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

特許第4811993号
(P4811993)

(45)発行日　平成23年11月9日(2011.11.9)　(24)登録日　平成23年9月2日(2011.9.2)

(51)Int.Cl.		F I		
G10L 11/00	(2006.01)	G10L	11/00	402A
G10L 21/04	(2006.01)	G10L	11/00	101D
G10L 15/20	(2006.01)	G10L	21/04	120B
		G10L	15/20	360Z

請求項の数 10　(全 54 頁)

(21)出願番号　特願2005-241264 (P2005-241264)
(22)出願日　平成17年8月23日 (2005.8.23)
(65)公開番号　特開2007-57692 (P2007-57692A)
(43)公開日　平成19年3月8日 (2007.3.8)
審査請求日　平成20年3月26日 (2008.3.26)

（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

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100115749
弁理士　谷川　英和
(72)発明者　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
(72)発明者　渡辺　秀行
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
(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
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】音声処理装置、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り、1以上の音韻毎のデータであり、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教師データを１以上格納している教師データ格納部と、
音声を受け付ける音声受付部と、
第一サンプリング周波数を格納している第一サンプリング周波数格納部と、
前記教師データのフォルマント周波数である教師データフォルマント周波数を格納している教師データフォルマント周波数格納部と、
前記音声受付部が受け付けた音声の話者である評価対象者のフォルマント周波数である評価対象者フォルマント周波数を格納している評価対象者フォルマント周波数格納部と、
第二サンプリング周波数「前記第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付部が受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理部と、
前記第二音声データを処理する音声処理部を具備し、
前記音声処理部は、
前記第二音声データを、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、
前記教師データと前記1以上のフレーム音声データに基づいて、前記音声受付部が受け付

けた音声の評定を行う評定手段と、
前記評定手段における評定結果を出力する出力手段を具備し、
前記評定手段は、
前記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と、
前記最適状態決定手段が決定した最適状態における確率値と、当該確率値に対応するフレームの全状態における確率値の総和との比率を用いて、事後確率を示す確率値である最適状態確率値を取得する最適状態確率値取得手段と、
前記最適状態確率値取得手段が取得した最適状態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を具備する音声処理装置。

【請求項２】
比較される対象の音声に関するデータであり、1以上の音韻毎のデータであり、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教師データを１以上格納している教師データ格納部と、
音声を受け付ける音声受付部と、
第一サンプリング周波数を格納している第一サンプリング周波数格納部と、
前記教師データのフォルマント周波数である教師データフォルマント周波数を格納している教師データフォルマント周波数格納部と、
前記音声受付部が受け付けた音声の話者である評価対象者のフォルマント周波数である評価対象者フォルマント周波数を格納している評価対象者フォルマント周波数格納部と、
第二サンプリング周波数「前記第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付部が受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理部と、
前記第二音声データを処理する音声処理部を具備し、
前記音声処理部は、
前記第二音声データを、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、
前記教師データと前記1以上のフレーム音声データに基づいて、前記音声受付部が受け付けた音声の評定を行う評定手段と、
前記評定手段における評定結果を出力する出力手段を具備し、
前記評定手段は、
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２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２以上の確率値の総和を、フレーム毎に算出し、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処理装置。

【請求項３】
前記第一サンプリング周波数で、前記音声受付部が受け付けた音声をサンプリングし、第一音声データを取得するサンプリング部をさらに具備し、
前記声道長正規化処理部は、
第二サンプリング周波数「前記第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記第一音声データに対して、リサンプリング処理を行い、第二音声データを得る請求項１または請求項２記載の音声処理装置。

【請求項４】
前記音声処理部は、
前記フレーム毎の入力音声データに基づいて、特殊な音声が入力されたことを検知する特

殊音声検知手段をさらに具備し、
前記評定手段は、
前記教師データと前記入力音声データと前記特殊音声検知手段における検知結果に基づいて、前記音声受付部が受け付けた音声の評定を行う請求項１から請求項３いずれか記載の音声処理装置。
【請求項５】
前記特殊音声検知手段は、
無音を示すＨＭＭに基づく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具備し、
前記評定手段は、
前記無音データの区間のデータを用いずに、音声の評定値を算出する請求項４記載の音声処理装置。
【請求項６】
前記特殊音声検知手段は、
一の音素の少なくとも最終フレームを含む後半部および当該音素の次の音素の少なくとも第一フレームを含む前半部の評定値が所定値より低い場合、または一の音素の少なくとも最終フレームを含む後半部および当該音素の次の音素の少なくとも第一フレームを含む前半部の評定値が所定値より低くかつ無音が挿入されていないと判断した場合、または、一の音素の少なくとも最終フレームを含む後半部および当該音素の次の音素の少なくとも第一フレームを含む前半部の評定値が所定値より低くかつ他の音韻ＨＭＭに対する確率値を算出し、所定の値より高い確率値を得た音韻を検出した場合、または評定値が所定値より低い音素が連続した場合に、音素の挿入があったと判断し、
前記評定手段は、
前記特殊音声検知手段が音素の挿入があったと判断した場合に、少なくとも音素の挿入があった旨を示す評定結果を構成する請求項４記載の音声処理装置。
【請求項７】
前記特殊音声検知手段は、
一の音素の評定値が所定の値より低くかつ当該音素の直前の音素および当該音素の直後の音素の評定値が所定の値より高い場合、または一の音素の評定値が所定の値より低く、かつ想定していない音素のＨＭＭに基づいて算出された評定値が所定の値より高い場合に、音韻の置換があったと判断し、
または、
一の音素の評定値が所定の値より低くかつ当該音素の直前の音素および当該音素の直後の音素の評定値が所定の値より高い場合、または一の音素の評定値が所定の値より低くかつ当該音素の直前の音素および当該音素の直後の音素の評定値が所定の値より高くかつ当該音素の区間長が所定の長さよりも短い場合、または直前の音素に対応する確率値、または直後の音素に対応する確率値が、当該一の音素の確率値より高い場合に、音韻の欠落があったと判断し、
前記評定手段は、
前記特殊音声検知手段が音素の置換があったと判断した場合に、少なくとも音素の置換があった旨を示す評定結果を構成し、
または、
前記特殊音声検知手段が音素の欠落があったと判断した場合に、少なくとも音素の欠落があった旨を示す評定結果を構成する請求項４記載の音声処理装置。
【請求項８】
前記音声処理装置は、カラオケ評価装置であって、
前記音声受付部は、
評価対象者の歌声の入力を受け付け、

前記音声処理部は、
前記歌声を評価する請求項１から請求項７いずれか記載の音声処理装置。
【請求項９】
第一サンプリング周波数を格納している第一サンプリング周波数格納部と、
前記教師データのフォルマント周波数である教師データフォルマント周波数を格納している教師データフォルマント周波数格納部と、
前記音声の話者である評価対象者のフォルマント周波数である評価対象者フォルマント周波数を格納している評価対象者フォルマント周波数格納部と、
第二サンプリング周波数「前記第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理部を具備し、
前記音声処理部は、
前記第二音声データを、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、
前記教師データと前記1以上のフレーム音声データに基づいて、前記音声受付部が受け付けた音声の評定を行う評定手段と、
前記評定手段における評定結果を出力する出力手段を具備し、
前記評定手段は、
前記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と、
前記最適状態決定手段が決定した最適状態における確率値と、当該確率値に対応するフレームの全状態における確率値の総和との比率を用いて、事後確率を示す確率値である最適状態確率値を取得する最適状態確率値取得手段と、
前記最適状態確率値取得手段が取得した最適状態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を具備するデジタルシグナルプロセッサ。
【請求項１０】
第一サンプリング周波数を格納している第一サンプリング周波数格納部と、
前記教師データのフォルマント周波数である教師データフォルマント周波数を格納している教師データフォルマント周波数格納部と、
前記音声の話者である評価対象者のフォルマント周波数である評価対象者フォルマント周波数を格納している評価対象者フォルマント周波数格納部と、
第二サンプリング周波数「前記第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理部を具備し、
前記音声処理部は、
前記第二音声データを、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、
前記教師データと前記1以上のフレーム音声データに基づいて、前記音声受付部が受け付けた音声の評定を行う評定手段と、
前記評定手段における評定結果を出力する出力手段を具備し、
前記評定手段は、
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２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２以上の確率値の総和を、フレーム毎に算出し、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デジタルシグナルプロセッサ。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、入力された音声を評価したり、入力された音声を認識したりする音声処理装置等に関するものである。

【背景技術】

【０００２】

従来の技術として、以下の音声処理装置がある（特許文献１参照）。本音声処理装置は、語学学習装置であり、当該語学学習装置は、学習者が選択した役割の発音をレファランスデータと比較して一致度によって点数化して表示し、点数によって適当な次の画面を自動に表示することにより、学習能率を向上させる装置である。本従来の音声処理装置は、入力された音声信号は音声認識技術により分析された後、学習者発音のスペクトルと抑揚とが学習者発音表示ボックスに表れるという構成になっている。そして、従来の技術においては、標準音データと学習者の発音のスペクトル、および抑揚が比較されて点数が表示される。

【０００３】

また、従来の技術として、以下の音声処理装置がある（特許文献２参照）。本音声処理装置は歌唱音声評価装置であり、本歌唱音声評価装置は、歌唱音声の周波数成分を抽出する抽出手段と、当該抽出された周波数成分から基本周波数成分と倍音周波数成分とをそれぞれ抽出する特定周波数成分抽出手段と、特定周波数成分抽出手段によって抽出された基本周波数成分に対する倍音周波数成分の比率に応じて、歌唱音声の評価を示す評価値を算出する評価手段とを備える。そして、本歌唱音声評価装置は、歌唱音声の周波数成分に基づいてその声質の良否を適正に評価し、これを歌唱音声の採点結果に反映させることにより、歌唱音声の採点をより人間の感性に近づけることを狙いとしている。

【０００４】

さらに、従来の技術として、以下の音声処理装置がある（特許文献３参照）。本音声処理装置は音声認識装置であり、入力音声パターンと標準パターンを、ＤＰ法を用いて照合し、最も照合距離の小さい標準パターンを認識結果とする音声認識装置であり、照合結果を用いて入力パターンを音素に分割し、各音素の継続時間と標準継続時間とのずれの分散を計算し、これを照合距離に付加することで距離を補正することを特徴とする。そして、分割部で照合結果を用いて音素に分割し、時間長ずれ計算部で標準継続時間とのずれの分散を計算し、距離補正部で照合距離を補正するように構成する。また、本音声認識装置は、時間長のずれを計算する対象音素を選択する音素選択部、距離補正する対象単語を選択する単語選択部を有し、単語の認識性能を高できる、というものである。

【特許文献１】特開２００３－２２８２７９（第１頁、第１図等）

【特許文献２】特開２００５－１０７０８８（第１頁、第１図等）

【特許文献３】特開平６－４０９６（第１頁、第１図等）

【発明の開示】

【発明が解決しようとする課題】

【０００５】

しかしながら、特許文献１や特許文献２の従来の技術においては、音声（歌声も含む）の話者である評価対象者の話者特性に応じた音声処理が行えず、その結果、精度の高い音声処理ができなかった。具体的には、従来の技術においては、例えば、評価対象者の声道長の違いにより、スペクトル包絡が高周波数域または低周波数域に伸縮するが、従来の発音評定装置や歌唱音声評価装置などの音声処理装置において、かかるスペクトル包絡の伸縮により、評価結果が異なる。つまり、従来の技術においては、同様の上手さの発音や歌唱でも、評価対象者の声道長の違いにより、発音や歌唱の評価結果が異なり、精度の高い評価ができなかった。

【０００６】

また、特許文献１の音声処理装置において、標準音データと学習者の発音のスペクトル、および抑揚が比較されて点数が表示される構成であるので、両者の類似度の評定の精度が低く、また、リアルタイムに高速に点数を表示するためには、処理能力が極めて高いＣＰＵ、多量のメモリが必要であった。

【０００７】

また、特許文献１の音声処理装置において、無音区間があれば、類似度が低く評価されると考えられ、評価の精度が低かった。また、音素の置換や挿入や欠落など、特殊な事象が発生していることを検知できなかった。

【０００８】

さらに、例えば、特許文献３に示すような音声認識処理を行う音声処理装置において、評価対象者の声道長の違いにより、スペクトル包絡の伸縮が生じるが、かかる評価対象者の話者特性に応じた音声認識処理を行っておらず、精度の高い音声認識ができなかった。

【課題を解決するための手段】

【０００９】

本第一の発明の音声処理装置は、比較される対象の音声に関するデータであり、1以上の音韻毎のデータである教師データを１以上格納している教師データ格納部と、音声を受け付ける音声受付部と、第一サンプリング周波数を格納している第一サンプリング周波数格納部と、前記第一サンプリング周波数で、前記音声受付部が受け付けた音声をサンプリングし、第一音声データを取得するサンプリング部と、前記教師データのフォルマント周波数である教師データフォルマント周波数を格納している教師データフォルマント周波数格納部と、前記音声受付部が受け付けた音声の話者である評価対象者のフォルマント周波数である評価対象者フォルマント周波数を格納している評価対象者フォルマント周波数格納部と、第二サンプリング周波数「前記第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付部が受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理部と、前記第二音声データを処理する音声処理部を具備する音声処理装置である。

かかる構成により、評価対象者の話者特性に応じた精度の高い音声処理ができる。

【００１０】

また、本第二の発明の音声処理装置は、第一の発明に対して、前記音声処理部は、前記第二音声データを、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、前記教師データと前記１以上のフレーム音声データに基づいて、前記音声受付部が受け付けた音声の評定を行う評定手段と、前記評定手段における評定結果を出力する出力手段を具備する音声処理装置である。

かかる構成により、評価対象者の話者特性に応じた精度の高い音声の評定ができる。

【００１１】

また、本第三の発明の音声処理装置は、第二の発明に対して、前記評定手段は、前記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と、前記最適状態決定手段が決定した最適状態における確率値を取得する最適状態確率値取得手段と、前記最適状態確率値取得手段が取得した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を具備する音声処理装置である。

かかる構成により、評価対象者の話者特性に応じた精度の高い音声の評定ができる。

【００１２】

また、本第四の発明の音声処理装置は、第二の発明に対して、前記評定手段は、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定値を算出する評定値算出手段を具備する音声処理装置である。

かかる構成により、評価対象者の話者特性に応じた、さらに精度の高い音声の評定ができる。

【００１３】

また、本第五の発明の音声処理装置は、第二の発明に対して、前記音声処理部は、前記フレーム毎の入力音声データに基づいて、特殊な音声が入力されたことを検知する特殊音声検知手段をさらに具備し、前記評定手段は、前記教師データと前記入力音声データと前記特殊音声検知手段における検知結果に基づいて、前記音声受付部が受け付けた音声の評定を行う音声処理装置である。

かかる構成により、評価対象者の話者特性に応じた精度の高い音声の評定ができ、かつ特殊音声を検知し、かかる特殊音声に対応した音声の評定ができる。

【００１４】

また、本第六の発明の音声処理装置は、第五の発明に対して、前記特殊音声検知手段は、無音を示すＨＭＭに基づく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具備する音声処理装置である。

かかる構成により、評価対象者の話者特性に応じた精度の高い音声の評定ができ、かつ無音区間を検知し、かかる無音区間に対応した音声の評定ができる。

【００１５】

また、本第七の発明の音声処理装置は、第五の発明に対して、前記特殊音声検知手段は、一の音素の後半部および当該音素の次の音素の前半部の評定値が所定の条件を満たすことを検知し、前記評定手段は、前記特殊音声検知手段が前記所定の条件を満たすことを検知した場合に、少なくとも音素の挿入があった旨を示す評定結果を構成する音声処理装置である。

かかる構成により、評価対象者の話者特性に応じた精度の高い音声の評定ができ、かつ音素の挿入を検知し、かかる音素の挿入に対応した音声の評定ができる。

【００１６】

また、本第八の発明の音声処理装置は、第七の発明に対して、前記特殊音声検知手段は、一の音素の評定値が所定の条件を満たすことを検知し、前記評定手段は、前記特殊音声検知手段が前記所定の条件を満たすことを検知した場合に、少なくとも音素の置換または欠落があった旨を示す評定結果を構成する音声処理装置である。

かかる構成により、評価対象者の話者特性に応じた精度の高い音声の評定ができ、かつ音素の置換または欠落を検知し、かかる音素の置換または欠落に対応した音声の評定ができる。

【００１７】

また、本第九の発明の音声処理装置は、第二から第八いずれかの発明に対して、前記音声処理装置は、カラオケ評価装置であって、前記音声受付部は、評価対象者の歌声の入力を受け付け、前記音声処理部は、前記歌声を評価する音声処理装置である。

かかる構成により、カラオケ評価装置として利用できる。

【００１８】

また、本第十の発明の音声処理装置は、第九の発明に対して、前記フレーム区分手段は、前記音声をフレームに区分し、かつ、前記第二音声データをフレームに区分し、前記フレーム音声データ取得手段は、前記音声が区分されたフレーム毎の音声データである第一フレーム音声データを１以上得て、かつ前記第二音声データが区分されたフレーム毎の音声データである第二フレーム音声データを１以上得、前記評定手段は、前記教師データと前記１以上の第一フレーム音声データに基づいて、前記音声受付部が受け付けた音声の評定を行う第一評定手段と、前記教師データと前記１以上の第二フレーム音声データに基づいて、前記音声受付部が受け付けた音声の評定を行う第二評定手段と、前記第一評定手段における評定結果と前記第二評定手段における評定結果に基づいて、最終的な評定結果を得る評定結果取得手段とを具備する音声処理装置である。

かかる構成により、優れたカラオケ評価装置として利用できる。

【００１９】
また、本第十一の発明の音声処理装置は、第九、第十いずれかの発明に対して、前記音声受付部は、所定の母音の音声を受け付けた後、評価対象者の歌声の入力を受け付け、前記サンプリング部は、前記第一サンプリング周波数で、前記母音の音声をもサンプリングし、前記サンプリングした母音の音声に基づいて、評価対象者のフォルマント周波数である評価対象者フォルマント周波数を取得する評価対象者フォルマント周波数取得部をさらに具備し、前記評価対象者フォルマント周波数格納部の評価対象者フォルマント周波数は、前記評価対象者フォルマント周波数取得部が取得した評価対象者フォルマント周波数である音声処理装置である。
かかる構成により、評価対象者の話者特性に応じた精度の高い音声の評定ができる。
また、本第十二の発明の音声処理装置は、第一の発明に対して、前記音声処理部は、前記第二音声データに基づいて、音声認識処理を行う音声処理装置である。
かかる構成により、評価対象者の話者特性に応じた精度の高い音声認識ができる。
【発明の効果】
【００２０】
本発明による音声処理装置によれば、評価対象者の話者特性に応じた精度の高い音声処理ができる。
【発明を実施するための最良の形態】
【００２１】
以下、音声処理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。
（実施の形態１）
【００２２】
本実施の形態において、比較対象の音声と入力音声の類似度の評定を精度高く、かつ高速にできる音声処理装置について説明する。本音声処理装置は、音声（歌唱を含む）を評価する発音評定装置である。特に、本音声処理装置は、入力音声のフレーム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装置をＤＡＰＳと呼ぶ。
【００２３】
また、本実施の形態における音声処理装置は、例えば、語学学習や物真似練習やカラオケ評定などに利用できる。図１は、本実施の形態における音声処理装置のブロック図である。本音声処理装置は、入力受付部１０１、教師データ格納部１０２、音声受付部１０３、教師データフォルマント周波数格納部１０４、第一サンプリング周波数格納部１０５、サンプリング部１０６、評価対象者フォルマント周波数取得部１０７、評価対象者フォルマント周波数格納部１０８、声道長正規化処理部１０９、音声処理部１１０を具備する。
音声処理部１１０は、フレーム区分手段１１０１、フレーム音声データ取得手段１１０２、評定手段１１０３、出力手段１１０４を具備する。
評定手段１１０３は、最適状態決定手段１１０３１、最適状態確率値取得手段１１０３２、評定値算出手段１１０３３を具備する。
【００２４】
なお、音声処理装置は、キーボード３４２、マウス３４３などの入力手段からの入力を受け付ける。また、音声処理装置は、マイク３４５などの音声入力手段から音声入力を受け付ける。さらに、音声処理装置は、ディスプレイ３４４などの出力デバイスに情報を出力する。
【００２５】
入力受付部１０１は、音声処理装置の動作開始を指示する動作開始指示や、入力した音声の評定結果の出力態様の変更を指示する出力態様変更指示や、処理を終了する終了指示などの入力を受け付ける。かかる指示等の入力手段は、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。入力受付部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ーや、メニュー画面の制御ソフトウェア等で実現され得る。

【００２６】

教師データ格納部１０２は、教師データとして比較される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（ＨＭＭ）に基づくデータを１以上格納している。教師データは、音韻毎の隠れマルコフモデル（ＨＭＭ）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、教師データは、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、教師データは、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。教師データは、全音素のＨＭＭの、単なる集合であっても良い。また、教師データは、必ずしもＨＭＭに基づくデータである必要はない。教師データは、単一ガウス分布モデルや、確率モデル（ＧＭＭ：ガウシャ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、複数の学習対象言語を母国語として話す外国人が発声した２以上のデータから学習した（推定した）モデルでも良い。教師データ格納部１０２は、ハードディスクやＲＯＭなどの不揮発性の記録媒体が好適であるが、ＲＡＭなどの揮発性の記録媒体でも実現可能である。

【００２７】

音声受付部１０３は、音声を受け付ける。音声受付部１０３は、例えば、マイク３４５のドライバーソフトで実現され得る。また、なお、音声受付部１０３は、マイク３４５とそのドライバーから実現されると考えても良い。音声は、マイク３４５から入力されても良いし、磁気テープやＣＤ－ＲＯＭなどの記録媒体から読み出すことにより入力されても良い。

【００２８】

教師データフォルマント周波数格納部１０４は、教師データのフォルマント周波数である教師データフォルマント周波数を格納している。教師データフォルマント周波数は、第一フォルマント周波数（Ｆ１）でも、第二フォルマント周波数（Ｆ２）でも、第三フォルマント周波数（Ｆ３）等でも良い。教師データフォルマント周波数格納部１０４の教師データフォルマント周波数は、予め格納されていても良いし、評価時に、動的に、教師データから取得しても良い。音声データからフォルマント周波数を取得する技術は、公知技術であるので説明を省略する。教師データフォルマント周波数格納部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。

【００２９】

第一サンプリング周波数格納部１０５は、第一のサンプリング周波数である第一サンプリング周波数を格納している。第一サンプリング周波数は、評価対象者の音声を、最初にサンプリングする場合のサンプリング周波数である。第一サンプリング周波数格納部１０５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。

【００３０】

サンプリング部１０６は、第一サンプリング周波数格納部１０５の第一サンプリング周波数で、音声受付部１０３が受け付けた音声をサンプリングし、第一音声データを取得する。なお、受け付けた音声をサンプリングする技術は公知技術であるので、詳細な説明を省略する。サンプリング部１０６は、通常、ＭＰＵやメモリ等から実現され得る。サンプリング部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。

【００３１】

評価対象者フォルマント周波数取得部１０７は、サンプリング部１０６が取得した第一音声データから、評価対象者のフォルマント周波数である評価対象者フォルマント周波数を取得する。評価対象者フォルマント周波数も、第一フォルマント周波数（Ｆ１）でも、

第二フォルマント周波数（Ｆ２）でも、第三フォルマント周波数（Ｆ３）でも良い。ただし、評価対象者フォルマント周波数と教師データフォルマント周波数は同一種のフォルマント周波数である。サンプリングして取得した第一音声データから、フォルマント周波数を取得する技術は公知技術であるので、詳細な説明を省略する。評価対象者フォルマント周波数取得部１０７は、通常、ＭＰＵやメモリ等から実現され得る。評価対象者フォルマント周波数取得部１０７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。

【００３２】

評価対象者フォルマント周波数格納部１０８は、音声受付部１０３が受け付けた音声の話者である評価対象者のフォルマント周波数である評価対象者フォルマント周波数を、少なくとも一時的に格納している。評価対象者フォルマント周波数格納部１０８の評価対象者フォルマント周波数は、通常、評価対象者フォルマント周波数取得部１０７が取得したフォルマント周波数であるが、予め評価対象者フォルマント周波数を格納していても良い。評価対象者フォルマント周波数格納部１０８に、予め評価対象者フォルマント周波数が格納されている場合、本音声処理装置において、評価対象者フォルマント周波数取得部１０７は不要である。評価対象者フォルマント周波数格納部１０８は、不揮発性の記録媒体でも、揮発性の記録媒体でも良い。

【００３３】

声道長正規化処理部１０９は、第二サンプリング周波数で、音声受付部１０３が受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る。第二サンプリング周波数は、「第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で算出されるサンプリング周波数である。声道長正規化処理部１０９は、音声受付部１０３が受け付けた音声をサンプリング処理して得られた第一音声データを、リサンプリング処理して第二音声データを得ることが好適であるが、音声受付部１０３が受け付けた音声をサンプリング処理し、直接的に第二音声データを得ても良い。直接的に第二音声データを得る場合、例えば、サンプリング処理を行うハードウェアが可変のサンプリング周波数でサンプリング処理を行えることが必要である。声道長正規化処理部１０９は、通常、演算「教師データフォルマント周波数／評価対象者フォルマント周波数」を行い、周波数スケール（「ｒ」とする）を得る。そして、声道長正規化処理部１０９は、第一サンプリング周波数格納部１０５の第一サンプリング周波数（Ｆｓ）と「ｒ」に基づいて、演算「Ｆｓ／ｒ」を行い、新しいサンプリング周波数（Ｆｓ／ｒ）を得る。この新しいサンプリング周波数（Ｆｓ／ｒ）が第二サンプリング周波数である。次に、声道長正規化処理部１０９は、第一音声データに対して、第二サンプリング周波数（Ｆｓ／ｒ）で、リサンプリング処理を行い、第二音声データを得る。声道長正規化処理部１０９は、通常、ＭＰＵやメモリ等から実現され得る。声道長正規化処理部１０９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。

【００３４】

音声処理部１１０は、第二音声データを処理する。音声処理部１１０は、ここでは、評定処理である。ただし、音声処理部１１０は、音声認識や音声出力などの他の音声処理を行っても良い。音声出力は、単に、リサンプリング処理された音声を出力する処理である。なお、本実施の形態において、音声処理部１１０は、評定処理を行うものとして、説明する。音声処理部１１０は、通常、ＭＰＵやメモリ等から実現され得る。音声処理部１１０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。

【００３５】

音声処理部１１０を構成しているフレーム区分手段１１０１は、第二音声データを、フレームに区分する。フレーム区分手段１１０１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１１０１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。

【００３６】

音声処理部１１０を構成しているフレーム音声データ取得手段１１０２は、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１１０２は、通常、ＭＰＵやメモリ等から実現され得る。フレーム音声データ取得手段１１０２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。

【００３７】

音声処理部１１０を構成している評定手段１１０３は、教師データ格納部１０２の教師データと1以上のフレーム音声データに基づいて、音声受付部１０３が受け付けた音声の評定を行う。評定方法の具体例は、後述する。「音声受付部１０３が受け付けた音声を評定する」の概念には、第二音声データを評定することも含まれることは言うまでもない。評定手段１１０３は、通常、ＭＰＵやメモリ等から実現され得る。評定手段１１０３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。

【００３８】

評定手段１１０３を構成している最適状態決定手段１１０３１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１１０３１は、例えば、全音韻ＨＭＭから、比較される対象（学習対象）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から、音素の順序で連結したデータ（比較される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。また、教師データは、上述の比較される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に基づくデータと考えても良いし、連結される前のデータであり、全音韻ＨＭＭのデータと考えても良い。

評定手段１１０３を構成している最適状態確率値取得手段１１０３２は、最適状態決定手段１１０３１が決定した最適状態における確率値を取得する。

【００３９】

評定手段１１０３を構成している評定値算出手段１１０３３は、最適状態確率値取得手段１１０３２が取得した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１１０３３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。評定値算出手段１１０３３は、例えば、最適状態確率値取得手段１１０３２が取得した確率値と、当該確率値に対応するフレームの全状態における確率値の総和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１１０３３は、ここでは、通常、フレームごとに評定値を算出する。

【００４０】

最適状態決定手段１１０３１、最適状態確率値取得手段１１０３２、評定値算出手段１１０３３は、通常、ＭＰＵやメモリ等から実現され得る。最適状態決定手段１１０３１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。

【００４１】

出力手段１１０４は、評定手段１１０３における評定結果を出力する。出力手段１１０４の出力態様は、種々考えられる。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。出力手段１１０４は、例えば、評定結果を視覚的に表示する。出力手段１１０４は、例えば、フレーム単位、または／および音素・単語単位、または／および発声全体の評定結果を視覚的に表示

する。出力手段１１０４は、ディスプレイ３４４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。出力手段１１０４は、出力デバイスのドライバーソフトまたは、出力デバイスのドライバーソフトと出力デバイス等で実現され得る。

次に、本音声処理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。

【００４２】

（ステップＳ２０１）入力受付部１０１は、音声処理装置の動作開始を指示する動作開始指示を受け付けたか否かを判断する。動作開始指示を受け付ければステップＳ２０２に行き、動作開始指示を受け付けなければステップＳ２１７に飛ぶ。

（ステップＳ２０２）音声受付部１０３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Ｓ２０３に行き、音声を受け付けなければステップＳ２１６に飛ぶ。

【００４３】

（ステップＳ２０３）サンプリング部１０６は、第一サンプリング周波数格納部１０５に格納されている第一サンプリング周波数を読み込み、当該第一サンプリング周波数で、音声受付部１０３が受け付けた音声をサンプリングし、第一音声データを得る。

【００４４】

（ステップＳ２０４）声道長正規化処理部１０９は、音声受付部１０３が受け付けた音声から、第二音声データを得る。かかる第二音声データを得る処理である声道長正規化処理の詳細については、図３のフローチャートを用いて、詳細に説明する。なお、声道長正規化処理は、個人差を吸収する評定のための前処理である。

（ステップＳ２０５）フレーム区分手段１１０１は、ステップＳ２０４で得た第二音声データを図示しないバッファに一時格納する。

【００４５】

（ステップＳ２０６）フレーム区分手段１１０１は、バッファに一時格納した第二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１１０１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１１０２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。

（ステップＳ２０７）フレーム音声データ取得手段１１０２は、カウンタｉに１を代入する。

【００４６】

（ステップＳ２０８）フレーム音声データ取得手段１１０２は、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Ｓ２０９に行き、ｉ番目のフレームが存在しなければステップＳ２１１に行く。

【００４７】

（ステップＳ２０９）フレーム音声データ取得手段１１０２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。

（ステップＳ２１０）フレーム音声データ取得手段１１０２は、カウンタｉを１、インクリメントする。ステップＳ２０８に戻る。

【００４８】

（ステップＳ２１１）最適状態決定手段１１０３１は、全フレームの最適状態を決定する。最適状態決定手段１１０３１が最適状態を決定するアルゴリズムは、例えば、Ｖｉｔ

ｅｒｂｉアルゴリズムによる。Ｖｉｔｅｒｂｉ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。

【００４９】

（ステップＳ２１２）最適状態確率値取得手段１１０３２は、全フレームの全状態の前向き尤度、および後向き尤度を算出する。最適状態確率値取得手段１１０３２は、例えば、全てのＨＭＭを用いて、フォワード・バックワードアルゴリズムにより、前向き尤度、および後向き尤度を算出する。

（ステップＳ２１３）最適状態確率値取得手段１１０３２は、ステップＳ２１２で取得した前向き尤度、および後向き尤度を用いて、最適状態の確率値（最適状態確率値）を、すべて算出する。

【００５０】

（ステップＳ２１４）評定値算出手段１１０３３は、ステップＳ２１３で算出した1以上の最適状態確率値から、1以上のフレームの音声の評定値を算出する。評定値算出手段１１０３３が評定値を算出する関数は問わない。評定値算出手段１１０３３は、例えば、取得した最適状態確率値と、当該最適状態確率値に対応するフレームの全状態における確率値の総和をパラメータとして音声の評定値を算出する。詳細については、後述する。

【００５１】

（ステップＳ２１５）出力手段１１０４は、ステップＳ２１４における評定結果（ここでは、音声の評定値）を、設定されている出力モードに従って、出力する。ステップＳ２０２に戻る。出力モードとは、評定値を数値で画面に表示するモード、評定値の遷移をグラフで画面に表示するモード、評定値を音声で出力するモード、評定値が所定の数値より低い場合に警告を示す情報を表示するモードなど、何でも良い。なお、ここでの出力モードは、ステップＳ２１８で設定されるモードである。

【００５２】

（ステップＳ２１６）音声受付部１０３は、タイムアウトか否かを判断する。つまり、音声受付部１０３は、所定の時間以上、音声の入力を受け付けなかったか否かを判断する。タイムアウトであればステップＳ２０１に戻り、タイムアウトでなければステップＳ２０２に戻る。

【００５３】

（ステップＳ２１７）入力受付部１０１は、出力態様変更指示を受け付けたか否かを判断する。出力態様変更指示を受け付ければステップＳ２１８に行き、出力態様変更指示を受け付なければステップＳ２１９に飛ぶ。出力態様変更指示は、上述した出力モードを有する情報である。

（ステップＳ２１８）出力手段１１０４は、ステップＳ２１７で受け付けた出力態様変更指示が有する出力モードを示す情報を書き込み、出力モードを設定する。ステップＳ２０１に戻る。

（ステップＳ２１９）入力受付部１０１は、終了指示を受け付けたか否かを判断する。終了指示を受け付ければ処理を終了し、終了指示を受け付なければステップＳ２０１に戻る。

なお、図２のフローチャートにおいて、本発音評定装置は、出力モードの設定機能を有しなくても良い。

次に、ステップＳ２０４における声道長正規化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。

【００５４】

（ステップＳ３０１）評価対象者フォルマント周波数取得部１０７は、サンプリング部１０６のサンプリング処理により得られた第一音声データから、評価対象者フォルマント周波数（Ｆｉ）を取得し、評価対象者フォルマント周波数格納部１０８に一時格納する。評価対象者フォルマント周波数は、例えば、第二フォルマント周波数（Ｆ２）である。

（ステップＳ３０２）声道長正規化処理部１０９は、第一サンプリング周波数格納部１０５の第一サンプリング周波数（Ｆｓ）を読み出す。

（ステップＳ３０３）声道長正規化処理部１０９は、教師データフォルマント周波数格納部１０４の教師データフォルマント周波数を読み出す。

【００５５】

（ステップＳ３０４）声道長正規化処理部１０９は、ステップＳ３０１で取得した評価対象者フォルマント周波数と、ステップＳ３０３で読み出した教師データフォルマント周波数から周波数スケールを算出する。具体的には、声道長正規化処理部１０９は、演算「教師データフォルマント周波数／評価対象者フォルマント周波数」を行い、周波数スケール（ｒ）を得る。

【００５６】

（ステップＳ３０５）声道長正規化処理部１０９は、ステップＳ３０２で読み出した第一サンプリング周波数（Ｆｓ）と周波数スケール（ｒ）に基づいて、演算「Ｆｓ／ｒ」を行い、第二サンプリング周波数（Ｆｓ／ｒ）を得る。

【００５７】

（ステップＳ３０６）声道長正規化処理部１０９は、サンプリング部１０６がサンプリングして得た第一音声データに対して、第二サンプリング周波数（Ｆｓ／ｒ）で、リサンプリング処理を行い、第二音声データを得る。なお、リサンプリング処理は公知技術であるので、詳細な説明を省略する。上位関数にリターンする。

【００５８】

なお、図２、図３のフローチャートにおいて、声道長正規化処理を行う対象の音声と、評価対象の音声が異なっても良い。つまり、例えば、音声受付部１０３は、所定の１以上の母音（例えば、「う」）の音声を受け付けた後、評価対象者の音声を受け付け、評価対象者フォルマント周波数取得部１０７は、当該１以上の母音の音声に基づいて、評価対象者フォルマント周波数を取得し、声道長正規化処理部１０９は、当該評価対象者フォルマント周波数をパラメータとして、声道長正規化処理を行う。そして、音声処理部１１０は、所定の母音の音声を受け付けた後に受け付けた音声を処理し、当該音声の評価を行っても良い。

以下、本実施の形態における音声処理装置の具体的な動作について説明する。本具体例において、音声処理装置が語学学習に利用される場合について説明する。

【００５９】

まず、本音声処理装置において、図示しない手段により、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の仕様の例について、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。

【００６０】

そして、図示しない手段により、学習したＬ種類の音韻ＨＭＭから、学習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教師データを構成する。そして、当該教師データを教師データ格納部１０２に保持しておく。ここでは、例えば、比較される対象の音声は、単語「ｒｉｇｈｔ」の音声である。また、ここでは、教師データを発生した者（教師）は、大人である、とする。

【００６１】

次に、学習者（評価対象者）が、語学学習の開始の指示である動作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。なお、ここでは、学習者は、例えば、子供（５歳から１１歳）である、とする。

【００６２】

まず、学習者は、母音「う」を発音する、とする。かかる場合、本音声処理装置は、学習に、「う」を発声するように促すことは好適である。「う」を発声するように促すために、音声処理装置は、例えば、「"う"と発声してください。」と画面出力しても良いし、

「"う"と発声してください。」と音声出力しても良い。また、母音「う」は、学習者の評価対象者フォルマント周波数を取得するために好適である。また、本音声処理装置は、第一サンプリング周波数として、「２２．０５ＫＨｚ」を保持している、とする。

そして、次に、サンプリング部１０６は、音声受付部１０３が受け付けた音声「う」をサンプリングし、「う」の第一音声データを得る。

【００６３】

次に、評価対象者フォルマント周波数取得部１０７は、サンプリング部１０６が音声「う」をサンプリングして得た第一音声データから、第二フォルマント周波数を取得する。そして、この第二フォルマント周波数を評価対象者フォルマント周波数（Ｆｉとする。今、このＦｉが「１７２５Ｈｚ」であった、とする。そして、評価対象者フォルマント周波数取得部１０７は、Ｆｉ（１７２５Ｈｚ）を、評価対象者フォルマント周波数格納部１０８に一時格納する。

【００６４】

次に、声道長正規化処理部１０９は、教師データフォルマント周波数格納部１０４の教師データフォルマント周波数を読み出す。教師データフォルマント周波数格納部１０４に格納されている教師データフォルマント周波数は、大人の第二フォルマント周波数であり、今、「１１８４Ｈｚ」である、とする。また、教師データフォルマント周波数は、例えば、教師データを構築する場合に、教師に、例えば、「う」と発声してもらい、当該音声「う」をサンプリング処理した後、取得した第二フォルマント周波数である。

【００６５】

なお、図５に、年齢層別、性別ごとの、「う」の第一フォルマント周波数（Ｆ１）、第二フォルマント周波数（Ｆ２）の計測結果を示す。図５により、年齢、性別により、第一フォルマント周波数（Ｆ１）、第二フォルマント周波数（Ｆ２）の値が大きく異なることが分る。

【００６６】

そして、次に、声道長正規化処理部１０９は、評価対象者フォルマント周波数「１７２５Ｈｚ」と教師データフォルマント周波数「１１８４Ｈｚ」から演算「教師データフォルマント周波数／評価対象者フォルマント周波数」を行い、周波数スケール（ｒ）を得る。具体的には、声道長正規化処理部１０９は、「１１８４／１７２５」により、周波数スケール「０．６８６」を得る。

【００６７】

次に、声道長正規化処理部１０９は、第一サンプリング周波数（Ｆｓ）と「ｒ」に基づいて、演算「Ｆｓ／ｒ」を行い、第二サンプリング周波数（Ｆｓ／ｒ）を得る。ここで、得た第二サンプリング周波数は、「２２．０５／０．６８６」により、「３２．１」である。そして、声道長正規化処理部１０９は、第二サンプリング周波数「３２．１ＫＨｚ」を一時格納する。

【００６８】

次に、学習者は、学習対象の音声「ｒｉｇｈｔ」を発音する。そして、音声受付部１０３は、学習者が発音した音声の入力を受け付ける。なお、音声処理装置は、学習者に「"ｒｉｇｈｔ"を発音してください。」などを表示、または音声出力するなどして、学習者に「ｒｉｇｈｔ」の発声を促すことは好適である。

【００６９】

次に、サンプリング部１０６は、受け付けた音声「ｒｉｇｈｔ」をサンプリング周波数「２２．０５ＫＨｚ」でサンプリング処理する。そして、サンプリング部１０６は、音声「ｒｉｇｈｔ」の第一音声データを得る。

次に、声道長正規化処理部１０９は、「ｒｉｇｈｔ」の第一音声データを第二サンプリング周波数「３２．１ＫＨｚ」でリサンプリング処理する。そして、声道長正規化処理部１０９は、第二音声データを得る。

次に、音声処理部１１０は、第二音声データを、以下のように処理する。

まず、フレーム区分手段１１０１は、第二音声データを、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。

【００７０】

そして、フレーム音声データ取得手段１１０２は、フレーム区分手段１１０１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$_2$，・・・，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、音声分析条件の例を図６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。また、音声分析の際のサンプリング周波数は、第一サンプリング周波数「２２．０５ＫＨｚ」である。

【００７１】

次に、最適状態決定手段１１０３１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）「q_t^*」を決定する。最適状態決定手段１１０３１が最適状態「q_t^*」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１１０３１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１１０３１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。

【００７２】

次に、最適状態確率値取得手段１１０３２は、以下の数式１により、最適状態における最適状態確率値（$\gamma_t(q_t^*)$）を算出する。なお、$\gamma_t(q_t^*)$は、状態ｊの事後確率関数$\gamma_t(j)$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$\gamma_t(j)$は、数式２を用いて算出される。この確率値（$\gamma_t(j)$）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。

【数１】

$$DAP(t) = \gamma_t(q_t^*)$$

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（$\gamma_t(j)$）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】

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

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

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

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

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

【００７３】

数式２において、「α t （ｊ）」「β 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。「α t （ｊ）」は前向き尤度、「β 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。

また、数式２において、Ｎは、全ＨＭＭに渡る状態の総数を示す。

【００７４】

なお、評定手段１１０３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、評定手段１１０３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。

【００７５】

次に、評定値算出手段１１０３３は、例えば、上記の取得した最適状態確率値と、当該最適状態確率値に対応するフレームの全状態における確率値の総和をパラメータとして音声の評定値を算出する。かかる場合、もし学習者のｔフレーム目に対応する発声が、教師データが示す発音（例えば、正しいネイティブな発音）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（評定値）が大きくなる。逆にその区間が、教師データが示す発音に近くなければ、評定値は小さくなる。なお、どのネイティブ発音にも近くないような場合は、評定値はほぼ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、この評定値は十分小さくなる。また、ここでは、評定値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、収音環境等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、学習者が正しい発音をしていれば、その変動が相殺され評定値が高いスコアを維持する。よって、評定値算出手段１１０３３は、最適状態確率値取得手段１１０３２が取得した確率値と、当該確率値に対応するフレームの全状態における確率値の総和をパラメータとして音声の評定値を算出することは、極めて好適である。

【００７６】

かかる評定値算出手段１１０３３が算出した評定値（「ＤＡＰスコア」とも言う。）を、図７、図８に示す。図７、図８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Ｖｉｔｅｒ

ｂｉアルゴリズムで求まったもの）を表わしており，図の上部に音素名を表記している。図７は、アメリカ人男性による英語「ｒｉｇｈｔ」の発音のＤＡＰスコアを示す。なお、評定値を示すグラフの横軸、縦軸は、後述するグラフにおいても同様である。

【００７７】

図８は、日本人男性による英語「ｒｉｇｈｔ」の発音のＤＡＰ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。なお、図７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。

【００７８】

そして、出力手段１１０４は、評定手段１１０３の評定結果を出力する。具体的には、例えば、出力手段１１０４は、図９に示すような態様で、評定結果を出力する。つまり、出力手段１１０４は、各フレームにおける発音の良さを表すスコア（スコアグラフ）として、各フレームの評定値を表示する。その他、出力手段１１０４は、学習対象の単語の表示（単語表示）、音素要素の表示（音素表示）、教師データの波形の表示（教師波形）、学習者の入力した発音の波形の表示（ユーザ波形）を表示しても良い。なお、図９において、「録音」ボタンを押下すれば、動作開始指示が入力されることとなり、「停止」ボタンを押下すれば、終了指示が入力されることとなる。また、音素要素の表示や波形の表示をする技術は公知技術であるので、その詳細説明を省略する。また、本音声処理装置は、学習対象の単語（図９の「ｗｏｒｄ１」など）や、音素（図９の「ｐ１」など）や、教師波形を出力されるためのデータを予め格納している、とする。

【００７９】

また、図９において、フレーム単位以外に、音素単位、単語単位、発声全体の評定結果を表示しても良い。上記の処理において、フレーム単位の評定値を算出するので、単語単位、発声全体の評定結果を得るためには、フレーム単位の１以上の評定値をパラメータとして、単語単位、発声全体の評定値を算出する必要がある。かかる算出式は問わないが、例えば、単語を構成するフレーム単位の１以上の評定値の平均値を単語単位の評定値とする、ことが考えられる。

【００８０】

なお、図９において、音声処理装置は、波形表示（教師波形またはユーザ波形）の箇所においてクリックを受け付けると、再生メニューを表示し、音素区間内ではその音素またはその区間が属する単語、波形全体を再生し、単語区間外（無音部）では波形全体のみを再生するようにしても良い。

また、出力手段１１０４の表示は、図１０に示すような態様でも良い。図１０において、音素ごとのスコア、単語のスコア、総合スコアが、数字で表示されている。

なお、出力手段１１０４の表示は、図７、図８のような表示でも良いことは言うまでもない。

【００８１】

以上、本実施の形態によれば、ユーザが入力した発音を、教師データに対して、如何に似ているかを示す類似度（評定値）を算出し、出力できる。また、かかる場合、本実施の形態によれば、個人差、特に声道長の違いに影響を受けない、精度の高い評定ができる。

【００８２】

また、本実施の形態によれば、連結されたＨＭＭである連結ＨＭＭを用いて最適状態を求め、評定値を算出するので、高速に評定値を求めることができる。したがって、上記の具体例で述べたように、リアルタイムに、フレームごと、音素ごと、単語ごとの評定値を出力できる。また、本実施の形態によれば、動的計画法に基づいた事後確率を確率値として算出するので、さらに高速に評定値を求めることができる。また、本実施の形態によれば、フレームごとに確率値を算出するので、上述したように、フレーム単位だけではなく、または／および音素・単語単位、または／および発声全体の評定結果を出力でき、出力態様の自由度が高い。

【００８３】

また、本実施の形態によれば、音声処理装置は、語学学習に利用することを主として説

明したが、物真似練習や、カラオケ評定や、歌唱評定などに利用できる。つまり、本音声処理装置は、比較される対象の音声に関するデータとの類似度を精度良く、高速に評定し、出力でき、そのアプリケーションは問わない。つまり、例えば、本音声処理装置は、カラオケ評価装置であって、音声受付部は、評価対象者の歌声の入力を受け付け、音声処理部は、前記歌声を評価する、という構成でも良い。かかることは、他の実施の形態においても同様である。

【００８４】

また、本実施の形態において、音声の入力を受け付けた後または停止ボタン操作後に、スコアリング処理を実行するかどうかをユーザに問い合わせ、スコアリング処理を行うとの指示を受け付けた場合のみ、図１０に示すような音素スコア、単語スコア、総合スコアを出力するようにしても良い。

【００８５】

また、本実施の形態において、教師データは、比較される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（ＨＭＭ）に基づくデータであるとして、主として説明したが、必ずしもＨＭＭに基づくデータである必要はない。教師データは、単一ガウス分布モデルや、確率モデル（ＧＭＭ：ガウシャ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。かかることは、他の実施の形態においても同様である。

【００８６】

また、本実施の形態の具体例において、学習者は、母音「う」を発音し、音声処理装置は、かかる音声から第二サンプリング周波数を得た。しかし、学習者は、例えば、母音「あいえお」等、1以上の母音を発音し、かかる母音の音声から、音声処理装置は、第二サンプリング周波数を得ても良い。つまり、第二サンプリング周波数を得るために、学習者が発音する音は「う」に限られない。

【００８７】

また、本実施の形態において、音声処理装置が行う下記の処理を、一のＤＳＰ（デジタルシグナルプロセッサ）で行っても良い。つまり、本ＤＳＰは、第一サンプリング周波数を格納している第一サンプリング周波数格納部と、前記第一サンプリング周波数で、受け付けた音声をサンプリングし、第一音声データを取得するサンプリング部と、前記教師データのフォルマント周波数である教師データフォルマント周波数を格納している教師データフォルマント周波数格納部と、前記音声の話者である評価対象者のフォルマント周波数である評価対象者フォルマント周波数を格納している評価対象者フォルマント周波数格納部と、第二サンプリング周波数「前記第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理部を具備するデジタルシグナルプロセッサ、である。かかることは、他の実施の形態でも同様である。

【００８８】

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音声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、第一サンプリング周波数で、受け付けた音声をサンプリングし、第一音声データを取得するサンプリングステップと、第二サンプリング周波数「第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付ステップで受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理ステップと、前記第二音声データを処理する音声処理ステップを実行させるためのプログラム、である。

【００８９】

また、上記プログラムにおいて、音声処理ステップは、前記第二音声データを、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、前記教師データと前記1以上のフレーム音声データに基づいて、前記受け付けた音声の評定を行う評定ステップと、前記評定ステップにおける評定結果を出力する出力ステップを具備する、ことは好適である。

【００９０】

さらに、上記プログラムにおいて、前記評定ステップは、前記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における確率値を取得する最適状態確率値取得ステップと、前記最適状態確率値取得ステップで取得した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を具備することは好適である。

（実施の形態２）

【００９１】

本実施の形態における音声処理装置は、実施の形態１の音声処理装置と比較して、評定部における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。本実施の形態における音声処理装置が算出する事後確率を、実施の形態１におけるＤＡＰに対してｔ-ｐ－ＤＡＰと呼ぶ。

【００９２】

図１１は、本実施の形態における音声処理装置のブロック図である。本音声処理装置は、入力受付部１０１、教師データ格納部１０２、音声受付部１０３、教師データフォルマント周波数格納部１０４、第一サンプリング周波数格納部１０５、サンプリング部１０６、評価対象者フォルマント周波数取得部１０７、評価対象者フォルマント周波数格納部１０８、声道長正規化処理部１０９、音声処理部１１１０、発声催促部１１０９を具備する。

音声処理部１１１０は、フレーム区分手段１１０１、フレーム音声データ取得手段１１０２、評定手段１１１０３、出力手段１１０４を具備する。

評定手段１１１０３は、最適状態決定手段１１０３１、発音区間フレーム音韻確率値取得手段１１１０３２、評定値算出手段１１１０３３を具備する。

発音区間フレーム音韻確率値取得手段１１１０３２は、最適状態決定手段１１０３１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。

【００９３】

評定値算出手段１１１０３３は、発音区間フレーム音韻確率値取得手段１１１０３２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１１１０３３は、例えば、最適状態決定手段１１０３１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。

【００９４】

発音区間フレーム音韻確率値取得手段１１１０３２、および評定値算出手段１１１０３３は、通常、ＭＰＵ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１１１０３２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。

【００９５】

発声催促部１１０９は、入力受付部１０１が、動作開始指示を受け付けた場合、第二サンプリング周波数を算出するために、評価対象者に発声を促す処理を行ったり、評価対象者の発音評定のために発声を促す処理を行ったりする。評価対象者に発声を促す処理は、

例えば、「～を発音してください。」とディスプレイに表示したり、「～を発音してください。」とスピーカーから音出力したりする処理である。発声催促部１１０９は、ディスプレイ３４４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。発声催促部１１０９は、出力デバイスのドライバーソフトまたは、出力デバイスのドライバーソフトと出力デバイス等で実現され得る。

次に、本音声処理装置の動作について図１２から図１４のフローチャートを用いて説明する。図１２等のフローチャートにおいて、図２、図３のフローチャートと異なるステップについてのみ説明する。

（ステップＳ１２０１）発声催促部１１０９は、第二サンプリング周波数算出用の発声を促すために、例えば、母音「う」と発声してください、とディスプレイに表示する。

（ステップＳ１２０２）音声受付部１０３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Ｓ１２０３に行き、音声を受け付けなければステップＳ２１３に行く。

【００９６】

（ステップＳ１２０３）サンプリング部１０６は、第一サンプリング周波数格納部１０５に格納されている第一サンプリング周波数を読み込み、当該第一サンプリング周波数で、ステップＳ１２０２で受け付けた音声をサンプリングし、第一音声データを得る。

【００９７】

（ステップＳ１２０４）声道長正規化処理部１０９は、ステップＳ１２０３で得た第一音声データから、第二サンプリング周波数を得る。かかる第二サンプリング周波数算出処理は、図１３のフローチャートを用いて説明する。

（ステップＳ１２０５）発声催促部１１０９は、評定用の発声を促すために、例えば、「ｒｉｇｈｔ」と発声してください、とディスプレイに表示する。

（ステップＳ１２０６）音声受付部１０３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Ｓ１２０７に行き、音声を受け付けなければステップＳ２１３に行く。

【００９８】

（ステップＳ１２０７）サンプリング部１０６は、第一サンプリング周波数格納部１０５に格納されている第一サンプリング周波数を読み込み、当該第一サンプリング周波数で、ステップＳ１２０６で受け付けた音声をサンプリングし、第一音声データを得る。

（ステップＳ１２０８）声道長正規化処理部１０９は、ステップＳ１２０７で得た第一音声データに対して、ステップＳ１２０４で得た第二サンプリング周波数で、リサンプリングし、第二音声データを得る。

（ステップＳ１２０９）音声処理部１１１０は、ステップＳ１２０８で得た第二音声データに対して、評定処理を行う。評定処理の詳細は、図１４のフローチャートを用いて説明する。ステップＳ１２０２に戻る。

なお、図１２のフローチャートにおいて、第二サンプリング周波数を算出するための音声と、評定するための音声が同一または包含されていても良い。

【００９９】

ステップＳ１２０４の第二サンプリング周波数算出処理について、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のフローチャートにおけるステップＳ３０１からステップＳ３０５の処理を行う。

ステップＳ１２０９の評定処理について、図１４のフローチャートを用いて説明する。

【０１００】

（ステップＳ１４０１）発音区間フレーム音韻確率値取得手段１１１０３２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、発音区間フレーム音韻確率値取得手段１１１０３２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。
（ステップＳ１４０２）発音区間フレーム音韻確率値取得手段１１１０３２は、全フレームの最適状態確率値を算出する。
（ステップＳ１４０３）発音区間フレーム音韻確率値取得手段１１１０３２は、ｊに１を代入する。
【０１０１】
（ステップＳ１４０４）発音区間フレーム音韻確率値取得手段１１１０３２は、次の評定対象の発音区間である、ｊ番目の発音区間が存在するか否かを判断する。ｊ番目の発音区間が存在すればステップＳ１４０３に行き、ｊ番目の発音区間が存在しなければ上位関数にリターンする。
（ステップＳ１４０５）発音区間フレーム音韻確率値取得手段１１１０３２は、カウンタｋに１を代入する。
【０１０２】
（ステップＳ１４０６）発音区間フレーム音韻確率値取得手段１１１０３２は、ｋ番目のフレームが、ｊ番目の発音区間に存在するか否かを判断する。ｋ番目のフレームが存在すればステップＳ１４０７に行き、ｋ番目のフレームが存在しなければステップＳ１４１０に飛ぶ。
（ステップＳ１４０７）発音区間フレーム音韻確率値取得手段１１１０３２は、ｋ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。
（ステップＳ１４０８）評定値算出手段１１１０３３は、ステップＳ１４０７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する。
（ステップＳ１４０９）発音区間フレーム音韻確率値取得手段１１１０３２は、ｋを１、インクメントする。ステップＳ１４０６に戻る。
【０１０３】
（ステップＳ１４１０）評定値算出手段１１１０３３は、ｊ番目の発音区間の評定値を算出する。評定値算出手段１１１０３３は、例えば、最適状態決定手段１１０３１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。
（ステップＳ１４１１）出力手段１１０４は、ステップＳ１４１０で算出した評定値を出力する。
（ステップＳ１４１２）発音区間フレーム音韻確率値取得手段１１１０３２は、ｊを１、インクメントする。ステップＳ１４０４に戻る。
以下、本実施の形態における音声処理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。
【０１０４】
まず、学習者（評価対象者）が、語学学習の開始の指示である動作開始指示を入力する。そして、音声処理装置は、当該動作開始指示を受け付け、次に、発声催促部１１０９は、例えば、「"う"と発声してください。」と画面出力する。
【０１０５】
なお、ここでも、例えば、学習者は、実施の形態１と同様に子供である。また、教師データを作成するために発声した教師は、ネイティブの大人である、とする。かかることは、他の実施の形態の具体例の記載においても同様である、とする。
そして、評価対象者は、"う"と発声し、音声処理装置は、当該発声から、第二ンプリング周波数「３２．１ＫＨｚ」を得る。かかる処理は、実施の形態１において説明した処理と同様である。
【０１０６】
次に、発声催促部１１０９は、例えば、「"ｒｉｇｈｔ"と発声してください。」と画面出力する。そして、学習者は、学習対象の音声「ｒｉｇｈｔ」を発音する。そして、音声

受付部１０３は、学習者が発音した音声の入力を受け付ける。
次に、サンプリング部１０６は、受け付けた音声「ｒｉｇｈｔ」をサンプリング周波数「２２．０５ＫＨｚ」でサンプリング処理する。そして、サンプリング部１０６は、「ｒｉｇｈｔ」の第一音声データを得る。

【０１０７】

次に、声道長正規化処理部１０９は、「ｒｉｇｈｔ」の第一音声データを第二サンプリング周波数「３２．１ＫＨｚ」でリサンプリング処理する。そして、声道長正規化処理部１０９は、第二音声データを得る。次に、音声処理部１１１０は、第二音声データを、以下のように処理する。
まず、フレーム区分手段１１０１は、「ｒｉｇｈｔ」の第二音声データを、短時間フレームに区分する。
そして、フレーム音声データ取得手段１１０２は、フレーム区分手段１１０１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「$O = o_1, o_2, \cdots, o_T$」を算出する。
次に、発音区間フレーム音韻確率値取得手段１１１０３２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。

【０１０８】

次に、最適状態決定手段１１０３１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１１０３１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）を算出する処理と、最適状態を決定する処理の処理順序は問わない。

【０１０９】

次に、発音区間フレーム音韻確率値取得手段１１１０３２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。そして、評定値算出手段１１１０３３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。そして、評定値算出手段１１１０３３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１１１０３３は、数式３により評定値を算出する。数式３において、ｐ－ＤＡＰ（τ）は、各フレームにおける、すべての音韻の中で最適な音韻の事後確率（確率値）を表すように算出される評定値であり、数式４で算出され得る。なお、数式３のｔ－ｐ－ＤＡＰ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、算出される評定値である。また、数式３において、Ｔ（q_t^*）は、状態q_t^*を含むＨＭＭが含まれる評定対象の発音区間である。｜Ｔ（q_t^*）｜は、Ｔ（q_t^*）の区間長である。また、数式４において、Ｐ（q_t^*）は、状態q_t^*を含むＨＭＭが有する全状態識別子の集合である。

【数３】

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

【数４】

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

【０１１０】

かかる評定値算出手段１１１０３３が算出した評定値（「ｔ－ｐ－ＤＡＰスコア」とも

言う。）を、図１５の表に示す。図１５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、ＤＡＰの代わりにｐ－ＤＡＰの時間平均を採用したものである。図１５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。

そして、出力手段１１０４は、算出した発音区間ごと（ここでは、音素毎）の評定値を、順次出力する。かかる出力例は、図１６である。

【０１１１】

以上、本実施の形態によれば、ユーザが入力した発音を、教師データに対して、如何に似ているかを示す類似度（評定値）を算出し、出力できる。また、かかる場合、本実施の形態によれば、個人差、特に声道長の違いに影響を受けない、精度の高い評定ができる。

【０１１２】

また、本実施の形態によれば、連結されたＨＭＭである連結ＨＭＭを用いて最適状態を求め、評定値を算出するので、高速に評定値を求めることができる。したがって、上記の具体例で述べたように、リアルタイムに、発音区間ごとの評定値を出力できる。また、本実施の形態によれば、動的計画法に基づいた事後確率を確率値として算出するので、さらに高速に評定値を求めることができる。

【０１１３】

また、本実施の形態によれば、評定値を、発音区間の単位で算出でき、実施の形態１におけるような状態単位のＤＡＰと比較して、本来、測定したい類似度（発音区間の類似度）を精度良く、安定して求めることができる。

【０１１４】

さらに、本実施の形態における音声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、第一サンプリング周波数で、受け付けた音声をサンプリングし、第一音声データを取得するサンプリングステップと、第二サンプリング周波数「第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付ステップで受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理ステップと、前記第二音声データを処理する音声処理ステップを実行させるためのプログラム、である。

【０１１５】

また、上記プログラムにおいて、音声処理ステップは、前記第二音声データを、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、前記教師データと前記1以上のフレーム音声データに基づいて、前記受け付けた音声の評定を行う評定ステップと、前記評定ステップにおける評定結果を出力する出力ステップを具備する、ことは好適である。

【０１１６】

また、上記プログラムにおいて、前記評定ステップは、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定値を算出する評定値算出ステップを具備する、ことは好適である。

（実施の形態３）

【０１１７】

本実施の形態において、比較対象の音声と入力音声の類似度を精度高く評定できる音声処理装置について説明する。特に、本音声処理装置は、無音区間を検知し、無音区間を考慮した類似度評定が可能な音声処理装置である。

【０１１８】

図１７は、本実施の形態における音声処理装置のブロック図である。本音声処理装置は、入力受付部１０１、教師データ格納部１０２、音声受付部１０３、教師データフォルマント周波数格納部１０４、第一サンプリング周波数格納部１０５、サンプリング部１０６、評価対象者フォルマント周波数取得部１０７、評価対象者フォルマント周波数格納部１０８、声道長正規化処理部１０９、音声処理部１７１０、発声催促部１１０９を具備する。

音声処理部１７１０は、フレーム区分手段１１０１、フレーム音声データ取得手段１１０２、特殊音声検知手段１７１０１、評定手段１７１０３、出力手段１１０４を具備する。

特殊音声検知手段１７１０１は、無音データ格納手段１７１０１１、無音区間検出手段１７１０１２を具備する。

評定手段１７１０３は、最適状態決定手段１１０３１、最適状態確率値取得手段１１０３２、評定値算出手段１７１０３３を具備する。

【０１１９】

特殊音声検知手段１７１０１は、フレーム毎の入力音声データに基づいて、特殊な音声が入力されたことを検知する。なお、ここで特殊な音声は、無音も含む。また、特殊音声検知手段１７１０１は、例えば、フレームの最適状態の確率値を、ある音素区間において取得し、ある音素区間の１以上の確率値の総和が所定の値より低い場合（想定されている音素ではない、と判断できる場合）、当該音素区間において特殊な音声が入力されたと、検知する。かかる検知の具体的なアルゴリズムの例は後述する。特殊音声検知手段１７１０１は、通常、ＭＰＵやメモリ等から実現され得る。特殊音声検知手段１７１０１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。

【０１２０】

無音データ格納手段１７１０１１は、無音を示すデータであり、ＨＭＭに基づく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１７１０１１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。

【０１２１】

無音区間検出手段１７１０１２は、フレーム音声データ取得手段１１０２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１７１０１１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１７１０１２は、フレーム音声データ取得手段１１０２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１７１０１２は、下記で述べる最適状態確率値取得手段１１０３２が取得した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１１０２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。無音区間検出手段１７１０１２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１７１０１２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。

【０１２２】

評定手段１７１０３は、教師データと入力音声データと特殊音声検知手段１７１０１における検知結果に基づいて、音声受付部１０３が受け付けた音声の評定を行う。「特殊音声検知手段１７１０１における検知結果に基づく」とは、例えば、特殊音声検知手段１７１０１が無音を検知した場合、当該無音の区間を無視することである。また、「特殊音声検知手段１７１０１における検知結果に基づく」とは、例えば、特殊音声検知手段１７１０１が置換や脱落などを検知した場合、当該置換や脱落などの検知により、評定値を所定数値分、減じて、評定値を算出することである。評定手段１７１０３は、通常、ＭＰＵやメモリ等から実現され得る。評定手段１７１０３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。

【０１２３】

評定値算出手段１７１０３３は、無音区間検出手段１７１０１２が検出した無音区間を除いて、かつ最適状態確率値取得手段１１０３２が取得した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１７１０３３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。評定値算出手段１７１０３３は、例えば、最適状態確率値取得手段１１０３２が取得した確率値と、当該確率値に対応するフレームの全状態における確率値の総和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２１０２３は、ここでは、通常、無音区間検出手段１７１０１２が検出した無音区間を除いて、フレームごとに評定値を算出する。なお、評定値算出手段１７１０３３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１７１０３３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。評定値算出手段１７１０３３は、通常、ＭＰＵやメモリ等から実現され得る。評定値算出手段１７１０３３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。

【０１２４】

次に、音声処理装置の動作について図１８、図１９のフローチャートを用いて説明する。なお、図１８のフローチャートは、図１２のフローチャートと比較して、ステップＳ１８０１の評定処理のみが異なるので、図１８のフローチャートの説明は省略する。ステップＳ１８０１の評定処理の詳細について、図１９のフローチャートを用いて説明する。

【０１２５】

（ステップＳ１９０１）評定手段１７１０３は、ＤＡＰの評定値を算出する。ＤＡＰの評定値を算出するアルゴリズムは、実施の形態１で説明済みであるので、詳細な説明は省略する。ＤＡＰの評定値を算出する処理は、図２のフローチャートの、ステップＳ２１１からＳ２１４の処理により行う。

【０１２６】

（ステップＳ１９０２）特殊音声検知手段１７１０１は、ステップＳ１９０１で算出した値が、所定の値より低いか否かを判断する。所定の値より低ければステップＳ１９０３に行き、所定の値より低くなければステップＳ１９０６に飛ぶ。

（ステップＳ１９０３）無音区間検出手段１７１０１２は、無音データと全教師データの確率値を取得する。

【０１２７】

（ステップＳ１９０４）無音区間検出手段１７１０１２は、ステップＳ１９０３で取得した確率値の中で、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（かかる場合、無音の区間であると判断する）ステップＳ１９０５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Ｓ１９０６に行く。

（ステップＳ１９０５）無音区間検出手段１７１０１２は、カウンタｉを１、インクリメントする。ステップＳ２０８に戻る。

（ステップＳ１９０６）出力手段１１０４は、ステップＳ１９０１で算出した評定値を出力する。

【０１２８】

なお、図１９のフローチャートにおいて、出力手段１１０４は、無音区間と判定した区間の評定値は出力しなかった（無音区間を無視した）が、特殊音声が検知された区間が無音区間である旨を明示したり、無音区間が存在する旨を明示したりする態様で出力しても良い。また、評定値算出手段１７１０３３は、発音区間や、それ以上の単位のスコアを算出する場合に、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、１／１０にして、発音区間や発音全体のスコアを算出するなどしても良い。評定手段１７１０３は、教師データと入力音声データと特殊音声検知手段１７１０１における検知結果に基づいて、音声受付部１０３が受け付けた音声の評定を行えばよい。

【０１２９】
また、図１９のフローチャートにおいて、特殊音声検知手段１７１０１は、ｉ番目のフレーム音声データのＤＡＰスコアに基づいて特殊音声を検知したが、例えば、実施の形態２で説明したｔ－ｐ－ＤＡＰスコアに基づいて特殊音声を検知しても良い。
【０１３０】
以下、本実施の形態における音声処理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、実施の形態２とは異なる。そこで、その異なる処理を中心に説明する。
まず、学習者（評価対象者）が、語学学習の開始の指示である動作開始指示を入力する。そして、音声処理装置は、当該動作開始指示を受け付け、次に、例えば、「"う"と発声してください。」と画面出力する。
そして、評価対象者は、"う"と発声し、音声処理装置は、当該発声から、第二ンプリング周波数「３２．１ＫＨｚ」を得る。かかる処理は、実施の形態１等において説明した処理と同様である。
【０１３１】
次に、発声催促部１１０９は、例えば、「"ｒｉｇｈｔ"と発声してください。」と画面出力する。そして、学習者は、学習対象の音声「ｒｉｇｈｔ」を発音する。そして、音声受付部１０３は、学習者が発音した音声の入力を受け付ける。
次に、サンプリング部１０６は、受け付けた音声「ｒｉｇｈｔ」をサンプリング周波数「２２．０５ＫＨｚ」でサンプリング処理する。そして、サンプリング部１０６は、「ｒｉｇｈｔ」の第一音声データを得る。
【０１３２】
次に、声道長正規化処理部１０９は、「ｒｉｇｈｔ」の第一音声データを、第二サンプリング周波数「３２．１ＫＨｚ」でリサンプリング処理する。そして、声道長正規化処理部１０９は、第二音声データを得る。次に、音声処理部１７１０は、第二音声データを、以下のように処理する。
まず、フレーム区分手段１１０１は、「ｒｉｇｈｔ」の第二音声データを、短時間フレームに区分する。
そして、フレーム音声データ取得手段１１０２は、フレーム区分手段１１０１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「$O=o_1, o_2, \cdots, o_T$」を算出する。
次に、最適状態決定手段１１０３１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。
次に、最適状態確率値取得手段１１０３２は、上述した数式１、２により、最適状態における確率値を算出する。
【０１３３】
次に、評定値算出手段１７１０３３は、例えば、最適状態決定手段１１０３１が決定した最適状態を有する音韻全体の状態における１以上の確率値を取得し、当該１以上の確率値の総和をパラメータとして音声の評定値を算出する。つまり、評定値算出手段１７１０３３は、例えば、ＤＡＰスコアをフレーム毎に算出する。
【０１３４】
そして、特殊音声検知手段１７１０１は、算出されたフレームに対応する評定値（ＤＡＰスコア）を用いて、特殊な音声が入力されたか否かを判断する。具体的には、特殊音声検知手段１７１０１は、例えば、評価対象のフレームに対して算出された評定値が、所定の数値より低ければ、特殊な音声が入力された、と判断する。なお、特殊音声検知手段１７１０１は、一のフレームに対応する評定値が小さいからといって、直ちに特殊な音声が入力された、と判断する必要はない。つまり、特殊音声検知手段１７１０１は、フレームに対応する評定値が小さいフレームが所定の数以上、連続する場合に、当該連続するフレーム群に対応する区間が特殊な音声が入力された区間と判断しても良い。

【０１３５】
特殊音声検知手段１７１０１が、特殊音声を検知する場合について説明する図を図２０に示す。図２０（ａ）の縦軸は、ＤＡＰスコアであり、横軸はフレームを示す。図２０（ａ）において、（Ｖ）は、Ｖｉｔｅｒｂｉアライメントを示す。図２０（ａ）において、網掛けのフレーム群のおけるＤＡＰスコアは、所定の値より低く、特殊音声の区間である、と判断される。

【０１３６】
次に、特殊な音声が入力された、と判断した場合、無音区間検出手段１７１０１２は、無音データ格納手段１７１０１１から無音データを取得し、当該フレーム群と無音データとの類似度を算定し、類似度が所定値以上であれば当該フレーム群に対応する音声データが、無音データであると判断する。図２０（ｂ）は、無音データとの比較の結果、当該無音データとの類似度を示す事後確率の値（「ＤＡＰスコア」）が高いことを示す。その結果、無音区間検出手段１７１０１２は、当該特殊音声の区間は、無音区間である、と判断する。なお、図２０（ａ）において、網掛けのフレーム群のおけるＤＡＰスコアは、所定の値より低く、特殊音声の区間である、と判断され、かつ、無音データとの比較の結果、ＤＡＰスコアが低い場合には、無音区間ではない、と判断される。そして、かかる区間において、例えば、単に、発音が上手くなく、低い評定値が出力される。なお、図２０（ａ）に示しているように、通常、無音区間は、第一のワード（「ｗｏｒｄ１」）の最終音素の後半部、および第一のワードに続く第二のワード（「ｗｏｒｄ２」）の第一音素の前半部のスコアが低い。
そして、出力手段１１０４は、出力する評定値から、無音データの区間の評定値を考慮しないように、無視する。
そして、出力手段１１０４は、各フレームに対応する評定値を出力する。この場合、例えば、無音データの区間の評定値は、出力されない。
かかる評定値の出力態様例は、例えば、図９、図１０である。
なお、出力手段１１０４が行う出力は、無音区間の存在を示すだけの出力でも良い。

【０１３７】
以上、本実施の形態によれば、ユーザが入力した発音を、教師データに対して、如何に似ているかを示す類似度（評定値）を算出し、出力できる。また、かかる場合、本実施の形態によれば、個人差、特に声道長の違いに影響を受けない、精度の高い評定ができる。さらに、本音声処理装置は、無音区間を考慮して類似度を評定するので、極めて正確な評定結果が得られる。

【０１３８】
なお、無音区間のデータは、無視して評定結果を算出することは好適である。ただし、本実施の形態において、例えば、無音区間の評価の影響を他の区間と比較して少なくするなど、無視する以外の方法で、無音区間のデータを考慮して、評定値を出力しても良い。

【０１３９】
また、本実施の形態の具体例によれば、ＤＡＰスコアを用いて、評定値を算出したが、無音の区間を考慮して評定値を算出すれば良く、上述した他のアルゴリズム（ｔ－ｐ－ＤＡＰ等）、または、本明細書では述べていない他のアルゴリズムにより評定値を算出しても良い。つまり、本実施の形態によれば、教師データと入力音声データと特殊音声検知手段における検知結果に基づいて、音声受付部が受け付けた音声の評定を行い、特に、無音データを考慮して、評定値を算出すれば良い。
また、本実施の形態によれば、まず、ＤＡＰスコアが低い区間を検出してから、無音区間の検出をした。しかし、ＤＡＰスコアが低い区間を検出せずに、無音データとの比較により、無音区間を検出しても良い。

【０１４０】
さらに、本実施の形態における音声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、第一サンプリング周波数で、受け付けた音声をサンプリングし、第一音声データを取得するサンプリングステッ

プと、第二サンプリング周波数「第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付ステップで受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理ステップと、前記第二音声データを処理する音声処理ステップを実行させるためのプログラム、である。

【０１４１】

また、上記プログラムにおいて、音声処理ステップは、前記第二音声データを、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、前記フレーム毎の入力音声データに基づいて、特殊な音声が入力されたことを検知する特殊音声検知ステップと、教師データと前記入力音声データと前記特殊音声検知ステップにおける検知結果に基づいて、前記受け付けた音声の評定を行う評定ステップと、前記評定ステップにおける評定結果を出力する出力ステップを具備する、ことは好適である。

また、上記プログラムにおいて、特殊音声検知ステップは、無音を示すＨＭＭに基づくデータである無音データと、前記入力音声データに基づいて、無音の区間を検出する、ことは好適である。

（実施の形態４）

【０１４２】

本実施の形態において、入力音声において、特殊音声を検知し、比較対象の音声と入力音声の類似度を精度高く評定できる音声処理装置について説明する。特に、本音声処理装置は、音韻の挿入を検知できる音声処理装置である。

【０１４３】

図２１は、本実施の形態における音声処理装置のブロック図である。本音声処理装置は、入力受付部１０１、教師データ格納部１０２、音声受付部１０３、教師データフォルマント周波数格納部１０４、第一サンプリング周波数格納部１０５、サンプリング部１０６、評価対象者フォルマント周波数取得部１０７、評価対象者フォルマント周波数格納部１０８、声道長正規化処理部１０９、音声処理部２１１０、発声催促部１１０９を具備する。

【０１４４】

音声処理部２１１０は、フレーム区分手段１１０１、フレーム音声データ取得手段１１０２、特殊音声検知手段２１１０１、評定手段２１１０３、出力手段２１１０４を具備する。なお、評定手段２１１０３は、最適状態決定手段１１０３１、最適状態確率値取得手段１１０３２を具備する。

【０１４５】

特殊音声検知手段２１１０１は、一の音素の後半部および当該音素の次の音素の前半部の評定値が所定の条件を満たすことを検知する。後半部、および前半部の長さは問わない。特殊音声検知手段２１１０１は、通常、ＭＰＵやメモリ等から実現され得る。特殊音声検知手段２１１０１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。

【０１４６】

評定手段２１１０３は、特殊音声検知手段２１１０１が所定の条件を満たすことを検知した場合に、少なくとも音素の挿入があった旨を示す評定結果を構成する。なお、評定手段２１１０３は、実施の形態３で述べたアルゴリズムにより、特殊音声検知手段２１１０１が所定の条件を満たすことを検知した区間に無音が挿入されたか否かを判断し、無音が挿入されていない場合に、他の音素が挿入されたと検知しても良い。また、評定手段２１１０３は、無音が挿入されていない場合に、他の音韻ＨＭＭに対する確率値を算出し、所定の値より高い確率値を得た音韻が挿入された、との評定結果を得ても良い。なお、実施の形態３で述べた無音区間の検知は、無音音素の挿入の検知である、とも言える。評定手段２１１０３は、通常、ＭＰＵやメモリ等から実現され得る。評定手段２１１０３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。

【０１４７】

出力手段２１１０４は、評定手段２１１０３における評定結果を出力する。ここでの評定結果は、音素の挿入があった旨を示す評定結果を含む。また、評定結果は、音素の挿入があった場合に、所定数値分、減じられて算出された評定値（スコア）のみでも良い。また、評定結果は、音素の挿入があった旨、および評定値（スコア）の両方であっても良い。なお、教師データにおいて想定されていない音素の挿入を検知した場合、通常、評定値は低くなる。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。出力手段２１１０４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。出力手段２１１０４は、出力デバイスのドライバーソフトまたは、出力デバイスのドライバーソフトと出力デバイス等で実現され得る。

【０１４８】

次に、音声処理装置の動作について、図２２、図２３のフローチャートを用いて説明する。なお、図２２のフローチャートは、図１２のフローチャートと比較して、ステップＳ２２０１の評定処理のみが異なるので、図２２のフローチャートの説明は省略する。ステップS２２０１の評定処理の詳細について、図２３のフローチャートを用いて説明する。図２３のフローチャートにおいて、図２、図１９のフローチャートの処理と同様の処理については、その説明を省略する。

【０１４９】

（ステップＳ２３０１）特殊音声検知手段２１１０１は、フレームに対応するデータを一時的に蓄積するバッファにデータが格納されているか否かを判断する。なお、格納されているデータは、ステップＳ１９０２で、所定の値より低い評定値と評価されたフレーム音声データ、または当該フレーム音声データから取得できるデータである。データが格納されていればステップＳ２３０７に行き、データが格納されていなければ上位関数にリターンする。

【０１５０】

（ステップＳ２３０２）特殊音声検知手段２１１０１は、バッファにデータが格納されているか否かを判断する。データが格納されていればステップＳ２３０７に行き、データが格納されていなければステップステップＳ２３０３に行く。

（ステップＳ２３０３）出力手段２１１０４は、ステップＳ１９０１で算出した評定値を出力する。

（ステップＳ２３０４）特殊音声検知手段２１１０１は、カウンタｉを１、インクリメントする。ステップＳ２０８に戻る。

（ステップＳ２３０５）特殊音声検知手段２１１０１は、バッファに、所定の値より低い評定値と評価されたフレーム音声データ、または当該フレーム音声データから取得できるデータを一時蓄積する。

（ステップＳ２３０６）特殊音声検知手段２１１０１は、カウンタｉを１、インクリメントする。ステップＳ２０８に戻る。

（ステップＳ２３０７）特殊音声検知手段２１１０１は、カウンタｊに１を代入する。

【０１５１】

（ステップＳ２３０８）特殊音声検知手段２１１０１は、ｊ番目のデータが、バッファに存在するか否かを判断する。ｊ番目のデータが存在すればステップＳ２３０９に行き、ｊ番目のデータが存在しなければステップＳ２３１５に飛ぶ。

（ステップＳ２３０９）特殊音声検知手段２１１０１は、ｊ番目のデータに対応する最適状態の音素を取得する。

（ステップＳ２３１０）特殊音声検知手段２１１０１は、ｊ番目のデータに対する全教師データの確率値を算出し、最大の確率値を持つ音素を取得する。

【０１５２】

（ステップＳ２３１１）特殊音声検知手段２１１０１は、ステップＳ２３０９で取得した音素とステップＳ２３１０で取得した音素が異なる音素であるか否かを判断する。異なる音素であればステップＳ２３１２に行き、異なる音素でなければステップＳ２３１４に飛ぶ。

（ステップＳ２３１２）評定手段２１１０３は、音素の挿入があった旨を示す評定結果を構成する。

（ステップＳ２３１３）特殊音声検知手段２１１０１は、カウンタｊを１、インクリメントする。ステップＳ２３０８に戻る。

（ステップＳ２３１４）出力手段２１１０４は、バッファ中の全データに対応する全評定値を出力する。ここで、全評定値とは、例えば、フレーム毎のＤＡＰスコアである。ステップＳ２３１３に行く。

【０１５３】

（ステップＳ２３１５）出力手段２１１０４は、評定結果に「挿入の旨」の情報が入っているか否かを判断する。「挿入の旨」の情報が入っていればステップＳ２３１６に行き、「挿入の旨」の情報が入っていなければステップＳ２３１７に行く。

（ステップＳ２３１６）出力手段２１１０４は、評定結果を出力する。

（ステップＳ２３１７）出力手段２１１０４は、バッファをクリアする。ステップＳ２０８に戻る。

【０１５４】

なお、図２３のフローチャートにおいて、評定値の低いフレームが２つの音素に渡って存在すれば、音素の挿入があったと判断した。つまり、一の音素の後半部（少なくとも最終フレーム）および当該音素の次の音素の第一フレームの評定値が所定値より低い場合に、音素の挿入があったと判断した。しかし、図２３のフローチャートにおいて、一の音素の所定区間以上の後半部、および当該音素の次の音素の所定区間以上の前半部の評定値が所定値よりすべて低い場合に、音素の挿入があったと判断するようにしても良い。

以下、本実施の形態における音声処理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、音素の挿入の検知を行う処理が実施の形態３等とは異なる。そこで、その異なる処理を中心に説明する。

まず、学習者（評価対象者）が、語学学習の開始の指示である動作開始指示を入力する。そして、音声処理装置は、当該動作開始指示を受け付け、次に、例えば、「"あ"と発声してください。」と画面出力する。

そして、学習者は、"あ"と発声し、音声処理装置は、当該発声から、第二ンプリング周波数「３２．1ＫＨｚ」を得る。かかる処理は、実施の形態１等において説明した処理と同様である。

【０１５５】

次に、発声催促部１１０９は、例えば、「"ｒｉｇｈｔ"と発声してください。」と画面出力する。そして、学習者は、学習対象の音声「ｒｉｇｈｔ」を発音する。そして、音声受付部１０３は、学習者が発音した音声の入力を受け付ける。

次に、サンプリング部１０６は、受け付けた音声「ｒｉｇｈｔ」をサンプリング周波数「２２．０５ＫＨｚ」でサンプリング処理する。そして、サンプリング部１０６は、「ｒｉｇｈｔ」の第一音声データを得る。

【０１５６】

次に、声道長正規化処理部１０９は、「ｒｉｇｈｔ」の第一音声データを第二サンプリング周波数「３２．１ＫＨｚ」でリサンプリング処理する。そして、声道長正規化処理部１０９は、第二音声データを得る。次に、音声処理部２１１０は、第二音声データを、以下のように処理する。

まず、フレーム区分手段１１０１は、「ｒｉｇｈｔ」の第二音声データを、短時間フレームに区分する。

そして、フレーム音声データ取得手段１１０２は、フレーム区分手段１１０１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，・・・，o_T

」を算出する。
次に、評定手段２１１０３の最適状態決定手段１１０３１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。
次に、最適状態確率値取得手段１１０３２は、上述した数式１、２により、最適状態における確率値を算出する。
【０１５７】
次に、評定手段２１１０３は、例えば、最適状態決定手段１１０３１が決定した最適状態を有する音韻全体の状態における１以上の確率値を取得し、当該１以上の確率値の総和をパラメータとして音声の評定値を算出する。つまり、評定手段２１１０３は、例えば、ＤＡＰスコアをフレーム毎に算出する。ここで、算出するスコアは、上述したｔ－ｐ－ＤＡＰスコア等でも良い。
【０１５８】
そして、特殊音声検知手段２１１０１は、算出されたフレームに対応する評定値を用いて、特殊な音声が入力されたか否かを判断する。つまり、評定値(例えば、ＤＡＰスコア)が、所定の値より低い区間が存在するか否かを判断する。
【０１５９】
次に、特殊音声検知手段２１１０１は、図２４に示すように、評定値(例えば、ＤＡＰスコア)が、所定の値より低い区間が、２つの音素に跨っているか否かを判断し、２つの音素に跨がっていれば、当該区間に音素が挿入された、と判断する。なお、かかる場合の詳細なアルゴリズムの例は、図２３で説明した。また、図２４において、斜線部が、予期しない音素が挿入された区間である。
【０１６０】
次に、評定手段２１１０３は、音素の挿入があった旨を示す評定結果（例えば、「予期しない音素が挿入されました。」）を構成する。なお、予期しない音素が挿入された場合、評定手段２１１０３は、例えば、所定数値分、減じて、評定値を算出することは好適である。そして、出力手段２１１０４は、構成した評定結果（評定値を含んでも良い）を出力する。図２５は、評定結果の出力例である。なお、出力手段２１１０４は、通常の入力音声に対しては、上述したように評定値を出力することが好適である。
【０１６１】
以上、本実施の形態によれば、ユーザが入力した発音を、教師データに対して、如何に似ているかを示す類似度（評定値）を算出し、出力できる。また、かかる場合、本実施の形態によれば、個人差、特に声道長の違いに影響を受けない、精度の高い評定ができる。さらに、本音声処理装置は、特殊音声、特に、予期せぬ音素の挿入を検知できるので、極めて精度の高い評定結果が得られる。
【０１６２】
なお、本実施の形態において、音素の挿入を検知できれば良く、評定値の算出アルゴリズムは問わない。評定値の算出アルゴリズムは、上述したアルゴリズム（ＤＡＰ、ｔ－ｐ－ＤＡＰ）でも良く、または、本明細書では述べていない他のアルゴリズムでも良い。
【０１６３】
さらに、本実施の形態における音声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、第一サンプリング周波数で、受け付けた音声をサンプリングし、第一音声データを取得するサンプリングステップと、第二サンプリング周波数「第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付ステップで受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理ステップと、前記第二音声データを処理する音声処理ステップを実行させるためのプログラム、である。
【０１６４】
また、上記プログラムにおいて、音声処理ステップは、前記第二音声データを、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、前記フレーム毎の入力音声データに基づいて、特殊な音声が入力されたことを検知する特殊音声検知ステップと、教師データと前記入力音声データと前記特殊音声検知ステップにおける検知結果に基づいて、前記受け付けた音声の評定を行う評定ステップと、前記評定ステップにおける評定結果を出力する出力ステップを具備する、ことは好適である。

また、上記プログラムにおいて、特殊音声検知ステップは、一の音素の後半部および当該音素の次の音素の前半部の評定値が所定の条件を満たすことを検知する、ことは好適である。

（実施の形態５）

【０１６５】

本実施の形態において、入力音声において、特殊音声を検知し、比較対象の音声と入力音声の類似度を精度高く評定できる音声処理装置について説明する。特に、本音声処理装置は、音韻の置換を検知できる音声処理装置である。

【０１６６】

図２６は、本実施の形態における音声処理装置のブロック図である。本音声処理装置は、入力受付部１０１、教師データ格納部１０２、音声受付部１０３、教師データフォルマント周波数格納部１０４、第一サンプリング周波数格納部１０５、サンプリング部１０６、評価対象者フォルマント周波数取得部１０７、評価対象者フォルマント周波数格納部１０８、声道長正規化処理部１０９、音声処理部２６１０、発声催促部１１０９を具備する。

【０１６７】

音声処理部２６１０は、フレーム区分手段１１０１、フレーム音声データ取得手段１１０２、特殊音声検知手段２６１０１、評定手段２６１０３、出力手段２１１０４を具備する。なお、評定手段２６１０３は、最適状態決定手段１１０３１、最適状態確率値取得手段１１０３２を具備する。なお、評定手段２６１０３は、最適状態決定手段１１０３１、最適状態確率値取得手段１１０３２を具備する。

【０１６８】

音声処理部２６１０は、第二音声データを処理する。音声処理部２６１０は、フレーム毎の入力音声データに基づいて、特殊な音声が入力されたことを検知する特殊音声検知手段２６１０１を具備する。音声処理部２６１０は、通常、ＭＰＵやメモリ等から実現され得る。音声処理部２６１０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。

【０１６９】

特殊音声検知手段２６１０１は、一の音素の評定値が所定の値より低いことを検知する。また、特殊音声検知手段２６１０１は、一の音素の評定値が所定の値より低く、かつ当該音素の直前の音素および当該音素の直後の音素の評定値が所定の値より高いことをも検知しても良い。また、特殊音声検知手段２６１０１は、一の音素の評定値が所定の値より低く、かつ、想定していない音素のＨＭＭに基づいて算出された評定値が所定の値より高いことを検知しても良い。つまり、特殊音声検知手段２６１０１は、所定のアルゴリズムで、音韻の置換を検知できれば良い。そのアルゴリズムは種々考えられる。特殊音声検知手段２６１０１は、通常、ＭＰＵやメモリ等から実現され得る。特殊音声検知手段２６１０１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。

【０１７０】

評定手段２６１０３は、特殊音声検知手段２６１０１が所定の条件を満たすことを検知した場合に、少なくとも音素の置換があった旨を示す評定結果を構成する。評定手段２６１０３は、音素の置換があった場合に、所定数値分、減じられて算出された評定値（スコア）を算出しても良い。評定手段２６１０３は、通常、ＭＰＵやメモリ等から実現され得

る。評定手段２６１０３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。

【０１７１】

次に、音声処理装置の動作について、図２７、図２８のフローチャートを用いて説明する。なお、図２７のフローチャートは、図１２のフローチャートと比較して、ステップＳ２７０１の評定処理のみが異なるので、図２７のフローチャートの説明は省略する。ステップＳ２７０１の評定処理の詳細について、図２８のフローチャートを用いて説明する。図２８のフローチャートにおいて、図２、図１９、図２３のフローチャートの処理と同様の処理については、その説明を省略する。

【０１７２】

（ステップＳ２８０１）特殊音声検知手段２６１０１は、バッファに蓄積されているデータに対応するフレーム音声データ群が一の音素に対応するか否かを判断する。一の音素であればステップＳ２８０２に行き、一の音素でなければステップＳ２８１０に行く。

【０１７３】

（ステップＳ２８０２）特殊音声検知手段２６１０１は、バッファに蓄積されているデータに対応するフレーム音声データ群の音素の直前の音素の評定値を算出する。かかる評定値は、例えば、上述したＤＡＰスコアである。なお、直前の音素とは、現在評定中の音素に対して直前の音素である。音素の区切りは、Ｖｉｔｅｒｂｉアルゴリズムにより算出できる。

【０１７４】

（ステップＳ２８０３）特殊音声検知手段２６１０１は、ステップＳ２８０２で算出した評定値が所定の値以上であるか否かを判断する。所定の値以上であればステップＳ２８０４に行き、所定の値より小さければステップＳ２８１０に行く。

（ステップＳ２８０４）特殊音声検知手段２６１０１は、直後の音素の評定値を算出する。かかる評定値は、例えば、上述したＤＡＰスコアである。直後の音素とは、現在評定中の音素に対して直後の音素である。

【０１７５】

（ステップＳ２８０５）特殊音声検知手段２６１０１は、ステップＳ２８０４で算出した評定値が所定の値以上であるか否かを判断する。所定の値以上であればステップＳ２８０６に行き、所定の値より小さければステップＳ２８１０に行く。

【０１７６】

（ステップＳ２８０６）特殊音声検知手段２６１０１は、予め格納されている音韻ＨＭＭ（予期する音韻のＨＭＭは除く）の中で、所定の値以上の評定値が得られる音韻ＨＭＭが一つ存在するか否かを判断する。所定の値以上の評定値が得られる音韻ＨＭＭが存在すればステップＳ２８０７に行き、所定の値以上の評定値が得られる音韻ＨＭＭが存在しなければステップＳ２８１０に行く。なお、予め格納されている音韻ＨＭＭは、通常、すべての音韻に対する多数の音韻ＨＭＭである。なお、本ステップにおいて、予め格納されている音韻ＨＭＭの確率値を算出し、最大の確率値を持つ音素を取得し、当該音素と最適状態の音素が異なるか否かを判断し、異なる場合に音素の置換があったと判断しても良い。

（ステップＳ２８０７）評定手段２６１０３は、音素の置換があった旨を示す評定結果を構成する。

（ステップＳ２８０８）出力手段２１１０４は、ステップＳ２８０７で構成した評定結果を出力する。

（ステップＳ２８０９）出力手段２１１０４は、バッファをクリアする。ステップＳ２０８に戻る。

（ステップＳ２８１０）出力手段２１１０４は、バッファ中の全データに対応する全評定値を出力する。ステップＳ２８０９に行く。

以下、本実施の形態における音声処理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、音素の置換の検知を行う処理が実施の形態４等とは異なる。そこで、その

異なる処理を中心に説明する。

【０１７７】

まず、学習者（評価対象者）が、語学学習の開始の指示である動作開始指示を入力する。そして、音声処理装置は、当該動作開始指示を受け付け、次に、例えば、「"う"と発声してください。」と画面出力する。

そして、評価対象者は、"う"と発声し、音声処理装置は、当該発声から、第二ンプリング周波数「３２．１ＫＨｚ」を得る。かかる処理は、実施の形態１等において説明した処理と同様である。

【０１７８】

次に、発声催促部１１０９は、例えば、「"ｒｉｇｈｔ"と発声してください。」と画面出力する。そして、学習者は、学習対象の音声「ｒｉｇｈｔ」を発音する。そして、音声受付部１０３は、学習者が発音した音声の入力を受け付ける。

次に、サンプリング部１０６は、受け付けた音声「ｒｉｇｈｔ」をサンプリング周波数「２２．０５ＫＨｚ」でサンプリング処理する。そして、サンプリング部１０６は、「ｒｉｇｈｔ」の第一音声データを得る。

【０１７９】

次に、声道長正規化処理部１０９は、「ｒｉｇｈｔ」の第一音声データを第二サンプリング周波数「３２．１ＫＨｚ」でリサンプリング処理する。そして、声道長正規化処理部１０９は、第二音声データを得る。次に、音声処理部２６１０は、第二音声データを、以下のように処理する。

まず、フレーム区分手段１１０１は、「ｒｉｇｈｔ」の第二音声データを、短時間フレームに区分する。

そして、フレーム音声データ取得手段１１０２は、フレーム区分手段１１０１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，・・・，o_T」を算出する。

次に、評定手段２６１０３の最適状態決定手段１１０３１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。

次に、最適状態確率値取得手段１１０３２は、上述した数式１、２により、最適状態における確率値を算出する。

【０１８０】

次に、評定手段２６１０３は、例えば、最適状態決定手段１１０３１が決定した最適状態を有する音韻全体の状態における１以上の確率値を取得し、当該１以上の確率値の総和をパラメータとして音声の評定値を算出する。つまり、評定手段２６１０３は、例えば、ＤＡＰスコアをフレーム毎に算出する。ここで、算出するスコアは、上述したｔ－ｐ－ＤＡＰスコア等でも良い。

【０１８１】

そして、特殊音声検知手段２６１０１は、算出されたフレームに対応する評定値を用いて、特殊な音声が入力されたか否かを判断する。つまり、評定値(例えば、ＤＡＰスコア)が、所定の値より低い区間が存在するか否かを判断する。

【０１８２】

次に、特殊音声検知手段２６１０１は、図２９に示すように、評定値(例えば、ＤＡＰスコア)が、所定の値より低い区間が、一つの音素内（ここでは音素２）であるか否かを判断する。そして、一つの音素内で評定値が低ければ、次に、特殊音声検知手段２６１０１は、直前の音素（音素１）および／または直後の音素（音素３）に対する評定値（例えば、ＤＡＰスコア)を算出し、当該評定値が所定の値より高ければ、音素の置換が発生している可能性があると判断する。次に、特殊音声検知手段２６１０１は、予め格納されている音韻ＨＭＭ（予期する音韻のＨＭＭは除く）の中で、所定の値以上の評定値が得られる音韻ＨＭＭが一つ存在すれば、音素の置換が発生していると判断する。なお、図２９において、音素２において、音素の置換が発生した区間である。なお、図２９において縦軸

は評定値であり、当該評定値は、ＤＡＰ、ｔ－ｐ－ＤＡＰ等、問わない。

【０１８３】

次に、評定手段２６１０３は、音素の置換があった旨を示す評定結果（例えば、「音素の置換が発生しました。」）を構成する。そして、出力手段２１１０４は、構成した評定結果を出力する。なお、出力手段２１１０４は、通常の入力音声に対しては、上述したように評定値を出力することが好適である。

【０１８４】

以上、本実施の形態によれば、ユーザが入力した発音を、教師データに対して、如何に似ているかを示す類似度（評定値）を算出し、出力できる。また、かかる場合、本実施の形態によれば、個人差、特に声道長の違いに影響を受けない、精度の高い評定ができる。さらに、本音声処理装置は、特殊音声、特に、音素の置換を検知できるので、極めて精度の高い評定結果が得られる。

【０１８５】

なお、本実施の形態において、音素の置換を検知できれば良く、評定値の算出アルゴリズムは問わない。評定値の算出アルゴリズムは、上述したアルゴリズム（ＤＡＰ、ｔ－ｐ－ＤＡＰ）でも良く、または、本明細書では述べていない他のアルゴリズムでも良い。

【０１８６】

また、本実施の形態において、音素の置換の検知アルゴリズムは、他のアルゴリズムでも良い。例えば、音素の置換の検知において、所定以上の長さの区間を有することを置換区間の検知で必須としても良い。その他、置換の検知アルゴリズムの詳細は種々考えられる。

【０１８７】

さらに、本実施の形態における音声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、第一サンプリング周波数で、受け付けた音声をサンプリングし、第一音声データを取得するサンプリングステップと、第二サンプリング周波数「第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付ステップで受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理ステップと、前記第二音声データを処理する音声処理ステップを実行させるためのプログラム、である。

【０１８８】

また、上記プログラムにおいて、音声処理ステップは、前記第二音声データを、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、前記フレーム毎の入力音声データに基づいて、特殊な音声が入力されたことを検知する特殊音声検知ステップと、教師データと前記入力音声データと前記特殊音声検知ステップにおける検知結果に基づいて、前記受け付けた音声の評定を行う評定ステップと、前記評定ステップにおける評定結果を出力する出力ステップを具備する、ことは好適である。

【０１８９】

また、上記プログラムにおいて、特殊音声検知ステップは、一の音素の評定値が所定の条件を満たすことを検知し、特殊音声検知ステップで前記所定の条件を満たすことを検知した場合に、少なくとも音素の置換があった旨を示す評定結果を構成する、ことは好適である。

（実施の形態６）

【０１９０】

本実施の形態において、入力音声において、特殊音声を検知し、比較対象の音声と入力音声の類似度を精度高く評定できる音声処理装置について説明する。特に、本音声処理装置は、音韻の欠落を検知できる音声処理装置である。

【０１９１】

図３０は、本実施の形態における音声処理装置のブロック図である。本音声処理装置は

、入力受付部１０１、教師データ格納部１０２、音声受付部１０３、教師データフォルマント周波数格納部１０４、第一サンプリング周波数格納部１０５、サンプリング部１０６、評価対象者フォルマント周波数取得部１０７、評価対象者フォルマント周波数格納部１０８、声道長正規化処理部１０９、音声処理部３０１０、発声催促部１１０９を具備する。

【０１９２】

音声処理部３０１０は、フレーム区分手段１１０１、フレーム音声データ取得手段１１０２、特殊音声検知手段３０１０１、評定手段３０１０３、出力手段２１１０４を具備する。なお、評定手段３０１０３は、最適状態決定手段１１０３１、最適状態確率値取得手段１１０３２を具備する。

【０１９３】

特殊音声検知手段３０１０１は、一の音素の評定値が所定の値より低く、かつ当該音素の直前の音素または当該音素の直後の音素の評定値が所定の値より高いことを検知する。また、特殊音声検知手段３０１０１は、一の音素の評定値が所定の値より低く、かつ当該音素の直前の音素または当該音素の直後の音素の評定値が所定の値より高く、かつ当該音素の区間長が所定の長さよりも短いことを検知しても良い。また、特殊音声検知手段３０１０１は、直前の音素に対応する確率値、または直後の音素に対応する確率値が、当該一の音素の確率値より高いことを検知しても良い。かかる場合に、特殊音声検知手段３０１０１は、音韻の欠落を検知することは好適である。さらに、音素の区間長が所定の長さよりも短いことを欠落の条件に含めることにより、音韻の欠落の検知の精度は向上する。特殊音声検知手段３０１０１は、通常、ＭＰＵやメモリ等から実現され得る。特殊音声検知手段３０１０１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。

【０１９４】

評定手段３０１０３は、特殊音声検知手段３０１０１が所定の条件を満たすことを検知した場合に、少なくとも音素の欠落があった旨を示す評定結果を構成する。評定手段３０１０３は、通常、ＭＰＵやメモリ等から実現され得る。評定手段３０１０３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。

【０１９５】

次に、音声処理装置の動作について、図３１、図３２のフローチャートを用いて説明する。なお、図３１のフローチャートは、図１２のフローチャートと比較して、ステップＳ３１０１の評定処理のみが異なるので、図３１のフローチャートの説明は省略する。ステップＳ３１０１の評定処理の詳細について、図３２のフローチャートを用いて説明する。図３２のフローチャートにおいて、図２、図１９、図２３、図２８のフローチャートの処理と同様の処理については、その説明を省略する。

【０１９６】

（ステップＳ３２０１）特殊音声検知手段３０１０１は、バッファに蓄積されているデータに対して、直前の音素に対応する教師データの確率値または、直後の音素に対応する教師データの確率値が、予定されている音素に対応する教師データの確率値より高いか否かを判断する。高ければステップＳ３２０２に行き、高くなければステップＳ２８１０に行く。なお、ステップＳ３２０２に行くための条件として、バッファに蓄積されているデータに対応するフレーム音声データ群の区間長が所定の長さ以下であることを付加しても良い。

（ステップＳ３２０２）評定手段３０１０３は、音素の欠落があった旨を示す評定結果を構成する。ステップＳ２８０８に行く。

なお、図３２のフローチャートにおいて、評定対象の音素（欠落したであろう音素）の区間長が、所定の長さ（例えば、３フレーム）よりも短いことを条件としても良いし、かかる条件は無くても良い。

以下、本実施の形態における音声処理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、音素の欠落の検知を行う処理が実施の形態５等とは異なる。そこで、その異なる処理を中心に説明する。

【０１９７】

まず、学習者（評価対象者）が、語学学習の開始の指示である動作開始指示を入力する。そして、音声処理装置は、当該動作開始指示を受け付け、次に、例えば、「"う"と発声してください。」と画面出力する。

そして、評価対象者は、"う"と発声し、音声処理装置は、当該発声から、第二ンプリング周波数「３２．１ＫＨｚ」を得る。かかる処理は、実施の形態１等において説明した処理と同様である。

【０１９８】

次に、発声催促部１１０９は、例えば、「"ｒｉｇｈｔ"と発声してください。」と画面出力する。そして、学習者は、学習対象の音声「ｒｉｇｈｔ」を発音する。そして、音声受付部１０３は、学習者が発音した音声の入力を受け付ける。

次に、サンプリング部１０６は、受け付けた音声「ｒｉｇｈｔ」をサンプリング周波数「２２．０５ＫＨｚ」でサンプリング処理する。そして、サンプリング部１０６は、「ｒｉｇｈｔ」の第一音声データを得る。

【０１９９】

次に、声道長正規化処理部１０９は、「ｒｉｇｈｔ」の第一音声データを第二サンプリング周波数「３２．１ＫＨｚ」でリサンプリング処理する。そして、声道長正規化処理部１０９は、第二音声データを得る。次に、音声処理部３０１０は、第二音声データを、以下のように処理する。

まず、フレーム区分手段１１０１は、「ｒｉｇｈｔ」の第二音声データを、短時間フレームに区分する。

そして、フレーム音声データ取得手段１１０２は、フレーム区分手段１１０１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「$O=o_1, o_2, \cdots, o_T$」を算出する。

次に、最適状態決定手段１１０３１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。

次に、最適状態確率値取得手段１１０３２は、上述した数式１、２により、最適状態における確率値を算出する。

【０２００】

次に、評定手段３０１０３は、例えば、最適状態決定手段１１０３１が決定した最適状態を有する音韻全体の状態における１以上の確率値を取得し、当該１以上の確率値の総和をパラメータとして音声の評定値を算出する。つまり、評定手段３０１０３は、例えば、ＤＡＰスコアをフレーム毎に算出する。ここで、算出するスコアは、上述したｔ－ｐ－ＤＡＰスコア等でも良い。

【０２０１】

そして、特殊音声検知手段３０１０１は、算出されたフレームに対応する評定値を用いて、特殊な音声が入力されたか否かを判断する。つまり、評定値(例えば、ＤＡＰスコア)が、所定の値より低い区間が存在するか否かを判断する。

【０２０２】

次に、特殊音声検知手段３０１０１は、図３３に示すように、評定値(例えば、ＤＡＰスコア)が、所定の値より低い区間が、一つの音素内（ここでは音素２）であるか否かを判断する。そして、一つの音素内で評定値が低ければ、特殊音声検知手段３０１０１は、直前の音素（音素１）または直後の音素（音素３）に対する評定値（例えば、ＤＡＰスコア)を算出し、当該評定値が所定の値より高ければ、音素の欠落が発生している可能性があると判断する。そして、当該区間長が、例えば、３フレーム以下の長さであれば、かかる音素は欠落したと判断する。なお、図３３において、音素２の欠落が発生したことを示

す。なお、図３３において縦軸は評定値であり、当該評定値は、ＤＡＰ、ｔ－ｐ－ＤＡＰ等、問わない。また、上記区間長の所定値は、「３フレーム以下」ではなく、「５フレーム以下」でも、「６フレーム以下」でも良い。

【０２０３】

次に、評定手段３０１０３は、音素の欠落があった旨を示す評定結果（例えば、「音素の欠落が発生しました。」）を構成する。そして、出力手段２１１０４は、構成した評定結果を出力する。なお、出力手段２１１０４は、通常の入力音声に対しては、上述したように評定値を出力することが好適である。

【０２０４】

以上、本実施の形態によれば、ユーザが入力した発音を、教師データに対して、如何に似ているかを示す類似度（評定値）を算出し、出力できる。また、かかる場合、本実施の形態によれば、個人差、特に声道長の違いに影響を受けない、精度の高い評定ができる。さらに、本音声処理装置は、特殊音声、特に、音素の欠落を検知できるので、極めて精度の高い評定結果が得られる。

【０２０５】

なお、本実施の形態において、音素の欠落を検知できれば良く、評定値の算出アルゴリズムは問わない。評定値の算出アルゴリズムは、上述したアルゴリズム（ＤＡＰ、ｔ－ｐ－ＤＡＰ）でも良く、または、本明細書では述べていない他のアルゴリズムでも良い。

【０２０６】

また、本実施の形態において、音素の欠落の検知アルゴリズムは、他のアルゴリズムでも良い。例えば、音素の欠落の検知において、所定長さ未満の区間であることを欠落区間の検知で必須としても良いし、区間長を考慮しなくても良い。

【０２０７】

さらに、本実施の形態における音声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、第一サンプリング周波数で、受け付けた音声をサンプリングし、第一音声データを取得するサンプリングステップと、第二サンプリング周波数「第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付ステップで受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理ステップと、前記第二音声データを処理する音声処理ステップを実行させるためのプログラム、である。

【０２０８】

また、上記プログラムにおいて、音声処理ステップは、前記第二音声データを、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、前記フレーム毎の入力音声データに基づいて、特殊な音声が入力されたことを検知する特殊音声検知ステップと、教師データと前記入力音声データと前記特殊音声検知ステップにおける検知結果に基づいて、前記受け付けた音声の評定を行う評定ステップと、前記評定ステップにおける評定結果を出力する出力ステップを具備する、ことは好適である。

【０２０９】

また、上記プログラムにおいて、特殊音声検知ステップは、一の音素の評定値が所定の条件を満たすことを検知し、特殊音声検知ステップで前記所定の条件を満たすことを検知した場合に、少なくとも音素の欠落があった旨を示す評定結果を構成する、ことは好適である。

（実施の形態７）

【０２１０】

本実施の形態における音声処理装置は、サンプリング周波数を変更し、リサンプリングを行わずに評定した場合の評定値と、リサンプリングを行って評定した場合の評定値とを取得し、２つの評定値に基づいて、最終的な評定値を算出する音声処理装置である。例えば、本音声処理装置は、２つの評定値の平均値を最終的な評定値としても良いし、２つの

評定値の最大値を最終的な評定値としても良い。また、本音声処理装置は、例えば、カラオケ評定装置である。

【０２１１】

図３４は、本実施の形態における音声処理装置のブロック図である。本音声処理装置は、入力受付部１０１、教師データ格納部１０２、音声受付部１０３、教師データフォルマント周波数格納部１０４、第一サンプリング周波数格納部１０５、サンプリング部１０６、評価対象者フォルマント周波数取得部１０７、評価対象者フォルマント周波数格納部１０８、声道長正規化処理部１０９、音声処理部３４１０、発声催促部１１０９を具備する。

音声処理部３４１０は、フレーム区分手段３４１０１、フレーム音声データ取得手段３４１０２、評定手段３４１０３、出力手段１１０４を具備する。

評定手段３４１０３は、第一評定手段３４１０３１、第二評定手段３４１０３２、評定結果取得手段３４１０３３を具備する。

フレーム区分手段３４１０１は、音声をフレームに区分し、かつ、前記第二音声データをフレームに区分する。

【０２１２】

フレーム音声データ取得手段３４１０２は、音声が区分されたフレーム毎の音声データである第一フレーム音声データを1以上得て、かつ前記第二音声データが区分されたフレーム毎の音声データである第二フレーム音声データを1以上得る。

【０２１３】

評定手段３４１０３は、教師データと1以上のフレーム音声データに基づいて、音声受付部１０３が受け付けた音声の評定を行う。評定手段３４１０３は、以下の第一評定手段３４１０３１の評定結果と、第二評定手段３４１０３２の評定結果に基づいて、最終的な評定結果を得る。

第一評定手段３４１０３１は、教師データと1以上の第一フレーム音声データに基づいて、音声受付部が受け付けた音声の評定を行う。

第二評定手段３４１０３２は、教師データと1以上の第二フレーム音声データに基づいて、音声受付部が受け付けた音声の評定を行う。

【０２１４】

評定結果取得手段３４１０３３は、第一評定手段３４１０３１における評定結果（以下、適宜「第一評定結果」という。）と第二評定手段３４１０３２における評定結果（以下、適宜「第二評定結果」という。）に基づいて、最終的な評定結果を得る。評定結果取得手段３４１０３３は、例えば、第一評定結果と第二評定結果の平均値を、最終的な評定結果としても良いし、第一評定結果と第二評定結果の大きい方の値を最終的な評定結果としても良いし、第一評定結果と第二評定結果の小さい方の値を最終的な評定結果としても良い。

【０２１５】

フレーム区分手段３４１０１、フレーム音声データ取得手段３４１０２、第一評定手段３４１０３１、第二評定手段３４１０３２、評定結果取得手段３４１０３３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３４１０１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。

次に、音声処理装置の動作について図３５のフローチャートを用いて説明する。図３５のフローチャートにおいて、図２、図１２のフローチャートと異なるステップについてのみ説明する。

【０２１６】

（ステップＳ３５０１）第一評定手段３４１０３１は、第一評定処理を行う。第一評定処理とは、教師データと1以上の第一フレーム音声データに基づいて、音声受付部１０３が受け付けた音声の評定を行う処理である。第一評定処理は、リサンプリングしない第一音声データを評定する処理である。第一評定処理における評定のアルゴリズムは、上記の

実施の形態１から実施の形態６で述べたいずれのアルゴリズム（ＤＡＰ、ｔ－ｐ－ＤＡＰ、無音区間考慮、挿入を考慮、置換を考慮、欠落を考慮など）または、それらを組み合わせたアルゴリズムでも良い。

【０２１７】

（ステップＳ３５０２）第二評定手段３４１０３２は、第二評定処理を行う。第二評定処理とは、教師データと1以上の第二フレーム音声データに基づいて、音声受付部１０３が受け付けた音声の評定を行う処理である。第二評定処理は、リサンプリングした第二音声データを評定する処理である。第二評定処理における評定のアルゴリズムは、上記の実施の形態１から実施の形態６で述べたいずれのアルゴリズム（ＤＡＰ、ｔ－ｐ－ＤＡＰ、無音区間考慮、挿入を考慮、置換を考慮、欠落を考慮など）または、それらを組み合わせたアルゴリズムでも良い。なお、第一評定処理と第二評定処理のアルゴリズムは、同一であることが好適である。

【０２１８】

（ステップＳ３５０３）評定結果取得手段３４１０３３は、第一評定手段３４１０３１における評定結果（第一評定結果）と第二評定手段３４１０３２における評定結果（第二評定結果）に基づいて、最終的な評定結果を得る。評定結果取得手段３４１０３３は、例えば、第一評定結果と第二評定結果の評定値のうち、高得点の方の評定値を最終的な評定結果とする。

【０２１９】

以上、本実施の形態によれば、ユーザが入力した発音を、教師データに対して、如何に似ているかを示す類似度（評定値）を算出し、出力できる。また、かかる場合、本実施の形態によれば、個人差を考慮した精度の高い評定ができる。さらに、本音声処理装置は、個人差を考慮した評定と、個人差を考慮しない評定の両方を利用した評定が行える。つまり、本実施の形態によれば、例えば、第一評定結果と第二評定結果の評定値のうち、高得点の方の評定値を最終的な評定結果とすることができ、カラオケ評定装置等として有効である。

（実施の形態８）

本実施の形態における音声処理装置の音声処理は、音声認識である。

図３６は、本実施の形態における音声処理装置のブロック図である。

【０２２０】

本音声処理装置は、入力受付部１０１、教師データ格納部１０２、音声受付部１０３、教師データフォルマント周波数格納部１０４、第一サンプリング周波数格納部１０５、サンプリング部１０６、評価対象者フォルマント周波数取得部１０７、評価対象者フォルマント周波数格納部１０８、声道長正規化処理部１０９、音声処理部３６１０、発声催促部１１０９を具備する。

音声処理部３６１０は、音声認識手段３６１０１、出力手段３６１０２を具備する。

【０２２１】

音声処理部３６１０の音声認識手段３６１０１は、第二音声データに基づいて、音声認識処理を行う。音声認識のアルゴリズムは、問わない。音声認識処理は、公知のアルゴリズムで良い。本実施の形態において、リサンプリングした第二音声データに基づいて音声認識することにより、精度の高い音声認識が可能である。音声処理部３６１０は、通常、ＭＰＵやメモリ等から実現され得る。音声処理部３６１０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。

【０２２２】

出力手段３６１０２は、音声認識結果を出力する。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。出力手段３６１０２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。出力手段３６１０２は、出力デバイスのドライバーソフトまたは、出力デバイスのドライバーソフトと出力デバイス等で実現され得る。

次に、音声処理装置の動作について図３７のフローチャートを用いて説明する。なお、図３７のフローチャートにおいて、図２、図１２のフローチャートの処理と同様の処理については、その説明を省略する。

【０２２３】

（ステップＳ３７０１）音声認識手段３６１０１は、ステップＳ１２０８でリサンプリング処理され、得られた第二音声データに基づいて、音声認識処理を行う。なお、音声認識手段３６１０１は、教師データとのマッチングを取り、教師データに近い音であると認識することにより、認識結果を得る。

（ステップＳ３７０２）出力手段３６１０２は、ステップＳ３７０１における音声認識結果を出力する。ステップＳ１２０６に戻る。

以上、本実施の形態によれば、精度高く音声認識できる。

【０２２４】

なお、本実施の形態における音声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、第一サンプリング周波数で、受け付けた音声をサンプリングし、第一音声データを取得するサンプリングステップと、第二サンプリング周波数「第一サンプリング周波数／（教師データフォルマント周波数／評価対象者フォルマント周波数）」で、前記音声受付ステップで受け付けた音声に対して、サンプリング処理を行い、第二音声データを得る声道長正規化処理ステップと、前記第二音声データに基づいて、音声認識処理を行う音声処理ステップを実行させるためのプログラム、である。

【０２２５】

また、上記の実施の形態において検出した特殊音声は、無音、挿入、置換、欠落であった。音声処理装置は、かかるすべての特殊音声について検知しても良いことはいうまでもない。また、音声処理装置は、主として、実施の形態１、実施の形態２において述べた評定値の算出アルゴリズムを利用して、特殊音声の検出を行ったが、他の評定値の算出アルゴリズムを利用しても良い。

【０２２６】

また、特殊音声は、無音、挿入、置換、欠落に限られない。例えば、特殊音声は、ｇａｒｂａｇｅ（雑音などの雑多な音素等）であっても良い。受け付けた音声にｇａｒｂａｇｅが混入している場合、その区間は類似度の計算対象から除外するのがしばしば望ましい。例えば、発音評定においては、学習者の発声には通常、息継ぎや無声区間などが数多く表れ、それらに対応する発声区間を評定対象から取り除くことが好適である。なお、無音は、一般に、ｇａｒｂａｇｅの一種である、と考える。

【０２２７】

そこで，どの音素にも属さない雑多な音素（ｇａｒｂａｇｅ音素）を設定し、ｇａｒｂａｇｅのＨＭＭをあらかじめ格納しておく。スコア低下区間において、ｇａｒｂａｇｅのＨＭＭに対する評定値（γ_t（ｊ））が所定の値より大きい場合，その区間はｇａｒｂａｇｅ区間と判定することは好適である。特に、発音評定において，ｇａｒｂａｇｅ区間が２つの単語にまたがっている場合、息継ぎなどが起こったものとして、評定値の計算対象から除外することは極めて好適である。

【０２２８】

また、図３８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処理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３８は、このコンピュータシステム３４０の概観図であり、図３９は、コンピュータシステム３４０のブロック図である。

【０２２９】

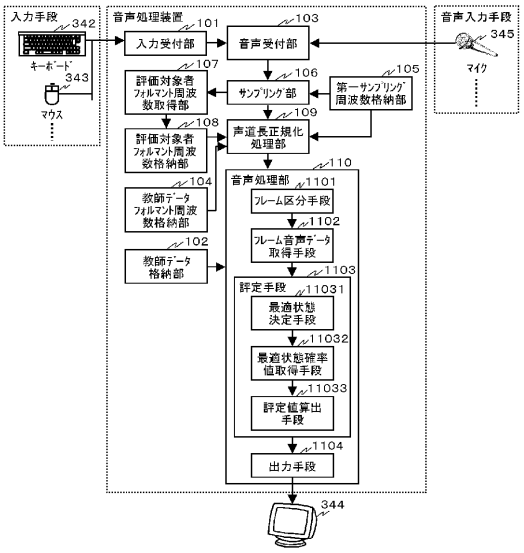
図３８において、コンピュータシステム３４０は、ＦＤ（Ｆｌｅｘｉｂｌｅ Ｄｉｓｋ）ドライブ、ＣＤ－ＲＯＭ（Ｃｏｍｐａｃｔ Ｄｉｓｋ Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）ドライブを含むコンピュータ３４１と、キーボード３４２と、マウス３４３と、モ

ニタ３４４と、マイク３４５とを含む。
【０２３０】
図３９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。
【０２３１】
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。
【０２３２】
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。
【０２３３】
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。
また、上記の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。
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】
【０２３４】
以上のように、本発明にかかる音声処理装置は、評価対象者の話者特性に応じた精度の高い音声処理ができるという効果を有し、発音評定装置やカラオケ評定装置や音声認識装置等として有用である。
【図面の簡単な説明】
【０２３５】
【図１】実施の形態１における音声処理装置のブロック図
【図２】同音声処理装置の動作について説明するフローチャート
【図３】同声道長正規化処理について説明するフローチャート
【図４】同ＨＭＭの仕様の例を示す図
【図５】同Ｆ１、Ｆ２の計測結果を示す図
【図６】同音声分析条件を示す図
【図７】同算出した評定値をグラフで表した例を示す図
【図８】同算出した評定値をグラフで表した例を示す図
【図９】同出力例を示す図

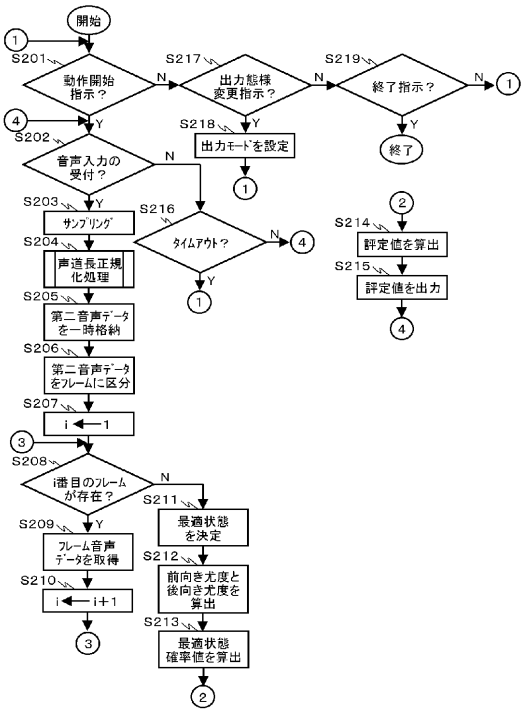
【図１０】同出力例を示す図
【図１１】実施の形態２における音声処理装置のブロック図
【図１２】同音声処理装置の動作について説明するフローチャート
【図１３】同第二サンプリング周波数算出処理について説明するフローチャート
【図１４】同評定処理について説明するフローチャート
【図１５】同評定結果（ｔ－ｐ－ＤＡＰスコア）を示す図
【図１６】同出力例を示す図
【図１７】実施の形態３における音声処理装置のブロック図
【図１８】同音声処理装置の動作について説明するフローチャート
【図１９】同評定処理について説明するフローチャート
【図２０】同無音データの検知について説明する図
【図２１】実施の形態４における音声処理装置のブロック図
【図２２】同音声処理装置の動作について説明するフローチャート
【図２３】同評定処理について説明するフローチャート
【図２４】同音素の挿入の検知について説明する図
【図２５】同出力例を示す図
【図２６】実施の形態５における音声処理装置のブロック図
【図２７】同音声処理装置の動作について説明するフローチャート
【図２８】同評定処理について説明するフローチャート
【図２９】同音素の置換の検知について説明する図
【図３０】実施の形態６における音声処理装置のブロック図
【図３１】同音声処理装置の動作について説明するフローチャート
【図３２】同評定処理について説明するフローチャート
【図３３】同音素の欠落の検知について説明する図
【図３４】実施の形態７における音声処理装置のブロック図
【図３５】同音声処理装置の動作について説明するフローチャート
【図３６】実施の形態８における音声処理装置のブロック図
【図３７】同音声処理装置の動作について説明するフローチャート
【図３８】同音声処理装置を構成するコンピュータシステムの概観図
【図３９】同音声処理装置を構成するコンピュータのブロック図
【符号の説明】
【０２３６】
１０１　入力受付部
１０２　教師データ格納部
１０３　音声受付部
１０４　教師データフォルマント周波数格納部
１０５　第一サンプリング周波数格納部
１０６　サンプリング部
１０７　評価対象者フォルマント周波数取得部
１０８　評価対象者フォルマント周波数格納部
１０９　声道長正規化処理部
１１０、１１１０、１７１０、２１１０、２６１０、３０１０、３４１０、３６１０　音声処理部
１１０１、３４１０１　フレーム区分手段
１１０２、３４１０２　フレーム音声データ取得手段
１１０３、１１１０３、１７１０３、２１１０３、２６１０３、３０１０３、３４１０３　評定手段
１１０４、２１１０４、３６１０２　出力手段
１１０９　発声催促部
１１０３１　最適状態決定手段

１１０３２　最適状態確率値取得手段
１１０３３、２１０２３、１１１０３３、１７１０３３　評定値算出手段
１７１０１、２１１０１、２６１０１、３０１０１　特殊音声検知手段
３６１０１　音声認識手段
１１１０３２　発音区間フレーム音韻確率値取得手段
１７１０１１　無音データ格納手段
１７１０１２　無音区間検出手段
３４１０３１　第一評定手段
３４１０３２　第二評定手段
３４１０３３　評定結果取得手段

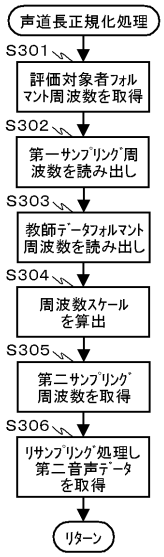
【図１】



【図２】



【図３】



【図４】

HMMﾀｲﾌﾟ	混合ｶﾞｳｽ型，mono phone（対角共分散行列）
状態数（ﾓﾃﾞﾙ毎）	3
混合数（状態毎）	8
HMM総数	44
総状態数	132

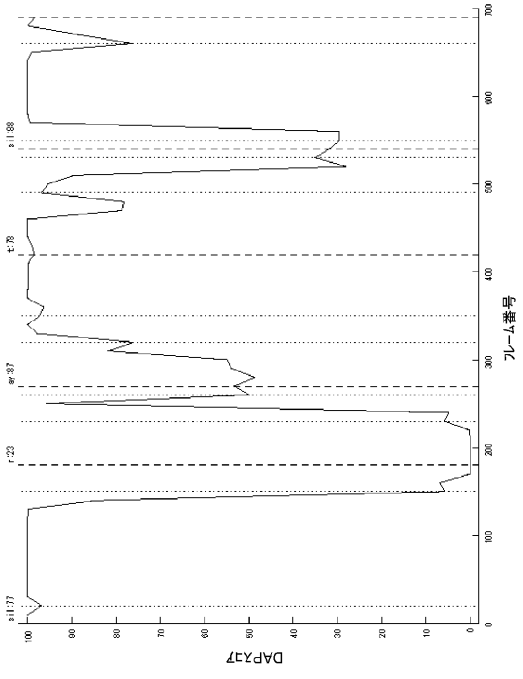
【図５】

	F1(Hz)	F2(Hz)
成人男性（139名の平均）	340	1184
成人女性（92名の平均）	370	1310
少年12-18歳（69名の平均）	402	1534
子供5-11歳（56名の平均）	510	1725

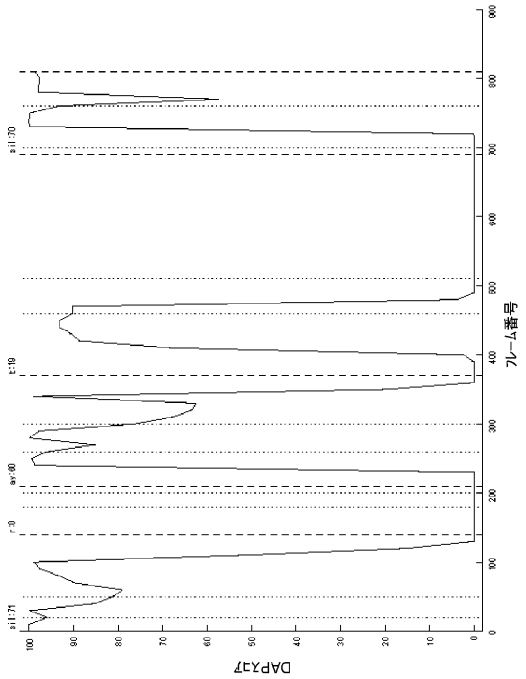
【図６】

ﾌﾟﾘｴﾝﾌｧｼｽ	$1-0.97z^{-1}$
窓関数	Hamming窓
分析ﾌﾚｰﾑ長	25msec
ﾌﾚｰﾑ周期	10msec
特徴ﾊﾟﾗﾒｰﾀ	MFCC（39次元）

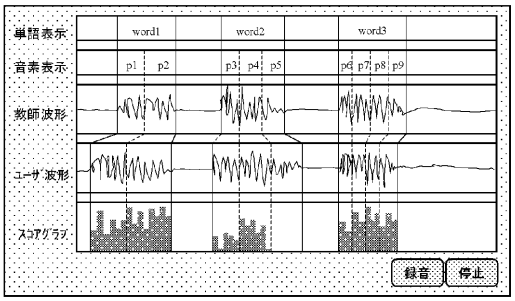
【図７】



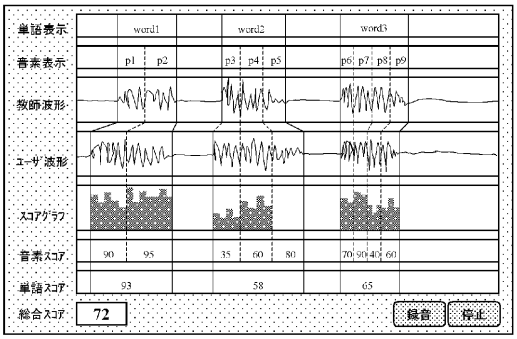
【図８】



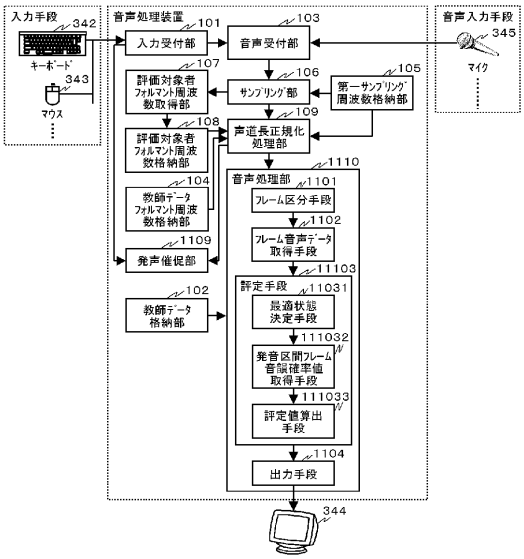
【図９】



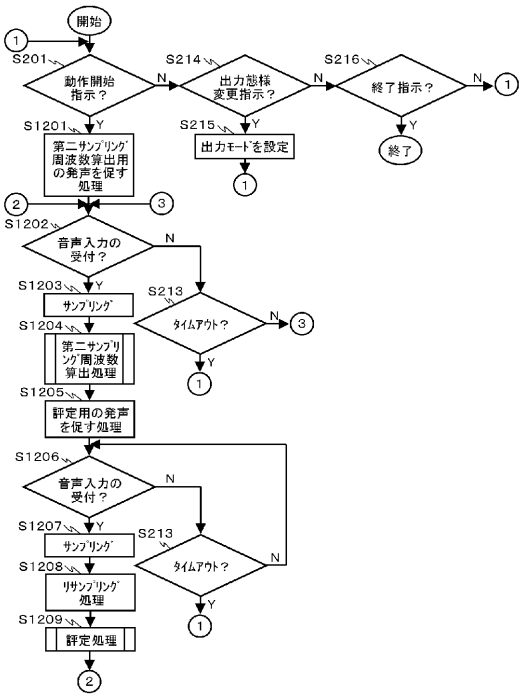
【図１０】



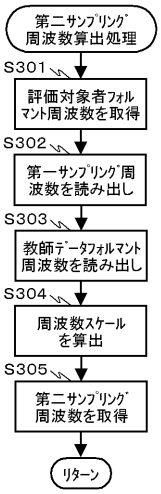
【図１１】



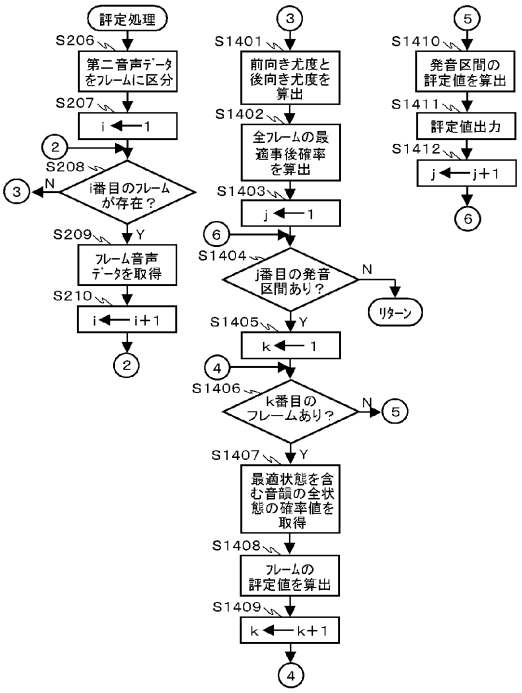
【図１２】



【図１３】



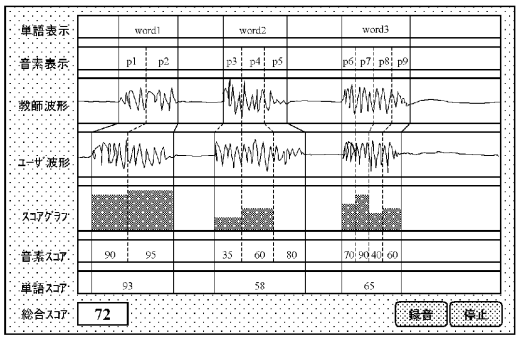
【図１４】



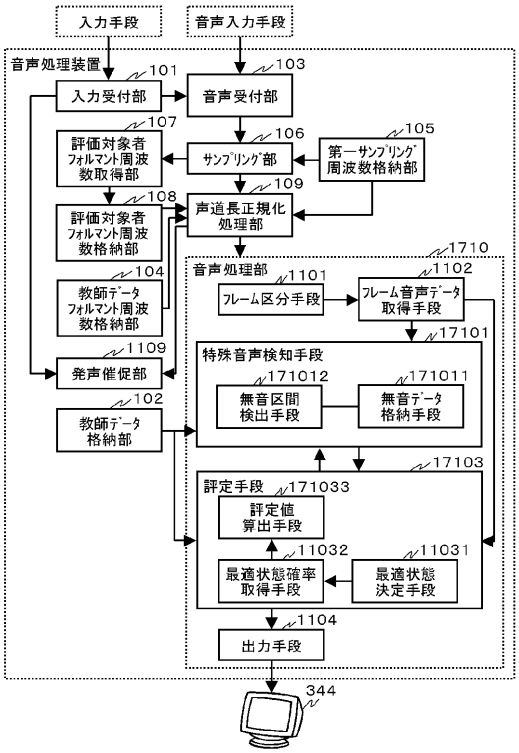
【図１５】

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

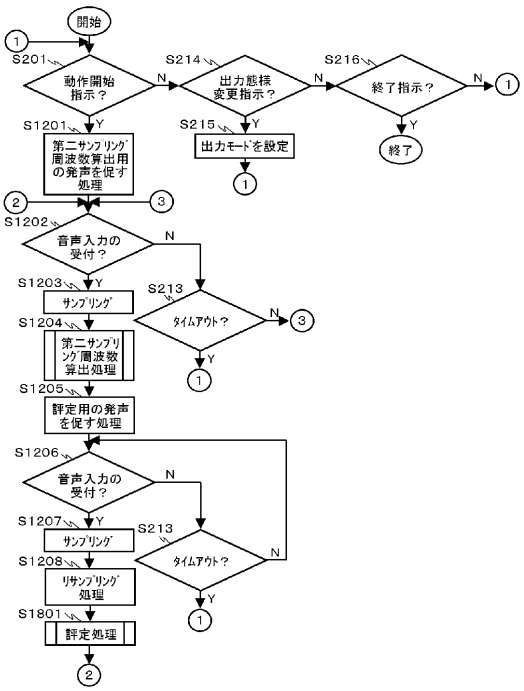
【図１６】



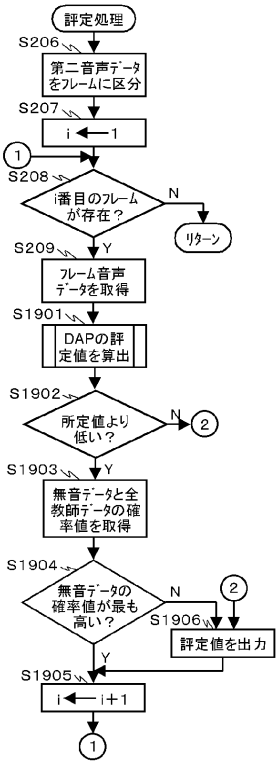
【図１７】



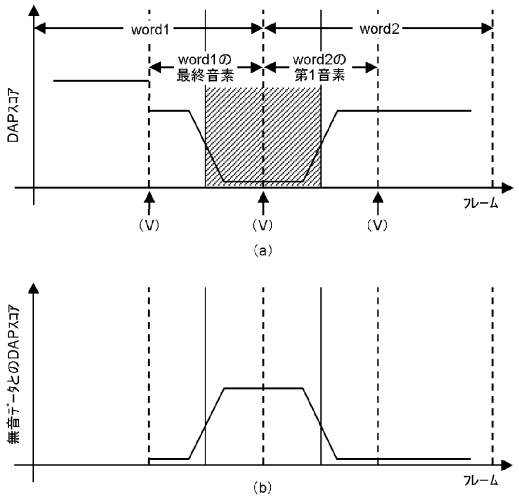
【図１８】



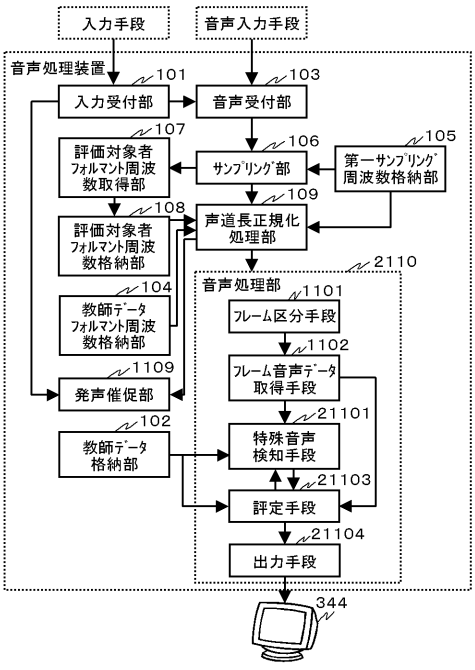
【図１９】



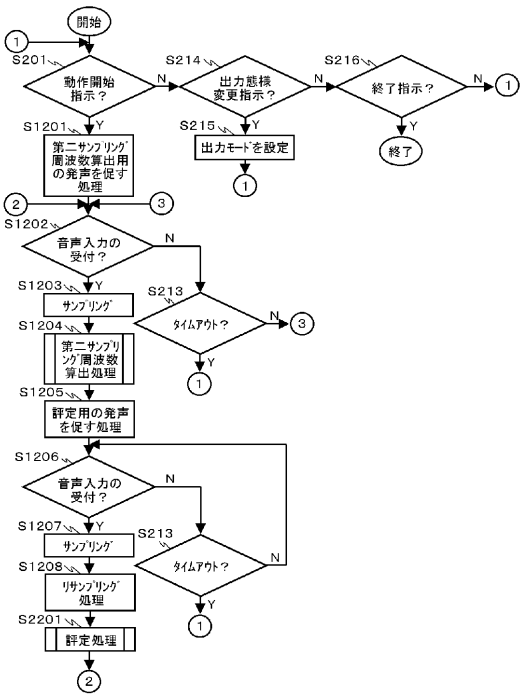
【図２０】



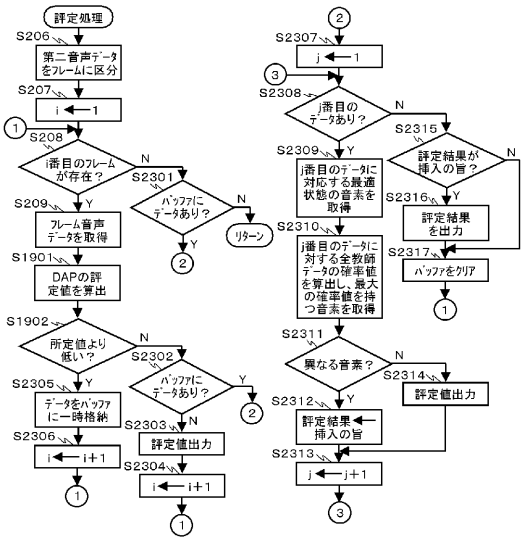
【図２１】



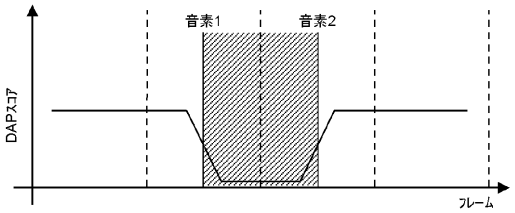
【図２２】



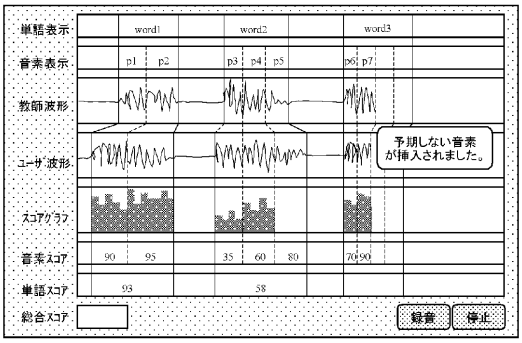
【図２３】



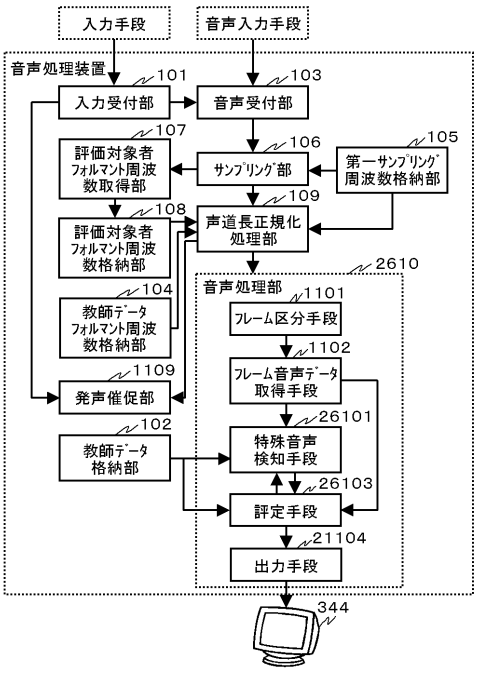
【図２４】



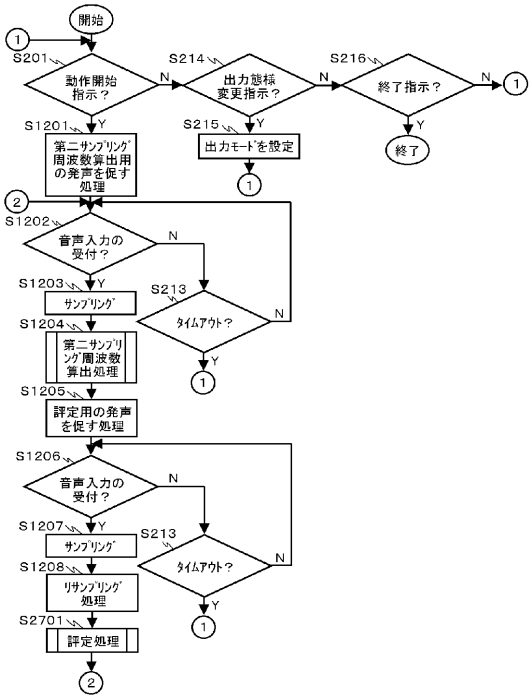
【図２５】



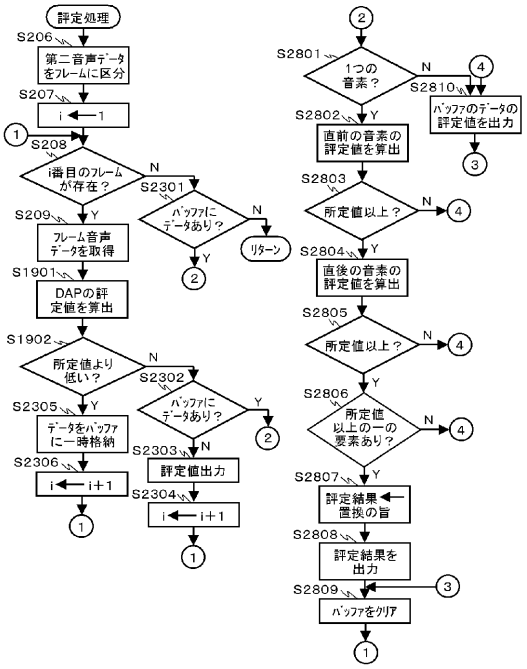
【図２６】



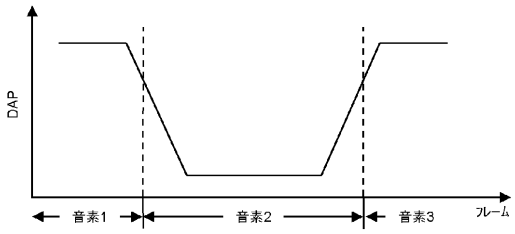
【図２７】



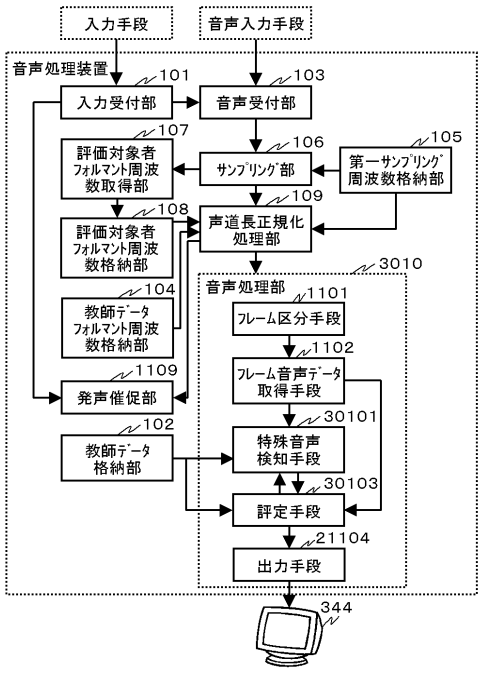
【図２８】



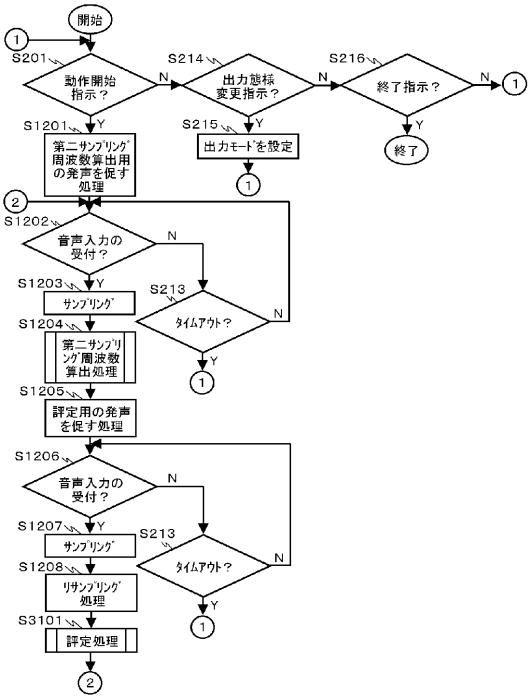
【図２９】



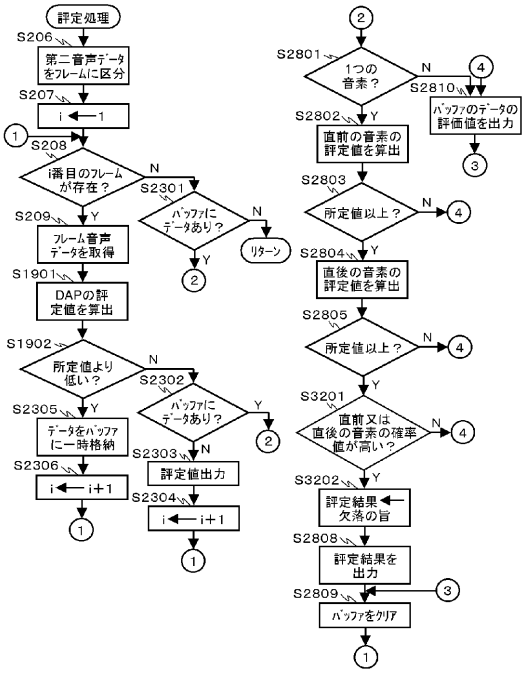
【図３０】



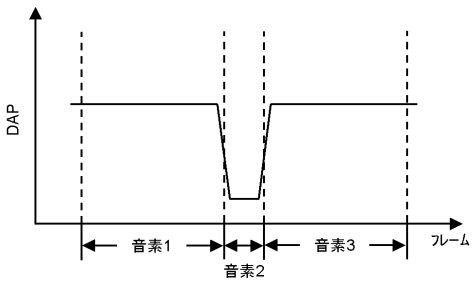
【図３１】



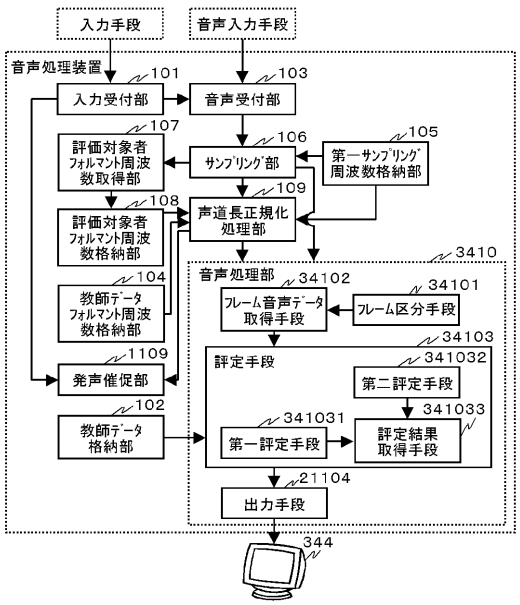
【図３２】



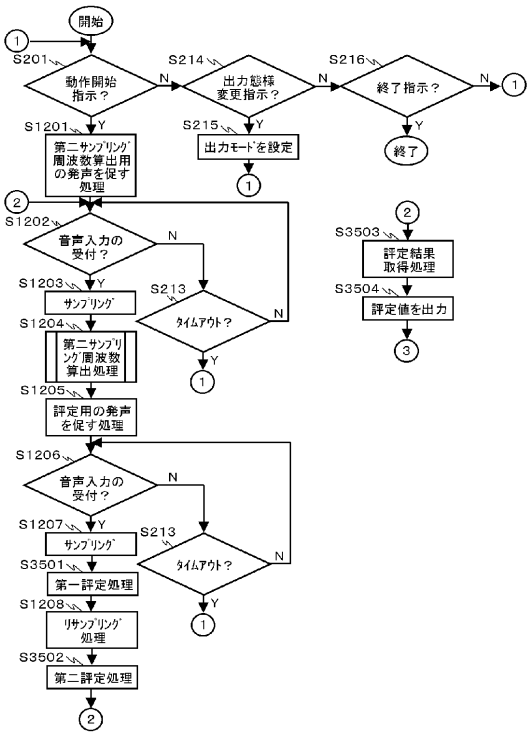
【図３３】



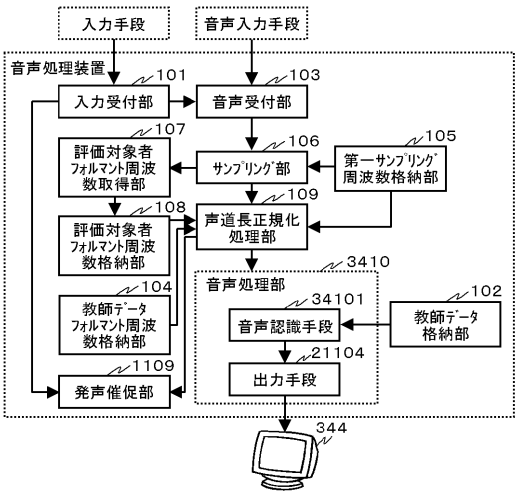
【図３４】



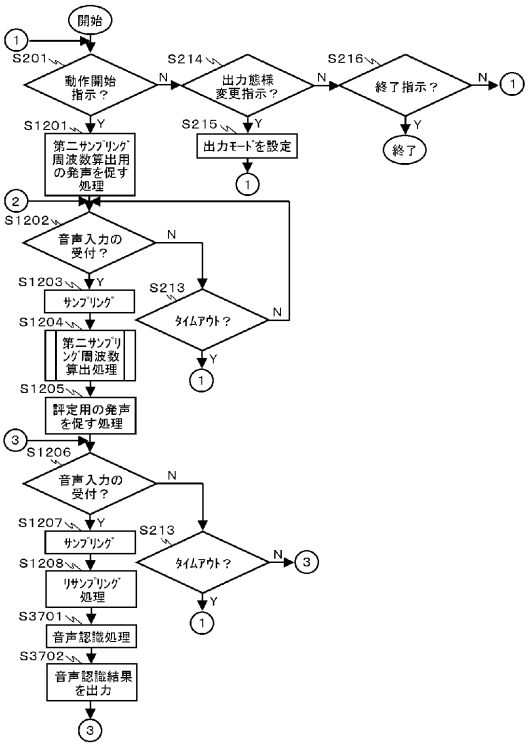
【図３５】



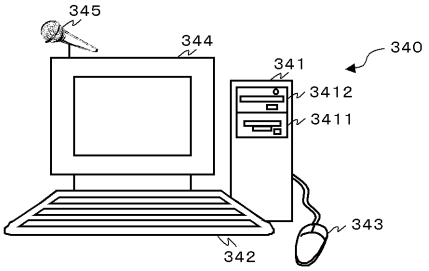
【図３６】



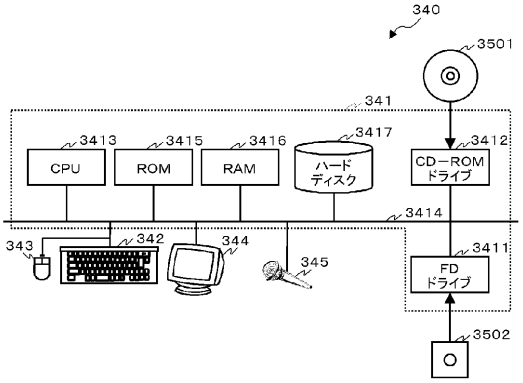
【図３７】



【図３８】



【図３９】



フロントページの続き

(72)発明者　加藤　宏明
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２００１－０４２８８９（ＪＰ，Ａ）
特開昭６２－１７４７９８（ＪＰ，Ａ）
特表２００２－５１５１３６（ＪＰ，Ａ）
特開２００１－２６５２１１（ＪＰ，Ａ）
特開平１０－２２２１９０（ＪＰ，Ａ）
特開平０６－１１０４９４（ＪＰ，Ａ）
特開２００６－２２７０３０（ＪＰ，Ａ）
特開２００１－１１７５９８（ＪＰ，Ａ）
特開平１１－２５９０８１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　　１１／００－２１／０６