在するものであり、従って、音素「a」及び「c」が前後関係として付随する音素「b」を表している。

【0031】

音素は音響モデル 235 によって表される。音素としてトライフォンを使用している場合には、音響モデル 235 は、トライフォンの各ノードを、複数のガウス型（正規分布型）確率密度関数（PDF）の混合によって表すものとなっている。例えば、トライフォン「abc」のノード「i」は、次の〔式 1〕に示した ab^ic で表される。

20

【0032】

【数 1】

$$ab^ic = \sum_k w_k N(\mu_k, c_k) \quad (1)$$

この〔式 1〕において、 w_k で表した各係数は混合重みであり、次の〔式 2〕が成り立つ。

30

【0033】

【数 2】

$$\sum_k w_k = 1 \quad (2)$$

また〔式 1〕において、 μ_k は、確率密度関数（PDF） N_k の平均ベクトルであり、 c_k は、確率密度関数（PDF） N_k の共分散行列である。フレーム・シーケンス中の各フレームが 24 個のパラメータを含んでいるのと同様に、各ベクトル μ_k も 24 個のパラメータを含んでいる。行列 c_k は、 24×24 行列である。トライフォンの各ノードは、最多では、例えば 16 個の PDF の混合として表される。

40

【0034】

1 つのディクテーション・ボキャブラリの全体は、アクティブ・ボキャブラリ 230 と、バックアップ・ボキャブラリ 245 とで構成されている。これらのうちバックアップ・ボキャブラリ 245 は、例えば、特定ユーザ用バックアップ・ボキャブラリ単語群と、全システムのバックアップ・ボキャブラリ単語群とを収容した複数のファイルで構成されるものである。全システムのバックアップ・ボキャブラリは、システムにとって既知の単語の全てを収容したものであり、ある時点でアクティブ・ボキャブラリに収容されている可能性のある単語は全て、この全システムのバックアップ・ボキャブラリに収容されている。

【0035】

50

誤り修正作業の実行中には、バックアップ・ボキャブラリを対象とした単語検索を行うが、その際には、先に、特定ユーザ用バックアップ・ボキャブラリを検索し、その後に、全システムのバックアップ・ボキャブラリを検索する。また、これらバックアップ・ボキャブラリの検索は、ユーザがキーボードから打ち込んだテキストに新単語が含まれていた場合にも実行される。

【0036】

認識部215は、プリフィルタ処理部240と並行して動作できるようにしてある。あるアタランスの処理を開始する際には、認識部215がプリフィルタ処理部240へ要求を発して、そのアタランスの先頭の単語として発声された可能性のある単語（即ち、そのアタランスの先頭の幾つかのフレームに対応する可能性のある単語）のリストを出力させる。プリフィルタ処理部240は、それら一連のフレームとアクティブ・ボキャブラリ230とを粗比較し、それによって、アクティブ・ボキャブラリ230に含まれている単語のうち、認識部215において、より精緻な比較を実行することが適当であると判断される単語だけを拾い出すことで、ボキャブラリの部分集合を編成する。

10

【0037】

制御／インターフェース・モジュール220は、音声認識ソフトウェアの動作を制御しており、また、その他のソフトウェアとの間のインターフェース、並びに、ユーザとの間のインターフェースの機能を担っている。制御／インターフェース・モジュール220は、認識部215から、各アタランスに対応した複数の認識候補を列挙したリストを受取る。それら認識候補は、ディクテーション・テキストに対応した認識候補であったり、音声認識コマンドに対応した認識候補であったり、外部コマンドに対応した認識候補であったりする。このリスト中の最良スコアの認識候補がディクテーション・テキストに対応した認識候補であったならば、制御／インターフェース・モジュール220は、そのディクテーション・テキストを、例えばワード・プロセッサ等の、その時点でアクティブ状態にあるアプリケーションへ転送する。制御／インターフェース・モジュール220は更に、その最良スコアの認識候補を、グラフィカル・ユーザ・インターフェースを介してユーザに表示することもできる。更に、制御／インターフェース・モジュール220は、音声認識コマンド（例えば「ウェイク・アップ（wake up）」コマンドや、「メイク・ザット（make that）」コマンド）に応答して、音声認識ソフトウェアの動作を制御し、また、外部コマンドを、そのコマンドが対応しているソフトウェアへ転送する。

20

30

【0038】

制御／インターフェース・モジュール220は更に、認識部215が使用するアクティブ・ボキャブラリ、音響モデル、及び制約グラマーの制御を実行する。例えば、音声認識ソフトウェアが、特定のアプリケーション（例えばMicrosoft Word等）に利用されているときには、制御／インターフェース・モジュール220は、アクティブ・ボキャブラリを更新して、そのアプリケーションに対応したコマンド単語を組み込み、また、そのアプリケーションに対応した制約グラマーをアクティブ状態にする。

【0039】

制御／インターフェース・モジュール220が担当しているその他の機能としては、ボキャブラリ・カスタマイザとしての機能、及びボキャブラリ・マネージャとしての機能がある。ボキャブラリ・カスタマイザは、ユーザから供給されたテキストをスキャンすることによって、特定のトピックの言語モデルを最適化するものである。また、ボキャブラリ・マネージャは、ボキャブラリ、グラマー、及びマクロのブラウズ（閲覧）及び操作を行うために用いられる開発ツールである。制御／インターフェース・モジュール220のこれらの機能の各々は、メイン音声認識ソフトウェアから分離した実行プログラムとして構成するようにしてもよい。同様に、制御／インターフェース・モジュール220それ自体も、独立した実行プログラムとして構成するようにしてもよい。

40

【0040】

制御／インターフェース・モジュール220は更に、エンロールメント・プログラムを実行する機能も担当している。エンロールメント・プログラムは、エンロール用テキストと

50

、このエンロール用テキストに対応したエンロール用グラマーとを使用して、音声認識ソフトウェアを、特定のユーザに適応するようにカスタマイズするためのプログラムである。エンロールメント・プログラムの動作モードとしては、エンロールメント作業をどのように行うべきかを、エンロールメント・プログラムがユーザに指示するインタラクティブ・モードと、コンピュータとは無関係にユーザに自由にエンロールメント作業を行わせるノンインタラクティブ・モードとの、両方のモードが利用可能である。インタラクティブ・モードでは、エンロールメント・プログラムがエンロール用テキストをユーザに表示して、ユーザはその表示されたテキストを読み上げる。ユーザがそのテキストを読み上げて行くのに合わせて、認識部 215 が、エンロール用グラマーを使用して、ユーザが次々と発声する一連のアタランスとエンロール用テキストのうちのそれらアタランスに対応した夫々の部分とをマッチングさせて行く。そして、認識部 215 が、両者をマッチングさせることができなかったときには、エンロールメント・プログラムがユーザへプロンプトを発して、エンロール用テキストの該当部分の読み上げを再度実行させる。認識部 215 は更に、ユーザが発声するアタランスから得られる音響情報に基づいて、エンロール用テキストのマッチング部分に対応した音響モデル 235 のトレーニング、即ち適応化を実行する。

10

【0041】

一方、ノンインタラクティブ・モードにおいては、ユーザは、コンピュータが発するプロンプトによって読み上げを指示されることなく、みずからの判断でエンロール用テキストの読み上げを行う。このモードでは、ユーザは、コンピュータがディスプレイ上に表示したテキストを読み上げるばかりでなく、コンピュータの動作とは全く無関係に、紙に印刷されたテキストを読み上げることによってエンロールメントを行うことも可能であるということが、このモードの大きな利点の1つである。従って、このモードを利用する場合、ユーザは、エンロール用テキストをポータブル録音装置に吹き込み、その録音した情報を後刻コンピュータにダウンロードして認識部 215 に処理を行わせるという方法を採用することもでき、更には、電話線を介してエンロール用テキストの読み上げを行うという方法を採用することもできる。また、このモードを利用する場合には、ユーザは、エンロール用テキストの全ての単語を読み上げる必要はなく、読み飛ばしたい単語やパラグラフは読み飛ばしてかまわない。更に、エンロール用テキストのうちのある部分を反復して読み上げてもかまわない。これらのことによって、エンロールメント作業の融通性が大いに高まっている。

20

30

【0042】

エンロールメント・プログラムは、複数のエンロール用テキストを列挙したリストを提示して、使用するエンロール用テキストをユーザに選択させることができ、それらエンロール用テキストの各々には、そのテキストに対応したエンロール用グラマーが付属している。これとは別の方法として、ユーザが、よそで入手したエンロール用テキストを入力することも可能である。この場合には、エンロールメント・プログラムが、その入力されたエンロール用テキストからエンロール用グラマーを生成するか、或いは、予め生成しておいたエンロール用グラマーを使用するようにすればよい。

【0043】

40

図3は、インタラクティブ・エンロールメント・プログラムのユーザ・インターフェース 1300 を示した図である。ユーザ・インターフェース 1300 は、エンロール用テキスト 1310 を表示して、ユーザが読み上げることができるようにすると共に、例えばそのエンロール用テキスト上を移動する矢印 1320 等によって、ユーザの現在位置を表示するようにしており、このユーザの現在位置のことを「エンロールメント・ポジション」という。図3のディスプレイでは、エンロールメント・ポジションが単語「for」の位置にあり、これによって、この単語「for」までの、ユーザによるエンロール用テキストの読み上げがエンロールメント・プログラムによって認識されたことが示されると共に、エンロールメント作業を続行するには、ユーザがこの単語「for」から先の読み上げを続行すべきことが示されている。一方、図4では、エンロールメント・ポジションが単語

50

「program」の位置にある。ユーザ・インターフェースがエンrollment・ポジションを表示する方法としては、このように矢印を使用する以外に、認識対象となっている部分のテキストをハイライト表示にするという方法もあり、また、エンrollment・ポジションにカーソルを表示するという方法もある。

【0044】

図3のユーザ・インターフェース1300を使用している場合に、ユーザがエンrollment・プログラムを開始するには、まず、音声コマンドを発するか、または、マウス等の入力デバイスを操作して、「Record（録音）」ボタン1330を選択する。続いてユーザは、表示されたテキスト1310の読み上げを、矢印1320で示されているエンrollment・ポジションから開始する。ユーザが読み上げて行くのに合わせて、認識部215が、エンrollment用グラマーを使用して、ユーザの音声とエンrollment用テキストとのマッチングを次々と調べて行くと共に、エンrollment・ポジションを（即ち矢印1320を）、次に読み上げるべき単語の先頭位置へ進めて行く。そして、ユーザ・アタランスとのマッチングの結果に基づいて、エンrollment用テキストに対応した一連の音響モデルを更新する。一般的に、認識部215は、エンrollment作業の実行中は、プリフィルタ処理部240を利用していない。

【0045】

図5に示したように、エンrollment・プログラムは、まず最初に、エンrollment用テキストの先頭にエンrollment・ポジションを設定し（ステップ1510）、続いてそのエンrollment・ポジションに矢印を表示する（ステップ1520）。次に、エンrollment・プログラムは、ユーザの音声を表している一連のデジタル・フレームのうちから、あるアタランスに相当するデジタル・フレームを取出す（ステップ1530）。次に、認識部215が、エンrollment用テキストに対応したエンrollment用グラマーを使用して、そのアタランスが、エンrollment用テキストのうちの、エンrollment・ポジションからはじまるある長さの部分とマッチングするか否かを判定する（ステップ540）。このときには、エンrollment用テキストのうちの、どれほどの長さの部分がそのアタランスに対応しているかが分からないため、認識部215は、エンrollment用グラマーを使用して、エンrollment用テキストのエンrollment・ポジション以降の様々な長さの部分のうち、そのアタランスに対して最良のマッチングを示す部分を識別する。

【0046】

もし、エンrollment用テキストのエンrollment・ポジション以降のどの部分も、そのアタランスに対して良好なマッチングを示していなかったならば、エンrollment・プログラムは、ユーザの音声を表している一連のデジタル・フレームのうちから、その次のユーザ・アタランスに相当する複数のデジタル・フレームを取出し、それらをエンrollment用テキストと比較する処理を再び実行する（ステップ1530、1540）。一方、あるユーザ・アタランスとエンrollment用テキストのいずれかの部分との間にマッチングが見いだされたならば、エンrollment・プログラムは、エンrollment用テキストのそのマッチング部分に対応した音響モデル235を、そのユーザ・アタランスに基づいて更新する（ステップ1550）。ただし、これに関する別の方法として、エンrollment・プログラムが、エンrollment用テキストのある特定のセグメント（例えばセンテンスやパラグラフ）の全体について認識作業を完了した後に、または、そのエンrollment用テキストの全体について認識作業を完了した後に、音響モデルの更新をまとめて一度に行うようにしてもよい。

【0047】

あるアタランスとマッチングしたエンrollment用テキストのマッチング部分が、エンrollment用テキストの末尾部分であって、それより後にはもうエンrollment用テキストが残っていなければ（ステップ1560）、このエンrollment用テキストを使用したエンrollment・プログラムは終了する（ステップ1570）。エンrollment・プログラムが、複数のエンrollment用テキストを用意している場合には、1つのエンrollment用テキストについての処理を終了したならば、自動的に次のエンrollment用テキストについての処理を開始し、そ

の新たなエンロール用テキストをディスプレイ上に表示して、そのエンロール用テキストの先頭にエンロールメント・ポジションを設定する（ステップ1510）。

【0048】

一方、ステップ1560において、エンロール用テキストにまだ残りの部分があると判定されたならば、そのエンロール用テキストのマッチング部分に続く、その次の単語へエンロールメント・ポジションを移動させた上で（ステップ1580）、矢印の表示を更新して、その更新後のエンロールメント・ポジションに矢印を表示し（ステップ1520）、エンロールメント作業を続行する。

【0049】

あるユーザ・アタランスがエンロール用テキストのいずれかの部分とマッチングするか否かの判定を行う（ステップ1540）ための方法には様々なものがある。例えば、ユーザが異なれば、エンロール用テキストの途中でポーズ（休止）を取る位置も異なるため、異なったユーザが発声したアタランスが、いずれもエンロール用テキストの同じ部分に対応するということではなく、この点を認識するようなエンロール用グラマーを設定するようにしてもよい。記述のごとく、ここに説明しているエンロールメント・グラマーは、エンロール用テキストの様々な長さの部分とユーザ・アタランスとの間のマッチングを調べることができるようにすることで、そのような状況に対処できるようにしたものである。

【0050】

また、ユーザ・アタランスがエンロール用テキストのいずれかの部分とマッチングしているか否かの判定に際しては、ユーザがエンロール用テキスト読み上げるときの読み上げ方が一通りではないため、様々な読み上げ方を認識して適切に対処できるようにエンロール用グラマーを設定するのがよい。様々な読み上げ方が予想される具体例としては、ディクテーションを行う際の、句読点の発声の仕方についての約束事がある。即ち、センテンスの末尾においてユーザが「ピリオド（period）」と発声する読み上げ方がある。これに対処するためには、センテンスの末尾で単語「period」が発声されたときには、それをオプション単語であると認識するようにエンロール用グラマーを設定しておけばよい。この場合、ユーザが「ピリオド」と発声したならば、エンロール用グラマーは、そのユーザの音声を「period」という単語にマッチングさせ、その音声に基づいて単語に対応した音響モデルの更新を行う。一方、ユーザが「ピリオド」と発声しなかったならば、エンロール用グラマーは、単語「period」はオプション単語であるものとして無視してしまい、エンロール用テキストの次の単語の処理に取りかかる。

【0051】

同様のルールを定めることで、段落書式に関するユーザの発声（例えば、ニュー・ライン（行替え）や、ニュー・パラグラフ（新段落）等）、それに、フォーマット方式に関するユーザの発声（例えば、キャピタライズ（大文字）、ボールド（太文字）、アンダーライン（下線）、イタリック等）にも対処することができる。例えば、図3中に示したテキストに対応したエンロール用グラマーは、拡張Backus-Naur形式（EBNF）で表したときに、次のように表されるものとすればよい。

<認識結果> ::= [キャピタライズ] dragon [キャピタライズ] system
is pleased to acknowledge [キャピタライズ] arther
[キャピタライズ] c [ピリオド] [キャピタライズ] clarkie and [キャピ
タライズ] harper [キャピタライズ] collins [キャピタライズ] publ
ishers for allowing us to use selections
from [キャピタライズ] arther [キャピタライズ] c [ピリオド] [キャピ
タライズ] {clarke's | clarke Apostrophe s} novel
threethousand [キャピタライズ] one [コロン] [キャピタライズ]
the [キャピタライズ] final [キャピタライズ] odyssey [コンマ] fo
r our training program [ピリオド]

このEBNF表記において、

[] は、オプションであることを表しており、

10

20

30

40

50

<>は、ルールであることを表しており、
|は、OR関数であり、

: :=は、定義であって、左右両辺が等しいことを表している。

このエンロール用グラマーには、ユーザがエンロール用テキスト上を前進して行くにつれて、変更が加えられて行く。例えばユーザが、図3に矢印1320で示したように、エンロール用テキスト上の単語「for」の位置にきたときには、このエンロール用グラマーは、EBNFで表した場合に、次のように表されるものになっている。

**<認識結果>: :=for our training
program [ピリオド]**

10

別法として、このエンロール用グラマーを、複数のルールから成るルール集合で表すことも可能である。添付図Aに、図3に示したエンロール用テキストに対応したルール型グラマーの可能な例を示した。ルール型グラマーの利点の1つは、ユーザがエンロール用テキスト上を前進して行くにつれて、変更を加えて行く必要がないことである。

【0052】

一般的に、エンロールメント・プログラムを実行することによって、音声認識システムの誤認識発生率を低下させることができる。それが可能であるのは、ユーザがエンロール用テキストを読み上げることによって、様々な音響モデルに対応したそのユーザの音声が集められ、それら発音に基づいて、それら音響モデルがそのユーザの音声に適応するようにカスタマイズされるからである。ただし、ユーザが、本来読み上げるはずのエンロール用テキスト以外の何か他の言葉を発した場合には、その発音に基づいて、本来読み上げるはずのエンロール用テキストに対応した音響モデルの更新を行う訳には行かない。このような状況が発生することがあるのは、例えば、ディスプレイ上に表示されているテキスト1210のうちの1行をユーザが読み飛ばしてしまった場合や、ディスプレイ上に表示されている単語をユーザが別の単語に読み違えてしまった場合、それに、ディスプレイ上に表示されているテキストを読み上げているときに話しかけられて、ユーザが他者と会話をしてしまった場合などである。また、このような状況は更に、音声認識システムが、ユーザの音声以外の咳やくしゃみ、それに交通騒音や犬の吠え声などの環境騒音を拾ってしまった場合にも発生することがある。

20

【0053】

30

エンロール用グラマーを使用することによって、認識部215は、エンロール用テキストのうちの、あるユーザ・アタランスに最も良く対応している部分を識別することができる。しかしながら、エンロール用グラマーを使用しても、認識部215は、そのユーザ・アタランスが実際にそのエンロール用テキストに対応しているか否かを検証することまではできない。その検証を可能にするために、ここでは、エンロールメント・プログラムが、エンロール用グラマーと併用する形でリジェクション・グラマーを使用するようにしており、これによって、認識部215が検出した音声で、本来ユーザが読み上げるはずのエンロール用テキストとマッチングしていないという状況を識別できるようにしている。

【0054】

40

図6は、図5のステップ1540のプロセスを更に詳細に示した図であり、このプロセスでは、エンロール用グラマーとリジェクション・グラマーとを併用することで、ユーザ・アタランスに基づいてエンロール用テキストの音響モデルの更新を行ってはいない状況を検出するようにしている。そのために、エンロールメント・プログラムは、エンロール用グラマーに対応した最良の認識候補を識別して、認識部215がその認識候補に付与したスコアを、エンロールメント・スコアに指定する(ステップ1541)。このとき、認識部215は、例えば音響モデル及び単語の前後関係についてのマッチングを行うことによって、その認識候補に付与するスコアを決定するようにすればよい。

【0055】

たとえ最良スコアの認識候補であっても、エンロール用テキストに良好にマッチングしていないことがある。例えば、図3に示すように、エンロールメント・ポジションが単語「

50

f o r」に設定されているにもかかわらず、ユーザがそれを見誤って発声したために、そのユーザ・アタランスが、エンロールメント・ポジションの1つ前の行に示されている「f r o m A r t h u r c . C l a r k e ' s n o v e l」になってしまったものとする。この場合、認識部215は、そのユーザ・アタランスと、エンロール用グラマーのうちの「f o r」以降の部分との間のマッチングを探索するため、認識部215が生成する最良のマッチングを示す認識候補は、エンロール用グラマーのうちの「f o r」以降の部分に対して最良のマッチングを示す認識候補になってしまう。

【0056】

エンロールメント・プログラムは更に、リジェクション・グラマーに対応した最良の認識候補を識別して、認識部215がその認識候補に付与したスコアを、リジェクション・スコアに指定する(ステップ1542)。一般的に、ユーザ・アタランスが、エンロール用テキストのうちの本来読み上げられるはずの部分に対応していない場合には、リジェクション・スコアの方がエンロールメント・スコアより良好なスコアになる。このリジェクション・グラマーをE B N Fで表すと、次のように表される。

<認識結果>::=<音素列>

このE B N F表記において、

<音素列>::=音素[<音素列>]、である。

また、ここで音素として表示しているのは、リジェクション・グラマー用音素集合に含まれている任意の音素である。リジェクション・グラマー用音素集合は、発声されることが予想される音声に含まれる殆どの音を、ある程度良好にモデル化できるように選定した音素から成る集合である。上のE B N F表記に示したように、リジェクション・グラマーを、それら音素のいかなる組合せをも許容するように定めてある場合には、あるアタランスに対して最良のマッチングを示す音素列を求めるために必要とされる処理量が膨大なものとなるおそれがある。そこで、リジェクション・グラマー音素集合は、認識部215が使用する音素集合のうちの部分集合に相当する、少数の音素しか含まないものにしておくのがよい。例えば、添付図Bに示した例では、全体の音素集合が50個の音素から成るのに対して、リジェクション・グラマー用音素集合は、その部分集合に相当する19個の音素しか含まない音素集合となっている。

【0057】

リジェクション・グラマーは、そのユーザ・アタランスを構成している複数の音響を、リジェクション・グラマー音素集合に含まれる音素から成る音素列にマッチングさせる。リジェクション・グラマー認識候補に付与されるスコアは、そのユーザ・アタランスが、良好にマッチングしていると判定された音素列にどれ程緊密に対応しているかの程度を反映しており、そのユーザ・アタランスがエンロール用テキストのどの部分にも良好にマッチングしていない場合には、リジェクション・スコアの方がエンロールメント・スコアより良いスコアになる確率が高くなるようにしてある。

【0058】

再び図3の具体例について説明する。この具体例では、エンロールメント・ポジションは「f o r」の位置にあるのに対して、ユーザ・アタランスは「f r o m A r t h u r C . C l a r k e ' s n o v e l」である。そのため、エンロール用グラマーを使用して得られる認識候補のスコアは、良好ではないマッチングを反映したスコアとなる可能性が大きい。一方、リジェクション・グラマーを使用して得られる認識候補は、たとえ正確に「f r o m A r t h u r C . C l a r k e ' s n o v e l」とはならないにしても、「f o r」からはじまるエンロールメント認識候補と比べれば、これに近いものとなる可能性が大きい。従って、リジェクション・グラマーを使用して得られる認識候補は、より良好なマッチングを示すものとなり、ひいてはより良いスコアを有するものとなる。

【0059】

エンロールメント・プログラムは、そのユーザ・アタランスがエンロールメント・アタランスとマッチングしているか否かを判定するために、エンロールメント・スコアとリジェ

10

20

30

40

50

クション・スコアとを比較する（ステップ1543）。この比較のための具体的な方法としては、様々な方法が利用可能である。例えば、簡単な方式の1つとして、エンロールメント・プログラムが、エンロールメント・スコアとリジェクション・スコアとを比較して、エンロールメント・スコアの方がリジェクション・スコアより良好なスコアであれば、そのユーザ・アタランスがエンロール用テキストのいずれかの部分にマッチングしていると判定するという方法がある（ステップ1544）。一方、上述の具体例のように、リジェクション・スコアの方が、エンロールメント・スコアよりも良好なスコアであるならば、エンロールメント・プログラムは、そのユーザ・アタランスが、エンロール用テキストのどの部分にもマッチングしていないものと判定する（ステップ1545）。

【0060】

以上に、リジェクション・グラマーをエンロール用グラマーと併用する場合について説明した。しかしながら、リジェクション・グラマーは、その他の使い方をすることもできる。即ち、リジェクション・グラマーは、一般的に、ラージ・ボキャブラリ・ディクテーション制約グラマーがアクティブ状態にないときに使用することができる。例えば、音声認識ソフトウェアが、リジェクション・グラマーと、「ウェイク・アップ（Wake Up）」コマンドを認識するためのグラマーとを除いた、その他全ての制約グラマーを非アクティブ状態にする「ゴー・ツー・スリープ（go to sleep）」コマンドを含んでいることがある。通常、このような「ゴー・ツー・スリープ」コマンドが実行されたならば、認識部215は、「ウェイク・アップ」コマンドしか認識しなくなる。このような状況において、制御／インターフェース・モジュール220は、リジェクション・グラマーを使用することで、「ウェイク・アップ」コマンドと、ノイズ等のその他の音とを区別することができる。

【0061】

図7は、このノンインタラクティブ方式のエンロールメント・プログラムが動作するときの処理手順1700を示したものである。エンロールメント・プログラムは、最初に、エンロール用テキストに対応したユーザの音声を収集する（ステップ1705）。このユーザの音声は、エンロール用テキストを読み上げたユーザの音声の全体であり、様々な方法で収集し得るものである。例えば、ユーザは、印刷されたテキストを読み上げて、アナログ録音装置に音声を録音するようにしてもよい。この場合、続いてユーザは、その録音した音声をポート165を介してコンピュータ125へ入力する。ここで使用する印刷されたテキストは、エンロールメント・プログラムが提供するようにしてもよい。

【0062】

また、ユーザが、デジタル録音装置を用いて音声を録音するようにしてもよい。この場合には、ユーザは、その録音した音声に対応したデジタル・ファイルをポート170を介してコンピュータ125へ入力すればよい。

【0063】

更に、ユーザが、マイクロホン105を使用して、印刷されたテキストまたはディスプレイ120上に表示されたテキストを読み上げるようにしてもよい。インタラクティブ方式のエンロールメント・プログラムとは対照的に、このノンインタラクティブ方式のエンロールメント・プログラムは、ユーザにプロンプトを発して、エンロール用テキスト上のユーザの位置を指示したりすることはない。即ち、このノンインタラクティブ方式のエンロールメント・プログラムは、例えばテキスト・エディタや、それに類似したプログラムを実行させて、エンロール用テキストの全体をディスプレイ上に表示し、ユーザが、そのエンロール用テキスト上のどこへでも自由に移動できるようにしておくだけでよい。また別法として、エンロールメント・プログラムが、エンロール用テキストを紙上に打ち出して、ユーザがそれを読み上げるようにしてもよい。

【0064】

ユーザがエンロール用テキストを読み上げたならば、プログラムはその音声を処理して、複数のユーザ・アタランスの夫々に対応した複数のフレーム集合を生成する（ステップ1710）。ポート165またはマイクロホン105を介して音声が入力される場合には、

10

20

30

40

50

エンロールメント・プログラムは、サウンド・カード140を介して、また上述の処理手順300に従って、その音声进行处理する。また、ポート170を介して音声が入力される場合には、一般的に、その入力に使用されるデジタル・ファイルの内容のフォーマット変更が必要である。またその場合に、そのデジタル・ファイルの内容を読みっているうちに十分な長さの沈黙（例えば30秒）に遭遇したならば、そのことをもって音声が終わったものとエンロールメント・プログラムが判定するようにしてもよい。

【0065】

次に、エンロールメント・プログラムは、エンロール用テキストのアクティブ部分を指定し、そのアクティブ部分に対応したエンロールメント制約グラマーを構築する（ステップ1715）。例えば、最初は、エンロール用テキストの先頭の3個ないし5個のパラグラフを、エンロール用テキストのアクティブ部分として指定してもよい。この場合、エンロールメント・プログラムは、それら指定した複数のパラグラフの各々に対応した夫々の制約グラマーを含むエンロール用グラマーを構築する。一般的に、それら制約グラマーは、融通性を有するものであり、ユーザが、エンロール用テキストに含まれている複数の単語を、それらがエンロール用テキストに出現する順序で発声していること以外は要求しないような制約グラマーである。具体的には、例えば、そのアタランスの先頭の単語及び末尾の単語は、エンロール用グラマーに含まれている複数の単語のうちのどの単語でもよく、単に、そのアタランスの先頭の単語から末尾の単語までの全ての単語がエンロール用グラマーの中に含まれていて、しかも、そのアタランスにおけるそれら単語の順序が、エンロール用グラマーにおけるそれら単語の順番と前後することがないということだけを、要求する。従って、このエンロール用グラマーは、ユーザが、単語やセンテンスを読み飛ばすことも、パラグラフをまるごと読み飛ばすことも許容しており、更には、それらを反復して読み上げることも許容している。

【0066】

エンロール用テキストの先頭の5個のパラグラフに対応したエンロール用グラマーをE B N Fで表すと次のように表される。

<認識結果>：：=<第1番パラグラフ> | <第2番パラグラフ> |
<第3番パラグラフ> | <第4番パラグラフ> | <第5番パラグラフ>

このE B N F表記において、

<第n番パラグラフ>：：=[単語¹ⁿ [単語²ⁿ [単語³ⁿ. . . 単語^{mn}]]]
| [単語²ⁿ [単語³ⁿ. . . 単語^{mn}]] | . . . [単語^{mn}]、であり、

また、

「単語^{i n}」は、第n番パラグラフの第i番の単語である。

既述のごとく、エンロール用グラマーは、ユーザ・アタランスがエンロール用テキストのいずれかの部分にマッチングしているか否かを判定する際に、ユーザがエンロール用テキストを読み上げるときの予想される読み上げ方のバリエーションを認識して、それら読み上げ方に適切に対処できるように定めるのがよい。そのため、上で説明したインタラクティブ・モードのエンロール用グラマーと同様に、このノンインタラクティブ・モードのエンロール用グラマーも、例えば句読点、段落書式、それにフォーマット等々を考慮に入れたものとしている。

【0067】

続いて、エンロールメント・プログラムは、1個のアタランスを取出し（ステップ1720）、そして認識部215に、アクティブ状態にあるエンロール用グラマーを使用してそのアタランスに対する認識処理を実行するよう指令する（ステップ1725）。この認識処理は、リジェクション・グラマーを使用して、インタラクティブ・モードのエンロールメント・プログラムに関連して上で説明した手順で実行される。

【0068】

認識部215が、そのアタランスの中に、エンロール用テキストにマッチングする単語を発見したならば（ステップ730）、エンロールメント・プログラムは、その発見した単語の音素に対応した音響モデルの更新を行う（ステップ1735）。この点に関する別法

として、全てのアタランスの解析が完了した後に、エンロールメント・プログラムが音響モデルの更新を行うようにしてもよい。

【0069】

音響モデルの更新を行った後に、エンロールメント・プログラムは、エンロール用テキストのアクティブ部分を更新する必要があるか否かを判定する（ステップ1740）。一般的に、エンロールメント・プログラムは、アクティブ部分の更新を行うことで、アクティブ状態にあるパラグラフで画成されるアクティブ・ウィンドウが、常に、エンロール用テキスト上のユーザ音声の現在ポジションを包含する位置にあるようにする。従って、例えば、エンロールメント・プログラムが常時、5個のパラグラフをアクティブ部分として維持するようにしている場合に、認識部215がエンロール用テキストの第4番パラグラフに含まれているテキストを識別した時点で、エンロール用テキストの第1番パラグラフ～第5番パラグラフがアクティブ部分として選択されていたならば、エンロールメント・プログラムが実行するアクティブ部分の更新処理においては、新たに第2番パラグラフ～第6番パラグラフをアクティブ部分として指定すればよい。

【0070】

エンロールメント・プログラムは、エンロール用テキストのアクティブ部分を更新する必要があるならば、そのアクティブ部分の更新を行うと共に、アクティブ・エンロール用文法の更新も行う（ステップ1745）。従って、上述の具体例についていえば、エンロールメント・プログラムは、その制約文法のうちの、エンロール用テキストの第1番パラグラフに対応する部分を削除し、第6番パラグラフに対応する部分を追加すればよい。こうして変更が加えられたエンロール用文法をEBNFで表すと次のように表される。

<認識結果>：：=<第2番パラグラフ> | <第3番パラグラフ> |

<第4番パラグラフ> | <第5番パラグラフ> | <第6番パラグラフ>

こうして処理を完了したアタランスが、最後のアタランスではなかったならば（ステップ1750）、エンロールメント・プログラムは次のアタランスを取出して（ステップ1720）、以上のプロセスを反復して実行する。

【0071】

最後のアタランスを処理するとき、エンロールメント・プログラムは、エンロールメント・パフォーマンス・レポートをディスプレイ上に表示する（ステップ1755）。このレポートには、例えば、ユーザの音声に含まれる単語のうちで認識できた単語のパーセンテージ表示等を含むものである。このレポートには更に、修正済音響モデルを使用した場合の認識パフォーマンスの予測値の表示を含めるようにしてもよい。ユーザは、このレポートの情報に基づいて、同じエンロールメント作業を再度実行することが効果的か否か、また、別のエンロール用テキストを使用してエンロールメント作業を行うことが効果的か否かを判断することができる。

【0072】

エンロールメント・プログラムは、各ユーザのアタランスに基づいて、そのエンロール用テキストに対応した、発声者独立の（即ち、発声者に依存しない）音響モデルに適応化を施して、特定のユーザに適応した発声者適応音響モデルを生成する。既述のごとく、発声者独立音響モデルは、トライフォンの各ノードを、複数のガウス型（正規分布型）確率密度関数（PDF） N_k の混合によって表したものであり、それら確率密度関数の平均ベクトルは μ_k で表され、共分散行列は c_k で表される。また、各PDFのトライフォンのノードへの寄与は、混合重み w_k で表される。発声者適応音響モデルは、その混合重み w_k 及び共分散行列 c_k の値が、発声者独立音響モデルのものと同一値であり、その平均ベクトルだけが、適応平均ベクトル μ_{kA} になっている。

【0073】

図8は、エンロールメント・プログラムが適応平均ベクトル μ_{kA} を生成する際の処理手順1800を示したものである。エンロールメント・プログラムは、発声者独立音響モデルの平均ベクトル μ_k を出発値とし（ステップ1805）、変換を利用した方式で、適応平

均ベクトル μ_{kA} を生成する。この変換を利用した方式は、適応平均ベクトル μ_{kA} が次の〔式 3〕で表されることを前提としている。

【0074】

【数 3】

$$\mu_{kA} = A\mu_k + b \quad (3)$$

この〔式 3〕において、A 及び b は変換である。特に、平均ベクトルが 24 個の成分を有するベクトルである場合には、A は 24×24 の変換行列であり、b は 24 個の成分を有する変換ベクトルである。

10

【0075】

次に、エンrollment・プログラムは、同一の変換を共有すると予測される PDF の集合を編成する（ステップ 1810）。1 つの集合 C は、音の響きが互いに類似したトライフォンを表す複数の PDF から成り、例えば、母音に関連した全ての PDF や、子音に関連した全ての PDF をもって集合とすることができる。集合を編成するための第 2 の方法は、平均ベクトル μ_k の値が互いに類似した PDF を同じグループに入れるというものである。この方法によれば、多数の集合を簡単に編成することができる。一般的に、このような集合の個数は、百個程度から数百個にも及ぶことがある。初歩的な実験を行った結果、第 1 の方法を用いることで、良好な結果が得られる（即ち、誤認識発生率が低いという結果が得られる）ことが判明した。一般的に、エンrollment・セッションを実行する前に、予め集合を定めておくのがよい。

20

【0076】

次に、エンrollment・プログラムは、各々の集合について、その集合に含まれている PDF に対応した全ての発声者データをアセンブルする（ステップ 1815）。この発声者データは、エンrollment・セッションの実行中に生成されたユーザ・アタランズに対応したデータである。

【0077】

次に、エンrollment・プログラムは、各々の PDF j について、発声者データのベクトル平均値 $y[j]$ を求める。ここで、 $f[n]$ を発声者データの第 n 番フレームのベクトルとし、 $p_j(n)$ を発声者データの第 n 番フレームが PDF j（第 j 番 PDF）に対応している確率（即ち、その発声者データの第 n 番フレームに対応したトライフォン・ノードの混合重み w_j ）とし、N をその発声者データのフレームの総数とすると、平均ベクトル $y[j]$ と、第 j 番フレームのフレーム個数 $N[j]$ とは、次の〔式 4〕及び〔式 5〕で表される。

30

【0078】

【数 4】

$$y[j] = \left(\sum_{n=1}^N p_j(n) \right)^{-1} \sum_{n=1}^N p_j(n) f[n] \quad (4)$$

40

【0079】

【数 5】

$$N[j] = \sum_{n=1}^N p_j(n) \quad (5)$$

これは、通常、Baum-Welch アルゴリズム、または EM アルゴリズムと呼ばれて

50

いるものである。

【0080】

次に、エンロールメント・プログラムは、各集合Cの変換（ A_C 及び b_C ）を求め、それには、発声者データの平均値と、発声者独立音響モデルのPDFとの間の関係を利用する（ステップ1825）。この関係は、次の〔式6〕で表される。

【0081】

【数6】

$$y[j] = A_C x[j] + b_C + e \quad (6) \quad 10$$

この〔式6〕において、 $x[j]$ は、第j番PDFの平均ベクトル μ_j に対応したベクトルであり、 e は誤差項である。この〔式6〕の関係に基づいて変換行列 A_C の転置行列 A_C^T を求めると、次の〔式7〕に示すようになる。

【0082】

【数7】

$$A_C^T = \left(\sum_{j=1}^{N_C} N(j) x[j] x^T[j] \right)^{-1} \sum_{j=1}^{N_C} N(j) x[j] y^T[j] \quad (7) \quad 20$$

この〔式7〕において、 $x^T[j]$ はベクトル $x[j]$ の転置ベクトルであり、 $y^T[j]$ はベクトル $y[j]$ の転置ベクトルであり、 N_C は集合Cの元の個数である。

【0083】

続いて、 A_C^T の転置行列を求めることによって、変換行列 A_C を求め、また、次の〔式8〕に従って変換ベクトル b_C を求める。

【0084】

【数8】

$$b_C = \left(\sum_{j=1}^{N_C} N(j) \right)^{-1} \sum_{j=1}^{N_C} N(j) (y[j] - A_C x[j]) \quad (8) \quad 30$$

続いて、エンロールメント・プログラムは、以上のようにして求めた変換行列及び変換ベクトルを用いて、集合Cの各成分の適応平均ベクトル μ_{jA} を求める（ステップ1830）。具体的には、適応平均ベクトル μ_{jA} は、次の〔式9〕によって求められる。

【0085】

【数9】

$$\mu_{jA} = A_C \mu_j + b_C \quad (9) \quad 40$$

続いて、エンロールメント・プログラムは、適応平均ベクトル μ_{jA} と、発声者識別情報とを組合せたものを、発声者適応音響モデルとして格納し、この音響モデルを、後刻そのユーザの音声を認識する際に使用する。以上のプロセスを、適応モデルを使用して数回反復して実行することで、フレーム個数（ $N[j]$ ）と、発声者データ（ $y[j]$ ）と、変換行列とを更新する。

【0086】

別の実施の形態として、発声者独立の平均値（ $x[j]$ ）と、特定発声者の平均値とを用いて、 $y[j]$ 及び $N[j]$ の値を次の〔式 10〕及び〔式 11〕に従って算出してもよい。

【0087】

【数 10】

$$y[j] = \left(r + \sum_{n=1}^N p_j(n) \right)^{-1} \left(\sum_{n=1}^N p_j(n) f[n] + r x[j] \right) \quad (10)$$

10

【0088】

【数 11】

$$N[j] = r + \sum_{n=1}^N p_j(n) \quad (11)$$

これらの〔式 10〕及び〔式 11〕において、係数 r は、発声者独立の平均値及び観測データの相対的重要度を制御する係数である。この係数 r の値の最適化の程度は、利用可能な適応データの量の関数となる。

20

【0089】

請求項に記載した範囲には、その他の実施の形態も含まれる。例えば、エンロールメント・グラマーとリジェクション・グラマーとを用いてユーザ・アタランズに夫々にスコアを付与するための具体的な方法は、様々な方法とすることができる。また同様に、それらスコアを評価して、そのユーザ・アタランズがエンロール用テキストのある部分にマッチングするか否かを判定する方法も、様々な方法とすることができる。

【0090】

以上に説明した様々な技法は、特定のハードウェアないしソフトウェアの形態に限定されるものではなく、音声認識を実行することのできるコンピュータ環境ないし処理環境であれば、いかなる環境においても適用可能な技法である。また、それら技法は、ハードウェアで実施することも、ソフトウェアで実施することも、更にはハードウェアとソフトウェアとを組合せて実施することも可能である。それら技法は、プログラム可能なコンピュータ上で動作するコンピュータ・プログラムによって実施することが好ましく、この場合のコンピュータは、プロセッサと、そのプロセッサで読出しが可能な格納媒体（揮発性ないし不揮発性のメモリや、格納装置等）と、少なくとも 1 つの入力デバイスと、少なくとも 1 つの出力デバイスとを備えたものであることが好ましい。更にその場合に、入力デバイスを使用して入力したデータに対して、プログラム・コードを適用することで、以上に説明した機能を実行し、出力情報を生成すればよい。そして、その出力情報を、1 つないし複数の出力デバイスに供給するようにすればよい。

30

【0091】

いずれのプログラムも、コンピュータ・システムとコミュニケーションすることのできるハイレベルの手続型またはオブジェクト指向のプログラミング言語で作成したプログラムとすることが好ましい。ただし場合によっては、それらプログラムをアセンブリ言語またはマシン語で作成することが好ましいことがあるかもしれず、そのような場合には、それら言語を使用すればよい。更に、いずれの場合も、それら言語をコンパイルして使用するようにしてもよく、或いは、インタープリター言語として使用するようにしてもよい。

40

【0092】

それらコンピュータ・プログラムは、いずれも、記録媒体ないし記録装置（例えば、CD-ROM、ハード・ディスク、または磁気ディスク）に格納しておき、それらに格納してあるものを、プログラム可能な汎用型ないし特定用途向けのコンピュータが読出し、

50

その読出したコンピュータ・プログラムによって、そのコンピュータのコンフィギュレーションを設定し、また、そのコンピュータを動作させ、それによって、本明細書に説明した処理手順を実行させるようにすることが好ましい。本発明の音声認識システムは更に、コンピュータによって読出しが可能な、コンピュータ・プログラムを設定した記録媒体として実現することもでき、この場合、コンピュータ・プログラムを設定したその記録媒体が、コンピュータを、以上に説明した特定の方式で動作させることになる。

【図面の簡単な説明】

【図 1】 音声認識システムのブロック図である。

【図 2】 図 1 のシステムの音声認識ソフトウェアのブロック図である。

【図 3】 エンロールメント・プログラムのためのユーザ・インターフェースを示した図である。

10

【図 4】 エンロールメント・プログラムのためのユーザ・インターフェースを示した図である。

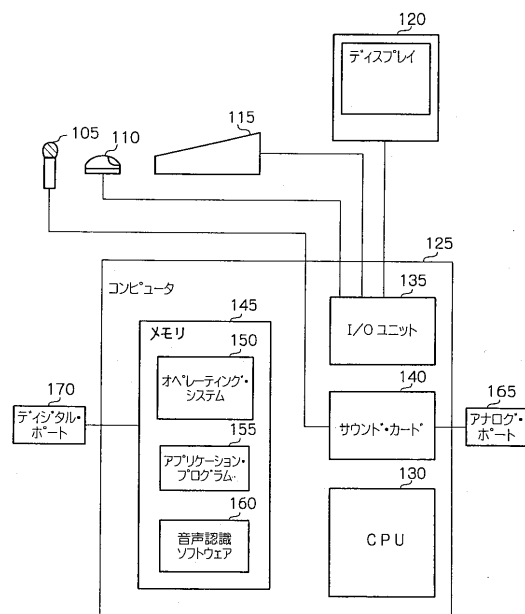
【図 5】 エンロールメント作業の手順を示したフローチャートである。

【図 6】 エンロールメント作業の手順を示したフローチャートである。

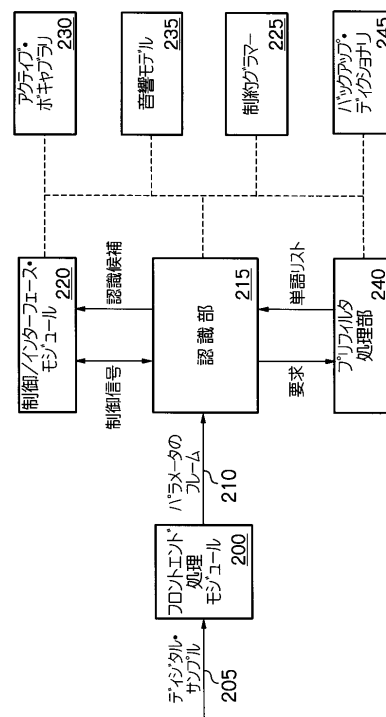
【図 7】 エンロールメント作業の手順を示したフローチャートである。

【図 8】 音響モデル適応化作業のフローチャートである。

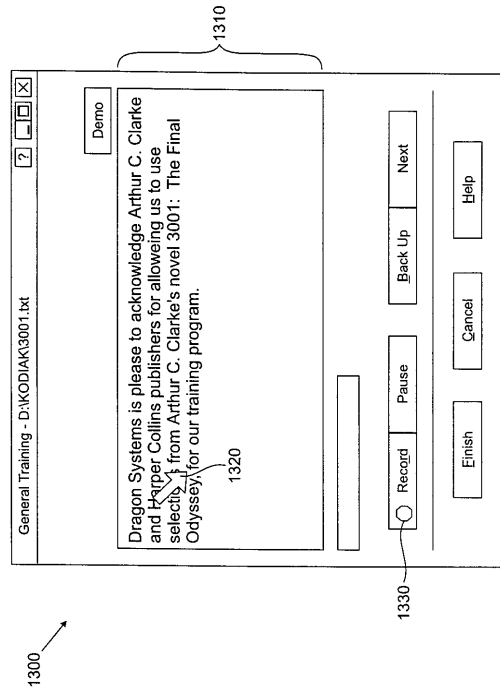
【図 1】



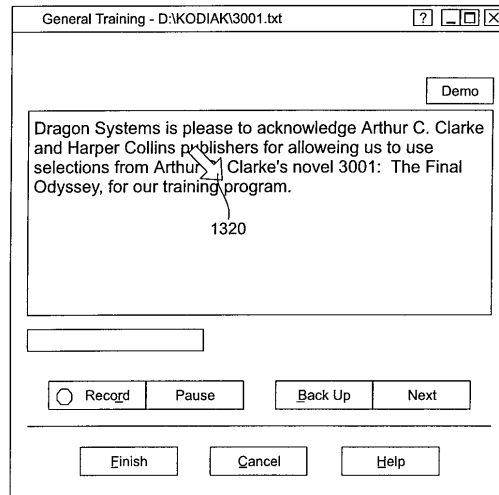
【図 2】



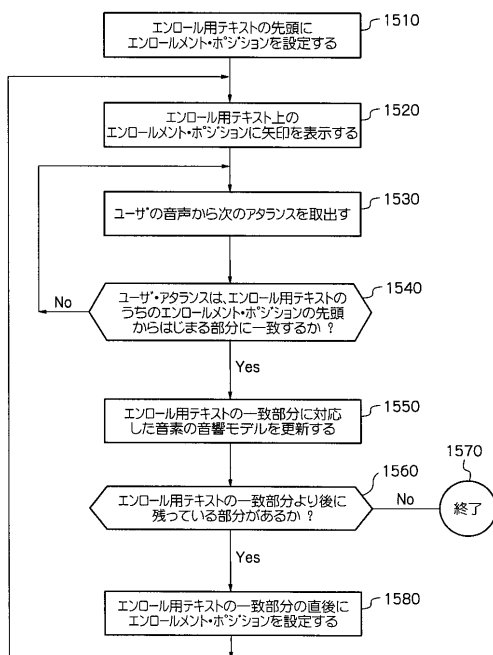
【図 3】



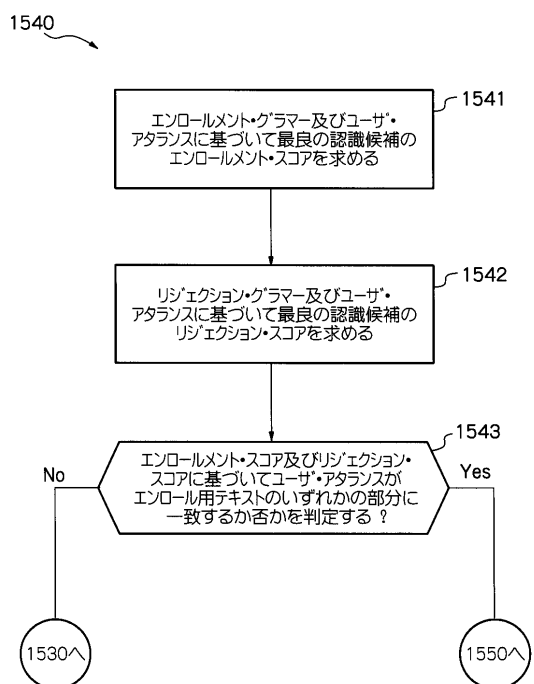
【図 4】



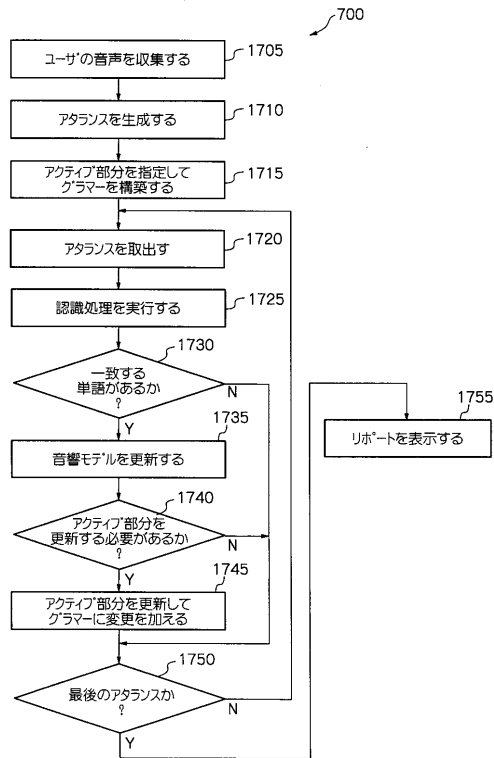
【図 5】



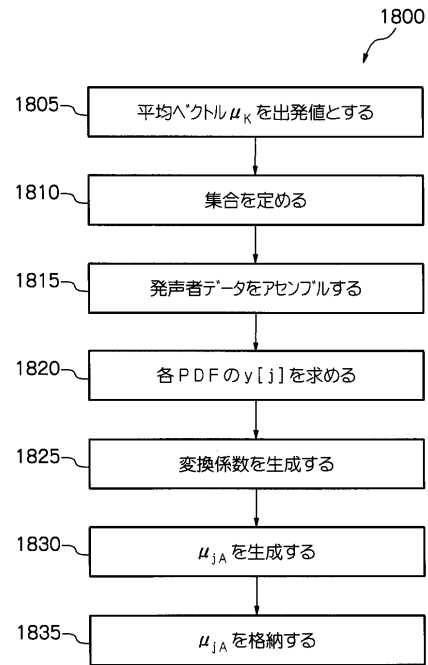
【図 6】



【図 7】



【図 8】



フロントページの続き

(74)代理人 100075270

弁理士 小林 泰

(72)発明者 ステファン・シャーウッド

アメリカ合衆国マサチューセッツ州02178, ベルモント, ビーチ・ストリート 146

(72)発明者 デービッド・ウィルスバーク・パーメンター

アメリカ合衆国マサチューセッツ州02158, ニュートン, ハムンウェル・アベニュー 165

(72)発明者 ジョエル・グールド

アメリカ合衆国マサチューセッツ州01890, ウィンチェスター, ソーンベリー・ロード 50

(72)発明者 トッフィー・エイ・アルビナ

アメリカ合衆国マサチューセッツ州02139, ケンブリッジ, レナード・アベニュー 13

(72)発明者 アラン・ゴールド

アメリカ合衆国マサチューセッツ州01720, アクトン, ジョン・スウィフト・ロード 22

審査官 山下 剛史

(56)参考文献 特開平02-193199 (JP, A)

特開平01-106098 (JP, A)

特開昭63-149699 (JP, A)

特開昭60-113298 (JP, A)

(58)調査した分野(Int.Cl., DB名)

G10L 15/00-17/00