(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4961565号
(P4961565)
(45)発行日　平成24年6月27日(2012.6.27)　(24)登録日　平成24年4月6日(2012.4.6)

(51)Int.Cl.　　F I
G10L 15/00 (2006.01)　　G10L 15/00　２００Ｔ
G10L 15/20 (2006.01)　　G10L 15/20　３６０Ｚ
G10L 11/00 (2006.01)　　G10L 11/00　４０２Ｂ

請求項の数 19　(全 29 頁)

(21)出願番号　特願2007-529275 (P2007-529275)
(86)(22)出願日　平成18年8月1日(2006.8.1)
(86)国際出願番号　PCT/JP2006/315228
(87)国際公開番号　WO2007/015489
(87)国際公開日　平成19年2月8日(2007.2.8)
審査請求日　平成21年6月5日(2009.6.5)
(31)優先権主張番号　特願2005-223155 (P2005-223155)
(32)優先日　平成17年8月1日(2005.8.1)
(33)優先権主張国　日本国(JP)

(73)特許権者　504174135
国立大学法人九州工業大学
福岡県北九州市戸畑区仙水町１番１号
(74)代理人　100121371
弁理士　石田　和人
(72)発明者　佐藤　寧
福岡県北九州市若松区ひびきの２－４　国立大学法人九州工業大学　ヒューマンライフＩＴ開発センター内

審査官　井上　健一

最終頁に続く

(54)【発明の名称】音声検索装置及び音声検索方法

(57)【特許請求の範囲】
【請求項１】
検索対象音声データの中から、クエリー音声データに一致又は類似する部分音声データを検索する音声検索装置であって、
　前記検索対象音声データの有声音のピッチ周期を等化したピッチ等化検索対象音声データの中から、音声の特徴量空間において、前記クエリー音声データの有声音のピッチ周期を等化したピッチ等化クエリー音声データに対する距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）である部分音声データを検索する部分音声検索手段
を備えていることを特徴とする音声検索装置。
【請求項２】
前記クエリー音声データの有声音のピッチ周期を等化することにより前記ピッチ等化クエリー音声データを生成するピッチ周期等化手段と、
　前記ピッチ等化クエリー音声データを特徴量の時系列データに変換したデータ（以下「クエリー特徴データ」という。）を生成する特徴データ生成手段と、
を備え、
　前記部分音声検索手段は、前記ピッチ等化検索対象音声データに含まれる部分音声データのうち、その特徴量が、前記クエリー特徴データとの間の距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）であるものを検索すること
を特徴とする請求項１記載の音声検索装置。
【請求項３】

前記部分音声検索手段は、
前記ピッチ等化検索対象音声データを特徴量の時系列データに変換した検索対象特徴データの中から、前記クエリー音声データと同じ音素長分の部分データ（以下「選択特徴データ」という。）を、選択位置を移動させながら順次選択する部分音声選択手段と、
前記各選択特徴データと前記クエリー特徴データとの間の距離尺度（又は類似尺度）を演算する特徴量尺度演算手段と、
前記距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）の場合、前記選択特徴データに対応する検索対象音声データ内の位置を出力する一致位置判定手段と、
を備えていることを特徴とする請求項１又は２記載の音声検索装置。
【請求項４】
前記検索対象特徴データを記憶する音声記憶手段
を備えていることを特徴とする請求項３記載の音声検索装置。
【請求項５】
前記検索対象音声データの有声音のピッチ周期を等化することにより前記ピッチ等化検索対象音声データを生成する第２のピッチ周期等化手段と、
前記ピッチ等化検索対象音声データを特徴量の時系列データに変換することにより、前記検索対象特徴データを生成する第２の特徴データ生成手段と、
を備えていることを特徴とする請求項３又は４記載の音声検索装置。
【請求項６】
前記ピッチ周期等化手段（又は第２のピッチ周期等化手段）は、
前記クエリー音声データ（又は前記検索対象音声データ）のピッチ周波数の検出を行うピッチ検出手段、
前記ピッチ周波数と所定の基準周波数との差分を演算する残差演算手段、
及び、前記差分が最小となるように、前記クエリー音声データ（又は前記検索対象音声データ）のピッチ周波数を等化する周波数シフタ
を具備することを特徴とする請求項２又は５記載の音声検索装置。
【請求項７】
前記検索対象特徴データ及び前記クエリー特徴データは、それぞれ、前記ピッチ等化検索対象音声データ及び前記ピッチ等化クエリー音声データを直交変換して得られるサブバンド・データの時系列であることを特徴とする請求項１乃至６の何れか一記載の音声検索装置。
【請求項８】
前記クエリー特徴データを、音素区間ごとに平均化し、平均値の時系列データに変換する第１の区間分割手段と、
前記検索対象特徴データを、音素区間ごとに平均化し、平均値の時系列データに変換する第２の区間分割手段と、
を備え、
前記特徴量尺度演算手段は、前記第１及び第２の区間分割手段が生成する平均値の時系列データの間の距離尺度（又は類似尺度）を演算すること
を特徴とする請求項２又は５記載の音声検索装置。
【請求項９】
前記クエリー音声データ（又は前記検索対象音声データ）に対して音素ラベリングを行うことによりクエリー音素列（又は検索対象音素列）を生成する音素ラベリング処理手段と、
前記前記選択特徴データに対応する前記検索対象音素列と前記クエリー音素列との距離尺度（又は類似尺度）を決定する音素列尺度演算手段と、
前記特徴量尺度演算手段が出力する特徴量の距離尺度（又は類似尺度）と、前記音素列尺度演算手段が出力する音素列の距離尺度（又は類似尺度）との線形和（以下「総合距離尺度（又は総合類似尺度）」という。）を算出する総合尺度演算手段と、
を備え、

前記一致位置判定手段は、前記総合距離尺度（又は総合類似尺度）が所定の閾値以下（又は所定の閾値以上）の場合、前記選択特徴データに対応する検索対象音声データ内の位置を出力すること
を特徴とする請求項１乃至８の何れか一記載の音声検索装置。
【請求項１０】
検索対象音声データの中から、クエリー音声データに一致又は類似する部分音声データを検索する音声検索方法であって、
前記検索対象音声データの有声音のピッチ周期を等化したピッチ等化検索対象音声データの中から、音声の特徴量空間において、前記クエリー音声データの有声音のピッチ周期を等化したピッチ等化クエリー音声データに対する距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）である部分音声データを検索する部分音声検索ステップ
を有することを特徴とする音声検索方法。
【請求項１１】
前記クエリー音声データの有声音のピッチ周期を等化することにより前記ピッチ等化クエリー音声データを生成するピッチ周期等化ステップと、
前記ピッチ等化クエリー音声データを特徴量の時系列データに変換したデータ（以下「クエリー特徴データ」という。）を生成する特徴データ生成ステップと、
を備え、
前記部分音声検索ステップにおいては、前記ピッチ等化検索対象音声データに含まれる部分音声データのうち、その特徴量が、前記クエリー特徴データとの間の距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）であるものを検索すること
を特徴とする請求項１０記載の音声検索方法。
【請求項１２】
前記部分音声検索ステップにおいては、
前記ピッチ等化検索対象音声データを特徴量の時系列データに変換した検索対象特徴データの中から、前記クエリー音声データと同じ音素長分の部分データ（以下「選択特徴データ」という。）を、選択位置を移動させながら順次選択する部分音声選択ステップと、
前記各選択特徴データと前記クエリー特徴データとの間の距離尺度（又は類似尺度）を演算する特徴量尺度演算ステップと、
前記距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）の場合、前記選択特徴データに対応する検索対象音声データ内の位置を出力する一致位置判定ステップと、
を有することを特徴とする請求項１０又は１１記載の音声検索方法。
【請求項１３】
前記検索対象特徴データを記憶する音声記憶ステップ
を備えていることを特徴とする請求項１２記載の音声検索方法。
【請求項１４】
前記検索対象音声データの有声音のピッチ周期を等化することにより前記ピッチ等化検索対象音声データを生成する第２のピッチ周期等化ステップと、
前記ピッチ等化検索対象音声データを特徴量の時系列データに変換することにより、前記検索対象特徴データを生成する第２の特徴データ生成ステップと、
を有することを特徴とする請求項１２又は１３記載の音声検索方法。
【請求項１５】
前記ピッチ周期等化ステップ（又は第２のピッチ周期等化ステップ）においては、
前記クエリー音声データ（又は前記検索対象音声データ）のピッチ周波数の検出を行うピッチ検出ステップと、
前記ピッチ周波数と所定の基準周波数との差分を演算する残差演算ステップと、
前記差分が最小となるように、前記クエリー音声データ（又は前記検索対象音声データ）のピッチ周波数を等化する周波数シフトステップと
を具備することを特徴とする請求項１１又は１４記載の音声検索方法。

【請求項１６】
前記検索対象特徴データ及び前記クエリー特徴データは、それぞれ、前記ピッチ等化検索対象音声データ及び前記ピッチ等化クエリー音声データを直交変換して得られるサブバンド・データの時系列であることを特徴とする請求項１０乃至１５の何れか一記載の音声検索方法。
【請求項１７】
前記クエリー特徴データを、音素区間ごとに平均化し、平均値の時系列データに変換する第１の区間分割ステップと、
前記検索対象特徴データを、音素区間ごとに平均化し、平均値の時系列データに変換する第２の区間分割ステップと、
を有し、
前記特徴量尺度演算ステップにおいては、前記第１及び第２の区間分割ステップにおいて生成される平均値の時系列データの間の距離尺度（又は類似尺度）を演算すること
を特徴とする請求項１１又は１４記載の音声検索方法。
【請求項１８】
前記クエリー音声データ（又は前記検索対象音声データ）に対して音素ラベリングを行うことによりクエリー音素列（又は検索対象音素列）を生成する音素ラベリングステップと、
前記選択特徴データに対応する前記検索対象音素列と前記クエリー音素列との距離尺度（又は類似尺度）を決定する音素列尺度演算ステップと、
前記特徴量尺度演算ステップにおいて出力される特徴量の距離尺度（又は類似尺度）と、前記音素列尺度演算ステップにおいて出力される音素列の距離尺度（又は類似尺度）との線形和（以下「総合距離尺度（又は総合類似尺度）」という。）を算出する総合尺度演算ステップと、
を備え、
前記一致位置判定ステップにおいては、前記総合距離尺度（又は総合類似尺度）が所定の閾値以下（又は所定の閾値以上）の場合、前記選択特徴データに対応する検索対象音声データ内の位置を出力すること
を特徴とする請求項１０乃至１７の何れか一記載の音声検索方法。
【請求項１９】
コンピュータに読み込んで実行することにより、コンピュータを請求項１乃至８の何れか一の音声検索装置として機能させることを特徴とするプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、蓄積された検索対象音声データの中から、所定の音声に合致する部分を検索するための音声検索装置に関する。
【背景技術】
【０００２】
近年、多くの蓄積映像・音声データの中から、視聴者が最も知りたい情報の部分だけを取り出すマルチメディア・データベースの要請が強まりつつある。代表的な例としては、蓄積された多くのニュース番組の中から、視聴者が最も知りたいニュースのみを取り出すニュース・オンデマンド（News On Demand：ＮＯＤ）・システムなどがある。
【０００３】
かかるマルチメディア・データベースを構築するためには、テレビニュースなどの蓄積された映像・音声データの中から、検索キーワードの音声（以下「クエリー音声」という。）に合致する部分を検索する音声検索技術が必要とされる。
【０００４】
検索対象音声データの中からクエリー音声に合致する部分を検索する音声検索装置としては、特許文献１に記載のものが公知である。

【０００５】
図１２は、特許文献１に記載の音声検索装置の構成を表す図である。この音声検索装置では、検索データ生成部１００の音声信号入力部１０２に音声信号が入力されると、当該音声信号は、検索対象音声データとして記録部２０１に記憶される。この際、映像検索インデックス生成部１０４が生成する映像検索インデックスが付加される。また、音声信号に同期して映像信号入力部１０１には映像信号が入力され、記録部２０１に蓄積映像データとして記憶される。一方、クエリー音声は、検索処理部２００のキーワード入力部２０３から入力され、キーワードパターン照合部２０５において検索対象音声データと照合され、もっとも一致する音声信号が音声信号出力部２０７から出力される。以下、これらの処理を概説する。

【０００６】
まず、音声信号入力部１０２に音声信号が入力されると、音声特徴パターン抽出部１０３は、入力音声を１０ｍｓｅｃの分析フレームに分割する。そして、各分析フレームについて、高速フーリエ変換を行い、発生周波数帯域の音響特性データを生成する。さらに、この音響特性データを、音響特徴量から構成されるＮ次元のベクトルデータ（以下「特徴パターン」という。）に変換する。ここで、音響特徴量としては、入力音声の発生周波数帯域における短時間スペクトル又はその対数値、入力音声の一定時間内における対数エネルギー等が用いられる。

【０００７】
次に、映像検索インデックス生成部１０４は、音声特徴パターン収納部１０５から第１番目の標準音声パターンを取り出す。

【０００８】
ここで、音声特徴パターン収納部１０５には、５００個の標準音声パターンが予め記憶されている。標準音声パターンとは、予め複数の話者から収集した発音を分析して、サブワード単位（#V, #CV, #CjV, CV, CjV, VC, QC, VQ, VV, V#：但し、Cは子音、Vは母音、jは拗音、Qは促音、#は無音。）で抽出した音声特徴パターンを統計処理して標準化したものである。

【０００９】
映像検索インデックス生成部１０４は、処理対象となる１つの音声区間に対して、第１番目の標準音声パターンと入力音声の音声特徴パターンとの類似度を、ＤＰ照合法やＨＭＭ（Hidden Markov Model）等の音声認識処理により計算される。そして、第１番目の標準音声パターンに対して最も高い類似度を示す区間を「サブワード区間」として検出する。以下、サブワード区間の類似度を「スコア」という。映像検索インデックス生成部１０４は、サブワード区間の音素記号、発声区間（始端時刻、終端時刻）、及びスコアの組を「映像検索インデックス」として出力する。

【００１０】
同様に、第２番目以降の標準音声パターンについてもサブワード区間を検出し、検出サブワード区間に関する映像検索インデックスを出力する。

【００１１】
当該音声区間において、すべての標準音声パターンに関して映像検索インデックスが生成されたならば、映像検索インデックス生成部１０４は、処理対象となる音声区間を隣接する次の音声区間に移し、同様の処理を実行する。そして、入力音声の全区間に亘って映像検索インデックスを作成したところで、処理を終了する。

【００１２】
入力音声の音声データと映像検索インデックスは、検索対象音声データとして記録部２０１に記憶される。図１３は記録部２０１に記憶された映像検索インデックスのラティス構造の一部を示す図である。図１３では、１０ｍｓｅｃ単位で分割した入力音声の各音声区間の終端を、その音声区間に対して生成した各映像検索インデックスの終端とし、同一音声区間における映像検索インデックスを生成された順番に配置している。このような映像検索インデックスのラティス構造を「音素類似度表」と呼ぶ。尚、「ラティス」とは、

連続する種々の音声区間に対して、複数の音素や単語の候補とその可能性を表の形で表したものをいう（非特許文献１，ｐ．１９８参照）。

【００１３】

クエリー音声を用いて映像シーンを検索する処理は次のように行われる。まず、キーワード入力部２０３に検索キーワードであるクエリー音声が入力される。キーワード変換部２０４は、クエリー音声をサブワードの時系列に変換する。次に、キーワードパターン照合部２０５は、音素類似度表の中から、クエリー音声を構成するサブワードだけをピックアップする。そして、ピックアップされた複数のラティス上のサブワードを、検索キーワードを変換したサブワードの系列順に隙間なく接続する。

【００１４】

例えば、クエリー音声としてキーワード入力部２０３に「空（そら）」が入力された場合、キーワード変換部２０４は、サブワードの系列「SO」，「OR」，「RA」を生成する。キーワードパターン照合部２０５は、音素類似度表からサブワード「SO」，「OR」，「RA」をピックアップして、これを隙間なく接続する。この場合、ある時刻のラティスからサブワード「RA」を取り出し、サブワード「RA」の始端時刻にあたるラティスからその前のサブワード「OR」を取り出し、さらにサブワード「OR」の始端時刻に当たるラティスからサブワード「SO」を取り出す。そして、最後のサブワード「RA」の終端を基準にして「SO」「OR」「RA」を連結する。

【００１５】

このようにサブワード（上記例では、「ＳＯ」「ＯＲ」「ＲＡ」）を連結することによって復元されたキーワードについて、その復元キーワードのスコアの総和を計算する。

【００１６】

以下同様に、サブワード「ＲＡ」の終端時刻をずらした復元キーワードをすべての時刻について順次作成し、各復元キーワードについてそのスコアを計算する（図１４参照）。

【００１７】

制御部２０２は、スコアが上位となる復元キーワードの先頭サブワードの始端時刻から対応する映像信号のタイムコードを算出する。そして、記憶部２０１に蓄積された蓄積映像データ・検索対象音声データの該当部分を再生する制御を行う。

【特許文献１】特開２０００－２３６４９４号公報（特許第３２５２２８２号公報）

【特許文献２】特開２００５－９１７０９号公報

【非特許文献１】古井貞熙，「音響・音声工学」，近代科学社，ｐｐ．１９４－２１０

【発明の開示】

【発明が解決しようとする課題】

【００１８】

上記従来の音声検索装置では、音声認識を行うにあたり、音声特徴パターン収納部１０５に格納された標準音声パターンを使用し、クエリー音声と標準音声パターンとの類似度によって音声認識を行う。この場合、認識精度を上げるためには標準音声パターンを多く用意する必要がある。しかし、標準音声パターンの数が増えると、類似度演算の処理時間が増大し又は演算回路の規模が大きくなる。また、標準音声パターンとして登録されていないクエリー音声が入力された場合には、正常に認識することができないため、音声検索機能が正常に働かない場合も考えられる。

【００１９】

また、通常、同じ音素に対する音声であっても男女間で周波数帯域が異なり、また同性でも個人間で周波数帯域が異なる。従って、標準音声パターンとクエリー音声との類似度に、これらの差異による影響が現れるため、認識精度に限界がある。

【００２０】

そこで、本発明の目的は、標準音声パターンを必要とせず、音声の個人差にも影響されず検索精度の高い音声検索装置を提供することにある。

【課題を解決するための手段】

【００２１】

本発明に係る音声検索装置の第１の構成は、検索対象音声データ（retrieval voice-data）の中から、クエリー音声データ（query voice-data）に一致又は類似する部分音声データ（partial voice-data）を検索する音声検索装置（voice retrieval device）であって、前記検索対象音声データの有声音（voiced sound）のピッチ周期（pitch period）を等化したピッチ等化検索対象音声データ（pitch-equalized retrieval voice-data）の中から、音声の特徴量空間において、前記クエリー音声データの有声音のピッチ周期を等化したピッチ等化クエリー音声データに対する距離尺度（distance measure）（又は類似尺度（likelihood measure））が所定の閾値以下（又は所定の閾値以上）である部分音声データを検索する部分音声検索手段を備えていることを特徴とする。

【００２２】

このように、検索対象音声データ及びクエリー音声データのピッチ周期を等化することによって、音声帯域の男女差や個人差が除去される。従って、ピッチ周期が等化された検索対象音声信号及びクエリー音声信号の特徴量空間における距離尺度や類似尺度は、音声帯域の男女差や個人差にほとんど影響されず、その音声が表す音素列に依存して定まる。故に、この距離尺度や類似尺度をマッチングの指標として用いることによって、高い精度で音声検索を行うことが可能となる。

【００２３】

ここで、「特徴量」とは、音声の発生周波数帯域における短時間スペクトル又はその対数値、一定時間内での対数エネルギーなどを用いることができる。特徴量として短時間スペクトルを用いる場合は、例えば、１０～３０チャンネル程度の帯域フィルタ群を用いて得られる各帯域の特徴データの時系列、短時間ＦＦＴを用いて直接的に計算されるスペクトル、ケプストラム変換により得られるケプストラム、相関関数により計算される相関データ列、ＬＰＣ分析を基礎として得られるＬＰＣ係数列、ＰＡＲＣＯＲ係数、ＬＳＰ周波数などが、特徴量として使用される。

【００２４】

「距離尺度」とは、特徴量に応じて種々の距離尺度を用いることができる。例えば、特徴量として短時間スペクトルを使用する場合、単純なユークリッド距離、聴覚の感度を考慮した重み付けを行った距離、判別分析，主成分分析などの統計的分析を行って低次元に射影した空間におけるユークリッド距離、マハラビノス距離、板倉・齋藤距離、COSH尺度、WLR尺度(重みつき尤度比)、PWLR尺度(パワー重みつき尤度比)、LPCケプストラム間ユークリッド距離、LPC重みつきケプストラム間ユークリッド距離などを用いることができる。

【００２５】

尚、特徴量（一般にベクトル量）ｘ，ｙの距離尺度ｄ（ｘ，ｙ）は、必ずしも数学的な意味での距離のように三角不等式を満たす必要はない。しかしながら、次式で定義される対称性と正値性を持つことが望ましく、また、ｄ（ｘ、ｙ）を効率よく計算するアルゴリズムが存在する必要がある。

【００２６】

【数１】

$$\text{(a)}\quad \text{対称性}\quad :\quad d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x}) \tag{1}$$

$$\text{(b)}\quad \text{正値性}\quad :\quad d(\mathbf{x},\mathbf{y}) > 0 \quad (\mathbf{x} \neq \mathbf{y})$$

$$d(\mathbf{x},\mathbf{y}) = 0 \quad (\mathbf{x} = \mathbf{y}) \tag{2}$$

【００２７】

「類似尺度」とは、二つの特徴量がどれだけ類似しているのかを示す尺度をいう。例えば、次式によって定義できる類似度等を用いることができる。ここで、ｘ，ｙは特徴量を表す。

【００２８】

【数２】

$$S(\mathbf{x},\mathbf{y}) = \frac{(\mathbf{x},\mathbf{y})}{||\mathbf{x}|| \cdot ||\mathbf{y}||} \qquad (3)$$

【００２９】

本発明に係る音声検索装置の第２の構成は、前記第１の構成において、前記クエリー音声データの有声音のピッチ周期を等化することにより前記ピッチ等化クエリー音声データを生成するピッチ周期等化手段と、前記ピッチ等化クエリー音声データを特徴量の時系列データに変換したデータ（以下「クエリー特徴データ（query feature-data）」という。）を生成する特徴データ生成手段と、を備え、前記部分音声検索手段は、前記ピッチ等化検索対象音声データに含まれる部分音声データのうち、その特徴量が、前記クエリー特徴データとの間の距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）であるものを検索することを特徴とする。

【００３０】

この構成により、クエリー音声データが入力されると、ピッチ周期等化手段が当該クエリー音声データの有声音のピッチ周期を等化する。そして、特徴データ生成手段は、ピッチ周期が等化されたクエリー音声データの特徴量を演算し、クエリー特徴データを生成する。これにより、部分音声検索手段は、ピッチ等化検索対象音声データの部分音声データとクエリー特徴データとの間の距離尺度（又は類似尺度）を閾値判定により抽出する。これにより、クエリー音声データに一致又は類似する音声データを、検索対象音声データの中から検索することが可能となる。

【００３１】

本発明に係る音声検索装置の第３の構成は、前記第１又は２の構成において、前記部分音声検索手段は、前記ピッチ等化検索対象音声データを特徴量の時系列データに変換した検索対象特徴データの中から、前記クエリー音声データと同じ音素長分の部分データ（以下「選択特徴データ」という。）を、選択位置を移動させながら順次選択する部分音声選択手段と、前記各選択特徴データと前記クエリー特徴データとの間の距離尺度（又は類似尺度）を演算する特徴量尺度演算手段と、前記距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）の場合、前記選択特徴データに対応する検索対象音声データ内の位置を出力する一致位置判定手段と、を備えていることを特徴とする。

【００３２】

この構成により、検索対象音声データの中から、特徴量空間におけるピッチ等化検索対象音声データとの（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）の部分音声データを抽出することが可能となる。

【００３３】

部分音声選択手段が「選択位置を移動」させる手順は、特に限定するものではない。例えば、部分音声データの開始位置を検索対象音声データの先頭から末尾に向かって逐次移動させる方法や、逆に、部分音声データの終端位置を検索対象音声データの末尾から先頭に向かって逐次移動させる方法などを採ることができる。

【００３４】

本発明に係る音声検索装置の第４の構成は、前記第３の構成において、前記検索対象特徴データを記憶する音声記憶手段を備えていることを特徴とする。

【００３５】

検索対象音声データを、検索対象特徴データとして、音声記憶手段に予め記憶させておくことにより、クエリー音声データに類似する部分音声データを素早く検索することが可能となる。

【００３６】

本発明に係る音声検索装置の第５の構成は、前記第３又は４の構成において、前記検索対象音声データの有声音のピッチ周期を等化することにより前記ピッチ等化検索対象音声データを生成する第２のピッチ周期等化手段と、前記ピッチ等化検索対象音声データを特

徴量の時系列データに変換することにより、前記検索対象特徴データを生成する第２の特徴データ生成手段と、を備えていることを特徴とする。

【００３７】

この構成により、音声データベース内の検索対象音声データが有声音のピッチ周期が等化されていない場合であっても、第２のピッチ周期等化手段によりピッチ周期を等化して第２の特徴データ生成手段により特徴量を算出することによって、ピッチ周期が等化された検索対象音声データの特徴量を得ることができる。

【００３８】

本発明に係る音声検索装置の第６の構成は、前記第２又は５の構成において、前記ピッチ周期等化手段（又は第２のピッチ周期等化手段）は、前記クエリー音声データ（又は前記検索対象音声データ）のピッチ周波数の検出を行うピッチ検出手段、前記ピッチ周波数と所定の基準周波数との差分を演算する残差演算手段、及び、前記差分が最小となるように、前記クエリー音声データ（又は前記検索対象音声データ）のピッチ周波数を等化する周波数シフタを具備することを特徴とする。

【００３９】

この構成により、ピッチ周期等化手段（又は第２のピッチ周期等化手段）は、クエリー音声データ（又は前記検索対象音声データ）のピッチ周波数を等化することができる。

【００４０】

本発明に係る音声検索装置の第７の構成は、前記第１乃至６の何れか一の構成において、前記検索対象特徴データ及び前記クエリー特徴データは、それぞれ、前記ピッチ等化検索対象音声データ及び前記ピッチ等化クエリー音声データを直交変換して得られるサブバンド・データの時系列であることを特徴とする。

【００４１】

このように特徴量としてサブバンドを使用することにより、簡単なフィルタバンクやＦＦＴ，ＤＦＴ等を使用して検索対象特徴データ及び前記クエリー特徴データを高速に求めることが可能となる。

【００４２】

本発明に係る音声検索装置の第８の構成は、前記第２又は５の構成において、前記クエリー特徴データを、音素区間ごとに平均化し、平均値の時系列データに変換する第１の区間分割手段と、前記検索対象特徴データを、音素区間ごとに平均化し、平均値の時系列データに変換する第２の区間分割手段と、を備え、前記特徴量尺度演算手段は、前記第１及び第２の区間分割手段が生成する平均値の時系列データの間の距離尺度（又は類似尺度）を演算することを特徴とする。

【００４３】

このように、音素区間で特徴量を平均化し、その平均値を用いてマッチング判定を行うことにより、ノイズや揺らぎの影響が低減され、検索精度が向上する。また、各特徴量は、音素区間ごとに時間的に離散化される。この際に、音声の伸縮の影響が除去される。従って、マッチング判定は単純な比較計算のみとなり、ＤＰマッチングのように計算量の多い方法を用いる必要がなく、装置構成の単純化、演算時間の高速化が図られる。

【００４４】

本発明に係る音声検索装置の第９の構成は、前記第１乃至８の何れか一の構成において、前記クエリー音声データ（又は前記検索対象音声データ）に対して音素ラベリングを行うことによりクエリー音素列（又は検索対象音素列）を生成する音素ラベリング処理手段と、前記前記選択特徴データに対応する前記検索対象音素列と前記クエリー音素列との距離尺度（又は類似尺度）を決定する音素列尺度演算手段と、前記特徴量尺度演算手段が出力する特徴量の距離尺度（又は類似尺度）と、前記音素列尺度演算手段が出力する音素列の距離尺度（又は類似尺度）との線形和（以下「総合距離尺度（又は総合類似尺度）」という。）を算出する総合尺度演算手段と、を備え、前記一致位置判定手段は、前記総合距離尺度（又は総合類似尺度）が所定の閾値以下（又は所定の閾値以上）の場合、前記選択特徴データに対応する検索対象音声データ内の位置を出力することを特徴とする。

【００４５】
このように、特徴量尺度に加えて音素列尺度をマッチング判定に考慮することにより、検索精度を高めることができる。
【００４６】
本発明に係る音声検索方法は、検索対象音声データの中から、クエリー音声データに一致又は類似する部分音声データを検索する音声検索方法であって、前記検索対象音声データの有声音のピッチ周期を等化したピッチ等化検索対象音声データの中から、音声の特徴量空間において、前記クエリー音声データの有声音のピッチ周期を等化したピッチ等化クエリー音声データに対する距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）である部分音声データを検索する部分音声検索ステップを有することを特徴とする。
【００４７】
本発明に係る音声検索方法の第２の構成は、前記第１の構成において、前記クエリー音声データの有声音のピッチ周期を等化することにより前記ピッチ等化クエリー音声データを生成するピッチ周期等化ステップと、前記ピッチ等化クエリー音声データを特徴量の時系列データに変換したデータ（以下「クエリー特徴データ」という。）を生成する特徴データ生成ステップと、を備え、前記部分音声検索ステップにおいては、前記ピッチ等化検索対象音声データに含まれる部分音声データのうち、その特徴量が、前記クエリー特徴データとの間の距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）であるものを検索することを特徴とする。
【００４８】
本発明に係る音声検索方法の第３の構成は、前記第１又は２の構成において、前記部分音声検索ステップにおいては、前記ピッチ等化検索対象音声データを特徴量の時系列データに変換した検索対象特徴データの中から、前記クエリー音声データと同じ音素長分の部分データ（以下「選択特徴データ」という。）を、選択位置を移動させながら順次選択する部分音声選択ステップと、前記各選択特徴データと前記クエリー特徴データとの間の距離尺度（又は類似尺度）を演算する特徴量尺度演算ステップと、前記距離尺度（又は類似尺度）が所定の閾値以下（又は所定の閾値以上）の場合、前記選択特徴データに対応する検索対象音声データ内の位置を出力する一致位置判定ステップと、を有することを特徴とする。
【００４９】
本発明に係る音声検索方法の第４の構成は、前記第３の構成において、前記検索対象特徴データを記憶する音声記憶ステップを備えていることを特徴とする。
【００５０】
本発明に係る音声検索方法の第５の構成は、前記第３又は４の構成において、前記検索対象音声データの有声音のピッチ周期を等化することにより前記ピッチ等化検索対象音声データを生成する第２のピッチ周期等化ステップと、前記ピッチ等化検索対象音声データを特徴量の時系列データに変換することにより、前記検索対象特徴データを生成する第２の特徴データ生成ステップとを有することを特徴とする。
【００５１】
本発明に係る音声検索方法の第６の構成は、前記第２又は５の構成において、前記ピッチ周期等化ステップ（又は第２のピッチ周期等化ステップ）においては、前記クエリー音声データ（又は前記検索対象音声データ）のピッチ周波数の検出を行うピッチ検出ステップと、前記ピッチ周波数と所定の基準周波数との差分を演算する残差演算ステップと、前記差分が最小となるように、前記クエリー音声データ（又は前記検索対象音声データ）のピッチ周波数を等化する周波数シフトステップとを具備することを特徴とする。
【００５２】
本発明に係る音声検索方法の第７の構成は、前記第１乃至６の何れか一の構成において、前記検索対象特徴データ及び前記クエリー特徴データは、それぞれ、前記ピッチ等化検索対象音声データ及び前記ピッチ等化クエリー音声データを直交変換して得られるサブバ

ンド・データの時系列であることを特徴とする。
【００５３】
本発明に係る音声検索方法の第８の構成は、前記第２又は５の構成において、前記クエリー特徴データを、音素区間ごとに平均化し、平均値の時系列データに変換する第１の区間分割ステップと、前記検索対象特徴データを、音素区間ごとに平均化し、平均値の時系列データに変換する第２の区間分割ステップと、を有し、前記特徴量尺度演算ステップにおいては、前記第１及び第２の区間分割ステップにおいて生成される平均値の時系列データの間の距離尺度（又は類似尺度）を演算することを特徴とする。
【００５４】
本発明に係る音声検索方法の第９の構成は、前記第１乃至８の何れか一の構成において、前記クエリー音声データ（又は前記検索対象音声データ）に対して音素ラベリングを行うことによりクエリー音素列（又は検索対象音素列）を生成する音素ラベリングステップと、前記選択特徴データに対応する前記検索対象音素列と前記クエリー音素列との距離尺度（又は類似尺度）を決定する音素列尺度演算ステップと、前記特徴量尺度演算ステップにおいて出力される特徴量の距離尺度（又は類似尺度）と、前記音素列尺度演算ステップにおいて出力される音素列の距離尺度（又は類似尺度）との線形和（以下「総合距離尺度（又は総合類似尺度）」という。）を算出する総合尺度演算ステップと、を備え、前記一致位置判定ステップにおいては、前記総合距離尺度（又は総合類似尺度）が所定の閾値以下（又は所定の閾値以上）の場合、前記選択特徴データに対応する検索対象音声データ内の位置を出力することを特徴とする。
【００５５】
本発明に係るプログラムは、コンピュータに読み込んで実行することにより、コンピュータを前記第１乃至８の何れか一の構成の音声検索装置として機能させることを特徴とする。
【発明の効果】
【００５６】
以上のように、本発明によれば、検索対象音声データ及びクエリー音声データのピッチ周期を等化することにより、音声帯域の男女差や個人差が除去した音声データを用いて、特徴量のマッチングにより音声検索を行うことで、音声帯域の男女差や個人差にほとんど影響されず、音声検索の精度を向上させることができる。
【００５７】
また、音素区間ごとにピッチ周期を等化した検索対象音声データ及びクエリー音声データの特徴量を平均化し、その特徴量の平均値の時間列のマッチング検査によって音声検索を行うことで、ノイズや揺らぎの影響が低減されるとともに、音声の伸縮による影響が除去される。その結果、音声検索の精度を向上させることができる。
【図面の簡単な説明】
【００５８】
【図１】本発明の実施例１に係る音声検索装置１の全体構成を表す図である。
【図２】図１の音声符号化器２の構成を表すブロック図である。
【図３】図２のピッチ周期等化手段１０の構成を表すブロック図である。
【図４】ピッチ検出手段２１及びピッチ平均手段２２における信号処理の概略を説明する図である。
【図５】有声音「あ」のフォルマント特性を示す図である。
【図６】無声音「す」の自己相関及びケプストラム波形並びに周波数特性を示す図である。
【図７】周波数シフタ２３の内部構成を表す図である。
【図８】周波数シフタ２３の内部構成の他の例を表す図である。
【図９】図１の音声復号器５の構成を表すブロック図である。
【図１０】図１の部分音声検索手段６の構成を表すブロック図である。
【図１１】量子化ビット数についての説明図である。

【図１２】特許文献１に記載の音声検索装置の構成を表す図である。
【図１３】記録部２０１に記憶された映像検索インデックスのラティス構造の一部を示す図である。
【図１４】各復元キーワードについてそのスコアを計算するために接続されたラティスの構造を表す図である。
【符号の説明】
【００５９】
１　音声検索装置
２　音声符号化器
３　音声記憶手段
４　データ読出手段
５　音声復号器
６　部分音声検索手段
１０　ピッチ周期等化手段
１１　特徴データ生成手段
１２ａ，１２ｂ　出力切替手段
１３　量子化器
１４　ピッチ等化波形符号化器
１５　差分ビット演算器
１６　ピッチ情報符号化器
１７　音素ラベリング処理手段
１８　リサンプラ
１９　アナライザ
２０　抵抗
２１　入力ピッチ検出手段
２２　ピッチ平均手段
２３　周波数シフタ
２４　出力ピッチ検出手段
２５　残差演算手段
２６　ＰＩＤコントローラ
２７　ピッチ検出手段
２８　ＢＰＦ
２９　周波数カウンタ
３１　ＢＰＦ
３２　周波数カウンタ
３４　アンプ
３６　コンデンサ
４１　発信器
４２　変調器
４３　ＢＰＦ
４４　ＶＣＯ
４５　復調器
５１　ピッチ等化波形復号器
５２　逆量子化器
５３　シンセサイザ
５４　ピッチ情報復号器
５５　ピッチ周波数検出手段
５６　差分器
５７　加算器
５８　周波数シフタ

５９　出力切替手段
６１　動作切替手段
６２　部分音声選択手段
６３，６４　区間分割手段
６５　特徴量尺度演算手段
６６　音素列尺度演算手段
６７　総合尺度演算手段
６８　一致位置判定手段
【発明を実施するための最良の形態】
【００６０】
以下、本発明を実施するための最良の形態について、図面を参照しながら説明する。
【実施例１】
【００６１】
図１は、本発明の実施例１に係る音声検索装置の全体構成を表す図である。実施例１の音声検索装置１は、音声符号化器２、音声記憶手段３、データ読出手段４、音声復号器５、及び部分音声検索手段６を備えている。
【００６２】
検索対象音声データやクエリー音声データは、入力音声データとして音声符号化器２に入力される。音声符号化器２は、入力音声データに対して有声音のピッチ周期を等化するとともに、特徴量の時系列データ（特徴データ）に変換する。この際、入力音声データのピッチ周期の情報は特徴データとは分離され、符号化されて符号化ピッチデータとして出力される。一方、特徴データは、サブバンド波形として出力される。またさらに、音声符号化器２は、特徴データを符号化し、符号化特徴データとして出力する。また、音声符号化器２は、入力音声データに対して音素ラベリング処理を行い、各音素の音素ラベル及び時間区間の情報からなる音素ラベルデータとして出力する。
【００６３】
音声記憶手段３は、音声符号化器２により符号化特徴データ，符号化ピッチデータ，及び音素ラベルデータに分解され符号化された検索対象音声データを記憶する。この音声記憶手段３に記憶された符号化特徴データ及び符号化ピッチデータが、符号化された検索対象特徴データである。
【００６４】
データ読出手段４は、データ選択信号に従って、音声記憶手段３内の符号化された検索対象音声データ（符号化特徴データ，符号化ピッチデータ，及び音素ラベルデータ）の部分データを読み出す。
【００６５】
音声復号器５は、データ読出手段４により読み出された符号化特徴データ及び符号化ピッチデータを復号し、特徴データ又は出力音声データとして出力する。
【００６６】
部分音声検索手段６は、音声記憶手段３に蓄積されている符号化された検索対象音声データから、クエリー音声データに一致又は類似する部分データを検索する。
【００６７】
図２は、図１の音声符号化器２の構成を表すブロック図である。音声符号化器２は、ピッチ周期等化手段１０、特徴データ生成手段１１、出力切替手段１２ａ，１２ｂ、量子化手段１３、ピッチ等化波形符号化器１４、差分ビット演算器１５、ピッチ情報符号化器１６、及び音素ラベリング手段１７を備えている。
【００６８】
ピッチ周期等化手段１０は、入力音声データx_{in}（ｔ）の有声音のピッチ周期を等化する。ピッチ周期が等化された入力音声データ（以下「ピッチ等化音声データ」という。）x_{out}（ｔ）は、出力端子Out_１から出力される。
【００６９】

特徴データ生成手段１１は、出力端子Out_１から出力されるピッチ等化音声データx_{out}（ｔ）を特徴量の時系列データに変換する。本実施例においては、特徴量として、短時間周波数スペクトルが用いられる。

【００７０】

特徴データ生成手段１１は、リサンプラ１８及びアナライザ（変形離散コサイン変換器（Modified Discrete Cosine Transformer：ＭＤＣＴ））１９から構成されている。

【００７１】

リサンプラ１８は、ピッチ周期等化手段１００の出力端子Out_１から出力されるピッチ等化音声データx_{out}（ｔ）の各ピッチ区間について、同一の標本化数となるように再標本化を行い、完全等化音声データx_{eq}（ｔ）として出力する。

【００７２】

アナライザ１９は、完全等化音声データx_{eq}（ｔ）について、一定のピッチ区間数で変形離散コサイン変換を行い、短時間周波数スペクトル（以下「特徴データ」という。）Ｘ（ｆ）を生成する。すなわち、本実施例においては、特徴データは、短時間周波数スペクトルからなるベクトル量の時系列（式（４））として与えられる。

【００７３】

【数３】

$$\mathbf{X}(\mathbf{f}) = (X_{f_1}(t), X_{f_2}(t), \cdots, X_{f_n}(t)) \qquad (4)$$

【００７４】

ここで、ｔは時刻、X_{fi}（ｔ）（ｉ＝１，２，…，ｎ）は時刻ｔにおける周波数f_iのサブバンドの短時間スペクトル値を表す。

【００７５】

出力切替手段１２ａは、部分音声検索手段６から入力される切替信号に従って、アナライザ１９が生成する特徴データＸ（ｆ）の出力先を、部分音声検索手段６又は音声記憶手段３に切り替える。具体的には、入力音声データとして、検索対象音声データが入力される場合には、特徴データＸ（ｆ）の出力先は音声記憶手段３に切り替えられる。入力音声データとして、クエリー音声データが入力される場合には、特徴データＸ（ｆ）の出力先は部分音声検索手段６に切り替えられる。

【００７６】

量子化器１３は、特徴データＸ（ｆ）を所定の量子化曲線に従って量子化する。ピッチ等化波形符号化器１４は、量子化器１３が出力する特徴データＸ（ｆ）を符号化し、符号化特徴データとして出力する。この符号化には、ハフマン符号化法や算術符号化法等のエントロピ符号化法が使用される。

【００７７】

差分ビット演算器１５は、ピッチ等化波形符号化器１４が出力する符号化特徴データの符号量から目的ビット数を減算し差分（以下「差分ビット数」という。）を出力する。量子化器１３は、この差分ビット数によって量子化曲線を平行移動させ、符号化特徴データの符号量が目的ビット数の範囲内となるように調整する。

【００７８】

ピッチ情報符号化器１６は、ピッチ周期等化手段１０が出力する残差周期信号ΔV_{pitch}及び基準周期信号AV_{pitch}を符号化し、符号化ピッチデータとして出力する。この符号化には、ハフマン符号化法や算術符号化法等のエントロピ符号化法が使用される。

【００７９】

音素ラベリング手段１７は、入力音声データを音素区間に区分するとともに、各音素区間に対して音素ラベリングを行う。そして、音素ラベル及び時間区間の情報からなる音素ラベルデータとして出力する。

【００８０】

出力切替手段１２ｂは、音素ラベリング処理手段１７が生成する音素ラベルデータの出力先を、部分音声検索手段６又は音声記憶手段３に切り替える。具体的には、入力音声デ

ータとして、検索対象音声データが入力される場合には、音素ラベルデータの出力先は音声記憶手段３に切り替えられる。入力音声データとして、クエリー音声データが入力される場合には、音素ラベルデータの出力先は部分音声検索手段６に切り替えられる。

【００８１】

図３は、図２のピッチ周期等化手段１０の構成を表すブロック図である。ピッチ周期等化手段１０は、入力ピッチ検出手段２１、ピッチ平均手段２２、周波数シフタ２３、出力ピッチ検出手段２４、残差演算手段２５、及びＰＩＤコントローラ２６を備えている。

【００８２】

入力ピッチ検出手段２１は、入力音声データx_{in}（ｔ）から、当該音声信号に含まれるピッチの基本周波数を検出する。ピッチの基本周波数を検出する方法は、現在までに種々の方法が考案されているが、本実施例ではその代表的なものを示す。この入力ピッチ検出手段２１は、ピッチ検出手段２７、バンドパスフィルタ（Band Pass Filter：以下「ＢＰＦ」という。）２８、及び周波数カウンタ２９を備えている。

【００８３】

ピッチ検出手段２７は、入力音声データx_{in}（ｔ）から、ピッチの基本周期$T_0 = 1/f_0$を検出する。例えば、入力音声データx_{in}（ｔ）が図４（ａ）のような波形であったとする。ピッチ検出手段２７は、まずこの波形に対して短時間フーリエ変換を行い、図４（ｂ）のようなスペクトル波形Ｘ（ｆ）を導出する。

【００８４】

通常、音声波形は、ピッチ以外にも多くの周波数成分を含み、ここで得られるスペクトル波形は、ピッチの基本周波数及びピッチの高調波成分以外にも、付加的に多くの周波数成分を有する。したがって、このスペクトル波形Ｘ（ｆ）からピッチの基本周波数f_0を抽出するのは一般に困難である。そこで、ピッチ検出手段２７は、このスペクトル波形Ｘ（ｆ）に対し再度フーリエ変換を行う。これにより、スペクトル波形Ｘ（ｆ）に含まれるピッチの高調波の間隔Δf_0の逆数$F_0 = 1/\Delta f_0$の点に鋭いピークを持つスペクトル波形が得られる（図４（ｃ）参照）。ピッチ検出手段２７は、このピークの位置F_0を検出することによって、ピッチの基本周波数$f_0 = \Delta f_0/2 = F_0/2$を検出する。

【００８５】

また、ピッチ検出手段２７は、スペクトル波形Ｘ（ｆ）から、入力音声データx_{in}（ｔ）が有声音か無声音かを判別する。有声音の場合には、ノイズフラグ信号V_{noise}として０を出力する。無声音の場合にはノイズフラグ信号V_{noise}として１を出力する。なお、有声音と無声音の判別は、スペクトル波形Ｘ（ｆ）の傾き検出によって行われる。図５は有声音「あ」のフォルマント特性を示す図であり、図６は無声音「す」の自己相関及びケプストラム波形並びに周波数特性を示す図である。有声音は、図５のように、スペクトル波形Ｘ（ｆ）は、全体的に低周波側が大きく高周波側に向かって小さくなるようなフォルマント特性を示す。それに対して、無声音は、図６のように、全体的に高周波側に向かって大きくなるような周波数特性を示す。したがって、スペクトル波形Ｘ（ｆ）の全体的な傾きを検出することによって、入力音声データx_{in}（ｔ）が有声音か無声音かを判別することができる。

【００８６】

尚、入力音声データx_{in}（ｔ）が無声音の場合、ピッチが存在しないので、ピッチ検出手段２７が出力するピッチの基本周波数f_0は無意味な値となる。

【００８７】

ＢＰＦ２８は、通過帯域を外部から設定可能な狭帯域のバンドパスフィルタが使用される。ＢＰＦ２８は、ピッチ検出手段２７により検出されるピッチの基本周波数f_0を通過帯域の中心周波数として設定する（図４（ｄ）参照）。そして、ＢＰＦ２８は、入力音声データx_{in}（ｔ）をフィルタリングし、ピッチの基本周波数f_0のほぼ正弦波状の波形を出力する（図４（ｅ）参照）。

【００８８】

周波数カウンタ２９は、ＢＰＦ２８が出力するほぼ正弦波状の波形のゼロクロス点の時

間間隔をカウントすることにより、ピッチの基本周期T_0＝１／f_0を出力する。この検出されたピッチの基本周期T_0が入力ピッチ検出手段２１の出力信号（以下「基本周波数信号」）として出力される（図４（ｆ）参照）。

【００８９】

ピッチ平均手段２２は、ピッチ検出手段２７が出力するピッチの基本周期信号T_0を平均化するものであり、通常のローパスフィルタ（Low Pass Filter：以下「ＬＰＦ」という。）が使用される。ピッチ平均手段２２により、基本周期信号V_{pitch}が平滑化され、音素内では時間的にほぼ一定の信号となる。この平滑化された基本周期が基準周期T_s（基準周波数f_s＝１／T_s）として使用される（図４（ｇ）参照）。

【００９０】

周波数シフタ２３は、入力音声データx_{in}（ｔ）のピッチ周波数を基準周波数f_0に近づける方向にシフトさせることにより、音声信号のピッチ周期を等化する。

【００９１】

出力ピッチ検出手段２４は、周波数シフタ２３より出力される出力音声データ（以下「ピッチ等化音声データ」という。）x_{out}（ｔ）から、当該ピッチ等化音声データx_{out}（ｔ）に含まれるピッチの基本周期T_0’を検出する。この出力ピッチ検出手段２４も、基本的に入力ピッチ検出手段２１と同様の構成とすることができる。本実施例の場合、出力ピッチ検出手段２４は、ＢＰＦ３１及び周波数カウンタ３２を備えている。

【００９２】

ＢＰＦ３１は、通過帯域を外部から設定可能な狭帯域のＢＰＦが使用される。ＢＰＦ３１は、ピッチ検出手段２７により検出されるピッチの基本周波数f_0を通過帯域の中心周波数として設定する。そして、ＢＰＦ３１は、ピッチ等化音声データx_{out}（ｔ）をフィルタリングし、ピッチの基本周波数f_0’のほぼ正弦波状の波形を出力する。周波数カウンタ３２は、ＢＰＦ３１が出力するほぼ正弦波状の波形のゼロクロス点の時間間隔をカウントすることにより、ピッチの基本周期T_0’＝１／f_0’を出力する。この検出されたピッチの基本周期T_0’が出力ピッチ検出手段２４の出力信号として出力される。

【００９３】

残差演算手段２５は、出力ピッチ検出手段２４が出力する基本周期T_0’からピッチ平均手段２２が出力する基準周期T_sを引いた残差周期ΔT_{pitch}を出力する。この残差周期ΔT_{pitch}は、ＰＩＤコントローラ２６を介して周波数シフタ２３に入力される。周波数シフタ２３は、残差周波数１／ΔT_{pitch}に比例して、入力音声データのピッチ周波数を基準周波数f_0に近づける方向にシフトさせる。

【００９４】

尚、ＰＩＤコントローラ２６は、直列接続されたアンプ３４及び抵抗２０、並びに、アンプ３４に対して並列接続されたコンデンサ３６から構成されている。このＰＩＤコントローラ２６は、周波数シフタ２３、出力ピッチ検出手段２４、及び残差演算手段２５からなるフィードバックループの発振を防止するためのものである。

【００９５】

尚、図３では、ＰＩＤコントローラ２６は、アナログ回路表示しているが、デジタル回路で構成してもよい。

【００９６】

図７は周波数シフタ２３の内部構成を表す図である。周波数シフタ２３は、発信器４１、変調器４２、ＢＰＦ４３、電圧制御発信器（Voltage Controlled Oscillator：以下「ＶＣＯ」という。）４４、及び復調器４５を備えている。

【００９７】

発信器４１は、入力音声データx_{in}（ｔ）の周波数変調を行うための一定周波数の変調キャリア信号C_1を出力する。通常、音声信号の帯域は８ｋＨｚ程度である（図７（ｉ）参照）。したがって、発信器４１が発生する変調キャリア信号C_1の周波数（以下「変調キャリア周波数」という。）としては、通常は２０ｋＨｚ程度のものが使用される。

【００９８】

変調器４２は、発信器４１が出力する変調キャリア信号C_1を入力音声データx_{in}（ｔ）で周波数変調し、被変調信号を生成する。この被変調信号は、変調キャリア周波数を中心として、その両側に音声信号の帯域と同じバンド幅の側波帯（上側波帯及び下側波帯）を有する信号である（図７（ｉｉ）参照）。

【００９９】

ＢＰＦ４３は、変調キャリア周波数を下限遮断周波数とし、入力音声データの帯域幅よりも大きいバンド幅の通過域を有するＢＰＦである。これにより、ＢＰＦ４３から出力される被変調信号は、上側波帯のみが切り出された信号となる（図７（ｉｉｉ）参照）。

【０１００】

ＶＣＯ４４は、発信器４１が出力する変調キャリア信号C_1と同じ周波数の信号を、ＰＩＤコントローラ２６を介して残差演算手段２５から入力される残差周期ΔT_{pitch}の信号（以下「残差周期信号」という。）ΔV_{pitch}により周波数を変調して得られる信号（以下「復調キャリア信号」という。）を出力する。

【０１０１】

復調器４５は、ＢＰＦ４３が出力する上側波帯のみの被変調信号を、ＶＣＯ４４が出力する復調キャリア信号により復調し、音声信号を復元する（図７（ｉｖ）参照）。このとき、復調キャリア信号は、残差周期信号で変調されている。そのため、被変調信号を復調する際に、入力音声データx_{in}（ｔ）のピッチ周波数の基準周波数f_sからのずれが消去される。すなわち、入力音声データx_{in}（ｔ）のピッチ周期は、基準周期T_sに等化される。

【０１０２】

図８は、周波数シフタ２３の内部構成の他の例を表す図である。図８においては、図７の発信器４１とＶＣＯ４４とを入れ替えた構成とされている。この構成によっても、図７の場合と同様に、入力音声データx_{in}（ｔ）のピッチ周期は、基準周期T_sに等化することができる。

【０１０３】

図９は、図１の音声復号器５の構成を表すブロック図である。音声復号器５は、音声符号化器２により符号化された音声信号を復号する装置である。音声復号器５は、ピッチ等化波形復号器５１、逆量子化器５２、シンセサイザ５３、ピッチ情報復号器５４、ピッチ周波数検出手段５５、差分器５６、加算器５７、周波数シフタ５８、及び出力切替手段５９を備えている。

【０１０４】

音声復号器５には、符号化特徴データ及び符号化ピッチデータが入力される。符号化特徴データは、図２のピッチ等化波形符号化器１４から出力される符号化特徴データである。符号化ピッチデータは、図２のピッチ情報符号化器１６から出力される符号化ピッチデータである。

【０１０５】

ピッチ等化波形復号器５１は、符号化特徴データを復号し、量子化後の各サブバンドの特徴データ（以下「量子化特徴データ」という。）を復元する。逆量子化器５２は、この量子化特徴データを逆量子化し、ｎ個のサブバンドの特徴データ$X(f)=\{X(f_1), X(f_2), \cdots, X(f_n)\}$を復元する。

【０１０６】

シンセサイザ５３は、特徴データＸ（ｆ）を逆変形離散コサイン変換（Inverse Modified Discrete Cosine Transform：以下「ＩＭＤＣＴ」という。）し、１ピッチ区間の時系列データ（以下「等化音声信号」という。）x_{eq}（ｔ）を生成する。ピッチ周波数検出手段５５は、この等化音声信号x_{eq}（ｔ）のピッチ周波数を検出し等化ピッチ周波数信号V_{eq}として出力する。

【０１０７】

一方、ピッチ情報復号器５４は、符号化ピッチデータを復号することにより、基準周波数信号AV_{pitch}及び残差周波数信号ΔV_{pitch}を復元する。差分器５６は、基準周波数信

号ΔV_{pitch}から等化ピッチ周波数信号V_{eq}を差し引いた差分を基準周波数変化信号$\Delta \Delta V_{pitch}$として出力する。加算器５７は、残差周波数信号ΔV_{pitch}と基準周波数変化信号$\Delta \Delta V_{pitch}$とを加算してこれを修正残差周波数信号ΔV_{pitch}”として出力する。

【０１０８】

周波数シフタ５８は、図７又は図８に示した周波数シフタ２３と同様の構成を有する。この場合、入力端子Inには等化音声信号x_{eq}（ｔ）が入力され、ＶＣＯ４４には修正残差周波数信号ΔV_{pitch}”が入力される。ＶＣＯ４４は発信器４１が出力する変調キャリア信号C_1と同じキャリア周波数の信号を、加算器５７から入力される修正残差周波数信号ΔV_{pitch}”により周波数変調して得られる信号（以下「復調キャリア信号」という。）を出力するが、この場合、復調キャリア信号の周波数は、キャリア周波数に残差周波数を加えた周波数となる。

【０１０９】

これにより、周波数シフタ５８において等化音声信号x_{eq}（ｔ）の各ピッチ区間のピッチ周期に揺らぎ成分が加えられ、音声信号x_{res}（ｔ）が復元される。

【０１１０】

出力切替手段５９は、部分音声検索手段６から入力される切替信号に従って、逆量子化器５２が生成する特徴データＸ（ｆ）の出力先を、シンセサイザ５３又は部分音声検索手段６に切り替える。具体的には、部分音声検索動作を行う場合には、特徴データＸ（ｆ）の出力先は部分音声検索手段６に切り替えられる。一方、検索対象音声データを外部に出力する場合には、特徴データＸ（ｆ）の出力先はシンセサイザ５３に切り替えられる。

【０１１１】

図１０は、図１の部分音声検索手段６の構成を表すブロック図である。部分音声検索手段６は、動作切替手段６１、部分音声選択手段６２、区間分割手段６３，６４、特徴量尺度演算手段６５、音素列尺度演算手段６６、総合尺度演算手段６７、及び一致位置判定手段６８を備えている。

【０１１２】

動作切替手段６１は、音声検索装置１の動作を、音声記憶手段３に対する検索対象音声データの入出力動作、又は部分音声検索手段６による部分音声検索動作に切り替える切替信号を出力する。

【０１１３】

部分音声選択手段６２は、音声記憶手段３に記憶されている検索対象特徴データ（正確には、符号化された検索対象特徴データ）の中から、部分音声データを選択するためのデータ選択信号を出力する。このデータ選択信号は、データ読出手段４に入力される。データ読出手段４は、データ選択信号に従って、音声記憶手段３に記憶されている検索対象特徴データを選択し読み出す。

【０１１４】

区間分割手段６３は、音声符号化器２のアナライザ１９から入力されるクエリー音声の特徴データ（サブバンド波形）を、音素ラベリング処理手段１７から入力されるクエリー音声の音素ラベルデータの時間区間の情報に従って、音素区間ごとに分割する。そして、それぞれの音素区間ごとに、特徴データを平均化し、平均値の時系列データとして特徴量尺度演算手段６５に出力する。

【０１１５】

区間分割手段６４は、音声復号器５の逆量子化器５２から入力される検索対象音声の特徴データ（サブバンド波形）を、データ読出手段４から入力される検索対象音声の音素ラベルデータの時間区間の情報に従って、音素区間ごとに分割する。そして、それぞれの音素区間ごとに、特徴データを平均化し、平均値の時系列データとして特徴量尺度演算手段６５に出力する。

【０１１６】

特徴量尺度演算手段６５は、区間分割手段６３，６４から入力される特徴データの間の距離尺度D_1（X_q，X_o）を演算する。ここで、距離尺度は、特徴データを構成する各

サブバンド波形の相関係数の線形和として表される。
すなわち、クエリー音声の特徴データを$X_q(f)$、検索対象音声の特徴データを$X_o(f)$とし、それぞれ式（５）（６）で表す。
【０１１７】
【数４】

$$\mathbf{X}_q(\mathbf{f}) = (X_{q,f_1}(t), X_{q,f_2}(t), \cdots, X_{q,f_n}(t)) \tag{5}$$

$$\mathbf{X}_o(\mathbf{f}) = (X_{o,f_1}(t), X_{o,f_2}(t), \cdots, X_{o,f_n}(t)) \tag{6}$$

【０１１８】
特徴データ$X_q(f)$，$X_o(f)$の各サブバンド要素の相関係数は式（７）により表される。ここで、t_jはｊ番目の音素区間を表す。また、$X_{q,f_i}(t_j)$は、ｊ番目の音素区間における特徴データ$X_{q,f_i}(t)$の時間平均値、$X_{o,f_i}(t_j)$は、ｊ番目の音素区間における特徴データ$X_{o,f_i}(t)$を時間平均値である。
【０１１９】
【数５】

$$R(X_{q,f_i}, X_{o,f_i}) = \frac{\sum_{t_j}(X_{q,f_i}(t_j) - \overline{X_{q,f_i}})(X_{o,f_i}(t_j) - \overline{X_{o,f_i}})}{N\sigma_{q,f_i}\sigma_{o,f_i}} \quad (i = 1, 2, \cdots, n) \tag{7}$$

$$\overline{X_{q,f_i}} = \frac{\sum_{j=1}^{N} X_{q,f_i}(t_j)}{N}, \quad \sigma_{q,f_i} = \sqrt{\frac{\sum_{j=1}^{N}(X_{q,f_i}(t_j) - \overline{X_{q,f_i}})^2}{N}} \tag{8}$$

$$\overline{X_{o,f_i}} = \frac{\sum_{j=1}^{N} X_{o,f_i}(t_j)}{N}, \quad \sigma_{o,f_i} = \sqrt{\frac{\sum_{j=1}^{N}(X_{o,f_i}(t_j) - \overline{X_{o,f_i}})^2}{N}} \tag{9}$$

【０１２０】
本実施例１においては、特徴データの間の距離尺度$D_1(X_q, X_o)$を式（１０）により定義する。
【０１２１】
【数６】

$$D_1(\mathbf{X}_q, \mathbf{X}_o) = \sum_{i=1}^{n} w_i R(X_{q,f_i}, X_{o,f_i}) \tag{10}$$

ここで、w_iは重み係数である。重み係数w_iは、適宜設定される。
【０１２２】
音素列尺度演算手段６６は、音声符号化器２の音素ラベリング処理手段１７からクエリー音声の音素ラベルデータが入力されるとともに、データ読出手段４から検索対象音声の音素ラベルデータが入力される。音素列尺度演算手段６６は、これらの音素ラベルデータの距離尺度D_2を所定の音素間距離尺度表を用いて演算する。ここで、音素間距離尺度表とは、すべての２つの音素の組み合わせに対して２つの音素間の距離尺度をテーブルとして表したものである。
【０１２３】
総合尺度演算手段６７は、特徴量尺度演算手段６５が算出する特徴データの間の距離尺度$D_1(X_q, X_o)$と音素列尺度演算手段６６が算出する音素ラベルデータの距離尺度D_2の線形和をとることによって、総合距離尺度Ｄを演算する。すなわち、総合距離尺度Ｄは、式（１１）により表される。
【０１２４】
【数７】

$$D = W_1 D_1(\mathbf{X}_q, \mathbf{X}_o) + W_2 D_2 \tag{11}$$

ここで、W_1，W_2は重み係数であり、適宜決められる。

【０１２５】

一致位置判定手段６８は、距離尺度Ｄが所定の閾値D_{th}以下であるか否かを判定し、$D \leqq D_{th}$の場合には、当該部分データを選択するデータ選択信号を出力する。

【０１２６】

以上のように構成された本実施例の音声検索装置１について、以下その動作を説明する。

【０１２７】

〔１〕検索対象音声データの蓄積動作

まず、検索対象音声データを音声記憶手段３に蓄積する際の動作について説明する。この場合、部分音声検索手段６の動作切替手段６１は、切替信号として検索対象音声データの入出力動作を表すレベル（例えばＨレベル）を出力する。これにより、音声符号化器２の出力切替手段１２ａは、アナライザ１９が生成する特徴データＸ（ｆ）を量子化器１３に出力する。音声符号化器２の出力切替手段１２ｂは、音素ラベリング処理手段１７が生成する音素ラベルデータを音声記憶手段３に出力する。また、音声復号器５の出力切替手段５９は、逆量子化器５２が生成する特徴データＸ（ｆ）をシンセサイザ５３に出力する。

【０１２８】

まず、検索対象音声データとして入力音声データx_{in}（ｔ）が音声符号化器２へ入力されると、ピッチ周期等化手段１０の入力ピッチ検出手段２１は、入力音声データx_{in}（ｔ）が有声音か無声音かを判別してノイズフラグ信号V_{noise}を出力端子OUT_４へ出力するとともに、入力音声データx_{in}（ｔ）からピッチ周波数を検出し、基本周波数信号V_{pitch}をピッチ平均手段２２に出力する。ピッチ平均手段２２は、基本周波数信号V_{pitch}を平均化し（この場合、ＬＰＦを使用するので加重平均となる。）、これを基準周波数信号AV_{pitch}として出力する。この基準周波数信号AV_{pitch}は、出力端子OUT_３から出力されるとともに、残差演算手段２５に入力される。

【０１２９】

一方、周波数シフタ２３は、入力音声データx_{in}（ｔ）の周波数をシフトさせ、ピッチ等化音声データx_{out}（ｔ）として出力端子Out_１へ出力する。初期状態においては、残差周波数信号ΔV_{pitch}は０（リセット状態）であり、周波数シフタ２３は、入力音声データx_{in}（ｔ）がそのままピッチ等化音声データx_{out}（ｔ）として出力端子Out_１へ出力される。

【０１３０】

次に、出力ピッチ検出手段２４は、周波数シフタ２３が出力する出力音声データのピッチ周波数f_0’を検出する。検出されたピッチ周波数f_0’は、ピッチ周波数信号V_{pitch}’として残差演算手段２５に入力される。

【０１３１】

残差演算手段２５は、ピッチ周波数信号V_{pitch}’から基準周波数信号AV_{pitch}を差し引くことにより、残差周波数信号ΔV_{pitch}を生成する。この残差周波数信号ΔV_{pitch}は、出力端子Out_２へ出力されるとともに、ＰＩＤコントローラ２６を介して周波数シフタ２３へ入力される。

【０１３２】

周波数シフタ２３は、ＰＩＤコントローラ２６を介して入力される残差周波数信号ΔV_{pitch}に比例して、周波数のシフト量を設定する。この場合、残差周波数信号ΔV_{pitch}が正値であれば、残差周波数信号ΔV_{pitch}に比例した量だけ周波数を下げるようにシフト量が設定される。残差周波数信号ΔV_{pitch}が負値であれば、残差周波数信号ΔV_{pitch}に比例した量だけ周波数を上げるようにシフト量が設定される。

【０１３３】

このようなフィードバック制御により、入力音声データx_{in}（ｔ）のピッチ周期は、常に基準周期$1/f_s$に維持され、ピッチ等化音声データx_{out}（ｔ）のピッチ周期は等化

される。

【０１３４】

このように、ピッチ周期等化手段１０において、入力音声データx_{in}（ｔ）に含まれる情報は、

（ａ）有声音か無声音かを示す情報；

（ｂ）１ピッチ区間の音声波形を表す情報；

（ｃ）基準ピッチ周波数の情報；

（ｄ）各ピッチ区間のピッチ周波数の基準ピッチ周波数からの偏倚量を表す残差周波数情報；

に分離される。（ａ）～（ｄ）の情報は、それぞれ、ノイズフラグ信号V_{noise}、ピッチ周期が基準周期１／f_s（入力音声データの過去のピッチ周波数の加重平均の逆数）に等化されたピッチ等化音声データx_{out}（ｔ）、基準周波数信号AV_{pitch}、及び残差周波数信号ΔV_{pitch}として出力される。ノイズフラグ信号V_{noise}は出力端子Out_４から出力され、ピッチ等化音声データx_{out}（ｔ）は出力端子Out_１から出力され、基準周波数信号AV_{pitch}は出力端子Out_３から出力され、残差周波数信号ΔV_{pitch}は出力端子Out_２から出力される。

【０１３５】

ピッチ等化音声データx_{out}（ｔ）は、男女差、個人差、音素、感情及び会話内容によって変化するピッチ周波数のジッタ成分や変化成分が除去された音声信号であり、抑揚のない平坦的・機械的な音声信号である。したがって、同じ有声音のピッチ等化音声データx_{out}（ｔ）は、男女差、個人差、音素、感情又は会話内容に無関係にほぼ同じ波形が得られるため、ピッチ等化音声データx_{out}（ｔ）を比較することによって有声音についてのマッチングを精度よく行うことが可能となる。

【０１３６】

また、有声音のピッチ等化音声データx_{out}（ｔ）はピッチ周期が基準周期１／f_sに等化されているので、一定数のピッチ区間でサブバンド符号化を行うことにより、ピッチ等化音声データx_{out}（ｔ）の周波数スペクトルX_{out}（ｆ）は、基準周波数の高調波成分のサブバンド成分に集約される。音声はピッチ間の波形相関が大きいので、各サブバンド成分のスペクトル強度の時間変化は緩やかである。したがって、各サブバンド成分を符号化し、その他の雑音成分を省略することにより、高効率の符号化が可能となる。また、基準周波数信号AV_{pitch}、及び残差周波数信号ΔV_{pitch}は、音声の性質上、同一音素内で狭レンジでしか変動しないため、高効率の符号化が可能である。したがって、全体として入力音声データx_{in}（ｔ）の有声音成分を高効率で符号化することが可能となる。

【０１３７】

次に、リサンプラ１８は、各ピッチ区間において、基準周波数信号AV_{pitch}を一定のリサンプリング数ｎで除算することによりリサンプリング周期を計算する。そして、ピッチ等化音声データx_{out}（ｔ）をそのリサンプリング周期によりリサンプリングし、等標本数音声データx_{eq}（ｔ）として出力する。これにより、ピッチ等化音声データx_{out}（ｔ）の１ピッチ区間の標本化数が一定の値とされる。

【０１３８】

次に、アナライザ１９は、等標本数音声データx_{eq}（ｔ）を、一定のピッチ区間数のサブフレームに区分する。そして、サブフレーム毎に変形離散コサイン変換を行うことによって周波数スペクトル信号Ｘ（ｆ）を生成する。

【０１３９】

ここで、１つのサブフレームの長さは、１ピッチ周期の整数倍とされる。本実施例では、サブフレームの長さは１ピッチ周期（標本化数ｎ）とする。従って、ｎ個の周波数スペクトル信号｛Ｘ（f_1），Ｘ（f_2），…，Ｘ（f_n）｝が出力される。周波数f_1は基準周波数の第１高調波、周波数f_2は基準周波数の第２高調波、周波数f_nは基準周波数の第ｎ高調波である。

【０１４０】

このように、１ピッチ周期の整数倍のサブフレームに分割して各サブフレームを直交変換することによりサブバンド符号化を行うことで、音声波形データの周波数スペクトル信号は基準周波数の高調波のスペクトルに集約される。そして、音声の性質上、同一の音素内における連続するピッチ区間の波形は類似する、従って、隣接するサブフレーム間で基準周波数の高調波成分のスペクトルは類似する、従って、符号化効率は高められる。

【０１４１】

次に、量子化器１３は、周波数スペクトル信号Ｘ（ｆ）を量子化する。ここで、量子化器１３はノイズフラグ信号V_{noise}を参照し、ノイズフラグ信号V_{noise}が０（有声音）の場合と１（無声音）の場合とで量子化曲線を切り換える。

【０１４２】

ノイズフラグ信号V_{noise}が０（有声音）の場合、量子化曲線は、図２（ａ）に示したように、周波数が高くなるに従って量子化ビット数が減少するような量子化曲線とされる。これは、有声音の周波数特性は、図５に示したように低周波数域で大きく高周波域になるに従って減少する特性を有することに対応させたものである。

【０１４３】

一方、ノイズフラグ信号V_{noise}が１（無声音）の場合、量子化曲線は、図２（ｂ）に示したように、周波数が高くなるに従って量子化ビット数が増加するような量子化曲線とされる。これは、無声音の周波数特性は、図６に示したように高周波域になるに従って増加する特性を有することに対応させたものである。

【０１４４】

この量子化曲線の切り換えにより、有声音か無声音かに対応して最適な量子化曲線が選択される。

【０１４５】

尚、補足として、量子化ビット数について説明する。量子化器１３による量子化のデータフォーマットは図１１（ａ）（ｂ）に示したように、小数点以下の実数部（ＦＬ）及び２の冪乗を表す指数部（ＥＸＰ）によって表現される。但し、０以外の数を表す場合において、実数部（ＦＬ）の先頭の１ビットは必ず１であるように指数部（ＥＸＰ）が調整されるものとする。

【０１４６】

例えば、実数部（ＦＬ）が４ビット、指数部（ＥＸＰ）が２ビットの場合において、４ビットで量子化する場合、及び２ビットで量子化する場合は、次のようになる（図１１（ｃ），（ｄ）参照）。

【０１４７】

（１）４ビットで量子化する場合

（例１）　Ｘ（ｆ）＝８＝［１０００］$_2$（但し、［　］$_2$は２進数表記を表す。）は、

　　ＦＬ＝［１０００］$_2$，ＥＸＰ＝［１００］$_2$

（例２）　Ｘ（ｆ）＝７＝［０１００］$_2$は、

　　ＦＬ＝［１１１０］$_2$，ＥＸＰ＝［０１１］$_2$

（例３）　Ｘ（ｆ）＝３＝［１０００］$_2$は、

　　ＦＬ＝［１１００］$_2$，ＥＸＰ＝［０１０］$_2$

【０１４８】

（２）２ビットで量子化する場合

（例１）　Ｘ（ｆ）＝８＝［１０００］$_2$は、

　　ＦＬ＝［１０００］$_2$，ＥＸＰ＝［１００］$_2$

（例２）　Ｘ（ｆ）＝７＝［０１００］$_2$は、

　　ＦＬ＝［１１００］$_2$，ＥＸＰ＝［０１１］$_2$

（例３）　Ｘ（ｆ）＝３＝［１０００］$_2$は、

　　ＦＬ＝［１１００］$_2$，ＥＸＰ＝［０１０］$_2$

【０１４９】

すなわち、ｎビットで量子化する場合は、実数部（ＦＬ）の先頭からｎビットを残し、

残りのビットは０とするものとする（図１１（ｄ）参照）。

【０１５０】

次に、ピッチ等化波形符号化器１４は、量子化器１３が出力する量子化された周波数スペクトル信号Ｘ（ｆ）をエントロピ符号化法により符号化し、符号化特徴データを出力する。また、ピッチ等化波形符号化器１４は、符号化特徴データの符号量（ビット数）を差分ビット演算器１５に出力する。差分ビット演算器１５は、符号化特徴データの符号量から所定の目的ビット数を減算し、差分ビット数を出力する。量子化器１３は、差分ビット数に応じて、有声音に対する量子化曲線を平行移動的に上下させる。

【０１５１】

例えば、｛f_1，f_2，f_3，f_4，f_5，f_6｝に対する量子化曲線が｛６，５，４，３，２，１｝であったとし、差分ビット数として２が入力されたとすると、量子化器１３は、量子化曲線を下方に２だけ平行移動する。その結果、量子化曲線は｛４，３，２，１，０，０｝となる。また、差分ビット数として－２が入力されたとすると、量子化器１３は、量子化曲線を上方に２だけ平行移動する。その結果、量子化曲線は｛８，７，６，５，４，３｝となる。

【０１５２】

このように有声音の量子化曲線を上下に変化させることによって、各サブフレームの符号化特徴データの符号量が目的ビット数程度に調整される。

【０１５３】

一方、これに並行して、ピッチ情報符号化器１６は、基準周波数信号AV_{pitch}及び残差周波数信号ΔV_{pitch}を符号化する。

【０１５４】

一方、音素ラベリング処理手段１７は、入力音声データx_{in}（ｔ）を音素区間に区分し、各音素区間に対して音素ラベリングを行う。音素区間の分割方法や音素ラベリングの方法に関しては、音声認識の分野において多くの技術が公知であり、ここではそれら公知の方法を用いることができる。音素ラベリング処理手段１７は、音素ラベリングにより得られた音素ラベルと各音素ラベルに対する時間区間を表す音素区間の情報を、音素ラベルデータとして出力する。

【０１５５】

以上のようにして生成された、符号化特徴データ，符号化ピッチデータ，及び音素ラベルデータは、音声記憶手段３に出力され、保存される。

【０１５６】

〔２〕音声復号器の動作

データ読出手段４が、音声記憶手段３から符号化特徴データ及び符号化ピッチデータを読み出すと、これらのデータは音声復号器５に入力される。

【０１５７】

音声復号器５のピッチ等化波形復号器５１は、符号化特徴データを復号し、量子化後の各サブバンドの周波数スペクトル信号（以下「量子化周波数スペクトル信号」という。）を復元する。逆量子化器５２は、この量子化周波数スペクトル信号を逆量子化し、ｎ個のサブバンドの周波数スペクトル信号Ｘ（ｆ）＝｛Ｘ（f_1），Ｘ（f_2），…，Ｘ（f_n）｝を復元する。

【０１５８】

シンセサイザ５３は、周波数スペクトル信号Ｘ（ｆ）を逆変形離散コサイン変換（Inverse Modified Discrete Cosine Transform：以下「ＩＭＤＣＴ」という。）し、１ピッチ区間の時系列データ（以下「等化音声信号」という。）x_{eq}（ｔ）を生成する。ピッチ周波数検出手段５５は、この等化音声信号x_{eq}（ｔ）のピッチ周波数を検出し等化ピッチ周波数信号V_{eq}として出力する。

【０１５９】

一方、ピッチ情報復号器５４は、符号化ピッチデータを復号することにより、基準周波数信号AV_{pitch}及び残差周波数信号ΔV_{pitch}を復元する。差分器５６は、基準周波数信

号ΔV_{pitch}から等化ピッチ周波数信号V_{eq}を差し引いた差分を基準周波数変化信号$\Delta \Delta V_{pitch}$として出力する。加算器５７は、残差周波数信号ΔV_{pitch}と基準周波数変化信号$\Delta \Delta V_{pitch}$とを加算してこれを修正残差周波数信号ΔV_{pitch}”として出力する。

【０１６０】

周波数シフタ５８は、図７又は図８に示した周波数シフタ２３と同様の構成を有する。この場合、入力端子Inには等化音声信号x_{eq}（ｔ）が入力され、ＶＣＯ４４には修正残差周波数信号ΔV_{pitch}”が入力される。ＶＣＯ４４は発信器４１が出力する変調キャリア信号C_1と同じキャリア周波数の信号を、加算器５７から入力される修正残差周波数信号ΔV_{pitch}”により周波数変調して得られる信号（以下「復調キャリア信号」という。）を出力するが、この場合、復調キャリア信号の周波数は、キャリア周波数に残差周波数を加えた周波数となる。

【０１６１】

これにより、周波数シフタ５８において等化音声信号x_{eq}（ｔ）の各ピッチ区間のピッチ周期に揺らぎ成分が加えられ、音声信号x_{res}（ｔ）が復元される。

【０１６２】

〔３〕クエリー音声データによる部分音声データの検索動作

次に、クエリー音声データによる部分音声データの検索動作について説明する。この場合、部分音声検索手段６の動作切替手段６１は、切替信号として部分音声検索動作を表すレベル（例えばＬレベル）を出力する。これにより、音声符号化器２の出力切替手段１２ａは、アナライザ１９が生成する特徴データＸ（ｆ）を部分音声検索手段６に出力する。音声符号化器２の出力切替手段１２ｂは、音素ラベリング処理手段１７が生成する音素ラベルデータを部分音声検索手段６に出力する。また、音声復号器５の出力切替手段５９は、逆量子化器５２が生成する特徴データＸ（ｆ）を部分音声検索手段６に出力する。

【０１６３】

まず、クエリー音声データは、入力音声データx_{in}（ｔ）として音声符号化器２に入力される。

【０１６４】

ピッチ周期等化手段１では、上述のように、入力音声データx_{in}（ｔ）の有声音のピッチ周期を等化し、ピッチ等化音声データx_{out}（ｔ）として出力端子Out_１から出力する。また、特徴データ生成手段１９は、上述のように、ピッチ等化音声データx_{out}（ｔ）を短時間スペクトルの時系列からなる特徴データＸ（ｆ）に変換する。特徴データＸ（ｆ）は、出力切替手段１２ａを介して部分音声検索手段６へ出力される。

【０１６５】

一方、音素ラベリング処理手段１７では、上述のように、入力音声データx_{in}（ｔ）を音素区間に区分し、各音素区間に対して音素ラベリングを行う。そして、音素ラベルと音素区間の情報を、音素ラベルデータとして出力する。

【０１６６】

次に、部分音声検索手段６の部分音声選択手段６２は、音声記憶手段３に記憶された符号化特徴データ，符号化ピッチデータ，及び音素ラベルデータを、データの先頭から順に順次読み出すためのデータ選択信号を出力する。このとき、読み出す部分データの長さは、クエリー音声データと同じ音素長の長さとされる。データ読出手段４は、データ選択信号に従って、音声記憶手段３から部分データを読み出す。

【０１６７】

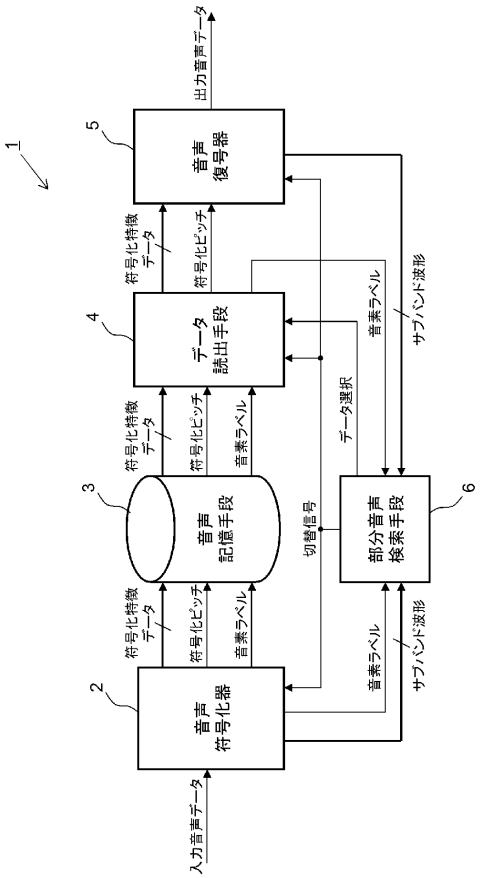
データ読出手段４により読み出された音素ラベルデータは、部分音声検索手段６に入力される。

【０１６８】

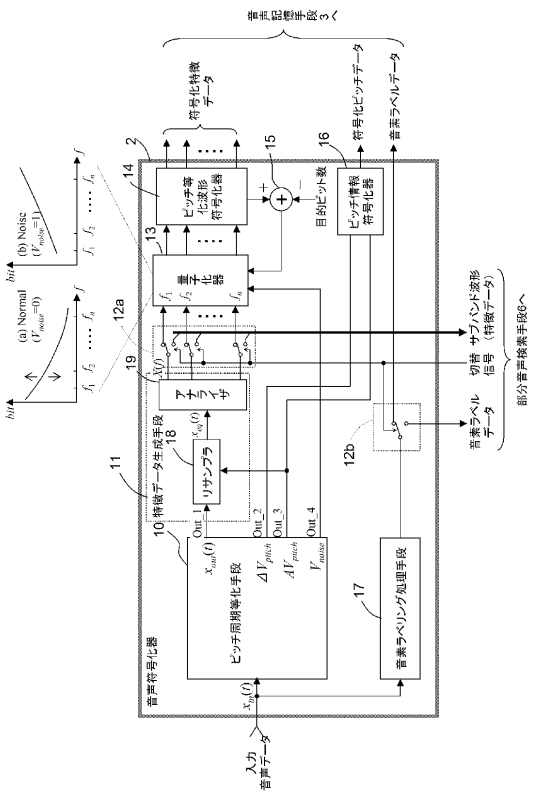
一方、データ読出手段４により読み出された符号化特徴データ及び符号化ピッチデータの部分データは、音声復号器５に入力される。音声復号器５では、上述のように、ピッチ等化波形復号器５１で符号化特徴データを復号し、逆量子化器５２で逆量子化を行うことにより、特徴データを生成し、部分音声検索手段６に出力する。

【０１６９】
以下、音声復号器５から部分音声検索手段６に入力される検索対象特徴データの部分データを「選択特徴データ」と呼ぶ。
【０１７０】
部分音声検索手段６においては、音声符号化器２からクエリー音声の特徴データ（以下「クエリー特徴データ」という。）及び音素ラベルデータが入力されると、区間分割手段６３は、クエリー特徴データを音素区間ごとに平均化し、平均値の時系列データに変換する。この場合、音素ラベルデータに含まれる音素区間の情報に基づき、クエリー特徴データを時間区間に区分し、各時間区間で平均値をとればよい。この平均値の時系列データは、特徴量尺度演算手段６５に入力される。
【０１７１】
また、音声復号器５及びデータ読出手段４から選択特徴データ及び音素ラベルデータが入力されると、区間分割手段６４は、選択特徴データを音素区間ごとに平均化し、平均値の時系列データに変換する。この平均値の時系列データは、特徴量尺度演算手段６５に入力される。
【０１７２】
特徴量尺度演算手段６５は、区間分割手段６３及び区間分割手段６４から入力される平均値の時系列データの間の距離尺度$D_1(X_q, X_o)$を式（１０）に従って算出する。
【０１７３】
一方、音素列尺度演算手段６６は、音声符号化器２から入力されるクエリー音声の音素ラベルデータとデータ読出手段から入力される検索対象音声の音素ラベルデータとの間の距離尺度D_2を音素間距離尺度表を用いて演算する。
【０１７４】
総合尺度演算手段６７は、特徴量尺度演算手段６５が算出する特徴データの間の距離尺度$D_1(X_q, X_o)$と音素列尺度演算手段６６が算出する音素ラベルデータの距離尺度D_2の線形和をとることによって、総合距離尺度Ｄを式（１１）により演算する。
【０１７５】
一致位置判定手段６８は、距離尺度Ｄが所定の閾値D_{th}以下であるか否かを判定し、$D \leqq D_{th}$の場合には、当該部分データを選択するデータ選択信号を出力する。そして、動作切替手段６１は、切替信号として部分音声検索動作を表すレベル（例えばＬレベル）を出力する。
【０１７６】
これにより、検索された検索対象データの部分データが、出力音声データとして出力される。
【０１７７】
尚、本実施例は、音声情報と映像とが一体として記録されたマルチメディア・データベースにおける情報の検索においても適用することができる。
【産業上の利用可能性】
【０１７８】
本発明は、音声データベースや音声情報を含むマルチメディア・データベース等において利用可能である。

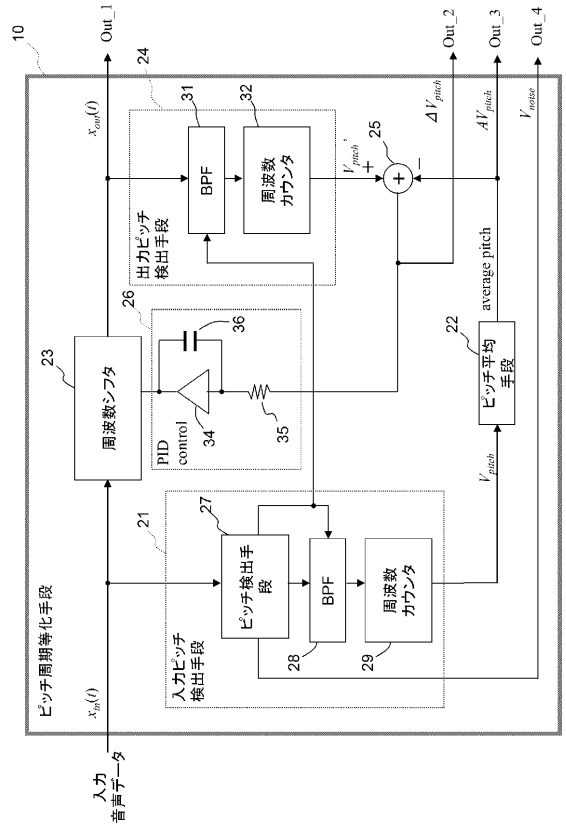
【図１】


1
2 音声符号化器
3 音声記憶手段
4 データ読出手段
5 音声復号器
6 部分音声検索手段
入力音声データ
出力音声データ
符号化特徴データ
符号化ピッチ
音素ラベル
サブバンド波形
切替信号
データ選択

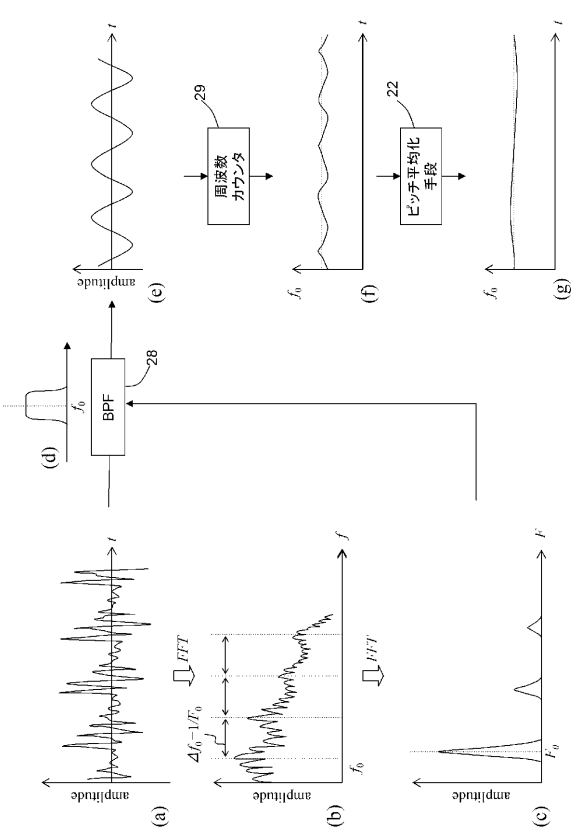
【図２】


音声記憶手段3へ
符号化特徴データ
符号化ピッチデータ
音素ラベルデータ
目的ビット数
ピッチ帯化波形符号化器
ピッチ情報符号化器
量子化器
(a) Normal
(b) Noise
12a
12b
19
アナライザ
18
リサンプラ
11
特徴データ生成手段
10
音声符号化器
ピッチ同期等化手段
17
音素ラベリング処理手段
切替信号
サブバンド波形〈特徴データ〉
音素ラベルデータ
部分音声検索手段6へ
入力音声データ

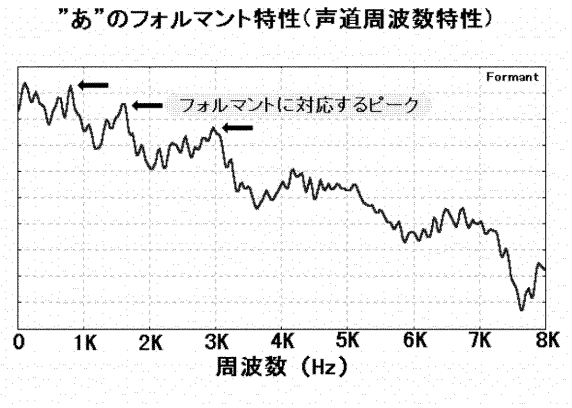
【図３】


10
Out_1
Out_2
Out_3
Out_4
ピッチ周期等化手段
出力ピッチ検出手段
入力ピッチ検出手段
BPF
周波数カウンタ
ピッチ検出手段
周波数シフタ
PID control
ピッチ平均手段
average pitch
入力音声データ

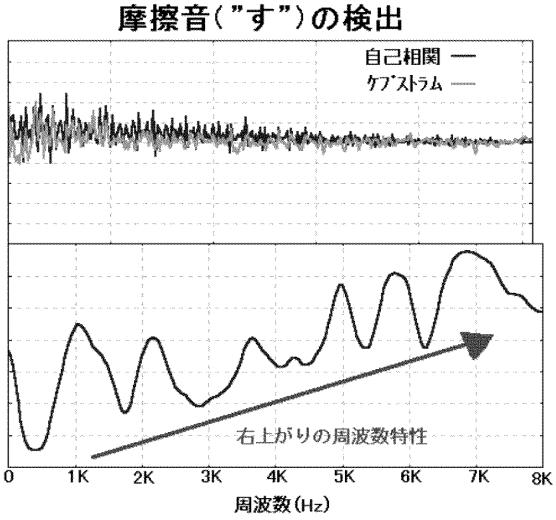
【図４】


(a)
(b)
(c)
(d)
(e)
(f)
(g)
FFT
IFFT
amplitude
BPF
周波数カウンタ
ピッチ平均化手段

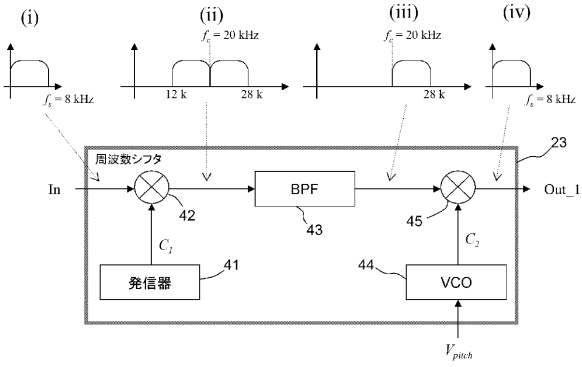
【図５】


"あ"のフォルマント特性（声道周波数特性）
Formant
フォルマントに対応するピーク
0
1K
2K
3K
4K
5K
6K
7K
8K
周波数（Hz）

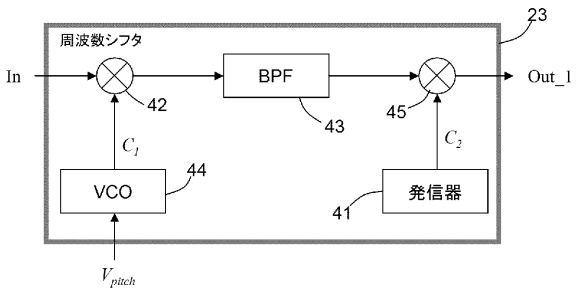
【図６】


摩擦音（"す"）の検出
自己相関
ケプストラム
右上がりの周波数特性
0
1K
2K
3K
4K
5K
6K
7K
8K
周波数(Hz)

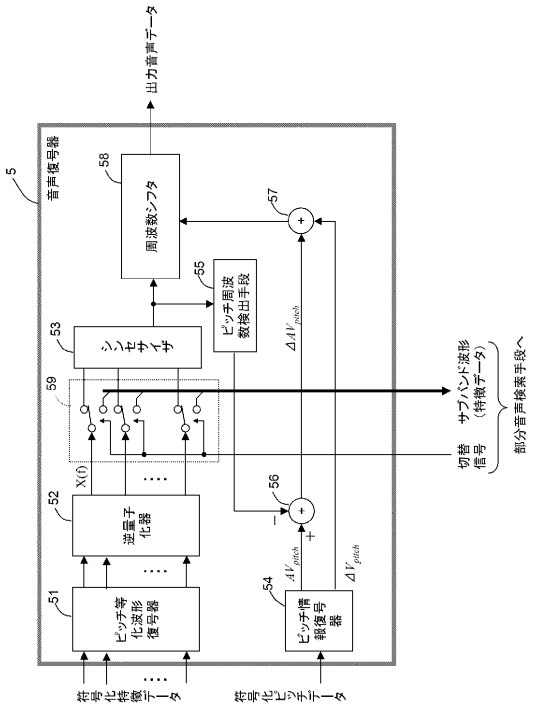
【図７】


(i)
(ii)
(iii)
(iv)
f_s = 8 kHz
f_s = 20 kHz
12 k
28 k
周波数シフタ
23
In
BPF
Out_1
42
43
45
C_1
C_2
発信器
41
44
VCO
V_{pitch}

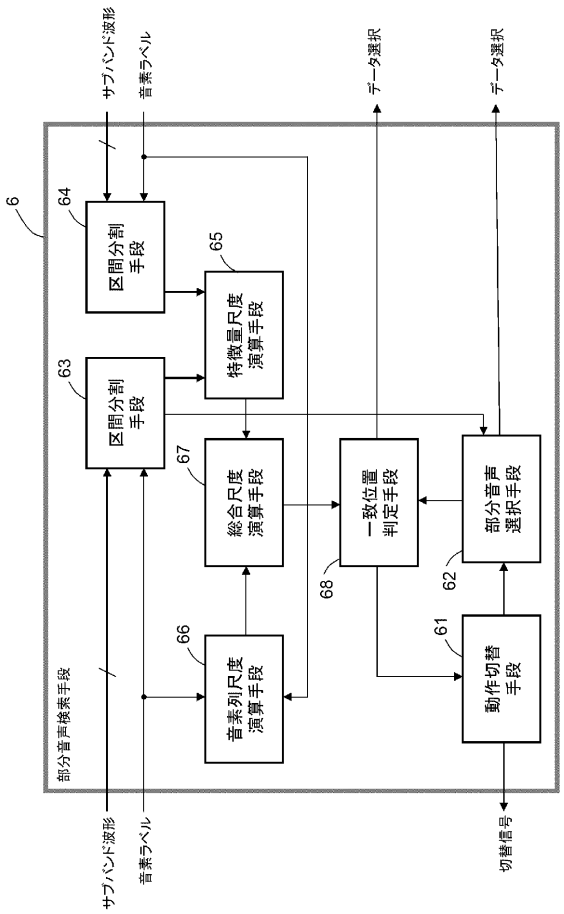
【図８】


周波数シフタ
23
In
BPF
Out_1
42
43
45
C_1
C_2
VCO
44
41
発信器
V_{pitch}

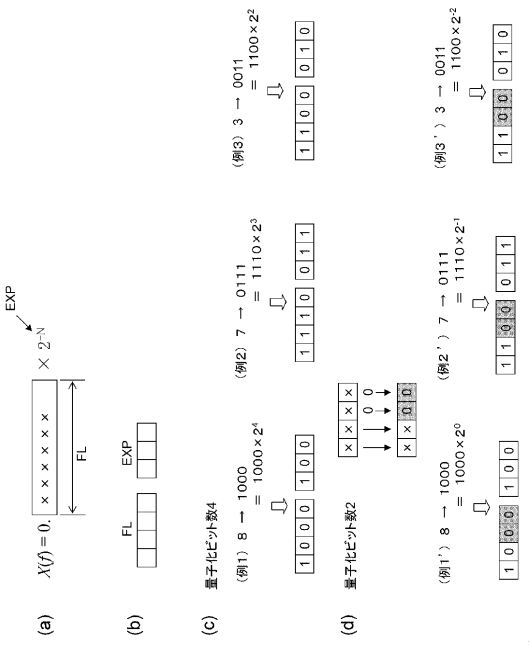
【図９】


出力音声データ
音声復号器
5
周波数シフタ
58
57
ピッチ周波数検出手段
55
ΔV_{pitch}
ふくらませ
53
59
サブバンド波形（特徴データ）
切替信号
部分音声検索手段へ
X(i)
逆量子化器
52
56
51
ピッチ等化波形復号器
54
ピッチ情報復号器
ΔV_{pitch}
ΔV_{pitch}

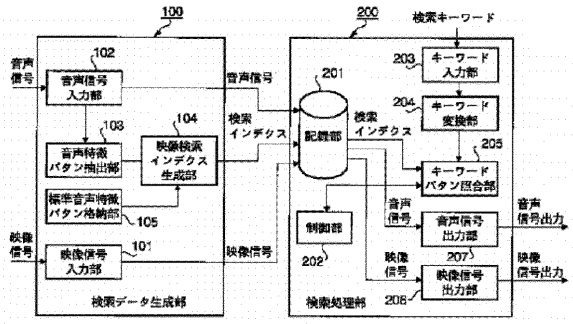
【図１０】



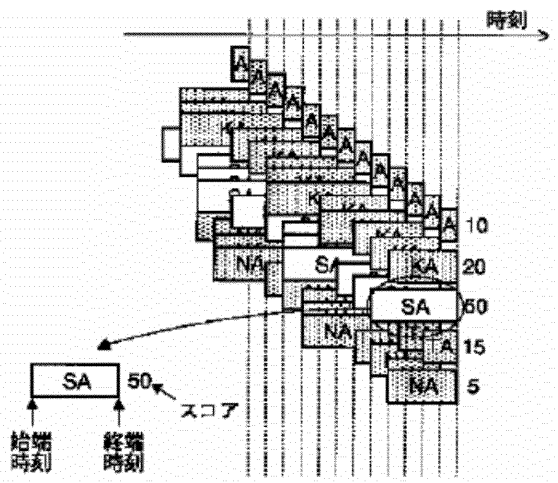
【図１１】



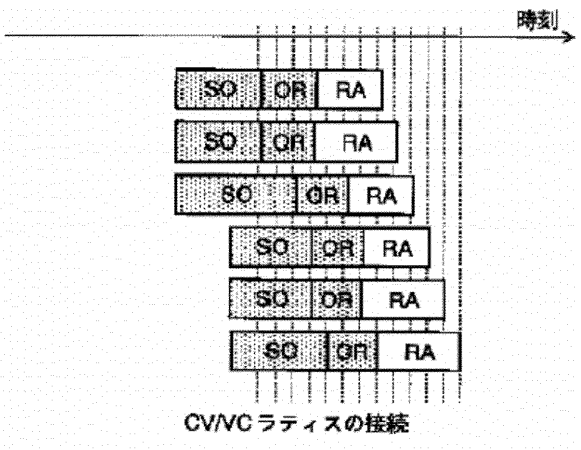
【図１２】



【図１３】



【図１４】



フロントページの続き

(56)参考文献 特許第２８３４４７１（ＪＰ，Ｂ２）
特許第３２５２２８２（ＪＰ，Ｂ２）
特開昭５９－９９５００（ＪＰ，Ａ）
Ｈ．Ｓｉｎｇｅｒ，Ｓ．Ｓａｇａｙａｍａ，Pitch dependent phone modelling for HMM-based speech recognition，Journal of the Acoustical Society of Japan (E)，１９９４年 3月

(58)調査した分野(Int.Cl.，ＤＢ名)
G10L 15/00
G10L 11/00
G10L 11/04
G10L 15/02
G10L 15/10
G10L 15/20