

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4798606号
(P4798606)

(45) 発行日 平成23年10月19日(2011.10.19)

(24) 登録日 平成23年8月12日(2011.8.12)

(51) Int.Cl.	F I
G 1 0 L 15/10 (2006.01)	G 1 0 L 15/10 3 0 0 G
G 1 0 L 15/22 (2006.01)	G 1 0 L 15/22 4 5 3

請求項の数 7 (全 28 頁)

(21) 出願番号	特願2005-319094 (P2005-319094)	(73) 特許権者	393031586
(22) 出願日	平成17年11月2日(2005.11.2)		株式会社国際電気通信基礎技術研究所
(65) 公開番号	特開2007-127738 (P2007-127738A)		京都府相楽郡精華町光台二丁目2番地2
(43) 公開日	平成19年5月24日(2007.5.24)	(74) 代理人	100115749
審査請求日	平成20年3月26日(2008.3.26)		弁理士 谷川 英和
(出願人による申告) 平成17年度独立行政法人情報通信研究機構、研究テーマ「人間情報コミュニケーションの研究開発」に関する委託研究、産業活力再生特別措置法第30条の適用を受ける特許出願		(72) 発明者	田川 博章
			京都府相楽郡精華町光台二丁目2番地2
		(72) 発明者	株式会社国際電気通信基礎技術研究所内
			足立 隆弘
			京都府相楽郡精華町光台二丁目2番地2
		(72) 発明者	株式会社国際電気通信基礎技術研究所内
			山田 玲子
			京都府相楽郡精華町光台二丁目2番地2
			株式会社国際電気通信基礎技術研究所内

最終頁に続く

(54) 【発明の名称】 音声認識装置、およびプログラム

(57) 【特許請求の範囲】

【請求項1】

比較される対象の音声に関するデータである音響データを格納している音響データ格納部と、

音声の入力を受け付ける音声入力部と、

前記音響データを用いて、前記音声入力部が受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する音声認識部と、

前記音声認識部が音声認識処理を行った後、前記音声入力部が受け付けた音声に対して、前記音響データを用いて、発音の良し悪しに関する評価を行う処理である発音評価処理を行い、発音評価結果を取得する発音評価部と、

前記発音評価結果が所定の閾値に対して所定の関係にあるか否かを判断する判断部と、

前記判断部が所定の関係にあると判断した場合のみ、前記音声認識部が取得した音声認識処理結果情報についての音声認識結果を出力する出力部を具備し、

前記音響データは、

状態を識別する2以上の状態識別子と状態間の遷移確率の情報を有するデータであり、

前記発音評価部は、

前記音声入力部が受け付けた音声を、フレームに区分するフレーム区分手段と、

前記区分されたフレーム毎の音声データであるフレーム音声データを1以上得るフレーム音声データ取得手段と、

前記1以上のフレーム音声データの前記遷移確率が最も高い状態である最適状態を決定す

10

20

る最適状態決定手段と、

前記最適状態決定手段が決定した最適状態の事後確率を示す確率値を、発音区間毎に取得する発音区間確率値取得手段と、

前記発音区間確率値取得手段が取得した1以上の発音区間毎の1以上の確率値の時間平均値を算出し、当該時間平均値を用いて音声の評定値を算出する評定値算出手段とを具備する音声認識装置。

【請求項2】

比較される対象の音声に関するデータである音響データを格納している音響データ格納部と、

音声の入力を受け付ける音声入力部と、

前記音響データを用いて、前記音声入力部が受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する音声認識部と、

前記音声認識部が音声認識処理を行った後、前記音声入力部が受け付けた音声に対して、前記音響データを用いて、発音の良し悪しに関する評価を行う処理である発音評定処理を行い、発音評定結果を取得する発音評定部と、

前記発音評定結果が所定の閾値に対して所定の関係にあるか否かを判断する判断部と、

前記判断部が所定の関係にあると判断した場合のみ、前記音声認識部が取得した音声認識処理結果情報についての音声認識結果を出力する出力部を具備し、

前記音響データは、

状態を識別する2以上の状態識別子と状態間の遷移確率の情報を有するデータであり、

前記発音評定部は、

前記音声入力部が受け付けた音声を、フレームに区分するフレーム区分手段と、

前記1以上のフレーム音声データの前記遷移確率が最も高い状態である最適状態を決定する最適状態決定手段と、

前記最適状態決定手段が決定した各フレームの最適状態を有する音韻全体の状態における1以上の事後確率を示す確率値を、発音区間毎に取得する発音区間フレーム音韻確率値取得手段と、

前記発音区間フレーム音韻確率値取得手段が取得した1以上の発音区間毎の1以上の確率値の総和に基づいて、発音区間毎の確率値の総和の時間平均値を1以上算出し、当該1以上の時間平均値を用いて音声の評定値を算出する評定値算出手段を具備する音声認識装置

。

【請求項3】

前記発音評定部は、

無音を示すデータである無音データを格納している無音データ格納手段と、

前記入力音声データおよび前記無音データに基づいて、無音の区間を検出する無音区間検出手段をさらに具備し、

前記最適状態決定手段は、

前記無音の区間についてのフレーム音声データを除いた前記1以上のフレーム音声データの最適状態を決定する請求項1または請求項2記載の音声認識装置。

【請求項4】

前記発音評定部は、

前記音声入力部が受け付けた音声に対して、前記音声認識部が取得した音声認識処理結果情報と前記音響データを用いて発音評定処理を行い、発音評定結果を取得する請求項1から請求項3いずれか記載の音声認識装置。

【請求項5】

音声認識処理結果情報は、

音素トランスクリプションであり、

前記発音評定部は、

前記音声認識部が取得した音素トランスクリプションに対応する音響データを用いて発音評定処理を行い、発音評定結果を取得する請求項4記載の音声認識装置。

【請求項 6】

コンピュータに、
音声の入力を受け付ける音声入力ステップと、
格納されている音響データを用いて、前記音声入力ステップで受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する音声認識ステップと、
前記音声認識処理の後、前記音声入力ステップで受け付けた音声に対して、前記音響データを用いて、発音の良し悪しに関する評価を行う処理である発音評価処理を行い、発音評価結果を取得する発音評価ステップと、
前記発音評価結果が所定の閾値に対して所定の関係にあるか否かを判断する判断ステップと、
前記判断ステップで所定の関係にあると判断した場合のみ、前記音声認識ステップで取得した音声認識処理結果情報についての音声認識結果を出力する出力ステップを実行させるためのプログラムであって、
前記音響データは、
状態を識別する 2 以上の状態識別子と状態間の遷移確率の情報を有するデータであり、
前記発音評価ステップは、
前記音声入力ステップで受け付けられた音声を、フレームに区分するフレーム区分ステップと、
前記区分されたフレーム毎の音声データであるフレーム音声データを 1 以上得るフレーム音声データ取得ステップと、
前記 1 以上のフレーム音声データの前記遷移確率が最も高い状態である最適状態を決定する最適状態決定ステップと、
前記最適状態決定ステップで決定された最適状態の事後確率を示す確率値を、発音区間毎に取得する発音区間確率値取得ステップと、
前記発音区間確率値取得ステップで取得された 1 以上の発音区間毎の 1 以上の確率値の時間平均値を算出し、当該時間平均値を用いて音声の評価値を算出する評価値算出ステップとを具備するプログラム。

【請求項 7】

コンピュータに、
音声の入力を受け付ける音声入力ステップと、
格納されている音響データを用いて、前記音声入力ステップで受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する音声認識ステップと、
前記音声認識処理の後、前記音声入力ステップで受け付けた音声に対して、前記音響データを用いて、発音の良し悪しに関する評価を行う処理である発音評価処理を行い、発音評価結果を取得する発音評価ステップと、
前記発音評価結果が所定の閾値に対して所定の関係にあるか否かを判断する判断ステップと、
前記判断ステップで所定の関係にあると判断した場合のみ、前記音声認識ステップで取得した音声認識処理結果情報についての音声認識結果を出力する出力ステップを実行させるためのプログラムであって、
前記音響データは、
状態を識別する 2 以上の状態識別子と状態間の遷移確率の情報を有するデータであり、
前記発音評価ステップは、
前記音声入力ステップで受け付けられた音声を、フレームに区分するフレーム区分ステップと、
前記 1 以上のフレーム音声データの前記遷移確率が最も高い状態である最適状態を決定する最適状態決定ステップと、
前記最適状態決定ステップで決定された各フレームの最適状態を有する音韻全体の状態に

おける 1 以上の事後確率を示す確率値を、発音区間毎に取得する発音区間フレーム音韻確率値取得ステップと、

前記発音区間フレーム音韻確率値取得ステップで取得された 1 以上の発音区間毎の 1 以上の確率値の総和に基づいて、発音区間毎の確率値の総和の時間平均値を 1 以上算出し、当該 1 以上の時間平均値を用いて音声の評定値を算出する評定値算出ステップとを具備するプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、音声認識処理を行う音声認識装置等に関するものである。

10

【背景技術】

【0002】

従来の音声認識装置において、認識精度を向上するための種々の音声認識アルゴリズムを用いた音声認識装置が存在する（例えば、特許文献 1、特許文献 2、特許文献 3、特許文献 4 参照）。

【特許文献 1】特開 2003 - 44079 号公報（第 1 頁、第 1 図等）

【特許文献 2】特開 2003 - 22091 号公報（第 1 頁、第 1 図等）

【特許文献 3】特開平 10 - 254350 号公報（第 1 頁、第 1 図等）

【特許文献 4】特開平 7 - 199989 号公報（第 1 頁、第 1 図等）

【発明の開示】

20

【発明が解決しようとする課題】

【0003】

しかしながら、従来の音声認識装置においては、たとえ認識精度の高いアルゴリズムで音声認識処理を行っても、誤った認識結果を出力する場合もあり、かかる場合には、その音声認識装置に対するユーザの信頼を構築できない、という課題があった。

【課題を解決するための手段】

【0004】

入力された音声の発音を評定して、発音評定結果が低い場合に、認識結果を出力しない音声認識装置により、誤った認識結果を出力する可能性を、大幅に減らすことができる。その結果、音声認識装置に対するユーザの信頼を担保できる。

30

【0005】

本音声認識装置は、具体的には、以下の音声認識装置である。

【0006】

本第一の発明の音声認識装置は、比較される対象の音声に関するデータである音響データを格納している音響データ格納部と、音声の入力を受け付ける音声入力部と、前記音響データを用いて、前記音声入力部が受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する音声認識部と、前記音声入力部が受け付けた音声に対して、前記音響データを用いて発音評定処理を行い、発音評定結果を取得する発音評定部と、前記発音評定結果が所定の閾値に対して所定の関係にあるか否かを判断する判断部と、前記判断部が所定の関係にあると判断した場合のみ、前記音声認識部が取得した音声認識処理結果情報についての音声認識結果を出力する出力部を具備する音声認識装置である。

40

【0007】

かかる構成により、誤った認識結果を出力する可能性を、大幅に減らすことができる。

【0008】

また、本第二の発明の音声認識装置は、第一の発明に対して、前記発音評定部は、前記音声入力部が受け付けた音声に対して、前記音声認識部が取得した音声認識処理結果情報と前記音響データを用いて発音評定処理を行い、発音評定結果を取得する音声認識装置である。

【0009】

50

かかる構成により、誤った認識結果を出力する可能性を、大幅に減らすことができる。

【0010】

また、本第三の発明の音声認識装置は、第一、第二いずれかの発明に対して、前記発音評価部は、前記音声受付部が受け付けた音声を、フレームに区分するフレーム区分手段と、前記区分されたフレーム毎の音声データであるフレーム音声データを1以上得るフレーム音声データ取得手段と、前記1以上のフレーム音声データの最適状態を決定する最適状態決定手段と、前記最適状態決定手段が決定した最適状態の確率値を、発音区間毎に取得する発音区間確率値取得手段と、前記発音区間確率値取得手段が取得した1以上の発音区間毎の1以上の確率値をパラメータとして、音声の評価値を算出する評価値算出手段とを具備する音声認識装置である。

10

【0011】

かかる構成により、さらに精度の高い発音評価が行え、誤った認識結果を出力する可能性を、さらに大幅に減らすことができる。

【0012】

また、本第四の発明の音声認識装置は、第一、第二いずれかの発明に対して、前記発音評価部は、前記音声受付部が受け付けた音声を、フレームに区分するフレーム区分手段と、前記1以上のフレーム音声データの最適状態を決定する最適状態決定手段と、前記最適状態決定手段が決定した各フレームの最適状態を有する音韻全体の状態における1以上の確率値を、発音区間毎に取得する発音区間フレーム音韻確率値取得手段と、前記発音区間フレーム音韻確率値取得手段が取得した1以上の発音区間毎の1以上の確率値をパラメータとして音声の評価値を算出する評価値算出手段を具備する音声認識装置である。

20

【0013】

かかる構成により、さらに精度の高い発音評価が行え、誤った認識結果を出力する可能性を、さらに大幅に減らすことができる。

【0014】

また、本第五の発明の音声認識装置は、第三、第四いずれかの発明に対して、前記発音評価部は、無音を示すデータである無音データを格納している無音データ格納手段と、前記入力音声データおよび前記無音データに基づいて、無音の区間を検出する無音区間検出手段をさらに具備し、前記最適状態決定手段は、前記無音の区間についてのフレーム音声データを除いた前記1以上のフレーム音声データの最適状態を決定する音声認識装置である。

30

【0015】

かかる構成により、無音データを検出することにより、さらに精度の高い発音評価が行え、その結果、誤った認識結果を出力する可能性を、さらに大幅に減らすことができる。

【発明の効果】

【0016】

本発明による音声認識装置によれば、誤った認識結果を出力することを防止できる。

【発明を実施するための最良の形態】

【0017】

以下、音声認識装置等の実施形態について図面を参照して説明する。なお、実施の形態において同じ符号を付した構成要素は同様の動作を行うので、再度の説明を省略する場合がある。

40

【0018】

(実施の形態1)

本実施の形態における音声認識装置は、受け付けた音声に対して発音の良し悪しを判断する発音評価処理を行い、発音評価処理の結果である評価値が所定の条件を満たす場合のみ音声認識結果を出力する音声認識装置である。また、本音声認識装置が行う発音評価処理における評価値は、発音区間ごとに算出される。

【0019】

なお、本音声認識装置が行う各フレームの発音評価処理について、入力音声のフレーム

50

に対する最適状態の事後確率を、動的計画法を用いて算出することから、当該事後確率を DAP (Dynamic A Posteriori Probability) と呼び、DAP に基づく類似度計算法および発音評定処理を行う装置を DAPS と呼ぶ。そして、本音声認識装置は、発音区間ごとに評定値を算出するので、本音声認識装置が算出する事後確率を、DAP に対して t-DAP と呼ぶ。

【0020】

図1は、本実施の形態における音声認識装置のブロック図である。

【0021】

音声認識装置は、受付部101、音声受付部102、音響データ格納部103、認識候補データ格納部104、音声認識部105、発音評定部106、判断部107、出力部108を具備する。

10

【0022】

発音評定部106は、フレーム区分手段1061、フレーム音声データ取得手段1062、最適状態決定手段1063、発音区間確率値取得手段1064、評定値算出手段1065を具備する。

【0023】

受付部101は、音声認識装置の処理を開始する指示である開始指示などを受け付ける。開始指示などの入力手段は、テンキーやキーボードやマウスやメニュー画面によるもの等、何でも良い。受付部101は、テンキーやキーボード等の入力手段のデバイスドライバや、メニュー画面の制御ソフトウェア等で実現され得る。

20

【0024】

音声受付部102は、音声の入力を受け付ける。音声は、マイク等から入力される。入力される音声は、音声認識の対象の情報である。音声受付部102は、マイクとドライバーソフト等で実現され得る。

【0025】

音響データ格納部103は、比較される対象の音声に関するデータである音響データを格納している。音響データは、例えば、音韻毎の隠れマルコフモデルを連結したHMMに基づくデータであることが好適である。また、音響データは、入力される音声を構成する音素に対応するHMMを、入力順序に従って連結されているHMMに基づくデータであることが好適である。ただし、音響データは、必ずしも、音韻毎のHMMを連結したHMMに基づくデータである必要はない。音響データは、全音素のHMMの、単なる集合であっても良い。また、音響データは、必ずしもHMMに基づくデータである必要はない。音響データは、単一ガウス分布モデルや、確率モデル(GMM: ガウシアンミクスチャモデル)や、統計モデルなど、他のモデルに基づくデータでも良い。HMMに基づくデータは、例えば、フレーム毎に、状態識別子と遷移確率の情報を有する。また、HMMに基づくデータは、例えば、複数の音声認識対象言語を母国語として話す人が発声した2以上のデータから学習した(推定した)モデルでも良い。音響データ格納部103は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。

30

【0026】

認識候補データ格納部104は、音声認識の候補を示すデータである認識候補データを1以上格納している。認識候補データは、例えば、単語の音素トランスクリプションである。なお、ここで音素トランスクリプションとは、音素を文字列に表現した情報である。音素トランスクリプションとは、例えば、「inu」「neko」「ume」「hitsuj i」などである。認識候補データ格納部104は、単語ではなく、一の音素(例えば、「u」)を文字列に表現した情報を格納していても良い。また、認識候補データは、例えば、音声認識の候補を識別する識別子でも良い。認識候補データが音声認識の候補を識別する識別子である場合、音声認識の候補を示すデータ自身は、例えば、外部の装置や、本音声認識装置が有する(図示しない)。認識候補データ格納部104は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。

40

【0027】

50

音声認識部 105 は、音響データ、または音響データと認識候補データを用いて、音声受付部 102 が受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する。音声認識処理結果情報は、例えば、音声認識処理結果の認識候補データ（例えば、音素トランスクリプション）や、音声認識処理結果の最適状態系列である。音声認識処理結果情報のデータ構造は問わない。音声認識部 105 は、通常、受付部 101 が開始指示を受け付けた後、音声認識処理を行う。音声認識部 105 は、音響データと認識候補データを用いて、特定の単語等を認識処理しても良いし、認識候補データを用いずに、音響データを利用して不特定の単語等を認識処理しても良い。また、音声認識処理のアルゴリズムは問わない。音声認識処理自体は種々の公知技術が存在するので、説明を省略する。音声認識部 105 は、通常、MPU やメモリ等から実現され得る。音声認識部 105 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM 等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

10

【0028】

発音評定部 106 は、音声受付部 102 が受け付けた音声に対して、音響データ格納部 103 の音響データを用いて発音評定処理を行い、発音評定結果を取得する。また、通常、発音評定部 106 は、音声受付部 102 が受け付けた音声に対して、音声認識部 105 が取得した音声認識処理結果情報と音響データを用いて発音評定処理を行い、発音評定結果を取得する。つまり、通常、発音評定部 106 は、音声認識部 105 における音声認識の結果得られた音声認識処理結果情報をも用いて、発音評定処理を行う。発音評定処理の詳細については後述するが、種々の発音評定処理アルゴリズムが考えられる。発音評定部 106 は、通常、MPU やメモリ等から実現され得る。発音評定部 106 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM 等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

20

【0029】

発音評定部 106 を構成するフレーム区分手段 1061 は、音声受付部 102 が受け付けた音声をフレームに区分する。フレーム区分手段 1061 の処理は公知技術による処理である。

【0030】

フレーム音声データ取得手段 1062 は、フレーム区分手段 1061 により区分されたフレーム毎の音声データであるフレーム音声データを1以上得る。

30

【0031】

最適状態決定手段 1063 は、1以上のフレーム音声データのうちの少なくとも一のフレーム音声データに対する最適状態を決定する。最適状態決定手段 1063 は、例えば、音響データ格納部 103 中の全音韻HMMから、認識対象の単語などの音声を構成する1以上の音素に対応するHMMを取得し、当該取得した1以上のHMMを、音素の順序で連結したデータ（認識対象の音声に関するデータであり、音韻毎の隠れマルコフモデルを連結したHMMに基づくデータ）を構成する。そして、構成した当該データ、および取得した特徴ベクトル系列を構成する各特徴ベクトル o_t に基づいて、所定のフレームの最適状態（特徴ベクトル o_t に対する最適状態）を決定する。なお、最適状態を決定するアルゴリズムは、例えば、Viterbi アルゴリズムである。

40

【0032】

発音区間確率値取得手段 1064 は、最適状態決定手段 1063 が決定した最適状態の確率値を、発音区間毎に取得する。ここで、発音区間とは、音韻、音節、単語など、発音の一まとまりを構成する区間である。

【0033】

評定値算出手段 1065 は、発音区間確率値取得手段 1064 が取得した1以上の発音区間毎の1以上の確率値をパラメータとして音声の評定値を算出する。評定値算出手段 1065 は、例えば、発音区間確率値取得手段 1064 が取得した各発音区間の1以上の確率値の時間平均値を、発音区間毎に算出し、1以上の時間平均値を得て、当該1以上の時

50

間平均値をパラメータとして音声の評定値を算出する。

【 0 0 3 4 】

フレーム区分手段 1 0 6 1、フレーム音声データ取得手段 1 0 6 2、最適状態決定手段 1 0 6 3、発音区間確率値取得手段 1 0 6 4、および評定値算出手段 1 0 6 5 は、通常、MPU やメモリ等から実現され得る。フレーム区分手段 1 0 6 1 等の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアは ROM 等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【 0 0 3 5 】

判断部 1 0 7 は、発音評定結果が閾値と所定の関係にあるか否かを判断する。ここで、発音評定結果とは、通常、発音評定のレベルを示す数値である。発音評定結果は、例えば、0 から 1 までの数値であったり、1 から 1 0 の 1 0 段階の数値である。また、閾値は、通常、予め格納されている数値である。また、所定の関係とは、例えば、発音評定結果（例えば、0 から 1 までの数値）が閾値（例えば、0 . 5 0 ）以上であることを言う。また、所定の関係とは、例えば、発音評定結果（例えば、0 から 1 までの数値）が閾値（例えば、0 . 5 0 ）より大きいことを言う。判断部 1 0 7 は、通常、MPU やメモリ等から実現され得る。判断部 1 0 7 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアは ROM 等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【 0 0 3 6 】

出力部 1 0 8 は、判断部 1 0 7 が所定の関係にあると判断した場合のみ、音声認識部 1 0 5 が取得した音声認識処理結果情報についての音声認識結果を出力する。音声認識結果は、音声認識処理結果情報と同じ情報でも良いし、音声認識処理結果情報から生成される情報等でも良い。音声認識処理結果情報が音素トランスクリプション（例えば、「neko」）の場合に、音声認識結果は、例えば、文字列（例えば、「ねこ」）である。また、出力部 1 0 8 は、判断部 1 0 7 が所定の関係にないと判断した場合、通常、音声認識できなかった旨の情報を出力する。ただし、かかる場合、何ら出力しなくても良い。ここで、出力とは、ディスプレイへの表示、プリンタへの印字、音出力、外部の装置への送信、記録媒体への蓄積等を含む概念である。出力部 1 0 8 は、ディスプレイやスピーカー等の出力デバイスを含むと考えても含まないと考えても良い。出力部 1 0 8 は、出力デバイスのドライバソフトまたは、出力デバイスのドライバソフトと出力デバイス等で実現され得る。

【 0 0 3 7 】

次に、音声認識装置の動作について図 2、図 3 のフローチャートを用いて説明する。

【 0 0 3 8 】

（ステップ S 2 0 1 ）受付部 1 0 1 は、開始指示を受け付けたか否かを判断する。開始指示を受け付ければステップ S 2 0 2 に行き、開始指示を受け付けなければステップ S 2 0 1 に戻る。

【 0 0 3 9 】

（ステップ S 2 0 2 ）音声受付部 1 0 2 は、音声を受け付けたか否かを判断する。音声を受け付ければステップ S 2 0 3 に行き、音声を受け付けなければステップ S 2 0 2 に戻る。本音声は、認識の対象となる音声である。

【 0 0 4 0 】

（ステップ S 2 0 3 ）音声認識部 1 0 5 は、ステップ S 2 0 2 で受け付けた音声に対して音声認識処理を行う。具体的には、音声認識部 1 0 5 は、音響データと認識候補データを用いて、ステップ S 2 0 2 で受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する。

【 0 0 4 1 】

（ステップ S 2 0 4 ）発音評定部 1 0 6 は、音声受付部 1 0 2 が受け付けた音声に対して、音響データ格納部 1 0 3 の音響データを用いて発音評定処理を行い、発音評定結果を取得する。発音評定部 1 0 6 は、通常、ステップ S 2 0 3 で得た音声認識処理結果情報か

10

20

30

40

50

ら、音響データを用いて発音評定処理を行い、発音評定結果を取得する。発音評定処理の詳細について、図3のフローチャートを用いて説明する。

【0042】

(ステップS205)判断部107は、発音評定結果が所定の閾値に対して所定の関係(例えば、発音評定結果>閾値)にあるか否かを判断する。所定の関係にあればステップS206に行き、所定の関係になればステップS207に行く。

【0043】

(ステップS206)出力部108は、音声認識処理結果情報についての音声認識結果(例えば、音声認識結果の文字列)を出力する。そして、ステップS202に戻る。

【0044】

(ステップS207)出力部108は、音声認識できなかった旨の情報を出力する。そして、ステップS202に戻る。

【0045】

なお、図2のフローチャートにおいて、発音評定処理を最初に行って、発音評定結果が閾値と所定の関係にある場合のみ、音声認識処理を行っても良い。

【0046】

また、図2のフローチャートにおいて、音声を受け付ける単位は、単語、文節、文、2以上の文等、問わない。また、音声認識を行う単位、発音評定を行う単位も問わない。図2のフローチャートにおいて、例えば、一単語の音声を受け付け、一単語の音声を認識し、一単語の音声の発音評定を行っている例である。音声を受け付ける単位が文である場合、例えば、ステップS203の音声認識処理、およびステップS204の発音評定処理は、発音区間の数分、繰り返し実行される。

【0047】

さらに、図2のフローチャートにおいて、電源オフや処理終了の割り込みにより処理は終了する。

【0048】

次に、図2のステップS204の発音評定処理について図3のフローチャートを用いて説明する。

【0049】

(ステップS301)フレーム区分手段1061は、音声受付部102が受け付けた音声データをフレームに区分する。かかる段階で、区分されたフレーム毎の音声データであるフレーム音声データが構成されている。フレーム区分手段1061が行うフレーム分割の処理は、例えば、フレーム音声データ取得手段1062がフレーム音声データを取り出す際の前処理であり、入力された音声のデータを、すべてのフレームに一度に分割するとは限らない。

【0050】

(ステップS302)フレーム音声データ取得手段1062は、カウンタ*i*に1を代入する。

【0051】

(ステップS303)フレーム音声データ取得手段1062は、ステップS301で区分した音声フレーム中、*i*番目のフレームが存在するか否かを判断する。*i*番目のフレームが存在すればステップS304に行き、*i*番目のフレームが存在しなければステップS306に行く。

【0052】

(ステップS304)フレーム音声データ取得手段1062は、*i*番目のフレーム音声データを取得する。フレーム音声データの取得とは、例えば、当該分割された音声データを音声分析し、特徴ベクトルデータを抽出し、メモリ上に配置することである。なお、フレーム音声データは、例えば、入力された音声データをフレーム分割されたデータである。また、フレーム音声データは、例えば、当該分割された音声データから音声分析され、抽出された特徴ベクトルデータを有する。本特徴ベクトルデータは、例えば、三角型フィ

10

20

30

40

50

ルタを用いたチャネル数 24 のフィルタバンク出力を離散コサイン変換した M F C C であり、その静的パラメータ、デルタパラメータおよびデルタデルタパラメータをそれぞれ 12 次元、さらに正規化されたパワーとデルタパワーおよびデルタデルタパワー (39 次元) を有する。

【0053】

(ステップ S 305) フレーム音声データ取得手段 1062 は、カウンタ i を 1、インクリメントする。そして、ステップ S 303 に戻る。

【0054】

(ステップ S 306) 発音区間確率値取得手段 1064 は、全フレームの全状態の前向き尤度と後向き尤度を算出する。そして、全フレーム、全状態の確率値を得る。具体的には、発音区間確率値取得手段 1064 は、例えば、各特徴ベクトルが対象の状態から生成された事後確率を算出する。この事後確率は、HMM の最尤推定における Baum - Welch アルゴリズムの中で現れる占有度数に対応する。Baum - Welch アルゴリズムは、公知のアルゴリズムであるので、説明は省略する。

10

【0055】

(ステップ S 307) 発音区間確率値取得手段 1064 は、全フレームの最適状態確率値を算出する。

【0056】

(ステップ S 308) 発音区間確率値取得手段 1064 は、発音区間 (例えば、単語の区間) に対応する 1 以上の最適状態の確率値をすべて読み出す。

20

【0057】

(ステップ S 309) 評定値算出手段 1065 は、ステップ S 308 で取得した 1 以上の発音区間毎の 1 以上の確率値をパラメータとして音声の評定値を算出する。例えば、ステップ S 308 で取得した 1 以上の確率値の平均値 (時間平均値) を算出する。そして、上位関数にリターンする。評定値の算出処理の詳細については、後述する。

【0058】

以下、本実施の形態における音声認識装置の具体的な動作について説明する。

【0059】

まず、本音声認識装置において、図示しない手段により、認識対象の単語 (文でも良い) の言語のネイティブ発音の音声データベースからネイティブ発音の音韻 HMM を学習しておく。ここで、音韻の種類数を L とし、 l 番目の音韻に対する HMM を λ_l とする。なお、かかる学習の処理については、公知技術であるので、詳細な説明は省略する。なお、HMM の仕様について、図 4 に示す。なお、HMM の仕様は、他の実施の形態における具体例の説明においても同様である。ただし、HMM の仕様が、他の仕様でも良いことは言うまでもない。

30

【0060】

そして、学習した L 種類の音韻 HMM から、音声認識対象の単語や文章などの音声を構成する 1 以上の音素に対応する HMM を取得し、当該取得した 1 以上の HMM を、音素の順序で連結した音響データを構成する。そして、当該音響データを音響データ格納部 103 に保持しておく。

40

【0061】

次に、ユーザ (ここでは、例えば、アメリカ人) が、音声認識の開始指示を入力する。かかる指示は、例えば、マウスで所定のボタンを押下することによりなされる。

【0062】

次に、ユーザは、音声「right」を発音する。そして、音声受付部 102 は、ユーザが発音した音声の入力を受け付ける。

【0063】

次に、音声認識部 105 は、音声認識処理を行い、認識結果の音素トランスクリプション「right」を得る、とする。

【0064】

50

次に、発音評価部 106 は、以下のように発音評価を行う。

【0065】

つまり、フレーム区分手段 1061 は、音声受付部 102 が受け付けた音声を、短時間フレームに区分する。なお、フレームの間隔は、予め決められている、とする。

【0066】

そして、フレーム音声データ取得手段 1062 は、フレーム区分手段 1061 が区分した音声データを、スペクトル分析し、特徴ベクトル系列「 $O = o_1, o_2, \dots, o_T$ 」を算出する。なお、 T は、系列長である。ここで、特徴ベクトル系列は、各フレームの特徴ベクトルの集合である。また、特徴ベクトルは、例えば、三角型フィルタを用いたチャンネル数 24 のフィルタバンク出力を離散コサイン変換した MFCC であり、その静的パラメータ、デルタパラメータおよびデルタデルタパラメータをそれぞれ 12 次元、さらに正規化されたパワーとデルタパワーおよびデルタデルタパワー（39 次元）を有する。また、スペクトル分析において、ケプストラム平均除去を施すことは好適である。なお、音声分析条件を図 5 の表に示す。なお、音声分析条件は、他の実施の形態における具体例の説明においても同様である。ただし、音声分析条件が、他の条件でも良いことは言うまでもない。

10

【0067】

次に、最適状態決定手段 1063 は、取得した特徴ベクトル系列を構成する各特徴ベクトル o_t に基づいて、所定のフレームの最適状態（特徴ベクトル o_t に対する最適状態）を決定する。最適状態決定手段 1063 が最適状態を決定するアルゴリズムは、例えば、Viterbi アルゴリズムによる。かかる場合、最適状態決定手段 1063 は、上記で連結した HMM を用いて最適状態を決定する。最適状態決定手段 1063 は、2 以上のフレームの最適状態である最適状態系列を求めることとなる。

20

【0068】

次に、発音区間確率値取得手段 1064 は、まず、以下の数式 1 により、最適状態（ q_t^* ）における最適状態確率値（ $y_t(q_t^*)$ ）を算出する。なお、 $y_t(q_t^*)$ は、状態 j の事後確率関数 $y_t(j)$ の j に q_t^* を代入した値である。そして、状態 j の事後確率関数 $y_t(j)$ は、数式 2 を用いて算出される。この確率値（ $y_t(j)$ ）は、 t 番目の特徴ベクトル o_t が状態 j から生成された事後確率であり、動的計画法を用いて算出される。なお、 j は、状態を識別する状態識別子である。

30

【数 1】

$$DAP(t) = r_t(q_t^*)$$

【0069】

数式 1 において、 q_t は、 o_t に対する状態識別子を表す。この確率値（ $y_t(j)$ ）は、HMM の最尤推定における Baum - Welch アルゴリズムの中で表れる占有度数に対応する。

【数 2】

$$\gamma_t(j) = \frac{\Pr(q_t=j, \mathbf{O} \mid \lambda^{\text{all}})}{\Pr(\mathbf{O} \mid \lambda^{\text{all}})}$$

$$= \frac{\Pr(q_t=j, \mathbf{O} \mid \lambda^{\text{all}})}{\sum_{k=1}^N \Pr(q_t=k, \mathbf{O} \mid \lambda^{\text{all}})} \quad (1)$$

$$= \frac{\alpha_t(j) \beta_t(j)}{\sum_{k=1}^N \alpha_t(k) \beta_t(k)} \quad (2)$$

10

$$\alpha_t(j) = \Pr(q_t=j, \{\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_t\} \mid \lambda^{\text{all}})$$

$$\beta_t(j) = \Pr(\{\mathbf{o}_{t+1}, \mathbf{o}_{t+2}, \dots, \mathbf{o}_T\} \mid q_t=j, \lambda^{\text{all}})$$

【0070】

数式 2 において、「 $\alpha_t(j)$ 」「 $\beta_t(j)$ 」は、全部の HMM を用いて、forward-backward アルゴリズムにより算出される。「 $\alpha_t(j)$ 」は前向き尤度、「 $\beta_t(j)$ 」は後向き尤度である。Baum-Welch アルゴリズム、forward-backward アルゴリズムは、公知のアルゴリズムであるので、詳細な説明は省略する。

20

【0071】

また、数式 2 において、N は、全 HMM に渡る状態の総数を示す。

【0072】

そして、発音区間確率値取得手段 1064 は、発音区間に対応する 1 以上の最適状態の確率値をすべて取得する。そして、評価値算出手段 1065 は、取得した 1 以上の確率値の平均値（時間平均値）を算出する。具体的には、評価値算出手段 1065 は、数式 3 により評価値を算出する。

30

【数 3】

$$\text{t-DAP}(m) = \frac{\sum_{\tau \in T(q_t^*)} \gamma_{\tau}(q_{\tau}^*)}{|T(q_t^*)|}$$

【0073】

なお、発音評価部 106 は、まず最適状態を求め、次に、最適状態の確率値（なお、確率値は、0 以上、1 以下である。）を求めても良いし、発音評価部 106 は、まず、全状態の確率値を求め、その後、特徴ベクトル系列の各特徴ベクトルに対する最適状態を求め、当該最適状態に対応する確率値を求めても良い。

40

【0074】

なお、もしユーザの t フレーム目に対応する発音が、音響データが示す発音（例えば、正しいネイティブの発音）に近ければ、数式 2 の（2）式の分子の値が、他の全ての可能な音韻の全ての状態と比較して大きくなり、結果的に最適状態の確率値（評価値）が大きくなる。逆にその区間が、音響データが示す発音に近くなければ、評価値は小さくなる。なお、どのネイティブ発音にも近くないような場合は、DAP 評価値（DAP スコアとも言う）はほぼ $1/N$ に等しくなる。N は全ての音韻 HMM における全ての状態の数であるから、通常、大きな値となり、DAP スコアは十分小さくなる。また、ここでは、DAP

50

スコアは最適状態における確率値と全ての可能な状態における確率値との比率で定義されている。したがって、話者性や収音環境の違いにより多少のスペクトルの変動があったとしても、ユーザが正しい発音をしていれば、その変動が相殺され評価値が高いスコアを維持する。

【 0 0 7 5 】

かかる評価値算出手段 1 0 6 5 が、中間に算出する D A P スコア（フレームごとのスコア）を、図 6、図 7 に示す。図 6、図 7 において、横軸は分析フレーム番号、縦軸はスコアを % で表わしたものである。太い破線は音素境界、細い点線は状態境界（いずれも V i t e r b i アルゴリズムで求めたもの）を表わしており、図の上部に音素名を表記している。図 6 は、アメリカ人男性による英語「 r i g h t 」の発音の D A P スコアを示す。10

【 0 0 7 6 】

図 7 は、日本人男性による英語「 r i g h t 」の発音の D A P スコアを示す。アメリカ人の発音は、日本人の発音と比較して、基本的にスコアが高い。また、図 6 において、状態の境界において所々スコアが落ち込んでいることがわかる。

【 0 0 7 7 】

そして、評価値算出手段 1 0 6 5 は、例えば、発音区間における発音評価結果（ t - D A P スコアとも言う）を「 6 8 」と、算出した、とする（図 8 参照）。

【 0 0 7 8 】

そして、次に、判断部 1 0 7 は、予め格納している閾値「 5 0 」と発音評価結果「 6 8 」20
を比較し、所定の条件を満たすか否かを判断する。所定の条件は、例えば、「発音評価結果 > 閾値」である。そして、「 6 8 > 5 0 」であるので、判断部 1 0 7 は、所定の条件を満たす、と判断する。

【 0 0 7 9 】

次に、出力部 1 0 8 は、音声認識部 1 0 5 が取得した音声認識処理結果情報「 r i g h t 」を出力する。かかる出力は、例えば、ディスプレイへの表示である。

【 0 0 8 0 】

さらに、次に、他のユーザ（日本人だとする）が、音声「 r i g h t 」を発音する。そして、音声受付部 1 0 2 は、ユーザが発音した音声の入力を受け付ける。

【 0 0 8 1 】

次に、音声認識部 1 0 5 は、音声認識処理を行い、認識結果の音素トランスクリプション「 l i g h t 」を得る、とする。30

【 0 0 8 2 】

次に、発音評価部 1 0 6 は、上述のアルゴリズムで、発音評価を行う。そして、発音評価結果「 2 9 」を得た、とする。

【 0 0 8 3 】

次に、判断部 1 0 7 は、予め格納している閾値「 5 0 」と発音評価結果「 2 9 」を比較し、所定の条件を満たすか否かを判断する。そして、「 5 0 > 2 9 」であるので、判断部 1 0 7 は、所定の条件を満たさない、と判断する。

【 0 0 8 4 】

そして、出力部 1 0 8 は、予め格納している情報「音声認識できませんでした。」を出力する。かかる出力は、例えば、スピーカーからの音声出力と、ディスプレイへの表示である。40

【 0 0 8 5 】

以上、本実施の形態によれば、発音評価を行って、所定の評価結果が得られた場合のみ（良好な場合のみ）、音声認識結果を出力する音声認識装置を提供できる。かかる処理により、誤った認識結果を出力する可能性を減らすことができ、信頼度の高い音声認識装置を提供できる。また、音声認識装置に対するユーザの信頼度を損なわず、ユーザは安心して装置を利用できる。

【 0 0 8 6 】

なお、本実施の形態によれば、音声認識処理を行った後、発音評定を行った。しかし、発音評定を行った後、評定結果が所定の条件を満たす場合のみ音声認識処理を行っても良い。かかることは、他の実施の形態においても同様である。

【0087】

また、本実施の形態によれば、音声認識部が行う音声認識のアルゴリズムは問わない。音声認識において、認識候補データ格納部104における音素トランスクリプションを用いても用いなくても良い。かかることも、他の実施の形態においても同様である。

【0088】

また、本実施の形態によれば、発音評定を行う評定値算出手段は、1以上の最適状態における最適状態確率値の時間平均値を算出した。しかし、評定値算出手段は、1以上の最適状態における最適状態確率値をパラメータとする関数であり、発音の良し悪しを示す解答が得られる関数であれば、どのような関数でスコアを算出しても良い。例えば、評定値算出手段は、1以上の最適状態における最適状態確率値のうち、中央値と中央値の前後10の値を取得し、当該11の値の平均値を、スコアとしても良い。

【0089】

また、本実施の形態によれば、発音評定した結果、文の一部のスコアが閾値より低い場合、文全体の認識結果を出力しないなど、発音評定の範囲と認識結果の出力範囲が完全一致するとは限らない。

【0090】

さらに、本実施の形態における処理は、ソフトウェアで実現しても良い。そして、このソフトウェアをソフトウェアダウンロード等により配布しても良い。また、このソフトウェアをCD-ROMなどの記録媒体に記録して流布しても良い。なお、このことは、本明細書における他の実施の形態においても該当する。なお、本実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータに、音声の入力を受け付ける音声入力ステップと、格納されている音響データを用いて、前記音声入力ステップで受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する音声認識ステップと、前記音声入力ステップで受け付けた音声に対して、前記音響データを用いて発音評定処理を行い、発音評定結果を取得する発音評定ステップと、前記発音評定結果が所定の閾値と所定の関係にあるか否かを判断する判断ステップと、前記判断ステップで所定の関係にあると判断した場合のみ、前記音声認識ステップで取得した音声認識処理結果情報についての音声認識結果を出力する出力ステップを実行させるためのプログラム、である。

【0091】

また、上記プログラムにおいて、前記発音評定ステップは、前記音声入力ステップで受け付けた音声に対して、前記音声認識ステップで取得した音声認識処理結果情報と前記音響データを用いて発音評定処理を行い、発音評定結果を取得することは好適である。

【0092】

さらに、上記プログラムにおいて、前記発音評定ステップは、前記音声受付ステップで受け付けた音声を、フレームに区分するフレーム区分ステップと、前記区分されたフレーム毎の音声データであるフレーム音声データを1以上得るフレーム音声データ取得ステップと、前記1以上のフレーム音声データの最適状態を決定する最適状態決定ステップと、前記最適状態決定ステップで決定した最適状態の確率値を、発音区間毎に取得する発音区間確率値取得ステップと、前記発音区間確率値取得ステップで取得した1以上の発音区間毎の1以上の確率値をパラメータとして、音声の評定値を算出する評定値算出ステップとを具備することは好適である。

【0093】

(実施の形態2)

本実施の形態における音声認識装置は、実施の形態1の音声認識装置と比較して、発音評定部における発音評定アルゴリズムが異なる。本実施の形態において、評定値は、最適状態を含む音韻の中の全状態の確率値を発音区間で評価して、算出される。本実施の形態

10

20

30

40

50

における発音評定装置が算出する事後確率を、実施の形態 1 で述べた t - D A P に対して t - p - D A P と呼ぶ。

【 0 0 9 4 】

図 9 は、本実施の形態における音声認識装置のブロック図である。

【 0 0 9 5 】

図 9 の音声認識装置は、図 1 の音声認識装置と比較して、発音評定部 9 0 6 のみが異なる。

【 0 0 9 6 】

発音評定部 9 0 6 は、フレーム区分手段 1 0 6 1、フレーム音声データ取得手段 1 0 6 2、最適状態決定手段 1 0 6 3、発音区間フレーム音韻確率値取得手段 9 0 6 1、評定値算出手段 9 0 6 2 を具備する。

10

【 0 0 9 7 】

発音評定部 9 0 6 を構成する発音区間フレーム音韻確率値取得手段 9 0 6 1 は、最適状態決定手段 1 0 6 3 が決定した各フレームの最適状態を有する音韻全体の状態における 1 以上の確率値を、発音区間毎に取得する。

【 0 0 9 8 】

評定値算出手段 9 0 6 2 は、発音区間フレーム音韻確率値取得手段 9 0 6 1 が取得した 1 以上の発音区間毎の 1 以上の確率値をパラメータとして音声の評定値を算出する。評定値算出手段 9 0 6 2 は、例えば、最適状態決定手段 1 0 6 3 が決定した各フレームの最適状態を有する音韻全体の状態における 1 以上の確率値の総和を、フレーム毎に得て、当該フレーム毎の確率値の総和に基づいて、発音区間毎の確率値の総和の時間平均値を 1 以上得て、当該 1 以上の時間平均値をパラメータとして音声の評定値を算出する。

20

【 0 0 9 9 】

発音区間フレーム音韻確率値取得手段 9 0 6 1、および評定値算出手段 9 0 6 2 は、通常、M P U やメモリ等から実現され得る。発音区間フレーム音韻確率値取得手段 9 0 6 1 等の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアは R O M 等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【 0 1 0 0 】

次に、本音声認識装置の動作について説明する。本音声認識装置は、実施の形態 1 の音声認識装置と比較して、ステップ S 2 0 4 の発音評定処理のみが異なる。本音声認識装置の発音評定処理について、図 1 0 のフローチャートを用いて説明する。図 1 0 のフローチャートにおいて、図 3 と異なるステップについてのみ説明する。

30

【 0 1 0 1 】

（ステップ S 1 0 0 1）発音区間フレーム音韻確率値取得手段 9 0 6 1 は、j に 1 を代入する。

【 0 1 0 2 】

（ステップ S 1 0 0 2）発音区間フレーム音韻確率値取得手段 9 0 6 1 は、j 番目のフレームが、本発音区間（例えば、単語）に存在するか否かを判断する。j 番目のフレームが存在すればステップ S 1 0 0 3 に行き、j 番目のフレームが存在しなければステップ S 1 0 0 6 に飛ぶ。

40

【 0 1 0 3 】

（ステップ S 1 0 0 3）発音区間フレーム音韻確率値取得手段 9 0 6 1 は、j 番目のフレームの最適状態を含む音韻の全ての確率値を取得する。

【 0 1 0 4 】

（ステップ S 1 0 0 4）評定値算出手段 9 0 6 2 は、ステップ S 1 0 0 3 で取得した 1 以上の確率値をパラメータとして、1 フレームの音声の評定値を算出し、メモリに一時的格納する。

【 0 1 0 5 】

（ステップ S 1 0 0 5）発音区間フレーム音韻確率値取得手段 9 0 6 1 は、j を 1、インクリメントする。そして、ステップ S 1 0 0 2 に戻る。

50

【 0 1 0 6 】

(ステップ S 1 0 0 6) 評定値算出手段 9 0 6 2 は、本発音区間の評定値を算出する。評定値算出手段 9 0 6 2 は、例えば、最適状態決定手段 1 0 6 3 が決定した各フレームの最適状態を有する音韻全体の状態における 1 以上の確率値の総和を、フレーム毎に得て、当該フレーム毎の確率値の総和に基づいて、発音区間の確率値の総和の時間平均値を、当該発音区間の音声の評定値として算出する。そして、上位関数にリターンする。

【 0 1 0 7 】

以下、本実施の形態における音声認識装置の具体的な動作について説明する。本実施の形態において、評定値の算出アルゴリズムが実施の形態 1 とは異なるので、その動作を中心に説明する。

10

【 0 1 0 8 】

まず、ユーザが、開始指示を入力した後、音声認識対象の音声（ここでは、「right」）を発音する。そして、音声受付部 1 0 2 は、ユーザが発音した音声の入力を受け付ける。

【 0 1 0 9 】

次に、フレーム区分手段 1 0 6 1 は、音声受付部 1 0 2 が受け付けた音声を、短時間フレームに区分する。

【 0 1 1 0 】

そして、フレーム音声データ取得手段 1 0 6 2 は、フレーム区分手段 1 0 6 1 が区分した音声データを、スペクトル分析し、特徴ベクトル系列「 $O = o_1, o_2, \dots, o_T$ 」を算出する。

20

【 0 1 1 1 】

次に、発音区間フレーム音韻確率値取得手段 9 0 6 1 は、各フレームの各状態の事後確率（確率値）を算出する。確率値の算出は、上述した数式 1、数式 2 により算出できる。

【 0 1 1 2 】

次に、最適状態決定手段 1 0 6 3 は、取得した特徴ベクトル系列を構成する各特徴ベクトル o_t に基づいて、各フレームの最適状態（特徴ベクトル o_t に対する最適状態）を決定する。つまり、最適状態決定手段 1 0 6 3 は、最適状態系列を得る。なお、各フレームの各状態の事後確率（確率値）を算出と、最適状態の決定の順序は問わない。

【 0 1 1 3 】

次に、発音区間フレーム音韻確率値取得手段 9 0 6 1 は、発音区間ごとに、当該発音区間に含まれる各フレームの最適状態を含む音韻の全ての確率値を取得する。

30

【 0 1 1 4 】

そして、評定値算出手段 9 0 6 2 は、各フレームの最適状態を含む音韻の全ての確率値の総和を、フレーム毎に算出する。まず、評定値算出手段 9 0 6 2 は、フレームの最適状態を含む音韻の全ての確率値の総和（これを「p-DAP」と言う。）を、以下の数式 4 により算出する。

【数 4】

$$p\text{-DAP}(t) \triangleq \sum_{j \in P(q_t^*)} \gamma_t(j)$$

40

【 0 1 1 5 】

そして、評定値算出手段 9 0 6 2 は、フレーム毎に算出された確率値の総和を、発音区間毎に時間平均し、発音区間毎の評定値を算出する。具体的には、評定値算出手段 9 0 6 2 は、数式 5 により評定値を算出する。

【数 5】

$$t\text{-}p\text{-}DAP(m) \triangleq \frac{\sum_{\tau \in T(q_t^*)} p\text{-}DAP(\tau)}{|T(q_t^*)|}$$

【0116】

かかる評定値算出手段 9062 が算出した評定値（「t - p - DAP スコア」とも言う。）を、図 11 の表に示す。図 11 において、アメリカ人男性と日本人男性の評定結果を示す。Phoneme および Word は、t - p - DAP における時間平均の範囲を示す。ここでは、実施の形態 1 における DAP の代わりに、p - DAP の時間平均を採用したものである。図 11 において、アメリカ人男性の発音の評定値が日本人男性の発音の評定値より高く、良好な評定結果が得られている。

10

【0117】

次に、判断部 107 は、予め格納している閾値「60」と、アメリカ人男性の発声の場合の発音評定結果「84」を比較し、所定の条件を満たすか否かを判断する。そして、「84 > 60」であるので、判断部 107 は、所定の条件を満たす、と判断する。

【0118】

次に、出力部 108 は、音声認識部 105 が取得した音声認識処理結果情報「right」を出力する。

20

【0119】

また、判断部 107 は、予め格納している閾値「60」と、日本人男性の発声の場合の発音評定結果「33」を比較し、所定の条件を満たさない、と判断する。

【0120】

そして、出力部 108 は、予め格納している情報「音声認識できませんでした。」を出力する。

【0121】

以上、本実施の形態によれば、発音評定を行って、所定の評定結果が得られた場合のみ（良好な場合のみ）、音声認識結果を出力する音声認識装置を提供できる。かかる処理により、誤った認識結果を出力する可能性を減らすことができ、信頼度の高い音声認識装置を提供できる。また、音声認識装置に対するユーザの信頼度を損なわず、ユーザは安心して装置を利用できる。さらに、本実施の形態によれば、精度高く発音評定ができ、その結果、精度高く、正しい認識結果を出力することができる。

30

【0122】

なお、本実施の形態によれば、発音評定を行う評定値算出手段は、各フレームの最適状態を有する音韻全体の状態における 1 以上の確率値の総和の時間平均値を算出した。しかし、評定値算出手段は、各フレームの最適状態を有する音韻全体の状態における 1 以上の確率値の総和をパラメータとする関数であり、発音の良し悪しを示す解答が得られる関数であれば、どのような関数でスコアを算出しても良い。例えば、評定値算出手段は、音韻全体の状態における 1 以上の確率値の総和のうち、中央値と中央値の前後 30 の値を取得し、当該 31 の値の平均値を、スコアとしても良い。

40

【0123】

さらに、本実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータに、音声の入力を受け付ける音声入力ステップと、格納されている音響データを用いて、前記音声入力ステップで受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する音声認識ステップと、前記音声入力ステップで受け付けた音声に対して、前記音響データを用いて発音評定処理を行い、発音評定結果を取得する発音評定ステップと、前記発音評定結果が所定の閾値と所定の関係にあるか否かを判断する判断ステップと、前記判断ステップで所定の関係にあると判断した場合のみ、前記音声認識ステップ

50

で取得した音声認識処理結果情報についての音声認識結果を出力する出力ステップを実行させるためのプログラム、である。

【 0 1 2 4 】

また、上記プログラムにおいて、前記発音評価ステップは、前記音声入力ステップで受け付けた音声に対して、前記音声認識ステップで取得した音声認識処理結果情報と前記音響データを用いて発音評価処理を行い、発音評価結果を取得することは好適である。

【 0 1 2 5 】

さらに、上記プログラムにおいて、前記発音評価ステップは、前記音声受付部が受け付けた音声を、フレームに区分するフレーム区分ステップと、前記1以上のフレーム音声データの最適状態を決定する最適状態決定ステップと、前記最適状態決定ステップで決定した各フレームの最適状態を有する音韻全体の状態における1以上の確率値を、発音区間毎に取得する発音区間フレーム音韻確率値取得ステップと、前記発音区間フレーム音韻確率値取得ステップで取得した1以上の発音区間毎の1以上の確率値をパラメータとして音声の評価値を算出する評価値算出ステップとを具備することは好適である。

【 0 1 2 6 】

(実施の形態 3)

本実施の形態における音声認識装置は、実施の形態 1、2 の音声認識装置と比較して、発音評価部における発音評価アルゴリズムが異なる。本実施の形態において、本音声認識装置は、無音区間を検知し、無音区間を考慮した発音評価が可能な装置である。さらに具体的には、通常、本音声認識装置は、無音区間のフレームは考慮せずに、発音評価を行う。

【 0 1 2 7 】

図 1 2 は、本実施の形態における音声認識装置のブロック図である。

【 0 1 2 8 】

図 1 2 の音声認識装置は、図 1 の音声認識装置と比較して、発音評価部 1 2 0 6 のみが異なる。

【 0 1 2 9 】

発音評価部 1 2 0 6 は、フレーム区分手段 1 0 6 1、フレーム音声データ取得手段 1 0 6 2、最適状態決定手段 1 0 6 3、発音区間確率値取得手段 1 0 6 4、無音データ格納手段 1 2 0 6 1、無音区間検出手段 1 2 0 6 2、評価値算出手段 1 2 0 6 3 を具備する。

【 0 1 3 0 】

発音評価部 1 2 0 6 は、音声受付部 1 0 2 が受け付けた音声に対して、音響データ格納部 1 0 3 の音響データを用いて発音評価処理を行い、発音評価結果を取得する。また、通常、発音評価部 1 0 6 は、音声受付部 1 0 2 が受け付けた音声に対して、音声認識部 1 0 5 が取得した音声認識処理結果情報と音響データを用いて発音評価処理を行い、発音評価結果を取得する。

【 0 1 3 1 】

発音評価部 1 2 0 6 は、通常、MPU やメモリ等から実現され得る。発音評価部 1 2 0 6 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア(専用回路)で実現しても良い。

【 0 1 3 2 】

発音評価部 1 2 0 6 を構成する無音データ格納手段 1 2 0 6 1 は、無音を示すである無音データを格納している。無音データは、例えば、音韻毎の隠れマルコフモデルを連結したHMMに基づくデータであることが好適である。また、無音データは、必ずしもHMMに基づくデータである必要はない。無音データは、単一ガウス分布モデルや、確率モデル(GMM: ガウシアンミクスチャモデル)や、統計モデルなど、他のモデルに基づくデータでも良い。無音データ格納手段 1 2 0 6 1 は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。

【 0 1 3 3 】

無音区間検出手段 1 2 0 6 2 は、入力音声データ、および無音データ格納手段 1 2 0 6

10

20

30

40

50

1の無音データに基づいて、無音の区間を検出する。さらに具体的には、無音区間検出手段12062は、フレーム音声データ取得手段1062が取得したフレーム音声データ、および無音データ格納手段12061の無音データに基づいて、無音の区間を検出する。無音区間検出手段12062は、フレーム音声データ取得手段1062が取得したフレーム音声データと無音データの類似度が所定の値以上である場合に、当該フレーム音声データは無音区間のデータであると判断しても良い。また、無音区間検出手段12062は、下記で述べる発音区間確率値取得手段1064が取得する各フレームの確率値が所定の値以下であり、かつ、フレーム音声データ取得手段1062が取得したフレーム音声データと無音データの類似度が所定の値以上である場合に、当該フレーム音声データは無音区間のデータであると判断しても良い。

10

【0134】

評定値算出手段12063は、無音区間検出手段12062が検出した無音区間を除いて、かつ発音区間確率値取得手段1064が取得した1以上の発音区間毎の1以上の確率値をパラメータとして音声の評定値を算出する。なお、評定値算出手段12063は、上記確率値を如何に利用して、評定値を算出するかは問わない。ここでは、例えば、評定値算出手段12063は、t-DAPにより評定値を算出する、とする。また、評定値算出手段12063は、かならずしも無音区間を除いて、評定値を算出する必要はない。評定値算出手段12063は、無音区間の影響を少なくするように評定値を算出しても良い。

【0135】

無音区間検出手段12062、および評定値算出手段12063は、通常、MPUやメモリ等から実現され得る。無音区間検出手段12062等の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア(専用回路)で実現しても良い。

20

【0136】

次に、本音声認識装置の動作について説明する。本音声認識装置は、実施の形態1の音声認識装置と比較して、ステップS204の発音評定処理のみが異なる。本音声認識装置の発音評定処理について、図13のフローチャートを用いて説明する。図13のフローチャートにおいて、図3と異なるステップについてのみ説明する。

【0137】

(ステップS1301)発音区間確率値取得手段1064は、カウンタ*i*に1を代入する。

30

【0138】

(ステップS1302)発音区間確率値取得手段1064は、*i*番目のフレーム音声データが存在するか否かを判断する。*i*番目のフレーム音声データが存在すればステップS1303に行き、*i*番目のフレーム音声データが存在しなければステップS1309に行く。なお、*i*番目のフレーム音声データは、ステップS304で取得されたフレーム音声データの中の*i*番目のデータである。

【0139】

(ステップS1303)無音区間検出手段12062は、*i*番目のフレーム音声データの最適状態の確率値を取得する。

40

【0140】

(ステップS1304)無音区間検出手段12062は、ステップS1303で取得された確率値が、閾値より低い(または閾値以下)であるか否かを判断する。閾値より低い(または閾値以下)であればステップS1305に行き、閾値より低い(または閾値以下)でなければステップS1307に行く。

【0141】

(ステップS1305)無音区間検出手段12062は、無音データ格納手段12061の無音データ、音響データ格納部103の全音響データを取得する。

【0142】

(ステップS1306)無音区間検出手段12062は、*i*番目のフレーム音声データ

50

が無音データの確率値が最も高いか否かを判断する。無音データの確率値が最も高ければステップS 1 3 0 8に行き、無音データの確率値が最も高くなければステップS 1 3 0 7に行く。

【0 1 4 3】

(ステップS 1 3 0 7) 発音区間確率値取得手段1 0 6 4は、ステップS 1 3 0 3で取得した確率値を一時格納する。

【0 1 4 4】

(ステップS 1 3 0 8) 発音区間確率値取得手段1 0 6 4は、カウンタ*i*を1、インクリメントし、ステップS 1 3 0 2に行く。

【0 1 4 5】

(ステップS 1 3 0 9) 評定値算出手段1 2 0 6 3は、ステップS 1 3 0 7で一時格納された1以上の確率値をパラメータとして音声の評定値を算出し、上位関数にリターンする。

【0 1 4 6】

なお、図13のフローチャートにおいて、評定値算出手段1 2 0 6 3は、上記のように、無音区間の評定値を無視して、スコアを算出することが好適であるが、無音区間の評定値の影響を、例えば、1 / 1 0にして、スコアを算出しても良い。

【0 1 4 7】

以下、本実施の形態における音声認識装置の具体的な動作について説明する。本実施の形態において、無音区間を考慮して評定値を算出するので、評定値の算出アルゴリズムが実施の形態1等とは異なる。そこで、その異なる処理を中心に説明する。

【0 1 4 8】

まず、ユーザが、開始指示を入力する。次に、ユーザは、例えば、認識対象の音声を発音する。そして、音声受付部1 0 2は、ユーザが発音した音声の入力を受け付ける。

【0 1 4 9】

次に、フレーム区分手段1 0 6 1は、音声受付部1 0 2が受け付けた音声を、短時間フレームに区分する。

【0 1 5 0】

そして、フレーム音声データ取得手段1 0 6 2は、フレーム区分手段1 0 6 1が区分した音声データを、スペクトル分析し、特徴ベクトル系列「 $O = o_1, o_2, \dots, o_T$ 」を算出する。

【0 1 5 1】

次に、最適状態決定手段1 0 6 3は、取得した特徴ベクトル系列を構成する各特徴ベクトル o_t に基づいて、所定のフレームの最適状態(特徴ベクトル o_t に対する最適状態)を決定する。

【0 1 5 2】

次に、発音区間確率値取得手段1 0 6 4は、全フレームの全状態の前向き尤度と後向き尤度を算出する。そして、次に、発音区間確率値取得手段1 0 6 4は、数式1により、最適状態における最適状態確率値($y_t(q_t^*)$)を、全フレームについて、フレーム毎に算出する。

【0 1 5 3】

さらに、無音区間検出手段1 2 0 6 2は、フレーム音声データごとに、最適状態の確率値を取得し、取得した確率値が所定の値より低いかな否かを判断する。

【0 1 5 4】

そして、無音区間検出手段1 2 0 6 2は、取得した確率値が所定の値より低いと判断した場合、当該フレーム音声データと無音データとの類似度を示す事後確率の値(DAPスコア)が、当該フレーム音声データと他のデータとの類似度を示す事後確率の値のいずれよりも高いかな否かを判断する。そして、フレーム音声データと無音データとの類似度を示す事後確率の値が最も高い場合、当該フレームは無音区間のフレームであるとして、発音評定部1 2 0 6は無視する。そして、評定値算出手段1 2 0 6 3は、無音データの区間を

10

20

30

40

50

除いて、 t -DAPスコアを算出する。ここで、評定値算出手段12063は、「75」と t -DAPスコアを算出した、とする。

【0155】

次に、判断部107は、予め格納している閾値「60」と、発音評定結果「75」を比較し、所定の条件を満たすか否かを判断する。そして、「 $75 > 60$ 」であるので、判断部107は、所定の条件を満たす、と判断する。

【0156】

次に、出力部108は、音声認識部105が取得した音声認識処理結果情報（例えば、「right」）を出力する。

【0157】

かかる出力処理は、実施の形態1における処理と同様である。

【0158】

以上、本実施の形態によれば、発音評定を行って、所定の評定結果が得られた場合のみ（良好な場合のみ）、音声認識結果を出力する音声認識装置を提供できる。かかる処理により、誤った認識結果を出力する可能性を減らすことができ、信頼度の高い音声認識装置を提供できる。また、音声認識装置に対するユーザの信頼度を損なわず、ユーザは安心して装置を利用できる。さらに、本実施の形態によれば、無音区間を除いて発音評定した結果に基づいて、音声認識結果を出力しなり、出力しなかったりするので、さらに精度高く正しい認識結果を出力することができる。

【0159】

なお、本実施の形態によれば、評定値算出手段における発音評定のアルゴリズムは問わない。発音評定のアルゴリズムは、 t -DAP、 t -p-DAP、その他のアルゴリズムでも良い。

【0160】

さらに、本実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータに、音声の入力を受け付ける音声入力ステップと、格納されている音響データを用いて、前記音声入力ステップで受け付けた音声に対して音声認識処理を行い、当該音声認識処理の結果である音声認識処理結果情報を取得する音声認識ステップと、前記音声入力ステップで受け付けた音声に対して、前記音響データを用いて発音評定処理を行い、発音評定結果を取得する発音評定ステップと、前記発音評定結果が所定の閾値と所定の関係にあるか否かを判断する判断ステップと、前記判断ステップで所定の関係にあると判断した場合のみ、前記音声認識ステップで取得した音声認識処理結果情報についての音声認識結果を出力する出力ステップを実行させるためのプログラム、である。

【0161】

また、上記プログラムの発音評定ステップにおいて、前記入力音声データおよび格納している無音データに基づいて、無音の区間を検出する無音区間検出ステップをさらに具備し、前記最適状態決定ステップは、前記無音の区間についてのフレーム音声データを除いた前記1以上のフレーム音声データの最適状態を決定することは好適である。

【0162】

また、上記各実施の形態において、各処理（各機能）は、単一の装置（システム）によって集中処理されることによって実現されてもよく、あるいは、複数の装置によって分散処理されることによって実現されてもよい。

【0163】

また、図14は、本明細書で述べたプログラムを実行して、上述した種々の実施の形態の音声認識装置を実現するコンピュータの外観を示す。上述の実施の形態は、コンピュータハードウェア及びその上で実行されるコンピュータプログラムで実現され得る。図14は、このコンピュータシステム340の概観図であり、図15は、コンピュータシステム340のブロック図である。

【0164】

図14において、コンピュータシステム340は、FD(Flexible Disk)ドライブ、CD-ROM(Compact Disk Read Only Memory)ドライブを含むコンピュータ341と、キーボード342と、マウス343と、モニタ344と、マイク345とを含む。

【0165】

図15において、コンピュータ341は、FDドライブ3411、CD-ROMドライブ3412に加えて、CPU(Central Processing Unit)3413と、CPU3413、CD-ROMドライブ3412及びFDドライブ3411に接続されたバス3414と、ブートアッププログラム等のプログラムを記憶するためのROM(Read-Only Memory)3415と、CPU3413に接続され、アプリケーションプログラムの命令を一時的に記憶するとともに一時記憶空間を提供するためのRAM(Random Access Memory)3416と、アプリケーションプログラム、システムプログラム、及びデータを記憶するためのハードディスク3417とを含む。ここでは、図示しないが、コンピュータ341は、さらに、LANへの接続を提供するネットワークカードを含んでも良い。

10

【0166】

コンピュータシステム340に、上述した実施の形態の音声処理装置の機能を実行させるプログラムは、CD-ROM3501、またはFD3502に記憶されて、CD-ROMドライブ3412またはFDドライブ3411に挿入され、さらにハードディスク3417に転送されても良い。これに代えて、プログラムは、図示しないネットワークを介してコンピュータ341に送信され、ハードディスク3417に記憶されても良い。プログラムは実行の際にRAM3416にロードされる。プログラムは、CD-ROM3501、FD3502またはネットワークから直接、ロードされても良い。

20

【0167】

プログラムは、コンピュータ341に、上述した実施の形態の音声処理装置の機能を実行させるオペレーティングシステム(OS)、またはサードパーティープログラム等は、必ずしも含まなくても良い。プログラムは、制御された態様で適切な機能(モジュール)を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいれば良い。コンピュータシステム340がどのように動作するかは周知であり、詳細な説明は省略する。

【0168】

また、上記プログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

30

【0169】

本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

【産業上の利用可能性】

【0170】

以上のように、本発明にかかる音声認識装置は、誤った認識結果を出力することを防止できるという効果を有し、高性能な音声認識装置等として有用である。

【図面の簡単な説明】

40

【0171】

【図1】実施の形態1における音声認識装置のブロック図

【図2】同音声認識装置の動作について説明するフローチャート

【図3】同発音評定処理について説明するフローチャート

【図4】同HMMの仕様を示す図

【図5】同音声分析条件を示す図

【図6】同DAPスコアを示す図

【図7】同DAPスコアを示す図

【図8】同t-DAPスコアの出力例を示す図

【図9】実施の形態2における音声認識装置のブロック図

50

【図 10】同発音評定処理の動作について説明するフローチャート

【図 11】同 t - p - DAP スコアを示す図

【図 12】実施の形態 3 における音声認識装置のブロック図

【図 13】同発音評定処理の動作について説明するフローチャート

【図 14】同音声認識装置を実現するコンピュータの外観図

【図 15】同音声認識装置を実現するコンピュータシステムのブロック図

【符号の説明】

【0172】

101 受付部

102 音声受付部

103 音響データ格納部

104 認識候補データ格納部

105 音声認識部

106、906、1206 発音評定部

107 判断部

108 出力部

1061 フレーム区分手段

1062 フレーム音声データ取得手段

1063 最適状態決定手段

1064 発音区間確率値取得手段

1065、9062、12063 評定値算出手段

9061 発音区間フレーム音韻確率値取得手段

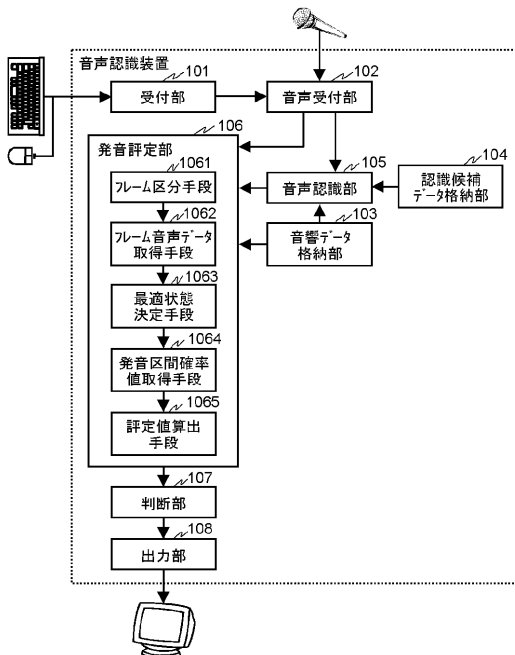
12061 無音データ格納手段

12062 無音区間検出手段

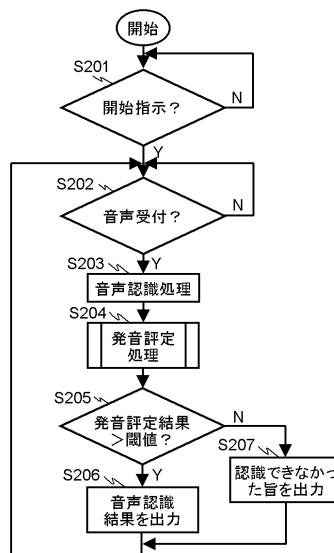
10

20

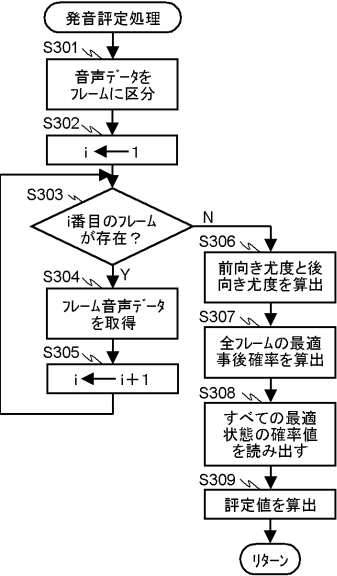
【図 1】



【図 2】



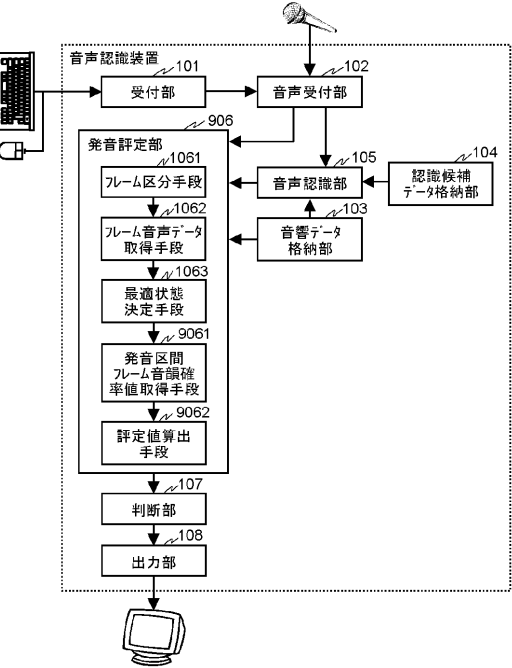
【図 3】



【図 4】

HMMタイプ	混合ガウス型, mono_phone (対角共分散行列)
状態数(モデル毎)	3
混合数(状態毎)	8
HMM総数	44
総状態数	132

【図 9】



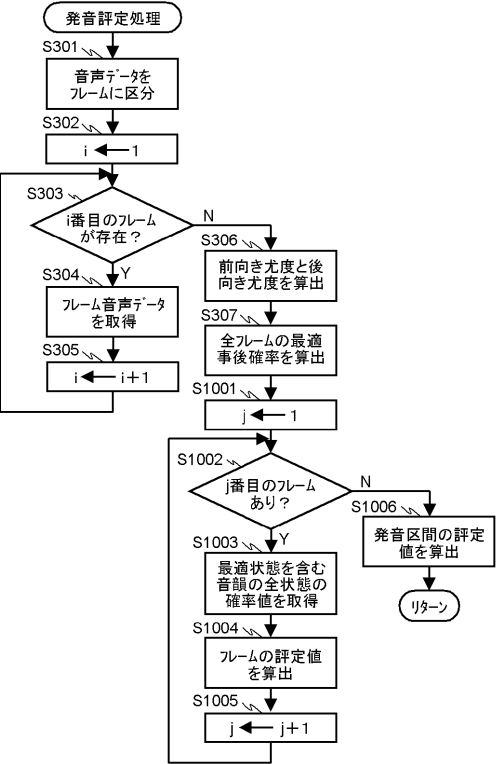
【図 5】

サンプリング周波数	22.05KHz
プリエンファシス	1-0.97 _z ⁻¹
窓関数	Hamming窓
分析フレーム長	25 msec
フレーム周期	10 msec
特徴パラメータ	MFCC(39次元)

【図 8】

	Phoneme			Word
	/r/	/ay/	/t/	/right/
アメリカ人男性	23	87	78	68
日本人男性	0	60	19	29

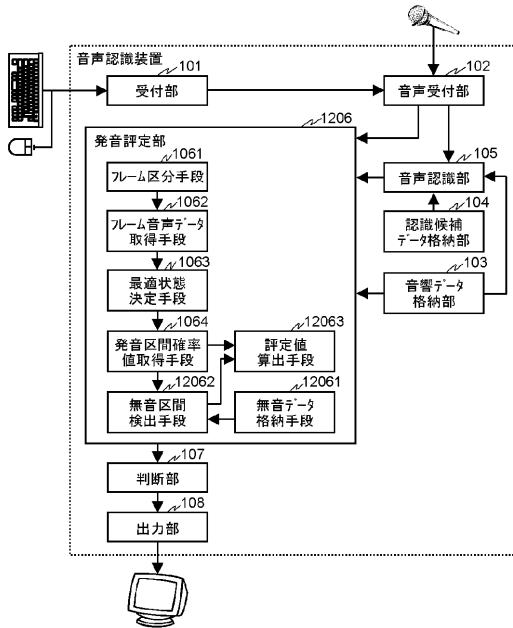
【図 10】



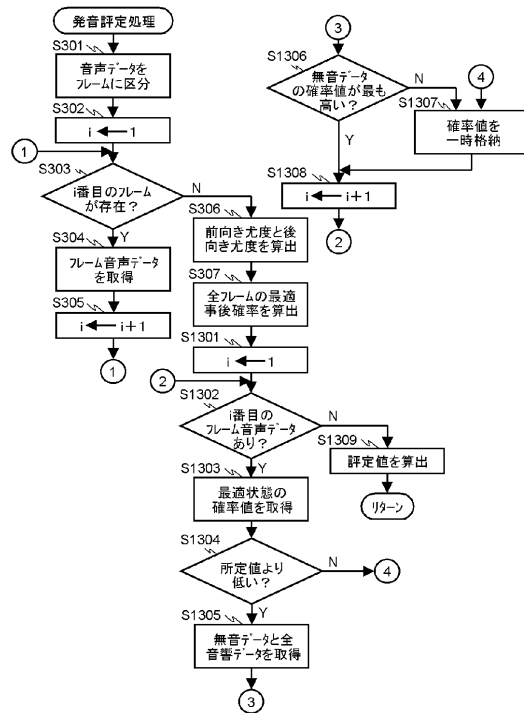
【 図 1 1 】

	Phoneme			Word
	/r/	/ay/	/t/	/right/
アメリカ人男性	39	99	100	84
日本人男性	0	70	22	33

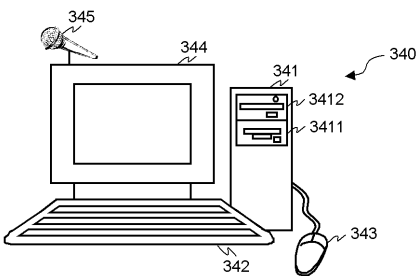
【 図 1 2 】



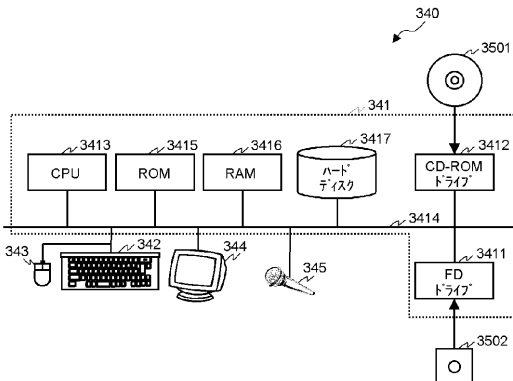
【 図 1 3 】



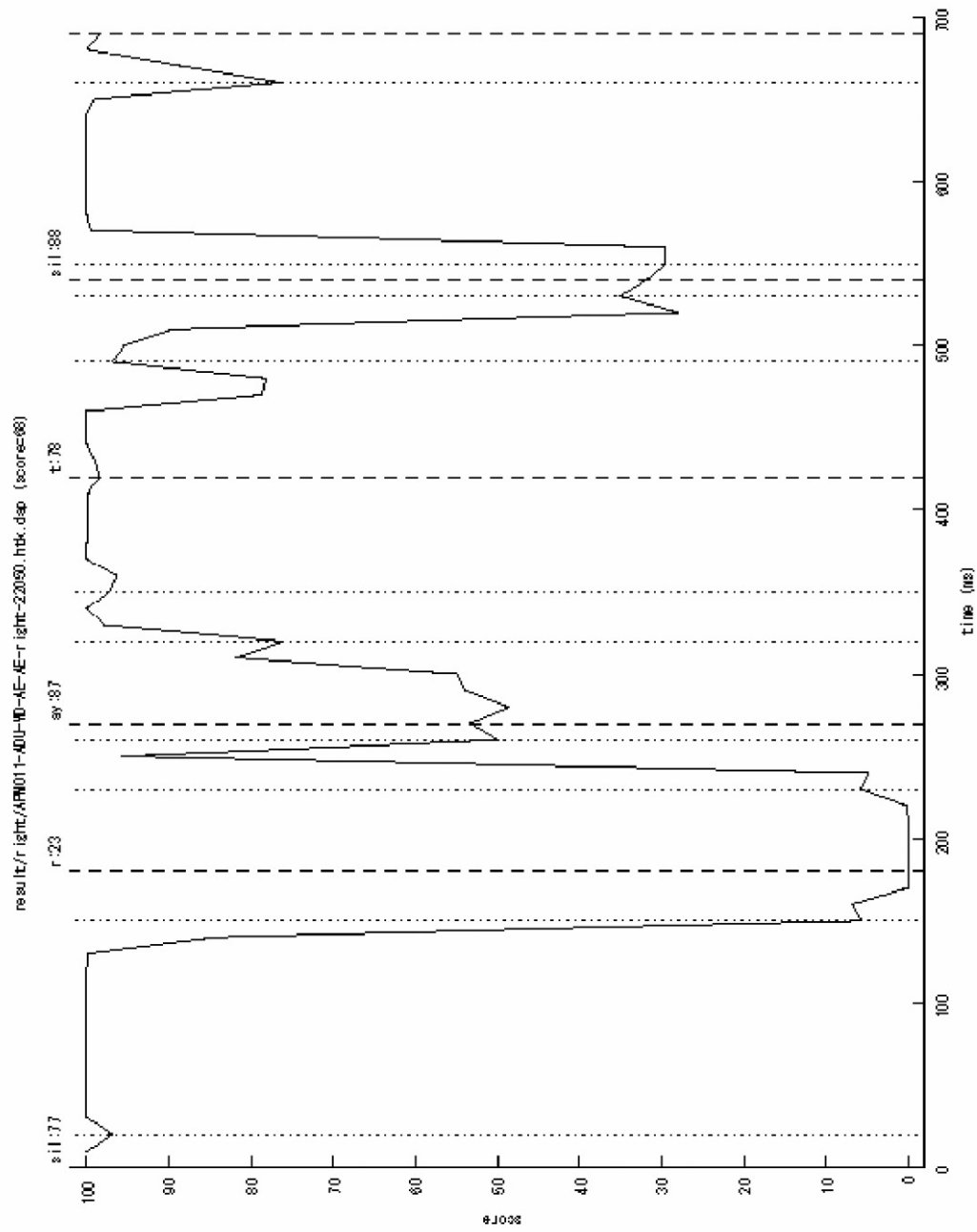
【 図 1 4 】



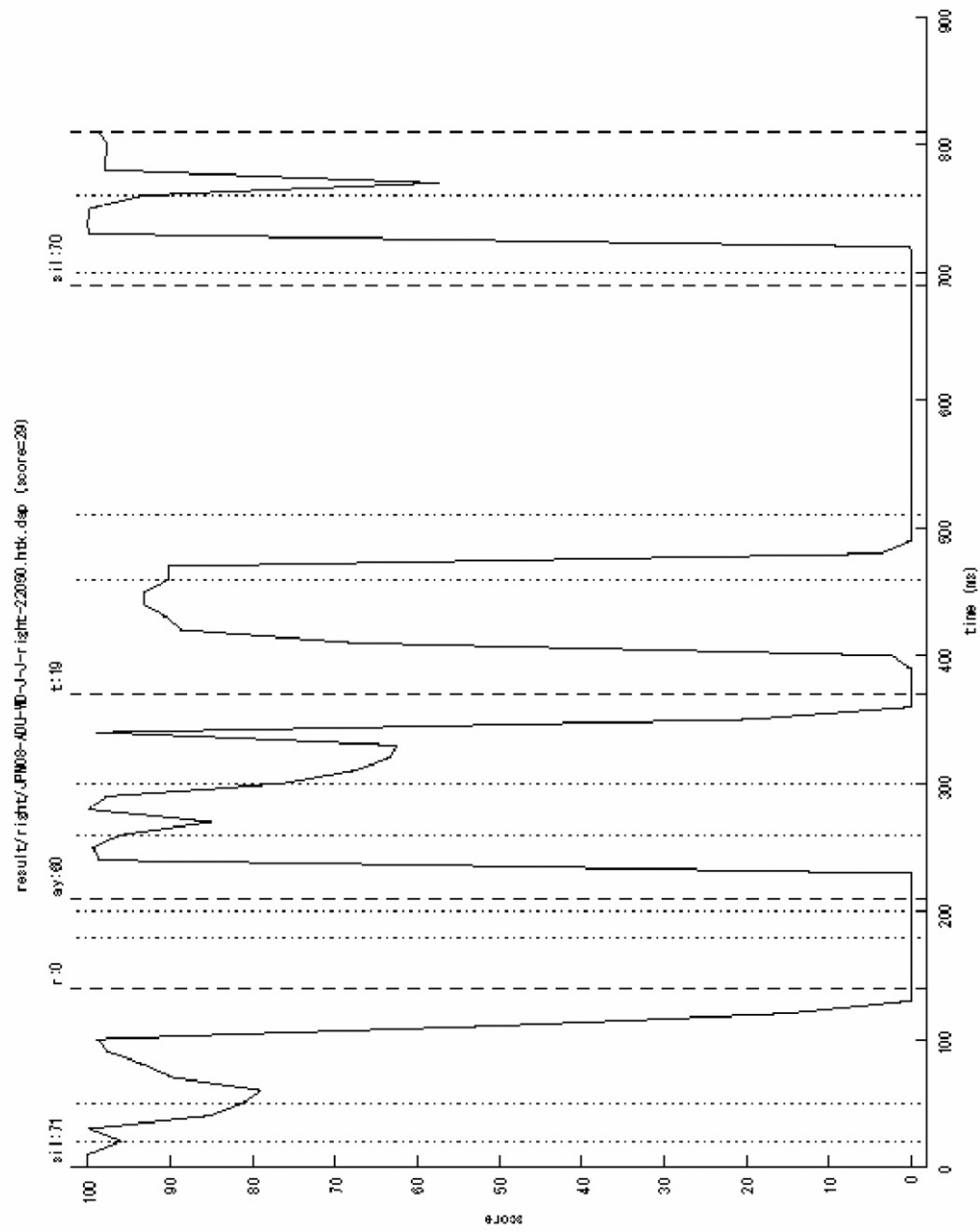
【 図 1 5 】



【図6】



【図 7】



フロントページの続き

(72)発明者 平原 達也

京都府相楽郡精華町光台二丁目2番地2 株式会社国際電気通信基礎技術研究所内

審査官 毛利 太郎

(56)参考文献 特開平08-297500(JP,A)

特開平11-085189(JP,A)

特開平08-016186(JP,A)

特開2005-031207(JP,A)

特開平11-052982(JP,A)

特開2000-099081(JP,A)

特開平09-097095(JP,A)

山田 玲子, 第二言語音の習得過程 - 知覚と生成の関係を中心に -, 電子情報通信学会技術研究報告, 日本, 社団法人電子情報通信学会, 1999年10月15日, Vol.99 No.353, p.37-42

(58)調査した分野(Int.Cl., DB名)

G10L 15/00 - 17/00