

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5028599号
(P5028599)

(45) 発行日 平成24年9月19日 (2012. 9. 19)

(24) 登録日 平成24年7月6日 (2012. 7. 6)

(51) Int. Cl.

F I

G 1 0 L 11/00 (2006. 01)
 G 1 0 L 15/06 (2006. 01)
 G 1 0 L 15/02 (2006. 01)
 G 1 0 K 15/04 (2006. 01)

G 1 0 L 11/00 4 0 2 A
 G 1 0 L 15/06 3 1 0 S
 G 1 0 L 15/02 3 0 0 D
 G 1 0 K 15/04 3 0 2 D

請求項の数 11 (全 83 頁)

(21) 出願番号 特願2006-161918 (P2006-161918)
 (22) 出願日 平成18年6月12日 (2006. 6. 12)
 (65) 公開番号 特開2007-199654 (P2007-199654A)
 (43) 公開日 平成19年8月9日 (2007. 8. 9)
 審査請求日 平成21年2月18日 (2009. 2. 18)
 (31) 優先権主張番号 特願2005-373089 (P2005-373089)
 (32) 優先日 平成17年12月26日 (2005. 12. 26)
 (33) 優先権主張国 日本国 (JP)

(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 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 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 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】

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、
 サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している
 声道長正規化パラメータ格納部と、
 前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出され
 る第二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と、
 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、
 前記教師データおよび前記音声受付部が受け付けた音声に基づいて、前記声道長正規化パ
 ラメータを算出する声道長正規化パラメータ算出部とを具備し、
 前記声道長正規化パラメータ格納部の声道長正規化パラメータは、前記声道長正規化パラ
 メータ算出部が算出した声道長正規化パラメータである音声処理装置であって、
 前記声道長正規化パラメータ算出部は、
音素平均ケプストラムベクトルを指定された発話内容に従って並べた音素平均ケプストラ
ムベクトル列である学習音響データを格納している学習音響データ格納手段と、
 前記音声受付部が受け付けた音声から、短区間分析により第二のケプストラムベクトル系

10

20

列を算出する第二ケプストラムベクトル系列算出手段と、
前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列
を用いて変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプス
トラム変換手段と、

ユーザが発声した発話音声構成する各フレームに対応する音素を識別する情報の列であ
る音素系列を取得する最適音素系列取得手段と、

前記学習音響データおよび前記第三のケプストラムベクトル系列および前記音素系列を用
いて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、
所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記最適音素
系列取得手段における処理、および前記ケプストラム変換パラメータ算出手段における処
理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ
取得手段と、

前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づ
いて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段とを具備
する音声処理装置。

【請求項 2】

前記声道長正規化パラメータ算出部は、請求項 1 記載の構成に代えて、
音素 HMM を指定された発話内容に従って連結した連結 HMM である学習音響データを格
納している学習音響データ格納手段と、

前記音声受付部が受け付けた音声から、短区間分析により第二のケプストラムベクトル系
列を算出する第二ケプストラムベクトル系列算出手段と、

前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列
を用いて変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプス
トラム変換手段と、

指定された発話内容に従って受け付けた音声の事後確率である占有度数を算出する占有度
算出手段と、

前記学習音響データおよび前記第三のケプストラムベクトル系列および前記占有度数を用
いて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、
所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記占有度算
出手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰
り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手
段と、

前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づ
いて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段とを具備
する請求項 1 記載の音声処理装置。

【請求項 3】

前記声道長正規化パラメータ算出手段は、
前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータから線
形周波数伸縮比を算出し、当該線形周波数伸縮比から前記声道長正規化パラメータを算出
する請求項 1 または請求項 2 記載の音声処理装置。

【請求項 4】

前記声道長正規化パラメータ算出部は、
周波数範囲を指定する情報である周波数範囲指定情報を格納している周波数範囲指定情報
格納手段をさらに具備し、
前記ケプストラム変換パラメータ算出手段は、
前記周波数範囲指定情報をも用いて、ケプストラム変換パラメータを算出する請求項 2 ま
たは請求項 3 記載の音声処理装置。

【請求項 5】

前記教師データは、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最適状態の事後確率を示す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音声の評定値を算出する評定値算出手段を具備する請求項1から請求項4いずれか記載の音声処理装置。

10

【請求項6】

前記評定手段は、
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2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を具備する請求項5記載の音声処理装置。

20

【請求項7】

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6いずれか記載の音声処理装置。

30

【請求項8】

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、
サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している声道長正規化パラメータ格納部と、
前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出される第二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と、
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、
前記教師データおよび前記音声受付部が受け付けた音声に基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出部とを具備し、
前記声道長正規化パラメータ格納部の声道長正規化パラメータは、前記声道長正規化パラメータ算出部が算出した声道長正規化パラメータであって、
前記声道長正規化パラメータ算出部は、
音素平均ケプストラムベクトルを指定された発話内容に従って並べた音素平均ケプストラムベクトル列である学習音響データを格納している学習音響データ格納手段と、
前記音声受付部が受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出手段と、
前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列

40

50

を用いて変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換手段と、

ユーザが発声した発話音声を構成する各フレームに対応する音素を識別する情報の列である音素系列を取得する最適音素系列取得手段と、

前記学習音響データおよび前記第三のケプストラムベクトル系列および前記音素系列を用いて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、

所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記最適音素系列取得手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、

前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段とを具備するデジタルシグナルプロセッサ。

【請求項 9】

前記声道長正規化パラメータ算出部は、請求項 8 記載の構成に代えて、

音素 HMM を指定された発話内容に従って連結した連結 HMM である学習音響データを格納している学習音響データ格納手段と、

前記音声受付部が受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出手段と、

前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列を用いて変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換手段と、

指定された発話内容に従って受け付けた音声の事後確率である占有度数を算出する占有度算出手段と、

前記学習音響データおよび前記第三のケプストラムベクトル系列および前記占有度数を用いて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、

所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記占有度算出手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、

前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段とを具備する請求項 8 記載のデジタルシグナルプロセッサ。

【請求項 10】

コンピュータを、

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、

サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している声道長正規化パラメータ格納部と、

前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出される第二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と、

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、

前記教師データおよび前記音声受付部が受け付けた音声に基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出部とを具備するものとして機能させるためのプログラムであって、

前記声道長正規化パラメータ格納部の声道長正規化パラメータは、前記声道長正規化パラメータ算出部が算出した声道長正規化パラメータであって、

10

20

30

40

50

前記声道長正規化パラメータ算出部は、
音素平均ケプストラムベクトルを指定された発話内容に従って並べた音素平均ケプストラムベクトル列である学習音響データを格納している学習音響データ格納手段と、
前記音声受付部が受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出手段と、
前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列を用いて変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換手段と、
ユーザが発声した発話音声を構成する各フレームに対応する音素を識別する情報の列である音素系列を取得する最適音素系列取得手段と、
前記学習音響データおよび前記第三のケプストラムベクトル系列および前記音素系列を用いて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、
所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記最適音素系列取得手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、
前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段とを具備するものとして機能させるためのプログラム。

10

【請求項 11】

20

前記声道長正規化パラメータ算出部は、請求項 10 記載の構成に代えて、
音素 HMM を指定された発話内容に従って連結した連結 HMM である学習音響データを格納している学習音響データ格納手段と、
前記音声受付部が受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出手段と、
前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列を用いて変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換手段と、
指定された発話内容に従って受け付けた音声の事後確率である占有度数を算出する占有度数算出手段と、
前記学習音響データおよび前記第三のケプストラムベクトル系列および前記占有度数を用いて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、
所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記占有度数算出手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、
前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段とを具備するものとして機能させるための請求項 10 記載のプログラム。

30

【発明の詳細な説明】

40

【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、

50

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。

【 0 0 0 3 】

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。

10

【 0 0 0 4 】

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。

20

【特許文献 1】特開 2 0 0 3 - 2 2 8 2 7 9（第 1 頁、第 1 図等）

【特許文献 2】特開 2 0 0 5 - 1 0 7 0 8 8（第 1 頁、第 1 図等）

【特許文献 3】特開平 6 - 4 0 9 6（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】

【 0 0 0 5 】

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。

30

【 0 0 0 6 】

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。

40

【 0 0 0 7 】

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。

【 0 0 0 8 】

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。

50

【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、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、サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している声道長正規化パラメータ格納部と、前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出される第二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。

【0010】

かかる構成により、評価対象者の話者特性に応じた音声処理ができる。

【0011】

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、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、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をさらに具備し、前記声道長正規化パラメータは、「評価対象者フォルマント周波数/教師データフォルマント周波数」により算出される値であり、前記声道長正規化処理部は、「前記第一サンプリング周波数×声道長正規化パラメータ」の演算式により、第二サンプリング周波数を算出し、当該第二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音声処理装置である。

【0012】

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。

【0013】

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、「前記第一サンプリング周波数×声道長正規化パラメータ」の演算式により、第二サンプリング周波数を算出し、当該第二サンプリング周波数で、前記第一音声データに対して、サンプリング処理を行い、第二音声データを得る音声処理装置である。

【0014】

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。

【0015】

また、本第四の発明の音声処理装置は、第一の発明に対して、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、前記教師データおよび前記音声受付部が受け付けた音声に基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出部とをさらに具備し、前記声道長正規化パラメータ格納部の声道長正規化パラメータは、前記声道長正規化パラメータ算出部が算出した声道長正規化パラメータである音声処理装置である。

【0016】

かかる構成により、動的に評価対象者ごとの声道長正規化パラメータで、かつ、話者特性に応じた精度の高い音声処理ができる。

【0017】

また、本第五の発明の音声処理装置は、第四の発明に対して、前記声道長正規化パラメータ算出部は、音素HMMを指定された発話内容に従って連結した連結HMMである学習音響データを格納している学習音響データ格納手段と、前記音声受付部が受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出手段と、前記第二のケプストラムベクトル系列を、ケプストラム変換パラ

10

20

30

40

50

メータを要素とする行列を用いて変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換手段と、指定された発話内容に従って受け付けた音声の事後確率である占有度数を算出する占有度算出手段と、前記学習音響データおよび前記第三のケプストラムベクトル系列および前記占有度数に基づいて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記占有度算出手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段を具備する音声処理装置である。

10

【0018】

かかる構成により、動的に評価対象者ごとの声道長正規化パラメータでき、かつ、話者特性に応じた精度の高い音声処理ができる。

【0019】

また、本第六の発明の音声処理装置は、第四の発明に対して、前記声道長正規化パラメータ算出部は、音素平均ケプストラムベクトルを指定された発話内容に従って並べた音素平均ケプストラムベクトル列である学習音響データを格納している学習音響データ格納手段と、前記音声受付部が受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出手段と、前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列を用いて変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換手段と、ユーザが発声した発話音声を構成する各フレームに対応する音素を識別する情報の列である音素系列を取得する最適音素系列取得手段と、前記学習音響データおよび前記第三のケプストラムベクトル系列および前記音素系列を用いて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記最適音素系列取得手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段を具備する音声処理装置である。

20

30

【0020】

かかる構成により、動的に評価対象者ごとの声道長正規化パラメータでき、かつ、話者特性に応じた精度の高い音声処理ができる。

【0021】

また、本第七の発明の音声処理装置は、第五、第六の発明に対して、前記声道長正規化パラメータ算出手段は、前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータから線形周波数伸縮比を算出し、当該線形周波数伸縮比から前記声道長正規化パラメータを算出する音声処理装置である。

【0022】

かかる構成により、動的に評価対象者ごとの声道長正規化パラメータでき、かつ、話者特性に応じた、精度の高い音声処理ができる。

40

【0023】

また、本第八の発明の音声処理装置は、第五、第六の発明に対して、前記声道長正規化パラメータ算出部は、周波数範囲を指定する情報である周波数範囲指定情報を格納している周波数範囲指定情報格納手段をさらに具備し、前記ケプストラム変換パラメータ算出手段は、前記周波数範囲指定情報を用いて、ケプストラム変換パラメータを算出する音声処理装置である。

【0024】

かかる構成により、動的に評価対象者ごとの声道長正規化パラメータでき、かつ、話者

50

特性に応じた、さらに精度の高い音声処理ができる。

【 0 0 2 5 】

また、本第九の発明の音声処理装置は、第一から第八いずれかの発明に対して、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。

【 0 0 2 6 】

かかる構成により、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。

10

【 0 0 2 7 】

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。

【 0 0 2 8 】

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。

【 0 0 2 9 】

20

また、本第十一の発明の音声処理装置は、第九の発明に対して、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。

【 0 0 3 0 】

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。

【 0 0 3 1 】

30

また、本第十二の発明の音声処理装置は、第九の発明に対して、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。

【 0 0 3 2 】

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。

【 0 0 3 3 】

また、本第十三の発明の音声処理装置は、第十二の発明に対して、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。

40

【 0 0 3 4 】

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。

【 0 0 3 5 】

また、本第十四の発明の音声処理装置は、第十二の発明に対して、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

50

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価結果を構成する音声処理装置である。

【 0 0 3 6 】

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価ができ、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価ができる。

【 0 0 3 7 】

また、本第十五の発明の音声処理装置は、第十二の発明に対して、前記特殊音声検知手段は、一の音素の評価値が所定の条件を満たすことを検知し、前記評価手段は、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価結果を構成する音声処理装置である。

10

【 0 0 3 8 】

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価ができ、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価ができる。

【 0 0 3 9 】

また、本第十六の発明の音声処理装置は、第一から第十五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。

【 0 0 4 0 】

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。

20

【 0 0 4 1 】

また、本第十七の発明の音声処理装置は、第十六の発明に対して、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。

【 0 0 4 2 】

30

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価ができる。

【 0 0 4 3 】

また、本第十八の発明の音声処理装置は、第一から第八いずれかの発明に対して、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。

【 0 0 4 4 】

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。

【 0 0 4 5 】

また、本第十九の発明は、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、サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している声道長正規化パラメータ格納部と、前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出される第二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。

40

【 0 0 4 6 】

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 DSP を提供できる。

【発明の効果】

【 0 0 4 7 】

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

50

理ができる。

【発明を実施するための最良の形態】

【0048】

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)

【0049】

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。

10

【0050】

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。

20

【0051】

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。

【0052】

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。

【0053】

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。

30

【0054】

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。

【0055】

40

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

50

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。

【0056】

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のドライバースoftwareで実現され得る。また、なお、音声受付部103は、マイク345とそのドライバースoftware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。

10

【0057】

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。

20

【0058】

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。

【0059】

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。

30

【0060】

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。

40

【0061】

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

50

。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。

【0062】

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、周波数軸変換率（「 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（ F_s ）と「 r 」に基づいて、演算「 F_s / r 」を行い、新しいサンプリング周波数（ F_s / r ）を得る。この新しいサンプリング周波数（ F_s / 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（ F_s / 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。なお、周波数軸変換率「 r 」の逆数「 $1 / r$ 」は、後述する声道長パラメータである、と言える。また、声道長パラメータは、声道長の正規化のためのパラメータであるので、周波数軸変換率「 r 」を声道長パラメータと考えても良い。

【0063】

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。

【0064】

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。

【0065】

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。

【0066】

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

評価を行う。評価方法の具体例は、後述する。「音声受付部 103 が受け付けた音声を評価する」の概念には、第二音声データを評価することも含まれることは言うまでもない。評価手段 1103 は、通常、MPU やメモリ等から実現され得る。評価手段 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。

【0067】

評価手段 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るても良いし、連結される前のデータであり、全音韻 HMM のデータと考えるても良い。

10

【0068】

評価手段 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。

20

【0069】

評価手段 1103 を構成している評価値算出手段 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価値を算出する。評価値算出手段 11033 は、上記確率値を如何に利用して、評価値を算出するかは問わない。評価値算出手段 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価値を算出する。評価値算出手段 11033 は、ここでは、通常、フレームごとに評価値を算出する。

【0070】

最適状態決定手段 11031、最適状態確率値取得手段 11032、評価値算出手段 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。

30

【0071】

出力手段 1104 は、評価手段 1103 における評価結果を出力する。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、評価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価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るても含まないと考えるても良い。出力手段 1104 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。

40

【0072】

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。

【0073】

（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。

【0074】

50

(ステップS 2 0 2) 音声受付部 1 0 3 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 1 6 に飛ぶ。

【 0 0 7 5 】

(ステップS 2 0 3) サンプリング部 1 0 6 は、第一サンプリング周波数格納部 1 0 5 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 1 0 3 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。

【 0 0 7 6 】

(ステップS 2 0 4) 声道長正規化処理部 1 0 9 は、音声受付部 1 0 3 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価のための前処理である。

10

【 0 0 7 7 】

(ステップS 2 0 5) フレーム区分手段 1 1 0 1 は、ステップS 2 0 4 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。

【 0 0 7 8 】

(ステップS 2 0 6) フレーム区分手段 1 1 0 1 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 1 1 0 1 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 1 1 0 2 がフレーム音声データを取り出す際の

20

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。

【 0 0 7 9 】

(ステップS 2 0 7) フレーム音声データ取得手段 1 1 0 2 は、カウンタ i に 1 を代入する。

【 0 0 8 0 】

(ステップS 2 0 8) フレーム音声データ取得手段 1 1 0 2 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S 2 0 9 に行き、 i 番目のフレームが存在しなければステップS 2 1 1 に行く。

【 0 0 8 1 】

30

(ステップS 2 0 9) フレーム音声データ取得手段 1 1 0 2 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ンネル数 2 4 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 1 2 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 (3 9 次元) を有する。

【 0 0 8 2 】

40

(ステップS 2 1 0) フレーム音声データ取得手段 1 1 0 2 は、カウンタ i を 1、インクリメントする。ステップS 2 0 8 に戻る。

【 0 0 8 3 】

(ステップS 2 1 1) 最適状態決定手段 1 1 0 3 1 は、全フレームの最適状態を決定する。最適状態決定手段 1 1 0 3 1 が最適状態を決定するアルゴリズムは、例えば、V i t e r b i アルゴリズムによる。V i t e r b i 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。

【 0 0 8 4 】

(ステップS 2 1 2) 最適状態確率値取得手段 1 1 0 3 2 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 1 1 0 3 2 は、例えば

50

、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。

【0085】

（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。

【0086】

（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。

【0087】

（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。

【0088】

（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。

【0089】

（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。

【0090】

（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。

【0091】

（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、終了指示を受け付ければステップS201に戻る。

【0092】

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。

【0093】

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。

【0094】

（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（F_i）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。

【0095】

（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1

10

20

30

40

50

05の第一サンプリング周波数(F_s)を読み出す。

【0096】

(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。

【0097】

(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軸変換率を算出する。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、周波数軸変換率(r)を得る。

10

【0098】

(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(F_s)と周波数軸変換率(r)に基づいて、演算「 F_s / r 」を行い、第二サンプリング周波数(F_s / r)を得る。

【0099】

(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(F_s / 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。

【0100】

20

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。

【0101】

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。

30

【0102】

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、 1 番目の音韻に対するHMMを λ_1 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。

【0103】

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、教師データを発声した者(教師)は、大人である、とする。

40

【0104】

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。

【0105】

50

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」を発音するように促すことは好適である。「う」を発音するように促すために、音声処理装置は、例えば、「"う"と発音してください。」と画面出力しても良いし、「"う"と発音してください。」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、「22.05 KHz」を保持している、とする。

【0106】

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。

【0107】

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(F_i)とする。今、このF_iが「1725 Hz」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、F_i(1725 Hz)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。

【0108】

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、「1184 Hz」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」と発音してもらい、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。

【0109】

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。

【0110】

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「1725 Hz」と教師データフォルマント周波数「1184 Hz」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、周波数軸変換率(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、周波数軸変換率「0.686」を得る。

【0111】

次に、声道長正規化処理部109は、第一サンプリング周波数(F_s)と「r」に基づいて、演算「F_s/r」を行い、第二サンプリング周波数(F_s/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「32.1 KHz」を一時格納する。

【0112】

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」の発音を促すことは好適である。

【0113】

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「22.05 KHz」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。

【0114】

10

20

30

40

50

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「32.1 KHz」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。

【0115】

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。

【0116】

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。

【0117】

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「 $O = o_1, o_2, \dots, o_T$ 」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「22.05 KHz」である。

【0118】

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。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、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。

【0119】

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（ 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。

【数 1】

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

【0120】

数式 2 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-Weilch 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

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

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

【0 1 2 1】

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

【0 1 2 2】

また、数式 2 において、N は、全 HMM に渡る状態の総数を示す。

【0 1 2 3】

なお、評価手段 1 1 0 3 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、評価手段 1 1 0 3 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。

30

【0 1 2 4】

次に、評価値算出手段 1 1 0 3 3 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価値を算出する。かかる場合、もし学習者の t フレーム目に対応する発音が、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（評価値）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、評価値はほぼ $1/N$ 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、この評価値は十分小さくなる。また、ここでは、評価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、收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価値が高いスコアを維持する。よって、評価値算出手段 1 1 0 3 3 は、最適状態確率値取得手段 1 1 0 3 2 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価値を算出することは、極めて好適である。

40

【0 1 2 5】

かかる評価値算出手段 1 1 0 3 3 が算出した評価値（「DAP スコア」とも言う。）の出力例を、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

50

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 7 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。なお、評価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。

【 0 1 2 6 】

図 8 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。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。

【 0 1 2 7 】

そして、出力手段 1 1 0 4 は、評価手段 1 1 0 3 の評価結果を出力する。具体的には、例えば、出力手段 1 1 0 4 は、図 9 に示すような態様で、評価結果を出力する。つまり、出力手段 1 1 0 4 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、各フレームの評価値を表示する。その他、出力手段 1 1 0 4 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「word 1」など）や、音素（図 9 の「p 1」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。

【 0 1 2 8 】

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、発声全体の評価結果を表示しても良い。上記の処理において、フレーム単位の評価値を算出するので、単語単位、発声全体の評価結果を得るためには、フレーム単位の 1 以上の評価値をパラメータとして、単語単位、発声全体の評価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 以上の評価値の平均値を単語単位の評価値とする、ことが考えられる。

【 0 1 2 9 】

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。

【 0 1 3 0 】

また、出力手段 1 1 0 4 の表示は、図 1 0 に示すような態様でも良い。図 1 0 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。

【 0 1 3 1 】

なお、出力手段 1 1 0 4 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。

【 0 1 3 2 】

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（評価値）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、精度の高い評価ができる。

【 0 1 3 3 】

また、本実施の形態によれば、連結された H M M である連結 H M M を用いて最適状態を求め、評価値を算出するので、高速に評価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、単語ごとの評価値を出力できる。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、さらに高速に評価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価結果を出力でき、出力

態様の自由度が高い。

【 0 1 3 4 】

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。

【 0 1 3 5 】

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。かかることも、他の実施の形態においても同様である。

【 0 1 3 6 】

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。かかることも、他の実施の形態においても同様である。

【 0 1 3 7 】

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。

【 0 1 3 8 】

また、本実施の形態において、音声処理装置は、声道長正規化処理部 109 において、声道長正規化パラメータ「 r 」を算出した。しかし、別途、声道長正規化パラメータ「 $1/r$ 」を算出してにおいて、かかる声道長正規化パラメータ「 $1/r$ 」を声道長正規化パラメータ格納部に格納していても良い。かかる場合、音声処理装置は、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、サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している声道長正規化パラメータ格納部と、前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出される第二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。かかる場合、音声処理装置は、教師データフォルマント周波数格納部 104、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 は、必須ではない。かかることも、他の実施の形態においても同様である。

【 0 1 3 9 】

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

10

20

30

40

50

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。かかることも、他の実施の形態でも同様である。

【0140】

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。

10

【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。

20

【0142】

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）

30

【0143】

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。

【0144】

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。

40

【0145】

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。

【0146】

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。

【0147】

50

発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 は、最適状態決定手段 1 1 0 3 1 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 1 4 8 】

評定値算出手段 1 1 1 0 3 3 は、発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 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 1 1 1 0 3 3 は、例えば、最適状態決定手段 1 1 0 3 1 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。

10

【 0 1 4 9 】

発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2、および評定値算出手段 1 1 1 0 3 3 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 1 1 1 0 3 2 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 1 5 0 】

発声催促部 1 1 0 9 は、入力受付部 1 0 1 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。」とスピーカから音出力したりする処理である。発声催促部 1 1 0 9 は、ディスプレイ 3 4 4 やスピーカ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 1 1 0 9 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。

20

【 0 1 5 1 】

次に、本音声処理装置の動作について図 1 2 から図 1 4 のフローチャートを用いて説明する。図 1 2 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。

【 0 1 5 2 】

（ステップ S 1 2 0 1）発声催促部 1 1 0 9 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。

30

【 0 1 5 3 】

（ステップ S 1 2 0 2）音声受付部 1 0 3 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 1 2 0 3 に行き、音声を受け付けなければステップ S 2 1 3 に行く。

【 0 1 5 4 】

（ステップ S 1 2 0 3）サンプリング部 1 0 6 は、第一サンプリング周波数格納部 1 0 5 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 S 1 2 0 2 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。

40

【 0 1 5 5 】

（ステップ S 1 2 0 4）声道長正規化処理部 1 0 9 は、ステップ S 1 2 0 3 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 1 3 のフローチャートを用いて説明する。

【 0 1 5 6 】

（ステップ S 1 2 0 5）発声催促部 1 1 0 9 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。

【 0 1 5 7 】

（ステップ S 1 2 0 6）音声受付部 1 0 3 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 1 2 0 7 に行き、音声を受け付けなければステップ S 2 1 3

50

に行く。

【 0 1 5 8 】

(ステップ S 1 2 0 7) サンプリング部 1 0 6 は、第一サンプリング周波数格納部 1 0 5 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 S 1 2 0 6 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。

【 0 1 5 9 】

(ステップ S 1 2 0 8) 声道長正規化処理部 1 0 9 は、ステップ S 1 2 0 7 で得た第一音声データに対して、ステップ S 1 2 0 4 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。

【 0 1 6 0 】

(ステップ S 1 2 0 9) 音声処理部 1 1 1 0 は、ステップ S 1 2 0 8 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 1 4 のフローチャートを用いて説明する。ステップ S 1 2 0 2 に戻る。

【 0 1 6 1 】

なお、図 1 2 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含まれていても良い。

【 0 1 6 2 】

ステップ S 1 2 0 4 の第二サンプリング周波数算出処理について、図 1 3 のフローチャートを用いて説明する。図 1 3 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 S 3 0 1 からステップ S 3 0 5 の処理を行う。

【 0 1 6 3 】

ステップ S 1 2 0 9 の評定処理について、図 1 4 のフローチャートを用いて説明する。

【 0 1 6 4 】

(ステップ S 1 4 0 1) 発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 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 1 1 1 0 3 2 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。

【 0 1 6 5 】

(ステップ S 1 4 0 2) 発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 は、全フレームの最適状態確率値を算出する。

【 0 1 6 6 】

(ステップ S 1 4 0 3) 発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 は、j に 1 を代入する。

【 0 1 6 7 】

(ステップ S 1 4 0 4) 発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 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 S 1 4 0 5 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。

【 0 1 6 8 】

(ステップ S 1 4 0 5) 発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 は、カウンタ k に 1 を代入する。

【 0 1 6 9 】

(ステップ S 1 4 0 6) 発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 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 S 1 4 0 7 に行き、k 番目のフレームが存在しなければステップ S 1 4 1 0 に飛ぶ。

【 0 1 7 0 】

10

20

30

40

50

(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。

【0171】

(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。

【0172】

(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CREMENTする。ステップS1406に戻る。

【0173】

(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。

10

【0174】

(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。

【0175】

(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、INCREMENTする。ステップS1404に戻る。

20

【0176】

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。

【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、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。

【0178】

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。

30

【0179】

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、第二ブリング周波数「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。

【0180】

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。

40

【0181】

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。

【0182】

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。

【0183】

50

まず、フレーム区分手段 1 1 0 1 は、「right」の第二音声データを、短時間フレームに区分する。

【0184】

そして、フレーム音声データ取得手段 1 1 0 2 は、フレーム区分手段 1 1 0 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。

【0185】

次に、発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 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。

10

【0186】

次に、最適状態決定手段 1 1 0 3 1 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 1 0 3 1 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。

【0187】

次に、発音区間フレーム音韻確率値取得手段 1 1 1 0 3 2 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、評価値算出手段 1 1 1 0 3 3 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、評価値算出手段 1 1 1 0 3 3 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、発音区間毎の評価値を算出する。具体的には、評価値算出手段 1 1 1 0 3 3 は、数式 3 により評価値を算出する。数式 3 において、 $p-DAP(\tau)$ 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価値であり、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、算出される評価値である。また、数式 3 において、 $T(q_t^*)$ は、状態 q_t^* を含む HMM が含まれる評価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。

20

30

【数 3】

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

【数 4】

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

40

【0188】

かかる評価値算出手段 1 1 1 0 3 3 が算出した評価値（「 $t-p-DAP$ スコア」とも言う。）の出力例を、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価結果を示す。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価値が日本人男性の発音の評価値より高く、良好な評価結果が得られている。

【0189】

そして、出力手段 1 1 0 4 は、算出した発音区間ごと（ここでは、音素毎）の評価値を

50

、順次出力する。かかる出力例は、図 16 である。

【0190】

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。

【0191】

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

10

【0192】

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。

【0193】

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。

20

【0194】

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。

30

【0195】

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）

【0196】

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。

40

【0197】

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。

50

【 0 1 9 8 】

音声処理部 1 7 1 0 は、フレーム区分手段 1 1 0 1、フレーム音声データ取得手段 1 1 0 2、特殊音声検知手段 1 7 1 0 1、評価手段 1 7 1 0 3、出力手段 1 1 0 4 を具備する。

【 0 1 9 9 】

特殊音声検知手段 1 7 1 0 1 は、無音データ格納手段 1 7 1 0 1 1、無音区間検出手段 1 7 1 0 1 2 を具備する。

【 0 2 0 0 】

評価手段 1 7 1 0 3 は、最適状態決定手段 1 1 0 3 1、最適状態確率値取得手段 1 1 0 3 2、評価値算出手段 1 7 1 0 3 3 を具備する。

10

【 0 2 0 1 】

特殊音声検知手段 1 7 1 0 1 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 1 7 1 0 1 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 1 7 1 0 1 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 1 7 1 0 1 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

【 0 2 0 2 】

無音データ格納手段 1 7 1 0 1 1 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 1 7 1 0 1 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 2 0 3 】

無音区間検出手段 1 7 1 0 1 2 は、フレーム音声データ取得手段 1 1 0 2 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 1 7 1 0 1 1 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 1 7 1 0 1 2 は、フレーム音声データ取得手段 1 1 0 2 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 1 7 1 0 1 2 は、下記で述べる最適状態確率値取得手段 1 1 0 3 2 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 1 1 0 2 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 1 7 1 0 1 2 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 1 7 1 0 1 2 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。

30

【 0 2 0 4 】

評価手段 1 7 1 0 3 は、教師データと入力音声データと特殊音声検知手段 1 7 1 0 1 における検知結果に基づいて、音声受付部 1 0 3 が受け付けた音声の評価を行う。「特殊音声検知手段 1 7 1 0 1 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 1 7 1 0 1 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 1 7 1 0 1 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 1 7 1 0 1 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、評価値を所定数値分、減じて、評価値を算出することである。評価手段 1 7 1 0 3 は、通常、MPU やメモリ等から実現され得る。評価手段 1 7 1 0 3 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。

40

【 0 2 0 5 】

評価値算出手段 1 7 1 0 3 3 は、無音区間検出手段 1 7 1 0 1 2 が検出した無音区間を

50

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。

10

【0206】

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。

【0207】

（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。

20

【0208】

（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なかを判断する。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。

【0209】

（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。

【0210】

30

（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なかを判断する。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。

【0211】

（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。

【0212】

（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。

40

【0213】

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。

50

【0214】

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。

【0215】

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価値を算出するので、評価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。

【0216】

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。

10

【0217】

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、第二サンプリング周波数「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。

【0218】

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。

20

【0219】

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。

【0220】

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。

【0221】

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。

30

【0222】

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「 $O = o_1, o_2, \dots, o_T$ 」を算出する。

【0223】

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。

40

【0224】

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。

【0225】

次に、評価値算出手段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価値を算出する。つまり、評価値算出手段171033は、例えば、DAPスコアをフレーム毎に算出する。

【0226】

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価値（DA

50

P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価値が、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価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価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。

【0227】

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、Viterbiaライメントを示す。図20(a)において、網掛けのフレーム群における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。

10

【0228】

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における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、低い評価値が出力される。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。

20

【0229】

そして、出力手段1104は、出力する評価値から、無音データの区間の評価値を考慮しないように、無視する。

30

【0230】

そして、出力手段1104は、各フレームに対応する評価値を出力する。この場合、例えば、無音データの区間の評価値は、出力されない。

【0231】

かかる評価値の出力態様例は、例えば、図9、図10である。

【0232】

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。

【0233】

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(評価値)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、精度の高い評価ができる。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価するので、極めて正確な評価結果が得られる。

40

【0234】

なお、無音区間のデータは、無視して評価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、評価値を出力しても良い。

【0235】

また、本実施の形態の具体例によれば、DAPスコアを用いて、評価値を算出したが、無音の区間を考慮して評価値を算出すれば良く、上述した他のアルゴリズム(t-p-D

50

A P 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価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価を行い、特に、無音データを考慮して、評価値を算出すれば良い。

【0236】

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。

【0237】

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。

【0238】

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価を行う評価ステップと、前記評価ステップにおける評価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。

【0239】

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)

【0240】

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価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。

【0241】

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。

【0242】

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、評価手段21103、出力手段21104を具備する。なお、評価手段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。

【0243】

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価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 2 1 1 0 1 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 2 4 4 】

評定手段 2 1 1 0 3 は、特殊音声検知手段 2 1 1 0 1 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 2 1 1 0 3 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 2 1 1 0 1 が所定の条件を満たすことを検知した区間に無音が入音されたか否かを判断し、無音が入音されていない場合に、他の音素が入音されたことを検知しても良い。また、評定手段 2 1 1 0 3 は、無音が入音されていない場合に、他の音韻 H M M 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入音された、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 2 1 1 0 3 は、通常、M P U やメモリ等から実現され得る。評定手段 2 1 1 0 3 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。

10

【 0 2 4 5 】

出力手段 2 1 1 0 4 は、評定手段 2 1 1 0 3 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 2 1 1 0 4 は、ディスプレイやスピーカ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 2 1 1 0 4 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。

20

【 0 2 4 6 】

次に、音声処理装置の動作について、図 2 2、図 2 3 のフローチャートを用いて説明する。なお、図 2 2 のフローチャートは、図 1 2 のフローチャートと比較して、ステップ S 2 2 0 1 の評定処理のみが異なるので、図 2 2 のフローチャートの説明は省略する。ステップ S 2 2 0 1 の評定処理の詳細について、図 2 3 のフローチャートを用いて説明する。図 2 3 のフローチャートにおいて、図 2、図 1 9 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。

30

【 0 2 4 7 】

（ステップ S 2 3 0 1）特殊音声検知手段 2 1 1 0 1 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 S 1 9 0 2 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 S 2 3 0 7 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。

40

【 0 2 4 8 】

（ステップ S 2 3 0 2）特殊音声検知手段 2 1 1 0 1 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 S 2 3 0 7 に行き、データが格納されていなければステップ S 2 3 0 3 に行く。

【 0 2 4 9 】

（ステップ S 2 3 0 3）出力手段 2 1 1 0 4 は、ステップ S 1 9 0 1 で算出した評定値を出力する。

【 0 2 5 0 】

（ステップ S 2 3 0 4）特殊音声検知手段 2 1 1 0 1 は、カウンタ i を 1、インクリメントする。ステップ S 2 0 8 に戻る。

50

【0251】

(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価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。

【0252】

(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。

【0253】

(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。

【0254】

(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。

【0255】

(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。

【0256】

(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。

【0257】

(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。

【0258】

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

【0259】

(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。

【0260】

(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価値を出力する。ここで、全評価値とは、例えば、フレーム毎のDAPスコアである。ステップS2313に行く。

【0261】

(ステップS2315) 出力手段21104は、評価結果に「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。

【0262】

(ステップS2316) 出力手段21104は、評価結果を出力する。

【0263】

(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。

【0264】

なお、図23のフローチャートにおいて、評価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価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価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。

10

20

30

40

50

【0265】

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。

【0266】

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。

【0267】

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、第二サンプリング周波数「32.1 KHz」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。

10

【0268】

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。

【0269】

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「22.05 KHz」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。

20

【0270】

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「32.1 KHz」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。

【0271】

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。

【0272】

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「 $O = o_1, o_2, \dots, o_T$ 」を算出する。

30

【0273】

次に、評価手段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。

【0274】

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。

【0275】

40

次に、評価手段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価値を算出する。つまり、評価手段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。

【0276】

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価値を用いて、特殊な音声が入力されたか否かを判断する。つまり、評価値（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。

【0277】

50

次に、特殊音声検知手段 2 1 1 0 1 は、図 2 4 に示すように、評定値（例えば、D A P 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 2 3 で説明した。また、図 2 4 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。

【 0 2 7 8 】

次に、評定手段 2 1 1 0 3 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 2 1 1 0 3 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 2 1 1 0 4 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 2 5 は、評定結果の出力例である。なお、出力手段 2 1 1 0 4 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。

10

【 0 2 7 9 】

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。

【 0 2 8 0 】

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（D A P、t - p - D A P）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。

20

【 0 2 8 1 】

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。

30

【 0 2 8 2 】

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。

【 0 2 8 3 】

40

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）

【 0 2 8 4 】

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。

【 0 2 8 5 】

図 2 6 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

50

、入力受付部 101、教師データ格納部 102、音声受付部 103、教師データフォー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、評価対象者フォーマント周波数取得部 107、評価対象者フォー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。

【0286】

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。

10

【0287】

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。

【0288】

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。

20

【0289】

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。

30

【0290】

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。

40

【0291】

（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。

【0292】

（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

50

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、V i t e r b iアルゴリズムにより算出できる。

【0293】

(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。

【0294】

(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。

10

【0295】

(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。

【0296】

(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。

20

【0297】

(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。

【0298】

(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。

30

【0299】

(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S2808に戻る。

【0300】

(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。

【0301】

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。

40

【0302】

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。

【0303】

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、第二ブリング周波数「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。

【0304】

次に、発声催促部1109は、例えば、「"r i g h t"と発声してください。」と画面

50

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。

【0305】

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。

【0306】

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。

10

【0307】

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。

【0308】

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「 $O = o_1, o_2, \dots, o_T$ 」を算出する。

【0309】

次に、評価手段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。

20

【0310】

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。

【0311】

次に、評価手段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価値を算出する。つまり、評価手段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。

30

【0312】

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価値を用いて、特殊な音声が入力されたか否かを判断する。つまり、評価値（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。

【0313】

次に、特殊音声検知手段26101は、図29に示すように、評価値（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、一つの音素内で評価値が低ければ、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）に対する評価値（例えば、DAPスコア）を算出し、当該評価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して判断する。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、所定の値以上の評価値が得られる音韻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において縦軸は評価値であり、当該評価値は、DAP、t - p - DAP等、問わない。

40

【0314】

次に、評価手段26103は、音素の置換があった旨を示す評価結果（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、構成した評価結果を出力する。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

50

うに評価値を出力することが好適である。

【0315】

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（評価値）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、精度の高い評価ができる。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価結果が得られる。

【0316】

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、評価値の算出アルゴリズムは問わない。評価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。

10

【0317】

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。

【0318】

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。

20

【0319】

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価を行う評価ステップと、前記評価ステップにおける評価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。

30

【0320】

また、上記プログラムにおいて、特殊音声検知ステップは、一の音素の評価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価結果を構成する、ことは好適である。

（実施の形態6）

【0321】

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価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。

40

【0322】

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。

50

【 0 3 2 3 】

音声処理部 3 0 1 0 は、フレーム区分手段 1 1 0 1、フレーム音声データ取得手段 1 1 0 2、特殊音声検知手段 3 0 1 0 1、評価手段 3 0 1 0 3、出力手段 2 1 1 0 4 を具備する。なお、評価手段 3 0 1 0 3 は、最適状態決定手段 1 1 0 3 1、最適状態確率値取得手段 1 1 0 3 2 を具備する。

【 0 3 2 4 】

特殊音声検知手段 3 0 1 0 1 は、一の音素の評価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価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 3 0 1 0 1 は、一の音素の評価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価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 3 0 1 0 1 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 3 0 1 0 1 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 3 0 1 0 1 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 3 0 1 0 1 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

【 0 3 2 5 】

評価手段 3 0 1 0 3 は、特殊音声検知手段 3 0 1 0 1 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価結果を構成する。評価手段 3 0 1 0 3 は、通常、MPU やメモリ等から実現され得る。評価手段 3 0 1 0 3 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

【 0 3 2 6 】

次に、音声処理装置の動作について、図 3 1、図 3 2 のフローチャートを用いて説明する。なお、図 3 1 のフローチャートは、図 1 2 のフローチャートと比較して、ステップ S 3 1 0 1 の評価処理のみが異なるので、図 3 1 のフローチャートの説明は省略する。ステップ S 3 1 0 1 の評価処理の詳細について、図 3 2 のフローチャートを用いて説明する。図 3 2 のフローチャートにおいて、図 2、図 1 9、図 2 3、図 2 8 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。

30

【 0 3 2 7 】

（ステップ S 3 2 0 1）特殊音声検知手段 3 0 1 0 1 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 S 3 2 0 2 に行き、高くなければステップ S 2 8 1 0 に行く。なお、ステップ S 3 2 0 2 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。

【 0 3 2 8 】

（ステップ S 3 2 0 2）評価手段 3 0 1 0 3 は、音素の欠落があった旨を示す評価結果を構成する。ステップ S 2 8 0 8 に行く。

40

【 0 3 2 9 】

なお、図 3 2 のフローチャートにおいて、評価対象の音素（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。

【 0 3 3 0 】

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。

50

【0331】

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。

【0332】

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、第二サンプリング周波数「32.1 KHz」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。

【0333】

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。

10

【0334】

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「22.05 KHz」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。

【0335】

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「32.1 KHz」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。

20

【0336】

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。

【0337】

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「 $O = o_1, o_2, \dots, o_T$ 」を算出する。

【0338】

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。

30

【0339】

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。

【0340】

次に、評価手段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価値を算出する。つまり、評価手段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。

40

【0341】

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価値を用いて、特殊な音声が入力されたか否かを判断する。つまり、評価値（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。

【0342】

次に、特殊音声検知手段30101は、図33に示すように、評価値（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、一つの音素内で評価値が低ければ、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）に対する評価値（例えば、DAPスコ

50

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。

【0343】

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。

10

【0344】

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。

【0345】

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。

20

【0346】

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。

【0347】

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。

30

【0348】

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。

40

【0349】

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）

【0350】

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。

50

【 0 3 5 1 】

図 3 4 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。

【 0 3 5 2 】

本音声処理装置は、入力受付部 1 0 1、教師データ格納部 1 0 2、音声受付部 1 0 3、教師データフォルマント周波数格納部 1 0 4、第一サンプリング周波数格納部 1 0 5、サンプリング部 1 0 6、評価対象者フォルマント周波数取得部 1 0 7、評価対象者フォルマント周波数格納部 1 0 8、声道長正規化処理部 1 0 9、音声処理部 3 4 1 0、発声催促部 1 1 0 9 を具備する。

【 0 3 5 3 】

音声処理部 3 4 1 0 は、音声認識手段 3 4 1 0 1、出力手段 3 4 1 0 2 を具備する。

10

【 0 3 5 4 】

音声処理部 3 4 1 0 の音声認識手段 3 4 1 0 1 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 3 4 1 0 は、通常、M P U やメモリ等から実現され得る。音声処理部 3 4 1 0 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 3 5 5 】

出力手段 3 4 1 0 2 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 3 4 1 0 2 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 3 4 1 0 2 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。

20

【 0 3 5 6 】

次に、音声処理装置の動作について図 3 5 のフローチャートを用いて説明する。なお、図 3 7 のフローチャートにおいて、図 2、図 1 2 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。

【 0 3 5 7 】

（ステップ S 3 5 0 1）音声認識手段 3 4 1 0 1 は、ステップ S 1 2 0 8 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 3 4 1 0 1 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。

30

【 0 3 5 8 】

（ステップ S 3 5 0 2）出力手段 3 4 1 0 2 は、ステップ S 3 5 0 1 における音声認識結果を出力する。ステップ S 1 2 0 6 に戻る。

【 0 3 5 9 】

以上、本実施の形態によれば、精度高く音声認識できる。

【 0 3 6 0 】

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。

40

（実施の形態 8）

【 0 3 6 1 】

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

50

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、主として、音声（歌唱を含む）を評価する発音評定装置である、として説明する。さらに、本実施の形態において、上記の実施の形態で記載した音声処理装置よりもさらに精度高く、評価対象者の話者特性に応じた発音評定が可能な音声処理装置について説明する。具体的には、本実施の形態において、最小自乗誤差基準に基づく、簡潔な声道長正規化法に基づいて、評価対象者の話者特性に左右されにくい音声処理装置について説明する。

【0362】

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。

【0363】

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。

【0364】

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、第一サンプリング周波数格納部105、サンプリング部106、声道長正規化パラメータ算出部3601、声道長正規化パラメータ格納部3602、声道長正規化処理部3609、音声処理部110を具備する。

【0365】

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。

【0366】

声道長正規化パラメータ算出部3601は、周波数範囲指定情報格納手段36011、長時間ケプストラム平均ベクトル格納手段36012、第二ケプストラムベクトル系列算出手段36013、ケプストラム変換手段36014、ケプストラム変換パラメータ算出手段36015、最終ケプストラム変換パラメータ取得手段36016、声道長正規化パラメータ算出手段36017を具備する。

【0367】

声道長正規化パラメータ算出部3601は、教師データ格納部102の教師データおよび音声受付部103が受け付けた音声に基づいて、声道長正規化パラメータを算出する。声道長正規化パラメータとは、評価対象者の話者特性（声道長の違い）を吸収するためのサンプリング周波数変換率である。なお、本実施の形態において、まず、算出されるのは、ケプストラム変換（ケプストラムワーピング）パラメータである。このケプストラム変換パラメータは、声道長の変換を直接表わすものではないため、声道長正規化パラメータ算出部3601は、後述の近似変換式を用いて最終的な声道長正規化パラメータ（サンプリング周波数変換率）を算出する。声道長正規化パラメータ算出部3601は、通常、MPUやメモリ等から実現され得る。声道長正規化パラメータ算出部の処理手順は、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。

【0368】

周波数範囲指定情報格納手段36011は、周波数範囲を指定する情報である周波数範囲指定情報（ W ）を格納している。周波数範囲指定情報（ W ）は、後述する最適なケプストラム変換パラメータ（ α ）を算出する場合に、1次オールパス関数等の双1次変換による周波数ワーピングが線形の周波数伸縮で近似できる範囲に、周波数範囲を限定するための情報である。かかる周波数範囲は、0周波数からナイキストレートの $1/3$ から $1/2$ 程度が好適である。ただし、双1次変換は、スペクトル領域ではなく、ケプストラム領域で行われるため、周波数範囲指定情報（ W ）は、例えば、以下に述べる行列の情報であることが好適である。周波数範囲指定情報は、例えば、図示しない周波数範囲指定情報算出手段により、以下のように算出される。サンプリング周波数を F_s （Hz）、指定される周波数範囲の最高周波数を F_{max} （Hz）とし、 N を十分大きな自然数（512や10

10

20

30

40

50

24など)、「 $N_m = N \times F_{max} / F_s$ 」において、ケプストラムベクトルに対する周波数範囲指定行列 W の (i, j) 成分を、周波数範囲指定情報算出手段は、以下の数式5に従って計算する。具体的には、周波数範囲指定情報算出手段は、コンピュータの記録媒体(図示しない)に格納されているサンプリング周波数(F_s)、最高周波数(F_{max})、予め決められた十分大きな自然数(N)を読み出す。そして、周波数範囲指定情報算出手段は、自ら保持している演算式の情報「 $N_m = N \times F_{max} / F_s$ 」を読み出し、読み出した F_s 、 F_{max} 、 N を演算式に代入し、 N_m を算出する。そして、周波数範囲指定情報算出手段は、格納している以下の数式5の情報を読み出し、 i 、 j を0から順に、1ずつインクリメントさせながら、ループ処理(2重ループの処理になる)により、 $\{W\}_{i,j}$ を算出する。そして、周波数範囲指定情報算出手段は、算出した $\{W\}_{i,j}$ のすべてを、少なくとも一時的に周波数範囲指定情報格納手段36011に格納する。なお、数式5において、「 k 」は、周波数インデクスであり、「 k 」の範囲は、「 $k = 0, 1, 2, \dots, N/2$ 」である。また、「 n 」は、離散時間インデクスであり、「 n 」の範囲は、「 $n = \dots - 2, -1, 0, 1, 2, \dots$ 」である。

【数5】

$$\{W\}_{i,j} = \begin{cases} w_i & (j=0) \\ w_{|i-j|} + w_{i+j} & (j>0) \end{cases} \quad (i, j=0, 1, 2, \dots, M)$$

$$w_n = \frac{2}{N} \sum_{k=0}^{N_m} \cos\left(\frac{2\pi kn}{N}\right) - \frac{1}{N}$$

【0369】

なお、周波数範囲指定情報のデータ構造は問わない。また、周波数範囲指定情報格納手段36011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。

【0370】

長時間ケプストラム平均ベクトル格納手段36012は、長時間ケプストラム平均ベクトル(μ)を格納している。長時間ケプストラム平均ベクトル(μ)は、教師データを構成するデータから、短区間分析により算出された第一のケプストラムベクトル系列の時間平均である。第一のケプストラムベクトル系列(x_t ($t = 1, 2, \dots, T_0$))は、通常、教師データを構成する単一音素(例えば、/u/)を短区間分析し、取得される。そして、長時間ケプストラム平均ベクトル(μ)は、ベクトル(x_t)から、以下の数式6により算出される。

【数6】

$$\mu = \frac{1}{T_0} \sum_{t=1}^{T_0} x_t$$

【0371】

なお、ケプストラムベクトルは0次係数も含めた $M+1$ 次元であり、ベクトル(x_t)およびベクトル(μ)は、それぞれ数式7、数式8で表わされる。

【数7】

$$X_t = [x_{t,0} \ x_{t,1} \ x_{t,2} \ \dots \ x_{t,M}]^T$$

【数8】

$$\mu = [\mu_0 \ \mu_1 \ \mu_2 \ \dots \ \mu_M]^T$$

【0372】

数式7、数式8において、 $(\dots)^T$ は行列またはベクトルの転置を表わす。

【0373】

10

20

30

40

50

また、第一のケプストラムベクトル系列 (x_t) は、図示しない第一ケプストラムベクトル系列算出手段が、教師データを構成するデータ (単一音素 (例えば、/u/)) から、短区間分析により算出して良い。

【0374】

また、図示しない長時間ケプストラム平均ベクトル取得手段が、第一のケプストラムベクトル系列 (x_t) の時間平均を、数式6により算出し、長時間ケプストラム平均ベクトル (μ) を取得しても良い。長時間ケプストラム平均ベクトル格納手段36012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。

【0375】

第二ケプストラムベクトル系列算出手段36013は、音声受付部103が受け付けた音声 (通常、単一音素 (例えば、/u/)) から、短区間分析により第二のケプストラムベクトル系列 (C_t) を算出する。第二ケプストラムベクトル系列算出手段36013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から、第二のケプストラムベクトル系列 (C_t) を算出して良く、かかる場合も音声受付部103が受け付けた音声から、短区間分析により第二のケプストラムベクトル系列 (C_t) を算出した、と言える。なお、音素を短区間分析し、ケプストラムベクトル系列を算出する処理は、公知技術による処理であるので、説明は省略する。第二ケプストラムベクトル系列算出手段36013は、通常、MPUやメモリ等から実現され得る。第二ケプストラムベクトル系列算出手段36013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。なお、第二ケプストラムベクトル系列算出手段36013が処理する音声であり音声受付部103が受け付ける音素 (例えば、/u/) は、長時間ケプストラム平均ベクトルの元になる音声と同一の音素 (例えば、/u/) である。また、この第二のケプストラムベクトル系列 (C_t) は、やはり0次係数も含めた $M+1$ 次元ベクトルとして数式9で表わされる。

【数9】

$$C_t = [c_{t,0} c_{t,1} c_{t,2} \dots c_{t,M}]^T$$

【0376】

また、第二ケプストラムベクトル系列算出手段36013は、後述するケプストラム変換パラメータ算出手段36015の処理のために、算出した第二のケプストラムベクトル系列 (C_t) に対して、以下の数式10の処理を行い、ベクトル (C_t^-) ($t=1, 2, \dots, T$) を取得することは好適である。

【数10】

$$\bar{C}_t = [\bar{c}_{t,0} \bar{c}_{t,1} \bar{c}_{t,2} \dots \bar{c}_{t,M}]^T$$

$$\bar{c}_{t,0} = -c_{t,1}$$

$$\bar{c}_{t,n} = (n-1)c_{t,n-1} - (n+1)c_{t,n+1} \quad (n=1, 2, \dots, M-1)$$

$$\bar{c}_{t,M} = (M-1)c_{t,M-1} - (M+1)c_{t,M}$$

(なお、 $\bar{C}_t$ は、 C_t^- と同じである。)

【0377】

ケプストラム変換手段36014は、第二のケプストラムベクトル系列 (C_t) を、ケプストラム変換パラメータ (α) を要素とする行列 ($F(\alpha)$) を用いて線形変換し、周波数ワープされた第三のケプストラムベクトル系列 (O_t) を算出する。具体的には、まず、ケプストラム変換手段38014は、ケプストラム変換パラメータ (α) の初期値 (α^-) を設定する。 (α^-) は α の最適値の近似値であることが望ましく、通常は $\alpha=0$ とおくが、例えば、最適値が「 $\alpha > 0$ 」であると予想できる場合は、小さな正の値の α でも良い。なお、初期値 (α^-) は、ケプストラム変換手段38014が、予め記憶媒体や

メモリ等に格納している。次に、ケプストラム変換手段 38014 は、与えられたケプストラム変換パラメータ (α) をパラメータとして、以下の数式 11 により、周波数ワープされた第三のケプストラムベクトル系列 (O_t) を算出する。なお、ケプストラム変換手段 38014 は、最初に、($O_t(\alpha^*)$) を算出する。

【数 11】

$$O_t(\alpha) = F(\alpha)C_t$$

【0378】

ケプストラム変換手段 36014 は、通常、MPU やメモリ等から実現され得る。ケプストラム変換手段 36014 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

【0379】

ケプストラム変換パラメータ算出手段 36015 は、長時間ケプストラム平均ベクトル (μ) および第三のケプストラムベクトル系列 (O_t) に基づいて、ケプストラム変換パラメータ (α) を算出する。ケプストラム変換パラメータ算出手段 36015 は、さらに好ましくは、周波数範囲指定情報 (W) で示される周波数範囲における長時間ケプストラム平均ベクトルおよび第三のケプストラムベクトル系列 (O_t) に基づいて、ケプストラム変換パラメータを算出する。

【0380】

20

具体的には、まず、ケプストラム変換パラメータ算出手段 36015 は、以下の数式 12 により、ベクトル ($u_t(\alpha)$) を算出する。そして、次に、ケプストラム変換パラメータ算出手段 36015 は、以下の数式 13 により、 α の最適値 (α^*) を算出する。なお、 α の最適値 (α^*) は、現繰り返しステップにおける最適値である。

【数 12】

$$u_t(\alpha) = O_t(\alpha) + \alpha \bar{C}_t$$

【数 13】

$$\alpha^* = \frac{\sum_{t=1}^T (u_t(\alpha) - \mu)^T W^T W \bar{C}_t}{\sum_{t=1}^T \bar{C}_t^T W^T W \bar{C}_t}$$

30

【0381】

ケプストラム変換パラメータ算出手段 36015 は、通常、MPU やメモリ等から実現され得る。ケプストラム変換パラメータ算出手段 36015 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。

【0382】

最終ケプストラム変換パラメータ取得手段 36016 は、所定のルールに基づいて、ケプストラム変換手段 36014 における処理、およびケプストラム変換パラメータ算出手段 36015 における処理を繰り返えさせ、最終的な最適なケプストラム変換パラメータを得る。ここで、所定のルールとは、例えば、予め決められた所定の繰り返し回数だけ、処理を繰り返し行われたことである。また、所定のルールとは、例えば、 α の最適値 (α^*) の変化が所定の値より小さくなった (先の (α^*) と今回の (α^*) の差が閾値以下など) ことである。また、所定のルールとは、その他のルールでも良い。最終ケプストラム変換パラメータ取得手段 36016 は、通常、MPU やメモリ等から実現され得る。最終ケプストラム変換パラメータ取得手段 36016 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。

40

【0383】

50

声道長正規化パラメータ算出手段 36017 は、最終ケプストラム変換パラメータ取得手段 36016 が得たケプストラム変換パラメータに基づいて、声道長正規化パラメータ (γ) を算出する。さらに具体的には、声道長正規化パラメータ算出手段 36017 は、まず、最終ケプストラム変換パラメータ取得手段 36016 が得たケプストラム変換パラメータから、例えば、以下の数式 14 に従って、線形周波数伸縮比 (ρ) を算出する。

【数 14】

$$\rho = 1 + 2 \arctan \left(\frac{\sqrt{3} \alpha^*}{2 - \alpha^*} \right)$$

【0384】

10

次に、声道長正規化パラメータ算出手段 36017 は、当該線形周波数伸縮比 (ρ) から、例えば、以下の数式 15 に従って、声道長正規化パラメータ (γ) を算出する。

【数 15】

$$\gamma = \frac{1}{\rho}$$

【0385】

声道長正規化パラメータ算出手段 36017 は、通常、MPU やメモリ等から実現され得る。声道長正規化パラメータ算出手段 36017 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

【0386】

声道長正規化パラメータ格納部 3602 は、サンプリング周波数の変換率に関する情報である声道長正規化パラメータ (γ) を格納している。声道長正規化パラメータ (γ) は、声道長正規化パラメータ算出部 3601 が取得した声道長正規化パラメータである。声道長正規化パラメータ格納部 3602 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。なお、声道長正規化パラメータ算出部 3601 は、ここでは、算出した声道長正規化パラメータ (γ) を、声道長正規化パラメータ格納部 3602 に蓄積する処理も行う。

【0387】

30

声道長正規化処理部 3609 は、声道長正規化パラメータ (γ) と第一サンプリング周波数をパラメータとして算出される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。「音声受付部 103 が受け付けた音声に対して」とは、サンプリング部 106 がサンプリングして取得した第一音声データに対して、第二サンプリング周波数でリサンプリング処理を行い、第二音声データを得ることも含むし、サンプリング部 106 が第一音声データを取得することなく、直接的に音声受付部 103 が受け付けた音声に対して、第二サンプリング周波数でサンプリング処理し、第二音声データを得ることも含む。声道長正規化処理部 3609 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 3609 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。

40

【0388】

次に、本音声処理装置の動作について説明する。まず、本音声処理装置における、声道長正規化パラメータを算出する処理について、図 37 のフローチャートを用いて説明する。なお、本声道長正規化パラメータ算出処理は、必ずしも音声処理装置で行う必要はない。

【0389】

(ステップ S3701) 音声処理装置は、初期化処理を行う。初期化処理とは、例えば、ユーザ (評価対象者) に対して、「 / u / 」と発声するように促す処理 (例えば、ディスプレイに「 / う」と発音してください。 」と表示する処理) と、周波数範囲指定情報格

50

納手段 3 6 0 1 1 の周波数範囲指定情報、および長時間ケプストラム平均ベクトル格納手段 3 6 0 1 2 の長時間ケプストラム平均ベクトルを読み出す処理である。

【 0 3 9 0 】

(ステップ S 3 7 0 2) 音声受付部 1 0 3 が、評価対象者からの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 3 7 0 3 に行き、音声を受け付けなければステップ S 3 7 0 2 に戻る。

【 0 3 9 1 】

(ステップ S 3 7 0 3) サンプリング部 1 0 6 は、ステップ S 3 7 0 2 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て、メモリ上に、少なくとも一時格納する。なお、音声をサンプリングする処理は公知技術である。

10

【 0 3 9 2 】

(ステップ S 3 7 0 4) 第二ケプストラムベクトル系列算出手段 3 6 0 1 3 は、ステップ S 3 7 0 3 で得た第一音声データを取得し、当該第一音声データから、短区間分析により第二のケプストラムベクトル系列 (C_t) を算出し、第二のケプストラムベクトル系列 (C_t) をメモリ上に、少なくとも一時格納する。

【 0 3 9 3 】

(ステップ S 3 7 0 5) 第二ケプストラムベクトル系列算出手段 3 6 0 1 3 は、ステップ S 3 7 0 4 で算出した第二のケプストラムベクトル系列 (C_t) を取得し、当該第二のケプストラムベクトル系列 (C_t) から、ベクトル (C_t^*) ($t = 1, 2, \dots, T$) を取得し、ベクトル (C_t^*) をメモリ上に、少なくとも一時格納する。

20

【 0 3 9 4 】

(ステップ S 3 7 0 6) ケプストラム変換手段 3 6 0 1 4 は、予め格納しているケプストラム変換パラメータの初期値 (α^*) を読み出し、変数 (α) に (α^*) を設定する。

【 0 3 9 5 】

(ステップ S 3 7 0 7) ケプストラム変換手段 3 6 0 1 4 は、第二のケプストラムベクトル系列 (C_t) を読み出し、当該第二のケプストラムベクトル系列 (C_t) を、ケプストラム変換パラメータ (α) を要素とする行列を用いて線形変換し、周波数ワープされた第三のケプストラムベクトル系列 (O_t) を算出し、メモリ上に、少なくとも一時格納する。

【 0 3 9 6 】

(ステップ S 3 7 0 8) ケプストラム変換パラメータ算出手段 3 6 0 1 5 は、ステップ S 3 7 0 7 で算出した第三のケプストラムベクトル系列 (O_t)、およびステップ S 3 7 0 5 で算出したベクトル (C_t^*) を読み出し、当該第三のケプストラムベクトル系列 (O_t) およびベクトル (C_t^*) から、ベクトル ($u_t(\alpha)$) を算出する (数式 1 2 参照)。ベクトル ($u_t(\alpha)$) を算出する場合に、格納している数式 1 2 の情報を読み出して、演算することは言うまでもない。

30

【 0 3 9 7 】

(ステップ S 3 7 0 9) ケプストラム変換パラメータ算出手段 3 6 0 1 5 は、予め格納している数式 1 3 の情報を読み出し、当該数式 1 3 の数式に、ベクトル ($u_t(\alpha)$)、長時間ケプストラム平均ベクトル (μ)、ベクトル (C_t^*)、周波数範囲指定情報 (W) の情報 (W^T 等も含む) を与え、数式 1 3 を演算し、 α の最適値 (α^*) を算出する。この α の最適値 (α^*) は、本ループにおける最適値である。そして、ケプストラム変換パラメータ算出手段 3 6 0 1 5 は、 α の最適値 (α^*) を、少なくともメモリに一時格納する。

40

【 0 3 9 8 】

(ステップ S 3 7 1 0) 最終ケプストラム変換パラメータ取得手段 3 6 0 1 6 は、予め決められた所定のルール (ルールの情報は、予め格納されている) に合致するか否かを判断する。ルールに合致すればステップ S 3 7 1 1 に行き、ルールに合致しなければステップ S 3 7 1 3 に行く。なお、ルールとは、上述したように、例えば、予め決められた所定の繰り返し回数 (この回数の情報は、予めメモリ等に格納されている) だけ、本ループ処

50

理 (α の最適値 (α^*) を算出し、 α^* を α に代入する処理) が繰り返し行われたことである。

【0399】

(ステップ S3711) 声道長正規化パラメータ算出手段 36017 は、最終ケプストラム変換パラメータ取得手段 36016 が得たケプストラム変換パラメータ (最終の α^*) を取得し、当該パラメータを数式 14 の演算式に与え、線形周波数伸縮比 (ρ) を算出する。なお、数式 14 の演算式の情報、声道長正規化パラメータ算出手段 36017 が予め格納している。また、声道長正規化パラメータ算出手段 36017 は、算出した線形周波数伸縮比 (ρ) を、少なくともメモリに一時格納する。

【0400】

(ステップ S3712) 声道長正規化パラメータ算出手段 36017 は、線形周波数伸縮比 (ρ) を読み出し、当該線形周波数伸縮比 (ρ) から、声道長正規化パラメータ (γ) を算出し、少なくともメモリに一時格納する。声道長正規化パラメータ算出手段 36017 は、例えば、予め格納している数式 15 の情報を読み出し、線形周波数伸縮比 (ρ) を代入し、数式 15 の演算式を実行し、声道長正規化パラメータ (γ) を得る。

【0401】

(ステップ S3713) 最終ケプストラム変換パラメータ取得手段 36016 は、ステップ S3709 で算出した (α^*) を α に代入する。そして、ステップ S3707 に戻る。

【0402】

本音声処理装置の動作 (音声の評定処理) について、図 38、図 39 のフローチャートを用いて説明する。図 38 のフローチャートにおいて、図 2 において説明した音声処理装置の動作と比較して、ステップ S204 の声道長正規化処理の内容のみが異なる。

【0403】

次に、図 39 のフローチャートを用いて声道長正規化処理の内容について説明する。

【0404】

(ステップ S3901) 声道長正規化処理部 3609 は、第一サンプリング周波数を、第一サンプリング周波数格納部 105 から読み出す。

【0405】

(ステップ S3902) 声道長正規化処理部 3609 は、声道長正規化パラメータ (γ) を、声道長正規化パラメータ格納部 3602 から読み出す。

【0406】

(ステップ S3903) 声道長正規化処理部 3609 は、予め決められた第二サンプリング周波数 (F_2) を算出する演算式の情報を読み出し、当該演算式の情報に、声道長正規化パラメータ (γ) と第一サンプリング周波数 (F_1) をパラメータとして代入し、演算式を実行し、第二サンプリング周波数 (F_2) を算出する。予め決められた第二サンプリング周波数 (F_2) を算出する演算式は、例えば、($F_2 = F_1 \times \gamma$) である。

【0407】

(ステップ S3904) 声道長正規化処理部 3609 は、第二サンプリング周波数で、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。

【0408】

なお、図 38 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。

【0409】

以下、図 40 の概念図を用いて、本実施の形態における音声処理装置における声道長正規化パラメータ算出処理の概念について説明する。声道長正規化パラメータの算出は、システム設計用に用いられる話者 (基準話者) のある音素 (教師データのある音素) の平均ケプストラムベクトル μ と、同じ音素のユーザ発話音声の変換されたケプストラムベクトル O_t との自乗誤差が最小になるように求められる。ただし、求まるパラメータ (α ?) はケプストラム変換 (ケプストラムワーピング) パラメータであり、このままでは声道長

10

20

30

40

50

の変換を直接表わすものではないため、近似変換式 ($1/\rho$) を用いて最終的な声道長正規化パラメータ (サンプリング周波数変換率) γ を計算する。なお、声道長正規化パラメータを算出するための数式は、上記の数式 5 から数式 15 により行われる。

【0410】

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、特に、精度の高い評定ができる。

【0411】

なお、本実施の形態によれば、音声処理装置が声道長正規化パラメータを算出する処理と、声道長正規化処理を行った。しかし、声道長正規化パラメータを算出する処理と声道長正規化処理を、異なる装置が行う構成でも良い。かかる場合、本音声処理装置は、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、サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している声道長正規化パラメータ格納部と、前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出される第二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。また、声道長正規化パラメータを算出する装置は、教師データを構成するデータから、短区間分析により算出された第一のケプストラムベクトル系列の時間平均である長時間ケプストラム平均ベクトルを格納している長時間ケプストラム平均ベクトル格納手段と、受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出手段と、前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列を用いて線形変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換手段と、前記長時間ケプストラム平均ベクトルおよび前記第三のケプストラムベクトル系列に基づいて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段を具備する装置である。

【0412】

また、本実施の形態によれば、ケプストラム変換パラメータの算出の際に、周波数範囲を限定したが、かかる処理は必須ではない。かかることは、実施の形態 9、実施の形態 10 においても同様である。

【0413】

また、本実施の形態によれば、音声処理部は、DAP に基づき発音評定を行った。しかし、他のアルゴリズムにより、発音評定を行っても良い。他のアルゴリズムとは、例えば、実施の形態 2 で述べた $t-p$ -DAP や、実施の形態 3 で述べた無音区間を考慮した類似度評定や、実施の形態 4 で述べた音韻の挿入を考慮した類似度評定や、実施の形態 5 で述べた音韻の置換を考慮した類似度評定や、実施の形態 6 で述べた音韻の欠落を考慮した類似度評定等である。かかることも、実施の形態 9、実施の形態 10 においても同様である。

【0414】

また、本実施の形態によれば、音声処理装置における音声処理部は、主として、発音評定を行ったが、音声処理部は、第二音声データに基づいて音声認識処理を行っても良い。かかることも、実施の形態 9、実施の形態 10 においても同様である。

【0415】

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

10

20

30

40

50

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、サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している声道長正規化パラメータ格納部と、前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出される第二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。かかることも、実施の形態9、実施の形態10においても同様である。

【0416】

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、音声を受け付ける音声受付ステップと、格納している声道長正規化パラメータと格納している第一サンプリング周波数をパラメータとして算出される第二サンプリング周波数で、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。

10

【0417】

また、上記プログラムにおいて、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以上格納しており、前記教師データおよび前記音声受付ステップで受け付けた音声に基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出ステップとをさらにコンピュータに実行させ、前記格納している声道長正規化パラメータは、前記声道長正規化パラメータ算出ステップで算出した声道長正規化パラメータであることは好適である。

20

(実施の形態9)

【0418】

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、主として、音声(歌唱を含む)を評価する発音評定装置である、として説明する。さらに、本実施の形態において、上記の実施の形態で記載した音声処理装置よりもさらに精度高く、評価対象者の話者特性に応じた発音評定が可能な音声処理装置について説明する。具体的には、本実施の形態において、出現確率最大化基準に基づく、簡潔な声道長正規化法に基づいて、評価対象者の話者特性に左右されにくい音声処理装置について説明する。なお、本実施の形態における音声処理装置は、実施の形態8の音声処理装置と比較して、声道長正規化パラメータの算出アルゴリズムが異なる。

30

【0419】

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。

【0420】

図41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。

【0421】

本音声処理装置は、図36における音声処理装置と比較して、声道長正規化パラメータ算出部4101が異なる。

40

【0422】

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。

【0423】

声道長正規化パラメータ算出部4101は、周波数範囲指定情報格納手段41011、学習音響データ格納手段41012、第二ケプストラムベクトル系列算出手段36013、ケプストラム変換手段36014、占有度数算出手段41013、ケプストラム変換パ

50

ラメータ算出手段 4 1 0 1 5、最終ケプストラム変換パラメータ取得手段 3 6 0 1 6、声道長正規化パラメータ算出手段 3 6 0 1 7を具備する。なお、ここでの最終ケプストラム変換パラメータ取得手段 3 6 0 1 6は、所定のルールに基づいて、ケプストラム変換手段 3 6 0 1 4における処理、ケプストラム変換パラメータ算出手段 4 1 0 1 5における処理だけでなく、占有度数算出手段 4 1 0 1 3における処理をも繰り返えさせ、最終的な最適なケプストラム変換パラメータを得る。

【 0 4 2 4 】

周波数範囲指定情報格納手段 4 1 0 1 1は、周波数範囲を指定する情報である周波数範囲指定情報 (W) を格納している。周波数範囲指定情報 (W) は、後述する最適なケプストラム変換パラメータ (α) を算出する場合に、1次オールパス関数等の双1次変換による周波数ワーピングが線形の周波数伸縮で近似できる範囲に、周波数範囲を限定するための情報である。かかる周波数範囲は、0周波数からナイキストレートの1/3から1/2程度が好適である。ただし、双1次変換は、スペクトル領域ではなく、ケプストラム領域で行われるため、周波数範囲指定情報 (W) は、例えば、以下に述べる行列の情報であることが好適である。周波数範囲指定情報 (W) は、例えば、図示しない周波数範囲指定情報算出手段により、以下のように算出される。サンプリング周波数を F_s (Hz)、指定される周波数範囲の最高周波数を F_{max} (Hz) とし、「 $\omega_{max} = 2\pi F_{max} / F_s$ 」と置いて、ケプストラムベクトルに対する周波数範囲指定行列 W の (i, j) 成分を、周波数範囲指定情報算出手段は、以下の数式 16 に従って計算する。具体的には、周波数範囲指定情報算出手段は、コンピュータの記録媒体 (図示しない) に格納されているサンプリング周波数 (F_s)、最高周波数 (F_{max}) を読み出す。そして、周波数範囲指定情報算出手段は、自ら保持している演算式の情報「 $\omega_{max} = 2\pi F_{max} / F_s$ 」を読み出し、 ω_{max} を算出する。そして、周波数範囲指定情報算出手段は、格納している以下の数式 16 の情報を読み出し、i、j を 0 から順に、1 ずつインクリメントさせながら、ループ処理 (2 重ループの処理になる) により、 $\{W\}_{i,j}$ を算出する。そして、周波数範囲指定情報算出手段は、算出した $\{W\}_{i,j}$ のすべてを、少なくとも一時的に周波数範囲指定情報格納手段 4 1 0 1 1 に格納する。

【 数 1 6 】

$$\{W\}_{ij} = \begin{cases} w_i & (j=0) \\ w_{|i-j|} + w_{i+j} & (j>0) \end{cases} \quad (i, j=0, 1, 2, \dots, M)$$

$$w_n = \begin{cases} \frac{\omega_{max}}{\pi} & (n=0) \\ \frac{\sin(n\omega_{max})}{n\pi} & (n \neq 0) \end{cases}$$

【 0 4 2 5 】

なお、周波数範囲指定情報のデータ構造は問わない。また、周波数範囲指定情報格納手段 4 1 0 1 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 4 2 6 】

学習音響データ格納手段 4 1 0 1 2 は、学習音響データを格納している。学習音響データとは、以下のように生成される。つまり、図示しない学習音響データ生成手段が、教師データを構成するデータ (例えば、教師データ格納部 1 0 2 のデータ) から、話者の音声のケプストラムベクトル系列を短区間分析で求め、それを用いて 2 以上の基準話者音素 HMM を取得し、メモリ上に、少なくとも一時格納する。そして、学習音響データ生成手段は、2 以上の基準話者音素 HMM を声道長正規化用の指定された発話内容 (例えば / あいうえお /) に従って連結し、連結 HMM を生成する。なお、ケプストラムベクトルは 0 次係数も含めた $M + 1$ 次元である。また、連結 HMM の j 番目状態における m 番目ガウス分

布成分の平均ベクトルおよび共分散行列を、それぞれ $\mu_{j,m}$ 、 $\Sigma_{j,m}$ とすると、 $\mu_{j,m}$ は、以下の数式 17 で表される。

【数 17】

$$\mu_{j,m} = [\mu_{j,m,0} \mu_{j,m,1} \mu_{j,m,2} \cdots \mu_{j,m,M}]^T$$

【0427】

なお、 $(\cdots)^T$ は、行列またはベクトルの転置を示す。

【0428】

そして、図示しない学習音響データ生成手段は、生成した学習音響データを、学習音響データ格納手段 41012 に蓄積する。

10

【0429】

なお、学習音響データは、連結 HMM が好適であるが、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。また、学習音響データ格納手段 41012 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。

【0430】

占有度数算出手段 41013 は、占有度数 ($y_t(j,m)$) を算出し、メモリ上に書き込み、少なくとも一時的に保持する。占有度数 ($y_t(j,m)$) とは、 t 番目のフレームが j 番目状態の m 番目ガウス成分から生成された事後確率である。 t 番目のフレームとは、例えば、音声受付部 103 が、声道長正規化用の指定された発話内容 (例えば / あいうえお /) の発声をユーザから受け付け、当該受け付けた音声から得られた t 番目のフレームである。占有度数算出手段 41013 は、ケプストラム変換手段 36014 が算出した第三のケプストラムベクトル系列 (O_t) と学習音響データ格納手段 41012 の学習音響データを用いて、占有度数 ($y_t(j,m)$) を算出する。占有度数算出手段 41013 は、さらに好ましくは、第三のケプストラムベクトル系列 (O_t) と学習音響データと周波数範囲指定情報格納手段 41011 の周波数範囲指定情報 (W) を用いて、占有度数 ($y_t(j,m)$) を算出する。占有度数算出手段 41013 は、占有度数 ($y_t(j,m)$) を、フォワード・バックワードアルゴリズムにより算出する。フォワード・バックワードアルゴリズムは、公知技術であるので、詳細な説明を省略する。

20

【0431】

さらに具体的には、占有度数算出手段 41013 は、以下の処理により、占有度数 ($y_t(j,m)$) を算出する。ここで、教師データの音声から学習される学習音素 HMM を、声道長正規化パラメータ推定のために指定した発話内容 (例えば / あいうえお /) にしたがって連結した連結 HMM を Λ とする。そして、ケプストラム変換パラメータ推定処理に使用される占有度数 ($y_t(j,m)$) は、第三ケプストラム系列を周波数範囲指定情報 W により周波数範囲指定されたケプストラムベクトル系列「 $Wo_1(\alpha)$, $Wo_2(\alpha)$, $\dots$, $Wo_T(\alpha)$ 」および Λ が与えられたときの t 番目フレームが Λ の j 番目状態の m 番目ガウス成分から生じた事後確率として定義されるので、占有度数 ($y_t(j,m)$) は、数式 18 で定義される。

30

【数 18】

$$\gamma_t(j,m) = P(s_t=j, m_t=m | Wo_1(\alpha), Wo_2(\alpha), \dots, Wo_T(\alpha), \Lambda)$$

40

【0432】

また、数式 18 における占有度数 ($y_t(j,m)$) は、前向き尤度 $A_t(j)$ および後向き尤度 $B_t(j)$ を使って数式 19 のように計算される。つまり、占有度数算出手段 41013 は、数式 19 で示される式の情報を保持しており、与えられるケプストラムベクトル系列、および Λ を用いて、占有度数 ($y_t(j,m)$) を算出する。

【数 19】

$$\gamma_t(j, m) = \frac{A_t(j)B_t(j)}{\sum_j A_t(j)} \cdot \frac{v_{j,m}b_{j,m}(Wo_t(\alpha))}{\sum_k v_{j,k}b_{j,k}(Wo_t(\alpha))}$$

$$A_t(j) = P(Wo_1(\alpha) \dots Wo_t(\alpha), s_t=j|\Lambda)$$

$$B_t(j) = P(Wo_{t+1}(\alpha) \dots Wo_T(\alpha) | s_t=j, \Lambda)$$

【0433】

なお、占有度数算出手段41013は、数式19の $A_t(j)$ および $B_t(j)$ を、ケプストラムベクトル系列「 $Wo_1(\alpha) \dots Wo_T(\alpha)$ 」および Λ から公知のforward-backwardアルゴリズムを用いて算出する。

10

【0434】

また、 $v_{j,m}$ は、 j 番目状態の m 番目ガウス成分に対する重み係数である。さらに、占有度数算出手段41013は、出力分布 $b_{j,m}(Wo_t(\alpha))$ を、数式20を用いて算出する。占有度数算出手段41013は、数式20の情報を予め保持しており、当該情報を読み出して、演算を実行し、出力分布 $b_{j,m}(Wo_t(\alpha))$ を得る。

【数 20】

$$b_{j,m}(Wo_t(\alpha)) = \frac{1}{\sqrt{(2\pi)^{M+1} | \sum_{j,m} |}} \exp \left\{ - \frac{(o_t(\alpha) - \mu_{j,m})^T W^T \sum_{j,m}^{-1} W(o_t(\alpha) - \mu_{j,m})}{2} \right\}$$

20

【0435】

なお、数式20において、 M はケプストラム次数、 $|A|$ は行列 A のdeterminantを表わす。

【0436】

占有度数算出手段41013は、通常、MPUやメモリ等から実現され得る。占有度数算出手段41013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。

【0437】

ケプストラム変換パラメータ算出手段41015は、学習音響データ格納手段41012の学習音響データを読み出し、およびメモリ上に一時格納されている第三のケプストラムベクトル系列(O_t)を読み出し、読み出した学習音響データおよび第三のケプストラムベクトル系列(O_t)に基づいて、ケプストラム変換パラメータ(α)を算出する。ケプストラム変換パラメータ算出手段41015は、さらに好ましくは、周波数範囲指定情報(W)で示される周波数範囲における学習音響データ、第三のケプストラムベクトル系列(O_t)および占有度数($\gamma_t(j, m)$)に基づいて、ケプストラム変換パラメータを算出する。なお、第三のケプストラムベクトル系列(O_t)は、ケプストラム変換手段36014が算出したデータである。

30

【0438】

具体的には、まず、ケプストラム変換パラメータ算出手段41015は、上述の数式12により、ベクトル($u_t(\alpha)$)を算出する。なお、ケプストラム変換パラメータ算出手段41015は、上述の数式12を示す情報を格納しており、かかる数式12の情報を読み出し、当該数式に、第三のケプストラムベクトル系列(O_t)、ケプストラム変換パラメータ(α)、ベクトル(C_t)の情報を与え、当該数式を演算する。そして、次に、ケプストラム変換パラメータ算出手段41015は、以下の数式21により、 α の最適値(α^*)を算出する。なお、 α の最適値(α^*)は、繰り返しステップ（ループ処理）における現在のループ内の処理での最適値である。また、ケプストラム変換パラメータ算出手段41015は、以下の数式21を示す情報を、予め格納している。

40

50

【数 2 1】

$$\alpha^* = \frac{\sum_t \sum_j \sum_m r_t(j, m) (u_t(\alpha) - \mu_{j, m})^T W^T \sum_{j, m}^{-1} W \bar{c}_t}{\sum_t \sum_j \sum_m r_t(j, m) \bar{c}_t^T W^T \sum_{j, m}^{-1} W \bar{c}_t}$$

【0 4 3 9】

ケプストラム変換パラメータ算出手段 4 1 0 1 5 は、通常、MPU やメモリ等から実現され得る。ケプストラム変換パラメータ算出手段 4 1 0 1 5 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

【0 4 4 0】

次に、本音声処理装置の動作について説明する。まず、本音声処理装置における、声道長正規化パラメータを算出する処理について、図 4 2 のフローチャートを用いて説明する。図 4 2 のフローチャートにおいて、図 3 7 のフローチャートと比較して差異のあるステップのみ説明する。なお、本声道長正規化パラメータ算出処理は、必ずしも音声処理装置で行う必要はない。また、図 4 2 における初期化処理は、例えば、ユーザ（評価対象者）に対して、「/ a i u e o /」と発声するように促す処理と、周波数範囲指定情報格納手段 4 1 0 1 1 の周波数範囲指定情報、および学習音響データ格納手段 4 1 0 1 2 の学習音響データを読み出す処理である。

【0 4 4 1】

20

（ステップ S 4 2 0 1）占有度数算出手段 4 1 0 1 3 は、ステップ S 3 7 0 7 で得られた第三のケプストラムベクトル系列（ O_t ）と学習音響データ格納手段 4 1 0 1 2 の学習音響データと周波数範囲指定情報（ W ）を用いて、占有度数（ $y_t(j, m)$ ）を算出し、メモリに一時記憶する。

【0 4 4 2】

（ステップ S 4 2 0 2）ケプストラム変換パラメータ算出手段 4 1 0 1 5 は、周波数範囲指定情報（ W ）で示される周波数範囲における学習音響データ、ベクトル（ $u_t(\alpha)$ ）および占有度数（ $y_t(j, m)$ ）に基づいて、本ループにおける最適なケプストラム変換パラメータ（ α^* ）を算出する。ケプストラム変換パラメータ算出手段 4 1 0 1 5 は、例えば、格納している数式 2 1 の情報を読み出し、また、メモリ上の学習音響データ、ベクトル（ $u_t(\alpha)$ ）および占有度数（ $y_t(j, m)$ ）を読み出し、数式 2 1 に代入し、ケプストラム変換パラメータ（ α^* ）を算出する。

30

【0 4 4 3】

次に、本音声処理装置の動作（音声の評定処理）について説明する。本音声処理装置の動作は、図 3 8、図 3 9 のフローチャートにおける動作と同様である。

【0 4 4 4】

以下、図 4 3 の概念図を用いて、本実施の形態における音声処理装置における声道長正規化パラメータ算出処理の概念について説明する。まず、システム設計用音声データベースから、指定された発話内容（例えば / あいうえお /）に従って音素 HMM（教師データが有する音素 HMM）を連結し連結 HMM（図 4 3 の基準話者 HMM）を構成する。そして、連結 HMM と、同じ音素列のユーザ発話音声の変換されたケプストラムベクトル O_t の、 Λ （ Λ は、ユーザ発話音声にしたがって連結された連結 HMM）に対する出現確率が最大になるように（図 4 3 の最適化に相当）パラメータを算出する。ただし、求まるパラメータ（ α^* ）はケプストラム変換（ケプストラムワーピング）パラメータであり、このままでは声道長の変換を直接表わすものではないため、近似変換式（ $1/p$ ）を用いて最終的な声道長正規化パラメータ（サンプリング周波数変換率） γ を計算する。なお、声道長正規化パラメータを算出するための数式は、上記の数式 1 6 から数式 2 1、数式 9 から数式 1 2、数式 1 4、数式 1 5 により行われる。

40

【0 4 4 5】

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

50

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、特に、精度の高い評定ができる。

【0446】

なお、本実施の形態によれば、音声処理装置が声道長正規化パラメータを算出する処理と、声道長正規化処理を行った。しかし、声道長正規化パラメータを算出する処理と声道長正規化処理を、異なる装置が行う構成でも良い。かかる場合、本音声処理装置は、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、サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している声道長正規化パラメータ格納部と、前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出される第二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。また、声道長正規化パラメータを算出する装置は、教師データを構成するデータから、話者の音声のケプストラムベクトル系列を短区間分析で求め、それを用いて2以上の基準話者音素HMMを取得し、当該2以上の基準話者音素HMMを声道長正規化用の指定された発話内容（例えば／あいうえお／）に従って連結して得られた連結HMMである学習音響データを格納している学習音響データ格納手段と、受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出手段と、前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列を用いて線形変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換手段と、指定された発話内容に従って受け付けた音声の事後確率である占有度数を算出する占有度算出手段と、前記学習音響データおよび前記第三のケプストラムベクトル系列および前記占有度数に基づいて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記占有度算出手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段を具備する装置である。

【0447】

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、音声を受け付ける音声受付ステップと、格納している声道長正規化パラメータと格納している第一サンプリング周波数をパラメータとして算出される第二サンプリング周波数で、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。

【0448】

また、上記プログラムにおいて、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以上格納しており、前記教師データおよび前記音声受付ステップで受け付けた音声に基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出ステップとをさらにコンピュータに実行させ、前記格納している声道長正規化パラメータは、前記声道長正規化パラメータ算出ステップで算出した声道長正規化パラメータであることは好適である。

【0449】

また、上記プログラムにおいて、音素HMMを指定された発話内容に従って連結した連結HMMである学習音響データを格納しており、前記声道長正規化パラメータ算出ステップは、前記音声受付ステップで受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出ステップと、前記第二のケ

プストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列を用いて線形変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換ステップと、指定された発話内容に従って受け付けた音声の事後確率である占有度数を算出する占有度算出ステップと、前記学習音響データおよび前記第三のケプストラムベクトル系列および前記占有度数に基づいて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出ステップと、所定のルールに基づいて、前記ケプストラム変換ステップにおける処理、および前記占有度算出ステップにおける処理、および前記ケプストラム変換パラメータ算出ステップにおける処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得ステップと、前記最終ケプストラム変換パラメータ取得ステップで得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出ステップを具備することは好適である。

10

【0450】

また、上記プログラムにおいて、前記声道長正規化パラメータ算出ステップにおいて、前記最終ケプストラム変換パラメータ取得ステップで得たケプストラム変換パラメータから線形周波数伸縮比を算出し、当該線形周波数伸縮比から前記声道長正規化パラメータを算出することは好適である。

【0451】

また、上記プログラムにおいて、周波数範囲を指定する情報である周波数範囲指定情報を格納しており、前記ケプストラム変換パラメータ算出ステップにおいて、前記周波数範囲指定情報で示される周波数範囲における学習音響データ、第三のケプストラムベクトル系列および占有度数に基づいて、ケプストラム変換パラメータを算出することは好適である。

20

(実施の形態10)

【0452】

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、主として、音声（歌唱を含む）を評価する発音評定装置である、として説明する。さらに、本実施の形態において、上記の実施の形態で記載した音声処理装置よりもさらに精度高く、評価対象者の話者特性に応じた発音評定が可能な音声処理装置について説明する。具体的には、本実施の形態において、複数の音素を用いた最小自乗誤差基準に基づく、簡潔な声道長正規化法に基づいて、評価対象者の話者特性に左右されにくい音声処理装置について説明する。

30

【0453】

さらに具体的には、本実施の形態において、基準話者（システム設計用の話者）の音声から、音素毎に基準話者平均ケプストラムベクトルを計算しておく。そして、ユーザ音声の周波数ワープされたケプストラムベクトル系列と上記の基準話者平均ケプストラムベクトルの系列との自乗誤差が最小になるように、最適なワーピングパラメータを求める。このとき、ユーザ音声は指定された発話内容にしたがって発声されたものであり、基準話者平均ケプストラムの系列は同じ発話内容にしたがって音素毎の基準話者平均ケプストラムを並べたものであることが好適である。ただし、ここで算出されるパラメータはケプストラムワーピングパラメータであり、声道長変換を直接表わすものではないため、予め決められた近似変換式を用いて最終的な声道長正規化パラメータ（サンプリング周波数変換率）を算出する。

40

【0454】

なお、本実施の形態における音声処理装置は、実施の形態8、9の音声処理装置と比較して、声道長正規化パラメータの算出アルゴリズムが異なる。

【0455】

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。

【0456】

50

図 4 4 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。

【 0 4 5 7 】

本音声処理装置は、図 4 1 における音声処理装置と比較して、声道長正規化パラメータ算出部 4 4 0 1 が異なる。

【 0 4 5 8 】

声道長正規化パラメータ算出部 4 4 0 1 は、周波数範囲指定情報格納手段 4 1 0 1 1、学習音響データ格納手段 4 4 0 1 2、第二ケプストラムベクトル系列算出手段 3 6 0 1 3、ケプストラム変換手段 3 6 0 1 4、最適音素系列取得手段 4 4 0 1 3、ケプストラム変換パラメータ算出手段 4 4 0 1 5、最終ケプストラム変換パラメータ取得手段 3 6 0 1 6、声道長正規化パラメータ算出手段 3 6 0 1 7 を具備する。なお、ここでの最終ケプストラム変換パラメータ取得手段 3 6 0 1 6 は、所定のルールに基づいて、ケプストラム変換手段 3 6 0 1 4 における処理、ケプストラム変換パラメータ算出手段 4 4 0 1 5 における処理だけではなく、最適音素系列取得手段 4 4 0 1 3 における処理をも繰り返えさせ、最終的な最適なケプストラム変換パラメータを得る。

【 0 4 5 9 】

学習音響データ格納手段 4 4 0 1 2 は、学習音響データを格納している。ここでの学習音響データは、音素平均ケプストラムベクトルを指定された発話内容に従って並べた音素平均ケプストラムベクトル列である。学習音響データは、例えば、以下のように生成される。つまり、図示しない学習音響データ生成手段が、教師データを構成するデータ（例えば、教師データ格納部 1 0 2 のデータ）から、各音素 l ($l = 1, 2, \dots, L$) に対して、基準話者の音声ケプストラムベクトル系列を短区間分析で求め、その時間平均 (μ_l) を算出する。なお、ケプストラムベクトルは 0 次係数も含めた $M + 1$ 次元である。そして、時間平均 (μ_l) は、以下の数式 2 2 で表される。

【数 2 2】

$$\mu_l = [\mu_{l,0} \ \mu_{l,1} \ \mu_{l,2} \ \cdots \ \mu_{l,M}]^T$$

【 0 4 6 0 】

なお、 $(\cdot \cdot \cdot)^T$ は、行列またはベクトルの転置を示す。

【 0 4 6 1 】

そして、図示しない学習音響データ生成手段は、生成した学習音響データを、学習音響データ格納手段 4 4 0 1 2 に蓄積する。

【 0 4 6 2 】

なお、学習音響データ格納手段 4 4 0 1 2 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 4 6 3 】

最適音素系列取得手段 4 4 0 1 3 は、ユーザが発声した発話音声を構成する各フレーム t に対応する音素番号の情報である音素系列 (s_t^* ($t = 1, 2, \dots, T$)) を取得する。最適音素系列取得手段 4 4 0 1 3 が音素系列 (s_t^*) を取得するアルゴリズムは問わない。最適音素系列取得手段 4 4 0 1 3 は、例えば、ユーザ（キーボードなど）からの音素系列 (s_t^*) の入力を受け付けても良い。また、最適音素系列取得手段 4 4 0 1 3 は、ユーザケプストラムベクトル系列 (C_t) からオートセグメンテーションにより、音素系列 (s_t^*) を算出しても良い。オートセグメンテーションは、公知技術であるので説明を省略する。また、最適音素系列取得手段 4 4 0 1 3 は、動的計画法 (DP マッチング) により、以下の数式 2 3 の J を最小にする音素系列 S_t ($t = 1, 2, \dots, T$) を取得し、その S_t を音素系列 (s_t^*) として取得しても良い。また、音素系列は、音素番号の列の情報でなくても良い。音素系列は、音素を識別する情報の列であれば良い。

【数 2 3】

$$J(\alpha; s_1, s_2, \dots, s_T) = \frac{1}{T} \sum_{t=1}^T \|W o_t(\alpha) - W \mu_{s_t}\|^2$$

【0 4 6 4】

ケプストラム変換パラメータ算出手段 4 4 0 1 5 は、学習音響データ格納手段 4 4 0 1 2 の学習音響データを読み出し、およびメモリ上に一時格納されている第三のケプストラムベクトル系列 (O_t) を読み出し、および最適音素系列取得手段 4 4 0 1 3 が取得した音素系列 (s_t^*) を読み出し、読み出した学習音響データおよび第三のケプストラムベクトル系列 (O_t) および音素系列 (s_t^*) を用いて、ケプストラム変換パラメータ (α) を算出する。ケプストラム変換パラメータ算出手段 4 4 0 1 5 は、さらに好ましくは、周波数範囲指定情報 (W) および読み出した学習音響データおよび第三のケプストラムベクトル系列 (O_t) および音素系列 (s_t^*) に基づいて、ケプストラム変換パラメータを算出する。

10

【0 4 6 5】

具体的には、まず、ケプストラム変換パラメータ算出手段 3 6 0 1 5 は、以下の数式 2 4 により、ベクトル ($u_t(\alpha)$) を算出する。そして、次に、ケプストラム変換パラメータ算出手段 4 4 0 1 5 は、以下の数式 2 5 により、 α の最適値 (α^*) を算出する。なお、 α の最適値 (α^*) は、現繰り返しステップにおける最適値である。

20

【数 2 4】

$$u_t(\alpha) = o_t(\alpha) + \alpha \bar{c}_t$$

【数 2 5】

$$\alpha^* = \frac{\sum_{t=1}^T (u_t(\alpha) - \mu_{s_t^*})^T W^T W \bar{c}_t}{\sum_{t=1}^T \|W \bar{c}_t\|^2}$$

【0 4 6 6】

なお、ケプストラム変換パラメータ算出手段 4 4 0 1 5 は、通常、上記の数式 2 4、2 5 の情報を格納しており、当該数式の情報を読み出し、演算を行う。ケプストラム変換パラメータ算出手段 4 4 0 1 5 は、通常、MPU やメモリ等から実現され得る。ケプストラム変換パラメータ算出手段 4 4 0 1 5 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。

30

【0 4 6 7】

なお、声道長正規化パラメータ算出手段 3 6 0 1 7 は、実施の形態 9 において、最終ケプストラム変換パラメータ取得手段 3 6 0 1 6 が得たケプストラム変換パラメータから、数式 1 4 に従って、線形周波数伸縮比 (ρ) を算出した。しかし、声道長正規化パラメータ算出手段 3 6 0 1 7 は、最終ケプストラム変換パラメータ取得手段 3 6 0 1 6 が得たケプストラム変換パラメータから、以下の数式 2 6 に従って、線形周波数伸縮比 (ρ) を算出して良い。

40

【数 2 6】

$$\rho = \frac{1 + K \alpha^*}{1 - \alpha^*} \quad (K \geq 0, K \simeq 1)$$

【0 4 6 8】

次に、本音声処理装置の動作について説明する。まず、本音声処理装置における、声道長正規化パラメータを算出する処理について、図 4 5 のフローチャートを用いて説明する。図 4 5 のフローチャートにおいて、図 3 7、図 4 2 のフローチャートと比較して差異のあるステップのみ説明する。なお、本声道長正規化パラメータ算出処理は、必ずしも音声

50

処理装置で行う必要はない。また、図 4 5 における初期化処理は、例えば、ユーザ（評価対象者）に対して、「 / あいうえお / 」と発声するように促す処理と、周波数範囲指定情報格納手段 4 1 0 1 1 の周波数範囲指定情報、および学習音響データ格納手段 4 4 0 1 2 の学習音響データを読み出す処理である。

【 0 4 6 9 】

（ステップ S 4 5 0 1）最適音素系列取得手段 4 4 0 1 3 は、ユーザが発声した発話音声構成する各フレーム t に対応する音素系列（ s^*_t （ $t = 1, 2, \dots, T$ ））を取得する。

【 0 4 7 0 】

（ステップ S 4 5 0 2）ケプストラム変換パラメータ算出手段 4 4 0 1 5 は、周波数範囲指定情報（ W ）で示される周波数範囲における学習音響データ、ベクトル（ $u_t(\alpha)$ ）およびステップ S 4 5 0 1 で取得した音素系列（ s^*_t ）に基づいて、本ループにおける最適なケプストラム変換パラメータ（ α^* ）を算出する。ケプストラム変換パラメータ算出手段 4 4 0 1 5 は、例えば、格納している数式 2 5 の情報を読み出し、また、メモリ上の学習音響データ、ベクトル（ $u_t(\alpha)$ ）および音素系列（ s^*_t ）を読み出し、数式 2 5 に代入し、ケプストラム変換パラメータ（ α^* ）を算出する。

【 0 4 7 1 】

次に、本音声処理装置の動作（音声の評定処理）について説明する。本音声処理装置の動作は、図 3 8、図 3 9 のフローチャートにおける動作と同様である。

【 0 4 7 2 】

以下、図 4 6 の概念図を用いて、本実施の形態における音声処理装置における声道長正規化パラメータ算出処理の概念について説明する。まず、基準話者音声データベースから算出された音素毎の平均ケプストラムベクトルを、指定された発話内容（例えば / あいうえお / ）の音素列に従って並べたベクトル列と、同じ音素列のユーザ発話音声の変換されたケプストラムベクトルとの自乗誤差が最小になるようにパラメータ（ α^* ）が求められる。ただし、求まるパラメータ（ α^* ）はケプストラム変換（ケプストラムワーピング）パラメータであり、このままでは声道長の変換を直接表わすものではないため、近似変換式（ $1/\rho$ ）を用いて最終的な声道長正規化パラメータ（サンプリング周波数変換率） γ を計算する。なお、声道長正規化パラメータを算出するための数式は、上記の数式 2 2 から数式 2 6、数式 9 から数式 1 1、数式 1 4 から数式 1 6 等により行われる。

【 0 4 7 3 】

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、特に、精度の高い評定ができる。

【 0 4 7 4 】

なお、本実施の形態によれば、音声処理装置が声道長正規化パラメータを算出する処理と、声道長正規化処理を行った。しかし、声道長正規化パラメータを算出する処理と声道長正規化処理を、異なる装置が行う構成でも良い。かかる場合、本音声処理装置は、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、サンプリング周波数の変換率に関する情報である声道長正規化パラメータを格納している声道長正規化パラメータ格納部と、前記声道長正規化パラメータと前記第一サンプリング周波数をパラメータとして算出される第二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。また、声道長正規化パラメータを算出する装置は、教師データを構成するデータから、話者の音声のケプストラムベクトル系列を短区間分析で求め、それを用いて 2 以上の基準話者音素平均ケプストラムベクトルを取得し、当該 2 以上の基準話者音素平均ケプストラムベクトルを声道長正規化用の指定された発話内容（例えば / あいうえお / ）に従って並べて得られた平均ケプストラムベクトル列である学習音響データを格納している学

10

20

30

40

50

習音響データ格納手段と、受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出手段と、前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列を用いて線形変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換手段と、ユーザが発声した発話音声を構成する各フレーム t に対応する音素を識別する情報の列である音素系列を取得する最適音素系列取得手段と、前記学習音響データおよび前記第三のケプストラムベクトル系列および前記音素系列に基づいて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出手段と、所定のルールに基づいて、前記ケプストラム変換手段における処理、および前記最適音素系列取得手段における処理、および前記ケプストラム変換パラメータ算出手段における処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、前記最終ケプストラム変換パラメータ取得手段が得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出手段を具備する装置である。

10

【0475】

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、音声を受け付ける音声受付ステップと、格納している声道長正規化パラメータと格納している第一サンプリング周波数をパラメータとして算出される第二サンプリング周波数で、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。

20

【0476】

また、上記プログラムにおいて、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以上格納しており、前記教師データおよび前記音声受付ステップで受け付けた音声に基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出ステップとをさらにコンピュータに実行させ、前記格納している声道長正規化パラメータは、前記声道長正規化パラメータ算出ステップで算出した声道長正規化パラメータであることは好適である。

【0477】

また、上記プログラムにおける声道長正規化パラメータ算出ステップは、前記音声受付ステップで受け付けた音声から、短区間分析により第二のケプストラムベクトル系列を算出する第二ケプストラムベクトル系列算出ステップと、前記第二のケプストラムベクトル系列を、ケプストラム変換パラメータを要素とする行列を用いて変換し、周波数ワープされた第三のケプストラムベクトル系列を算出するケプストラム変換ステップと、ユーザが発声した発話音声を構成する各フレーム t に対応する音素を識別する情報の列である音素系列を取得する最適音素系列取得ステップと、格納している学習音響データおよび前記第三のケプストラムベクトル系列および前記音素系列を用いて、ケプストラム変換パラメータを算出するケプストラム変換パラメータ算出ステップと、所定のルールに基づいて、前記ケプストラム変換ステップにおける処理、および前記最適音素系列取得ステップにおける処理、および前記ケプストラム変換パラメータ算出ステップにおける処理を繰り返えさせ、ケプストラム変換パラメータを得る最終ケプストラム変換パラメータ取得手段と、前記最終ケプストラム変換パラメータ取得ステップが得たケプストラム変換パラメータに基づいて、前記声道長正規化パラメータを算出する声道長正規化パラメータ算出サブステップを具備することは好適である。

30

40

【0478】

また、上記プログラムの声道長正規化パラメータ算出サブステップにおいて、前記最終ケプストラム変換パラメータ取得ステップで得たケプストラム変換パラメータから線形周波数伸縮比を算出し、当該線形周波数伸縮比から前記声道長正規化パラメータを算出することは好適である。

【0479】

50

また、上記プログラムにおいて、周波数範囲を指定する情報である周波数範囲指定情報を格納しており、前記ケプストラム変換パラメータ算出ステップにおいて、前記周波数範囲指定情報で示される周波数範囲における学習音響データ、第三のケプストラムベクトル系列および音素系列に基づいて、ケプストラム変換パラメータを算出することは好適である。

【0480】

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。

10

【0481】

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。

【0482】

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（ $y_t(j)$ ）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。

20

【0483】

また、図47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47は、このコンピュータシステム340の概観図であり、図48は、コンピュータシステム340のブロック図である。

30

【0484】

図47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。

【0485】

図48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。

40

【0486】

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

50

M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。

【0487】

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。

10

【0488】

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。

【0489】

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。

【0490】

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。

20

【産業上の利用可能性】

【0491】

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】

【0492】

【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

30

【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

40

【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

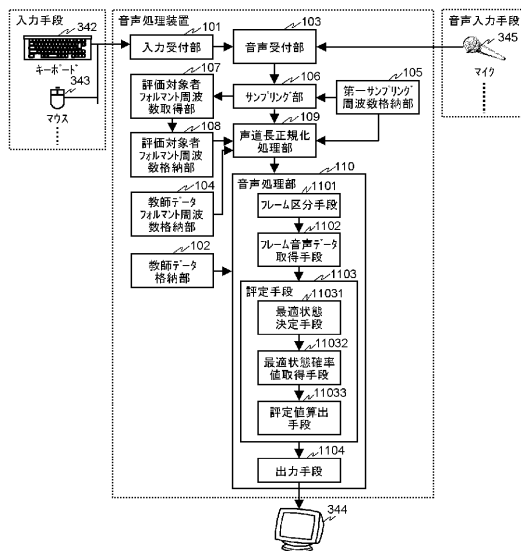
50

【図 2 3】同評価処理について説明するフローチャート	
【図 2 4】同音素の挿入の検知について説明する図	
【図 2 5】同出力例を示す図	
【図 2 6】実施の形態 5 における音声処理装置のブロック図	
【図 2 7】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 2 8】同評価処理について説明するフローチャート	
【図 2 9】同音素の置換の検知について説明する図	
【図 3 0】実施の形態 6 における音声処理装置のブロック図	
【図 3 1】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 2】同評価処理について説明するフローチャート	10
【図 3 3】同音素の欠落の検知について説明する図	
【図 3 4】実施の形態 7 における音声処理装置のブロック図	
【図 3 5】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 6】実施の形態 8 における音声処理装置のブロック図	
【図 3 7】同声道長正規化パラメータ算出処理について説明するフローチャート	
【図 3 8】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 9】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 0】同声道長正規化パラメータ算出処理の概念について説明する図	
【図 4 1】実施の形態 9 における音声処理装置のブロック図	
【図 4 2】同声道長正規化パラメータ算出処理について説明するフローチャート	20
【図 4 3】同声道長正規化パラメータ算出処理の概念について説明する図	
【図 4 4】実施の形態 10 における音声処理装置のブロック図	
【図 4 5】同声道長正規化パラメータ算出処理について説明するフローチャート	
【図 4 6】同声道長正規化パラメータ算出処理の概念について説明する図	
【図 4 7】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 4 8】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】	
【0 4 9 3】	
1 0 1 入力受付部	
1 0 2 教師データ格納部	30
1 0 3 音声受付部	
1 0 4 教師データフォーマント周波数格納部	
1 0 5 第一サンプリング周波数格納部	
1 0 6 サンプリング部	
1 0 7 評価対象者フォーマント周波数取得部	
1 0 8 評価対象者フォーマント周波数格納部	
1 0 9、3 6 0 9 声道長正規化処理部	
1 1 0、1 1 1 0、1 7 1 0、2 1 1 0、2 6 1 0、3 0 1 0、3 4 1 0 音声処理部	
1 1 0 1、3 4 1 0 1 フレーム区分手段	
1 1 0 2、3 4 1 0 2 フレーム音声データ取得手段	40
1 1 0 3、1 1 1 0 3、1 7 1 0 3、2 1 1 0 3、2 6 1 0 3、3 0 1 0 3 評価手段	
1 1 0 4、2 1 1 0 4、3 4 1 0 2 出力手段	
1 1 0 9 発声催促部	
3 6 0 1、4 1 0 1、4 4 0 1 声道長正規化パラメータ算出部	
3 6 0 2 声道長正規化パラメータ格納部	
1 1 0 3 1 最適状態決定手段	
1 1 0 3 2 最適状態確率値取得手段	
1 1 0 3 3、2 1 0 2 3、1 1 1 0 3 3、1 7 1 0 3 3 評価値算出手段	
1 7 1 0 1、2 1 1 0 1、2 6 1 0 1、3 0 1 0 1 特殊音声検知手段	
3 4 1 0 1 音声認識手段	50

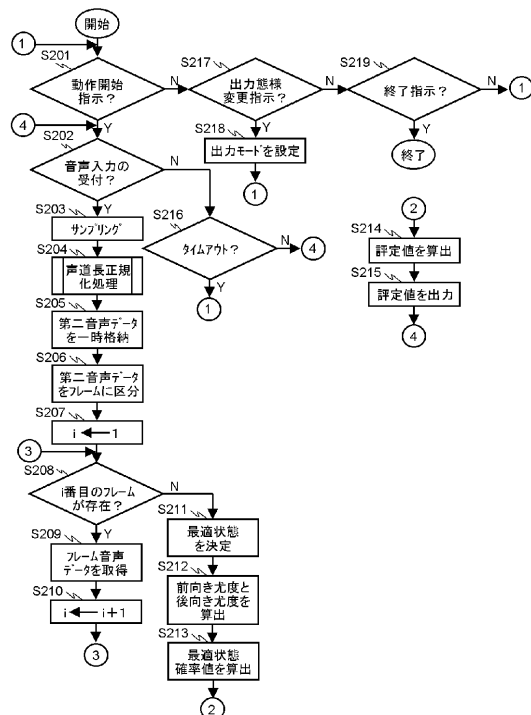
- 1 1 1 0 3 2 発音区間フレーム音韻確率値取得手段
 1 7 1 0 1 1 無音データ格納手段
 1 7 1 0 1 2 無音区間検出手段
 3 6 0 1 1、4 1 0 1 1 周波数範囲指定情報格納手段
 3 6 0 1 2 長時間ケプストラム平均ベクトル格納手段
 3 6 0 1 3 第二ケプストラムベクトル系列算出手段
 3 6 0 1 4 ケプストラム変換手段
 3 6 0 1 5、4 1 0 1 5、4 4 0 1 5 ケプストラム変換パラメータ算出手段
 3 6 0 1 6 最終ケプストラム変換パラメータ取得手段
 3 6 0 1 7 声道長正規化パラメータ算出手段
 4 1 0 1 2、4 4 0 1 2 学習音響データ格納手段
 4 1 0 1 3 占有度数算出手段
 4 4 0 1 3 最適音素系列取得手段

10

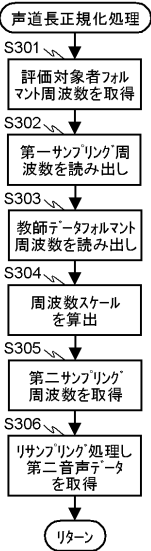
【図 1】



【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

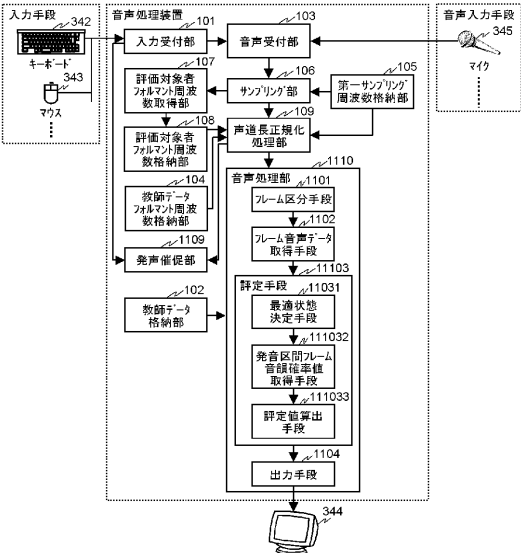
【図 5】

	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

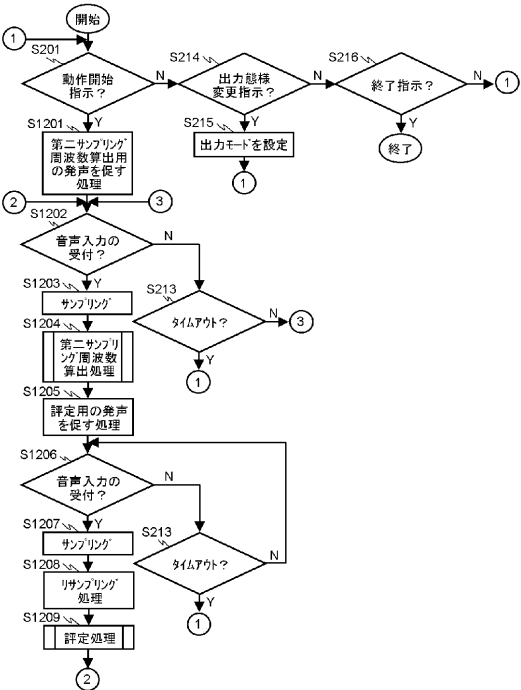
【図 6】

プリエンフィス	1-0.97 _z ⁻¹
窓関数	Hamming窓
分析フレーム長	25msec
フレーム周期	10msec
特徴パラメータ	MFCC(39次元)

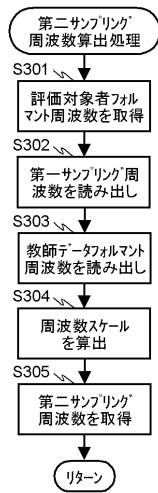
【図 1 1】



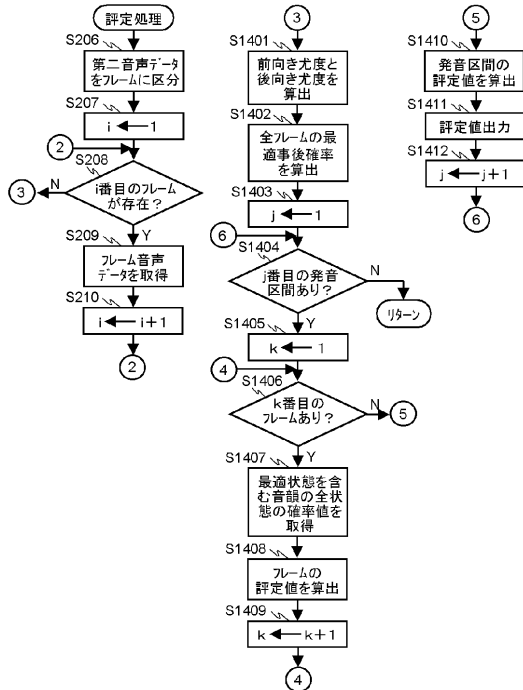
【図 1 2】



【図 13】



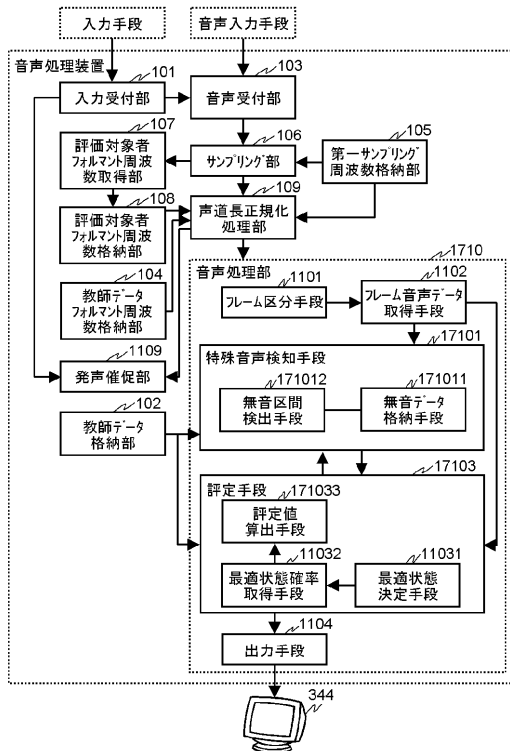
【図 14】



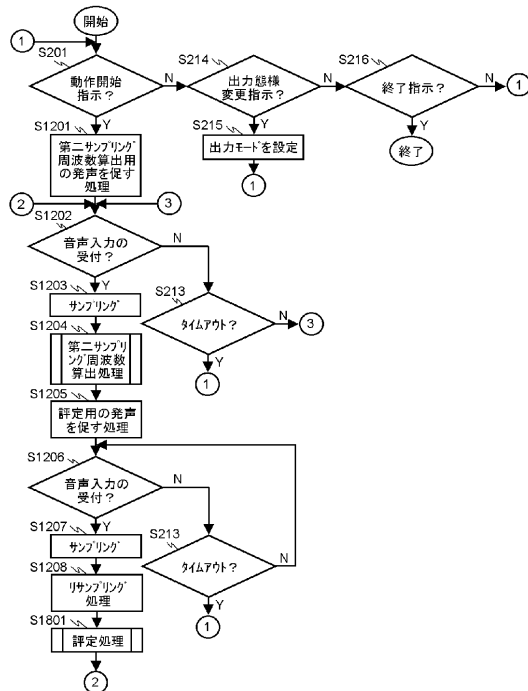
【図 15】

	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

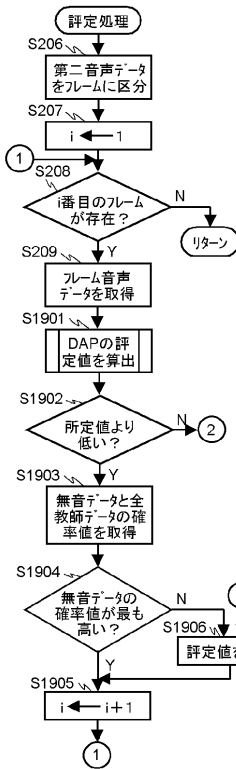
【図 17】



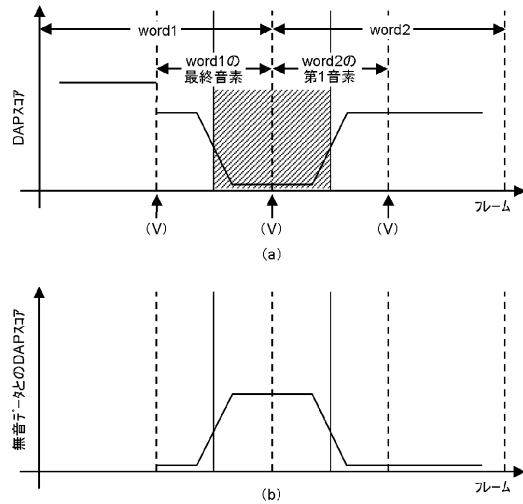
【図 18】



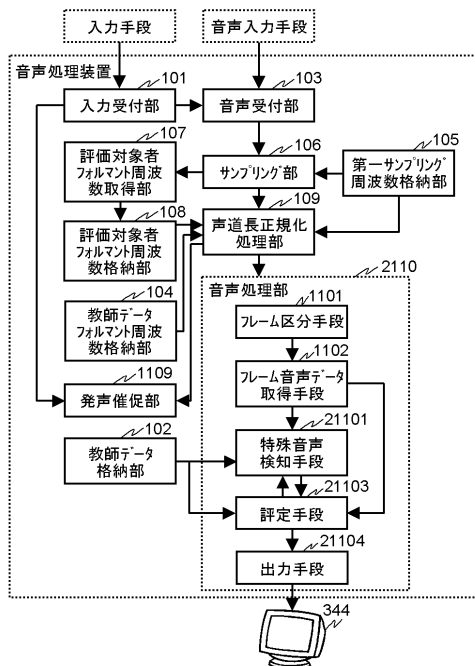
【図 19】



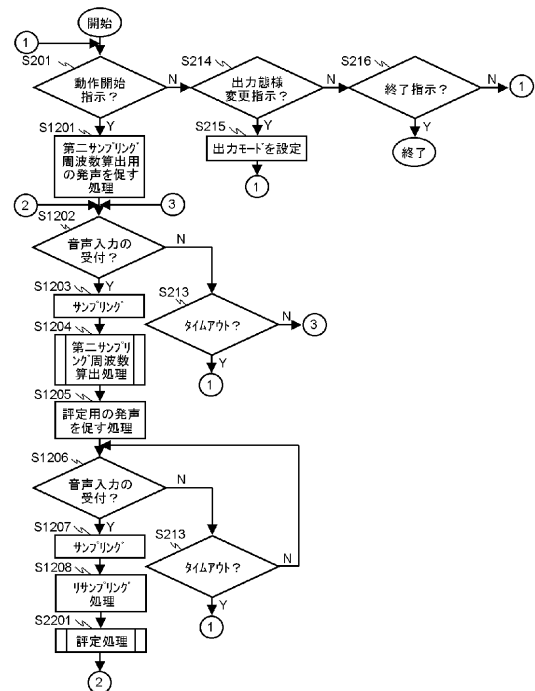
【図 20】



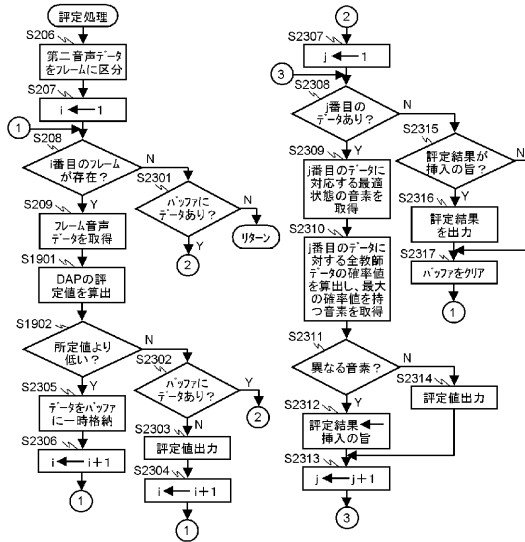
【図 21】



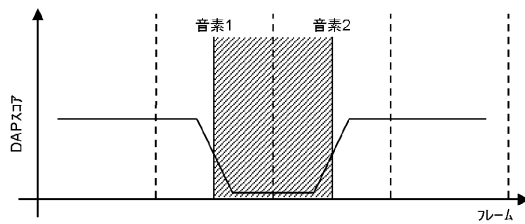
【図 22】



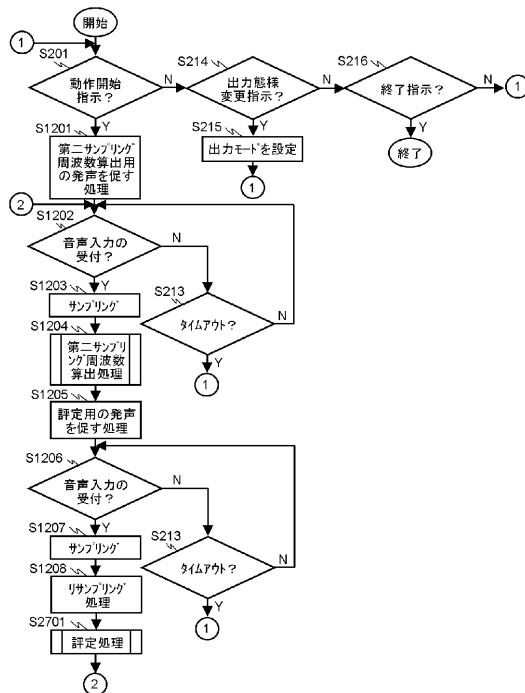
【図 23】



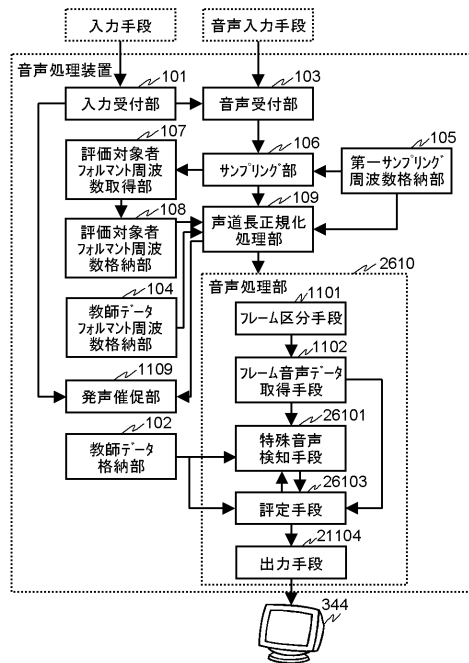
【図 24】



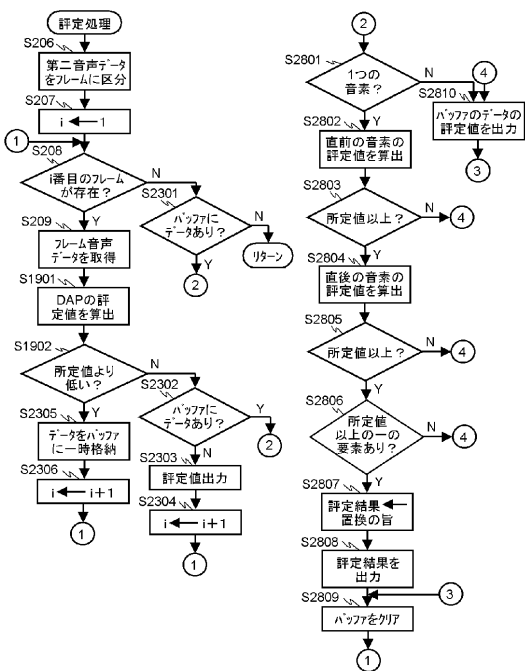
【図 27】



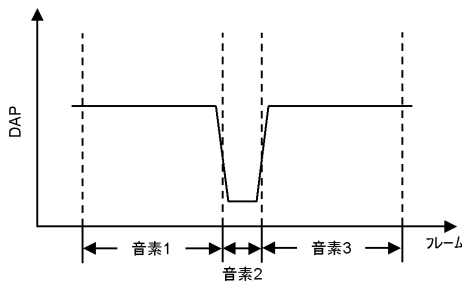
【図 26】



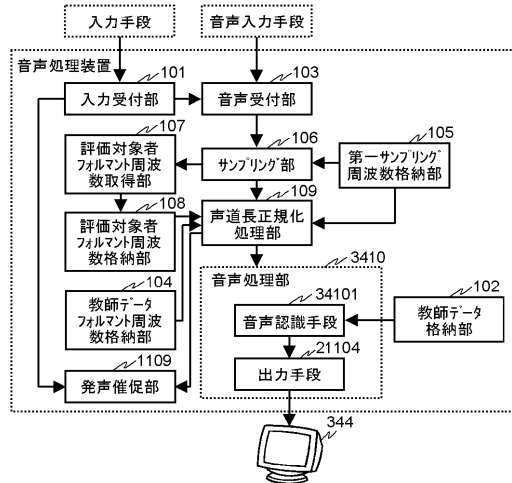
【図 28】



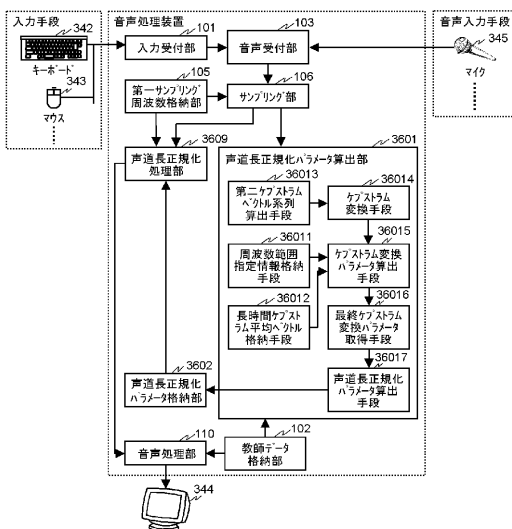
【図 3 3】



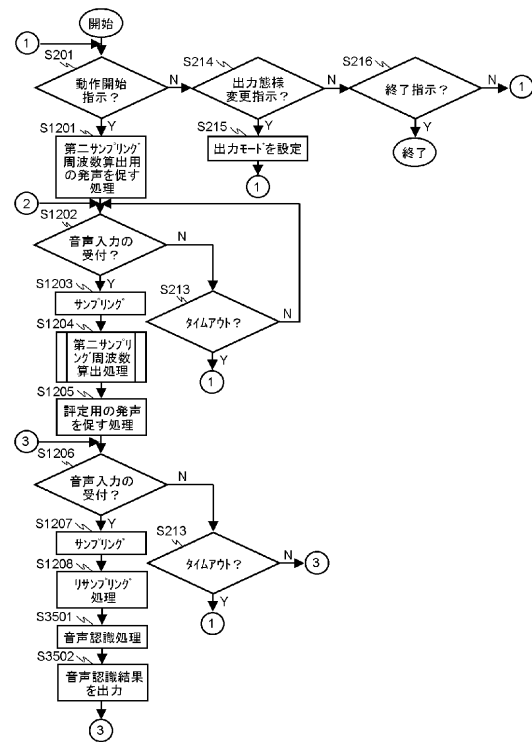
【図 3 4】



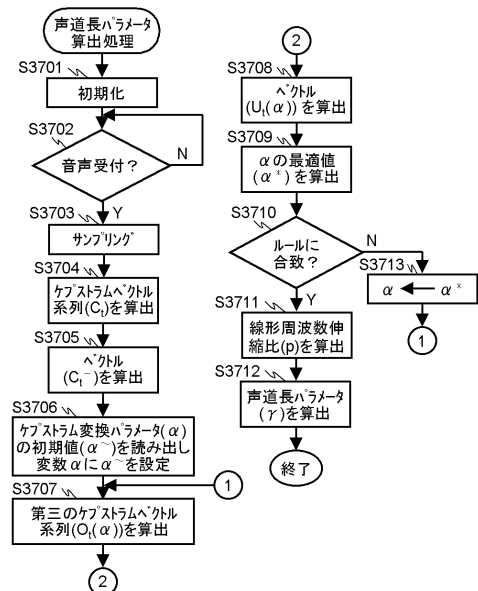
【図 3 6】



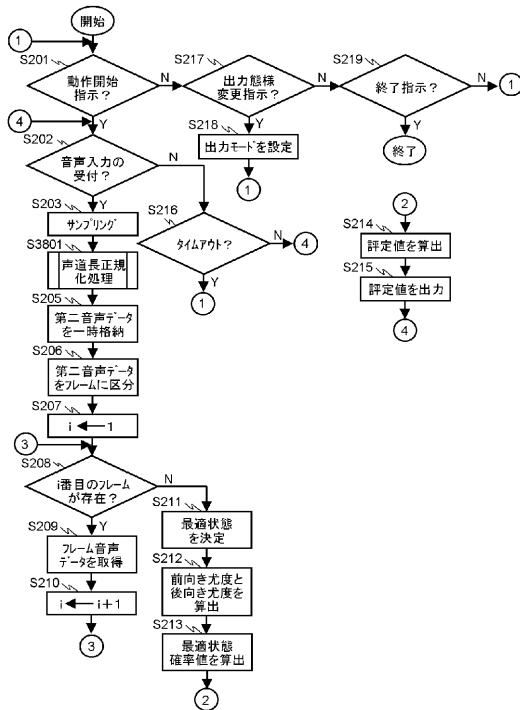
【図 3 5】



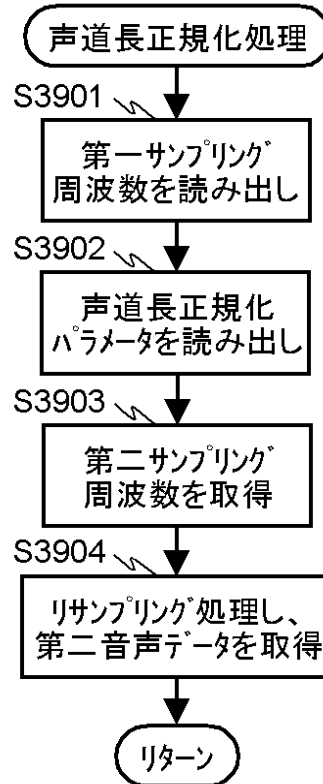
【図 3 7】



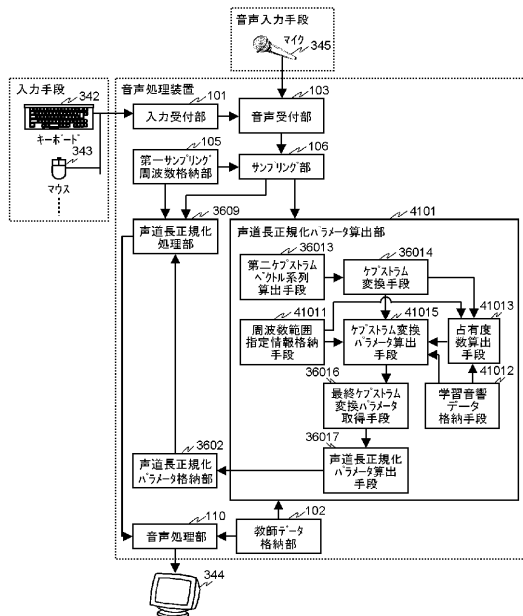
【 図 3 8 】



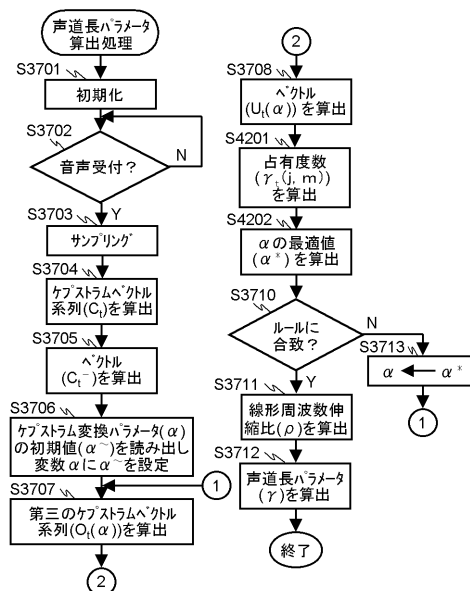
【 図 3 9 】



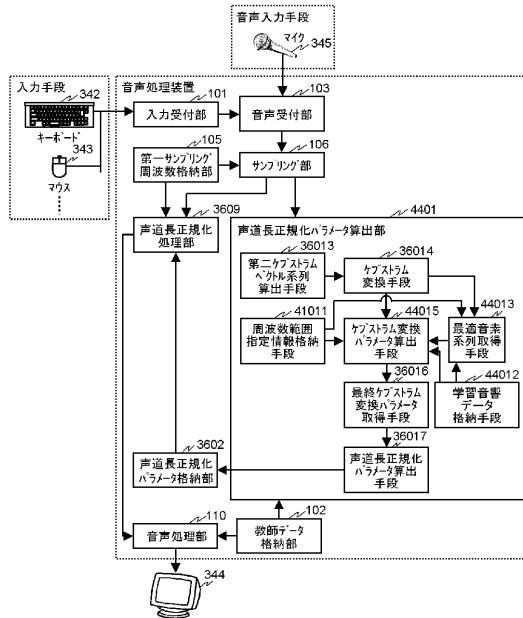
【 図 4 1 】



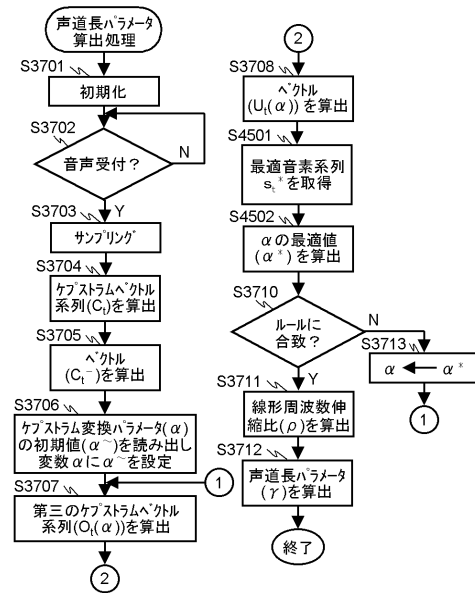
【 図 4 2 】



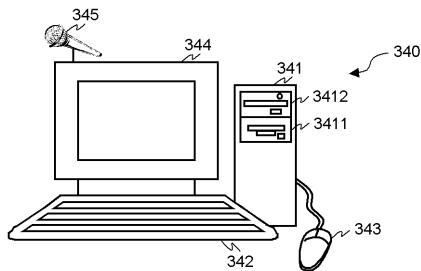
【 図 4 4 】



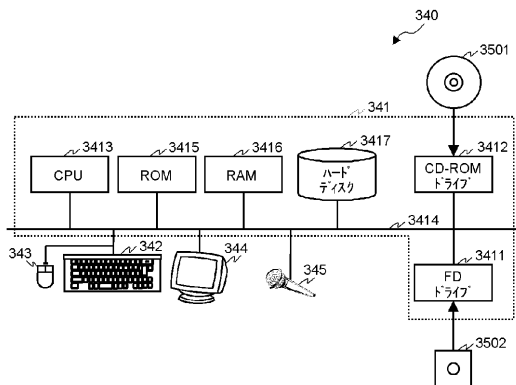
【 図 4 5 】



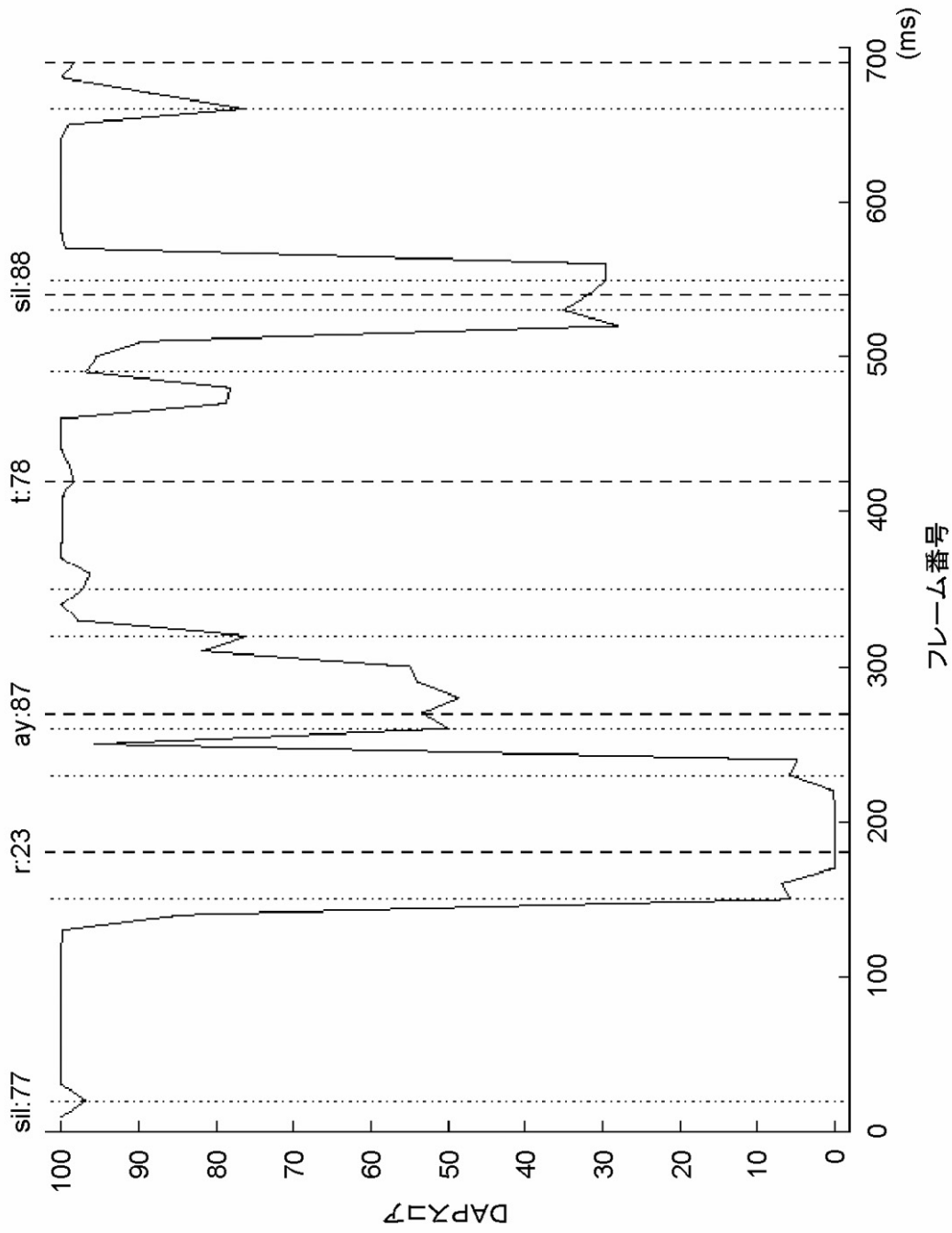
【 図 4 7 】



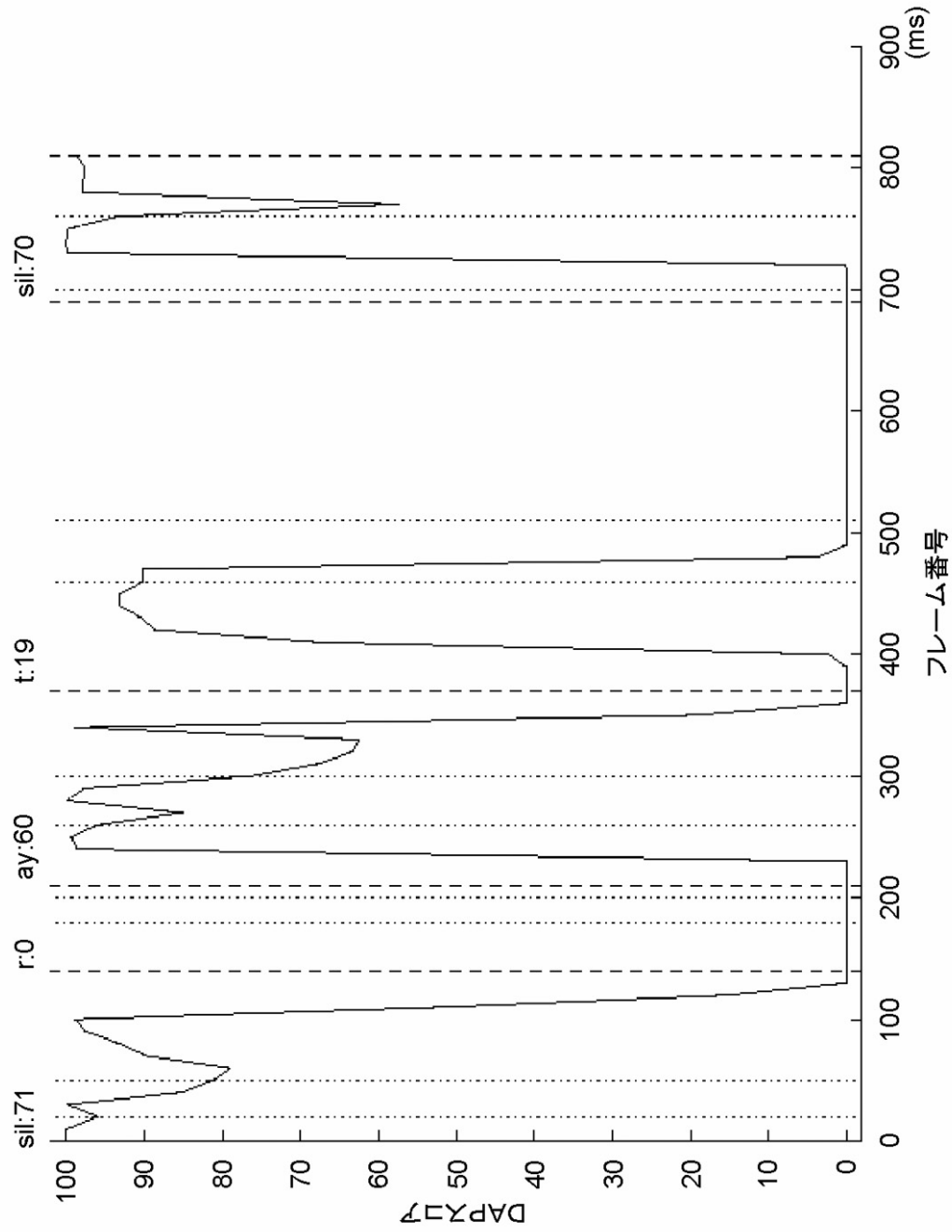
【 図 4 8 】



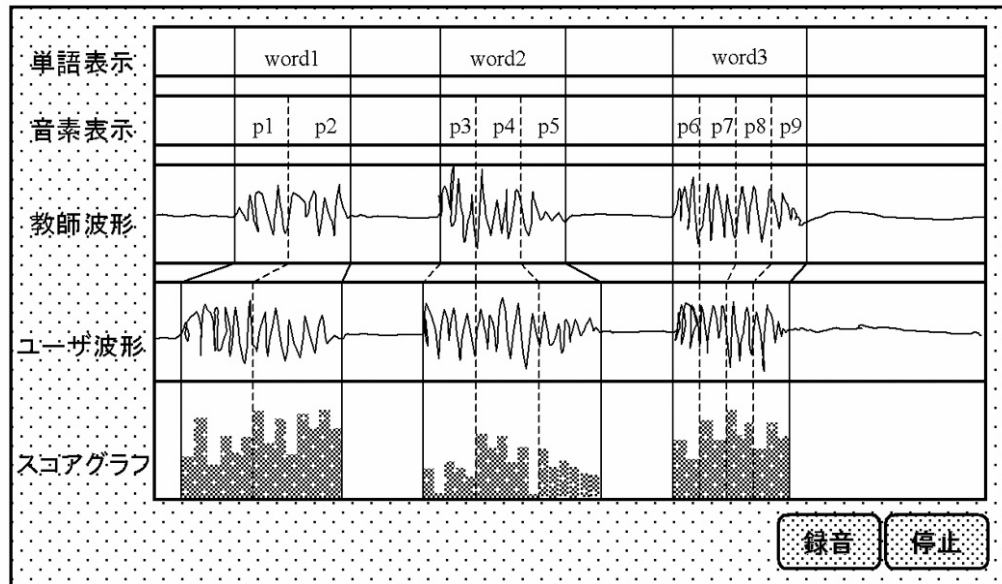
【図 7】



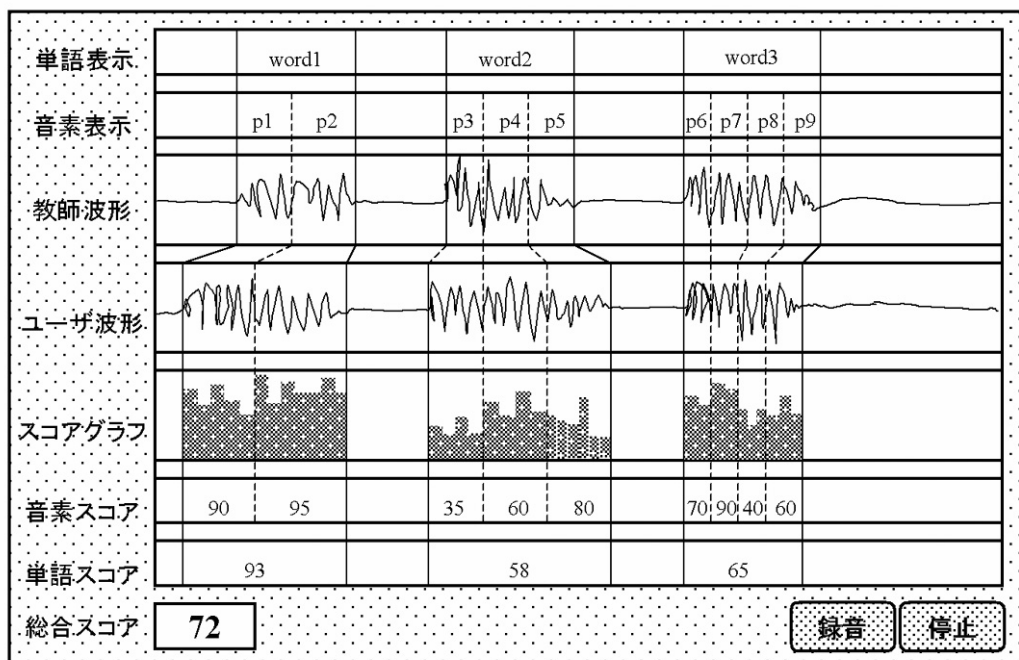
【図 8】



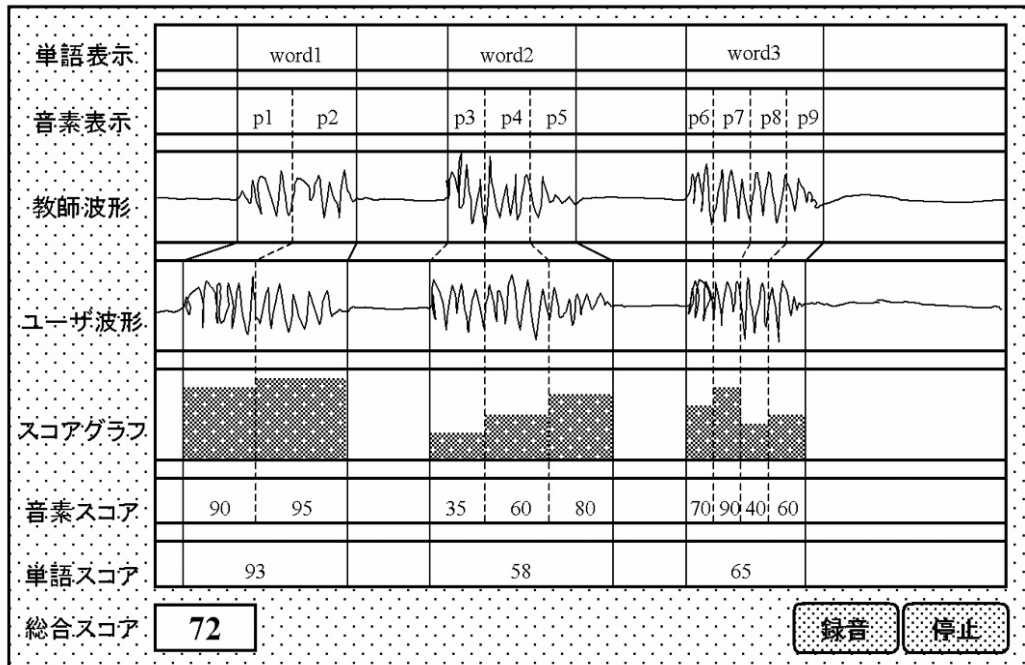
【図 9】



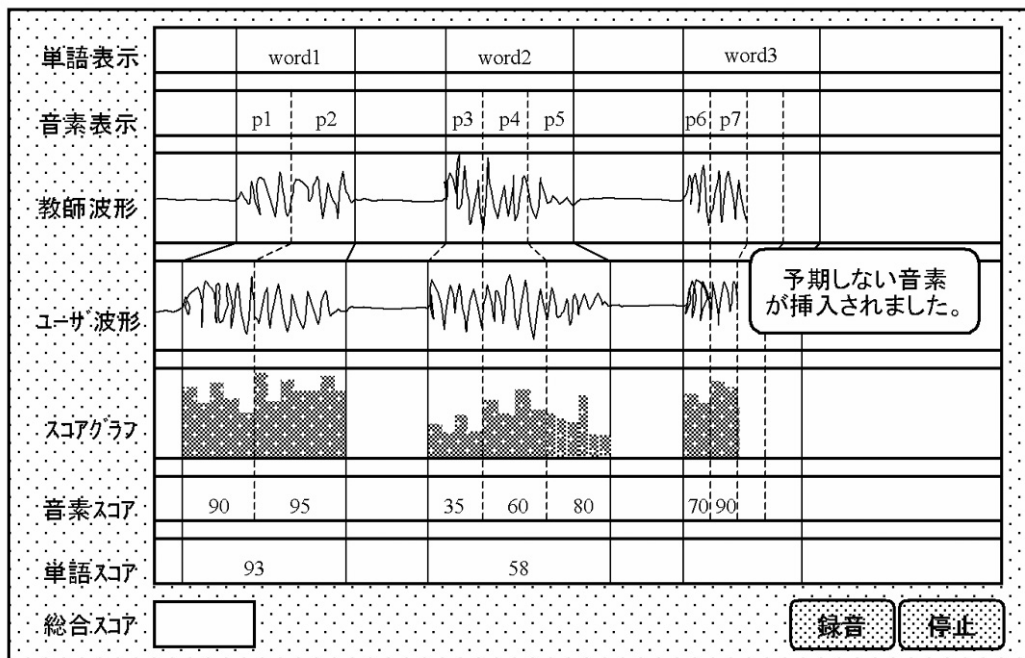
【図 10】



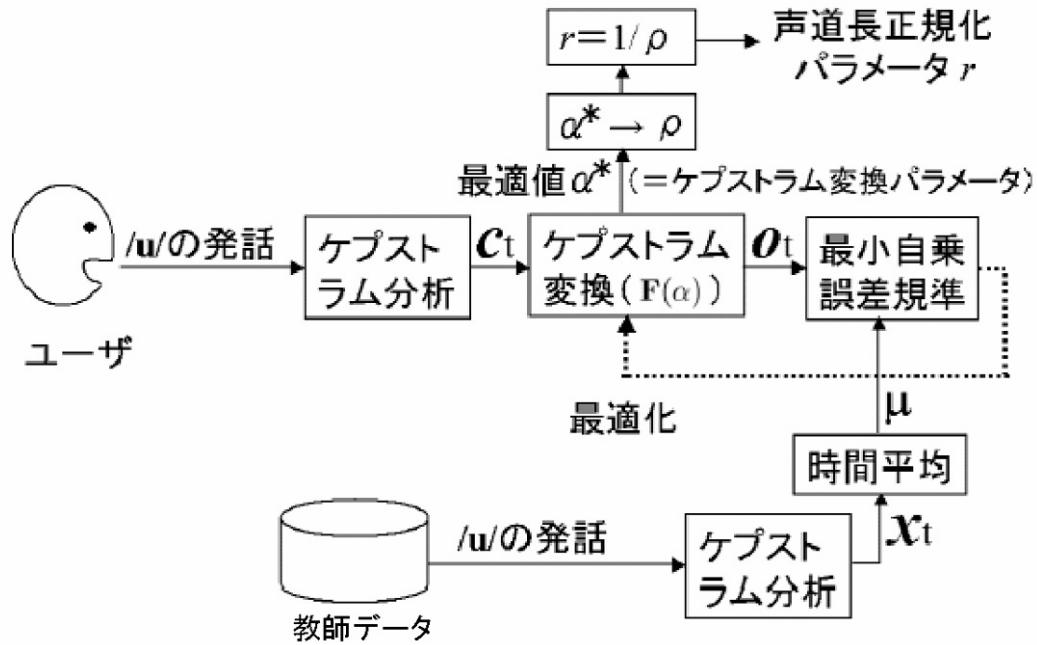
【図 16】



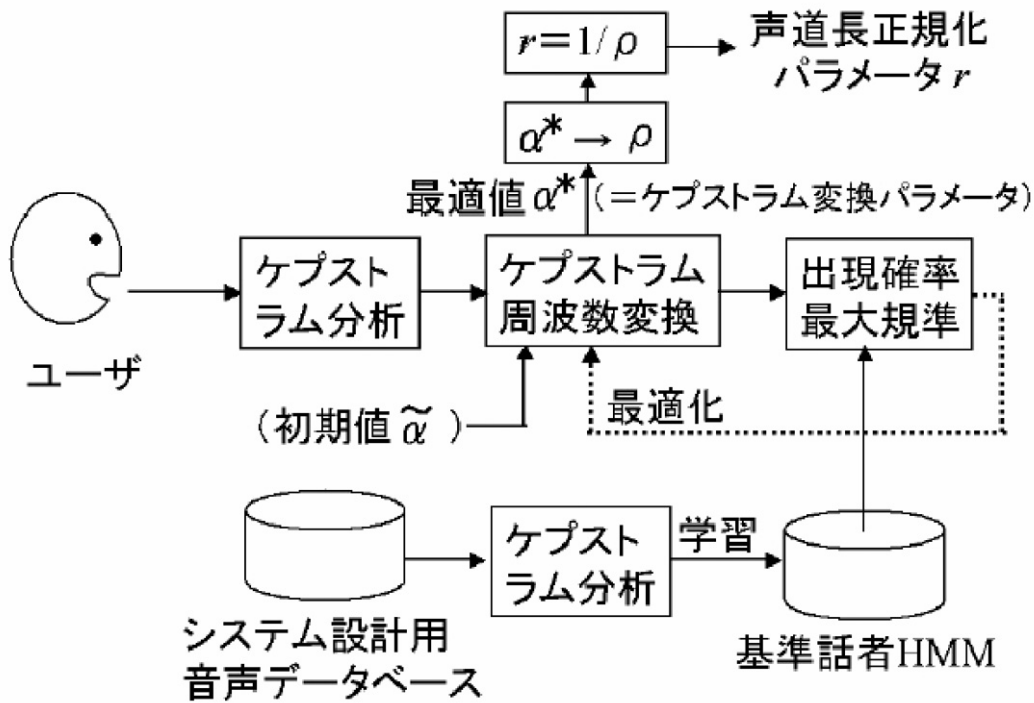
【図 25】



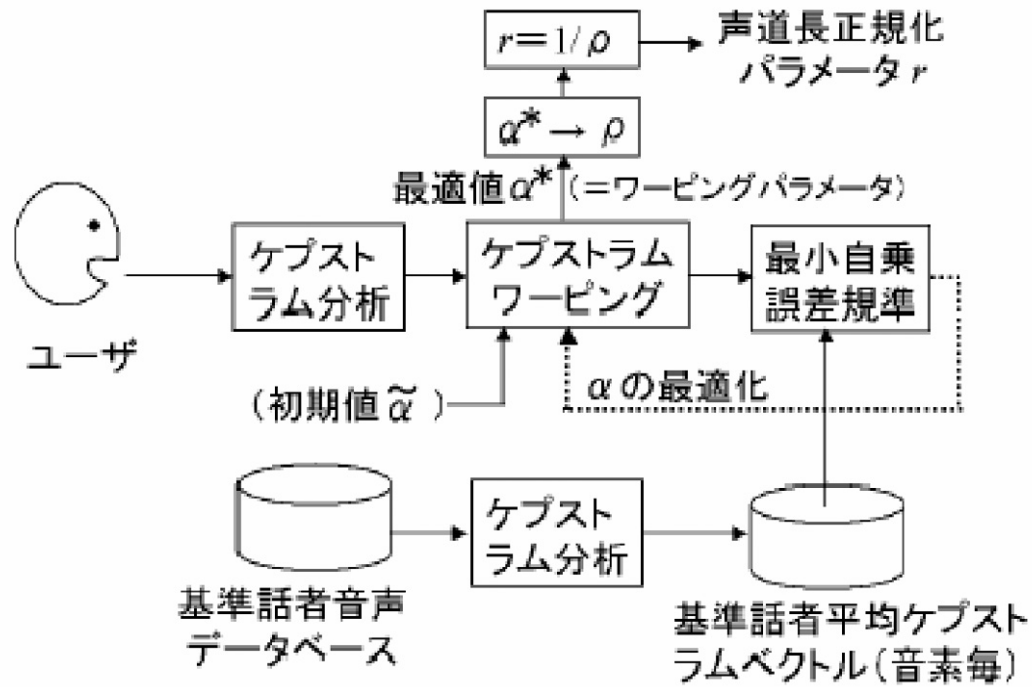
【図 4 0】



【図 4 3】



【図 4 6】



フロントページの続き

(72)発明者 足立 隆弘

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2007-057692(JP,A)

特開平11-327592(JP,A)

特開2001-117598(JP,A)

六井 淳 Jun ROKUI, 最尤推定に基づく線形変換を用いた声道長正規化 Vocal Tract Length Normalization Using Linear Transformation based on Maximum Likelihood Estimation, 情報処理学会研究報告 Vol. 2001 No. 123 IPSJ SIG Notes, 日本, 社団法人情報処理学会 Information Processing Society of Japan, 2001年12月20日

江森 正 Tadashi EMORI, 音声認識のための高速最尤推定を用いた声道長正規化 Vocal Tract Length Normalization Using Rapid Maximum-Likelihood Estimation for Speech Recognition, 電子情報通信学会論文誌 (J83-D-II) 第11号 THE TRANSACTIONS OF THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS D-II, 日本, 社団法人電子情報通信学会 THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS, 2000年12月

(58)調査した分野(Int.Cl., DB名)

G10L 11/00 - 11/06

G10L 15/00 - 17/00