

(19) 世界知的所有権機関
国際事務局(43) 国際公開日
2007 年 2 月 8 日 (08.02.2007)

PCT

(10) 国際公開番号
WO 2007/015489 A1

(51) 国際特許分類:

G10L 15/00 (2006.01) G10L 15/02 (2006.01)
G06F 17/30 (2006.01) G10L 15/10 (2006.01)
G10L 11/00 (2006.01) G10L 15/20 (2006.01)
G10L 11/04 (2006.01)

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 佐藤 寧 (SATO, Yasushi) [JP/JP]; 〒8080196 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 Fukuoka (JP).

(21) 国際出願番号: PCT/JP2006/315228

(22) 国際出願日: 2006 年 8 月 1 日 (01.08.2006)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:

特願2005-223155 2005 年 8 月 1 日 (01.08.2005) JP

(71) 出願人 (米国を除く全ての指定国について): 国立大学法人九州工業大学 (KYUSHU INSTITUTE OF TECHNOLOGY) [JP/JP]; 〒8048550 福岡県北九州市戸畑区仙水町 1-1 Fukuoka (JP).

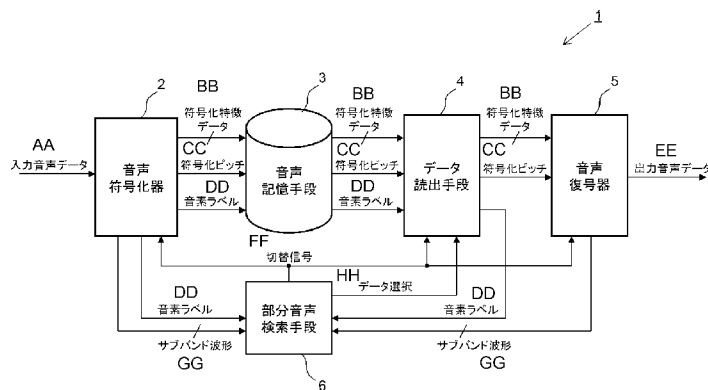
(74) 代理人: 石田 和人 (ISHIDA, Kazuto); 〒8080135 福岡県北九州市若松区ひびきの 2 番 1 号 北九州学術研究都市産学連携センター T-302 Fukuoka (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY,

[続葉有]

(54) Title: VOICE SEARCH DEVICE AND VOICE SEARCH METHOD

(54) 発明の名称: 音声検索装置及び音声検索方法



2 VOICE ENCODER
3 VOICE STORAGE MEANS
4 DATA READ OUT MEANS
5 AUDIO DECODER
6 PARTIAL VOICE SEARCH MEANS
AA INPUT VOICE DATA
BB ENCODED FEATURE DATA
CC ENCODED PITCH
DD PHONEME LABEL
EE OUTPUT VOICE DATA
FF SWITCHING SIGNAL
GG SUB BAND WAVEFORM
HH DATA SELECTION

(57) Abstract: It is possible to provide a voice search device having a high search accuracy not affected by a difference between individual voices without requiring a standard voice pattern. The voice search device includes partial voice search means for searching partial voice data in a voice feature amount space from pitch equalization search object voice data in which a pitch cycle of a voiced sound of search object voice data is equalized. The partial voice data has a distance scale is not greater than a predetermined threshold value for pitch equalization query voice data in which a pitch cycle of a voiced sound of query voice data is equalized. By equalizing the pitch cycle, it is possible to perform a voice search with a high accuracy not affected by a difference of voice range between a male and a female or individuals.

(57) 要約: 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。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、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が所定の閾値以下である部分音声データを検索する部分音声検索手段

[続葉有]

WO 2007/015489 A1



TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

規則4.17に規定する申立て:

— 発明者である旨の申立て (規則 4.17(iv))

添付公開書類:

— 国際調査報告書

2文字コード及び他の略語については、定期発行される各*PCT*ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

明 細 書

音声検索装置及び音声検索方法

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、その音声が表示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值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]

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

$$(b) \quad \text{正值性} \quad : \quad d(x, y) > 0 \quad (x \neq y) \\ d(x, y) = 0 \quad (x = y) \quad (2)$$

[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]

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

- [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ト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 リサンブラ

- 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]

$$X(f) = (X_{f_1}(t), X_{f_2}(t), \dots, X_{f_n}(t)) \quad (4)$$

[0074] ここで、 t は時刻、 $X_{fi}(t)$ ($i=1, 2, \dots, n$)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_{pi} 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 $1/\Delta T_{pitch}$ 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 $X(f) = \{X(f_1), X(f_2), \dots, X(f_n)\}$ 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 AV_{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 ΔAV_{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 ΔAV_{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]

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

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

- [0118] 特徴データ $\mathbf{X}_q(\mathbf{f})$, $\mathbf{X}_o(\mathbf{f})$ 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、 $X_{q,fi}(t_j)$ は、j番目の音素区間における特徴データ $X_{q,fi}(t)$ の時間平均値、 $X_{o,fi}(t_j)$ は、j番目の音素区間における特徴データ $X_{o,fi}(t)$ を時間平均値である。

- [0119] [数5]

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

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

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

- [0120] 本実施例1においては、特徴データの間の距離尺度 $D_1(\mathbf{X}_q, \mathbf{X}_o)$ を式(10)により定義する。

- [0121] [数6]

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

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 $D_1(\mathbf{X}_q, \mathbf{X}_o)$ 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]

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

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、 $D \leq D_{th}$ 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} が正值であれば、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 $\{X(f_1), X(f_2), \dots, X(f_n)\}$ 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 $X(f) = \{X(f_1), X(f_2), \dots, X(f_n)\}$ 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_q 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 AV_{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

ΔV_{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 ΔV_{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なる。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、 $D \leq D_{th}$ 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。

請求の範囲

- [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。
- [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。
- [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。
- [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。
- [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。
- [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。
- [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。
- [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。
- [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。

- [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。

- [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9記載の音声検索方法。

[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9又は10記載の音声検索方法。

[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11記載の音声検索方法。

[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11又は12記載の音声検索方法。

[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

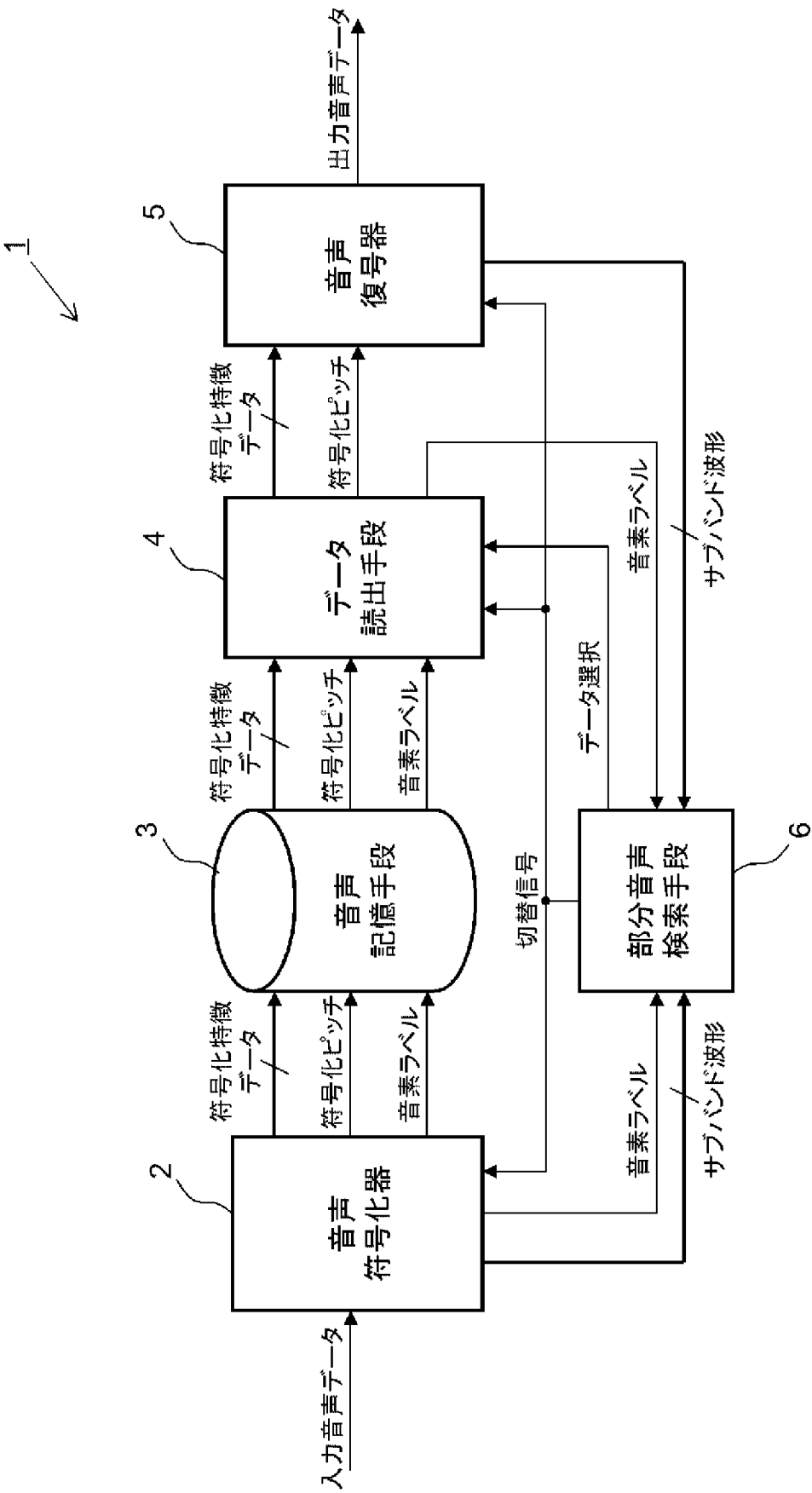
を具備することを特徴とする請求項11又は14記載の音声検索方法。

[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

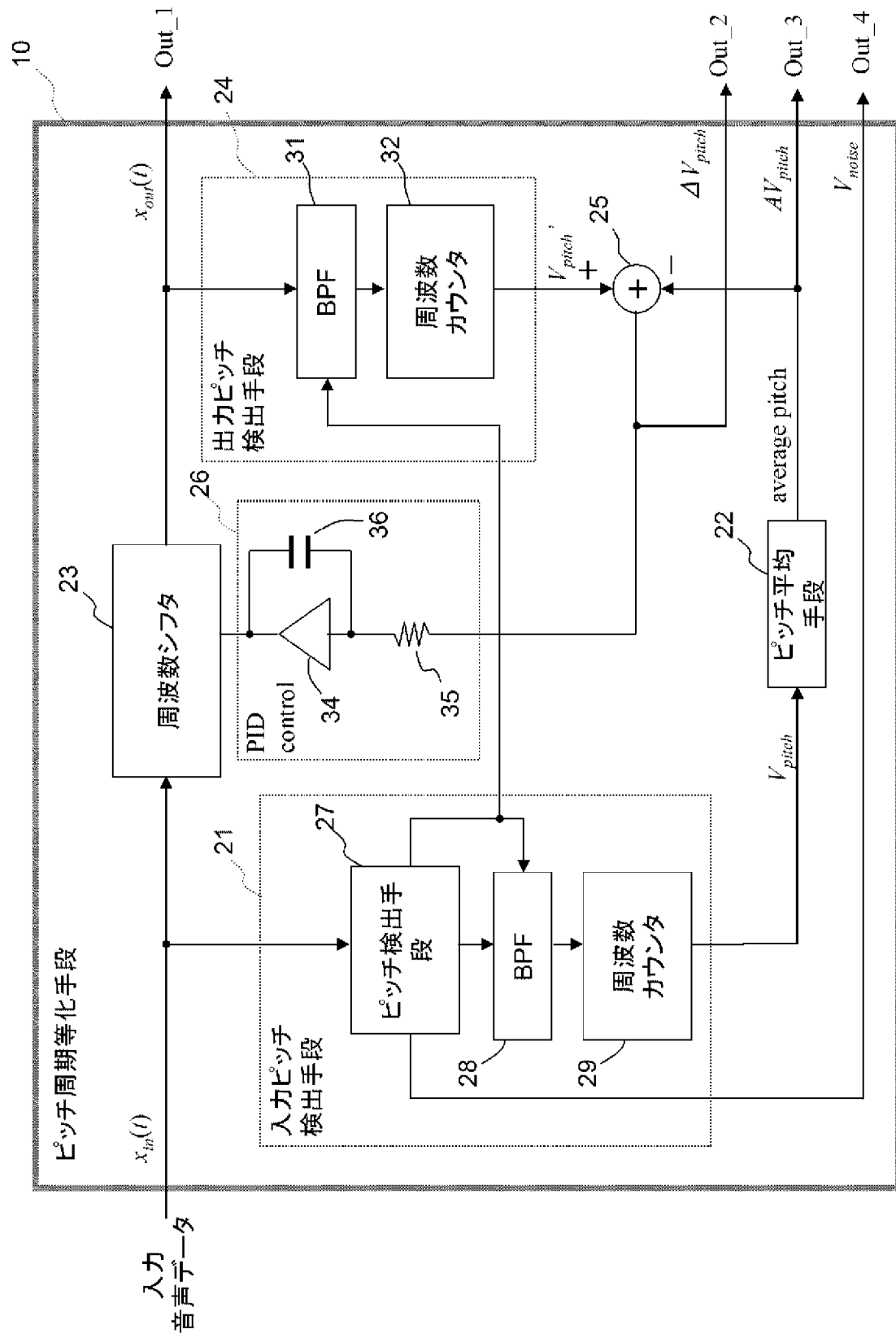
化検索対象音声データ及び前記ピッチ等化クエリー音声データを直交変換して得られるサブバンド・データの時系列であることを特徴とする請求項10乃至15の何れか一記載の音声検索方法。

- [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。
- [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。
- [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。

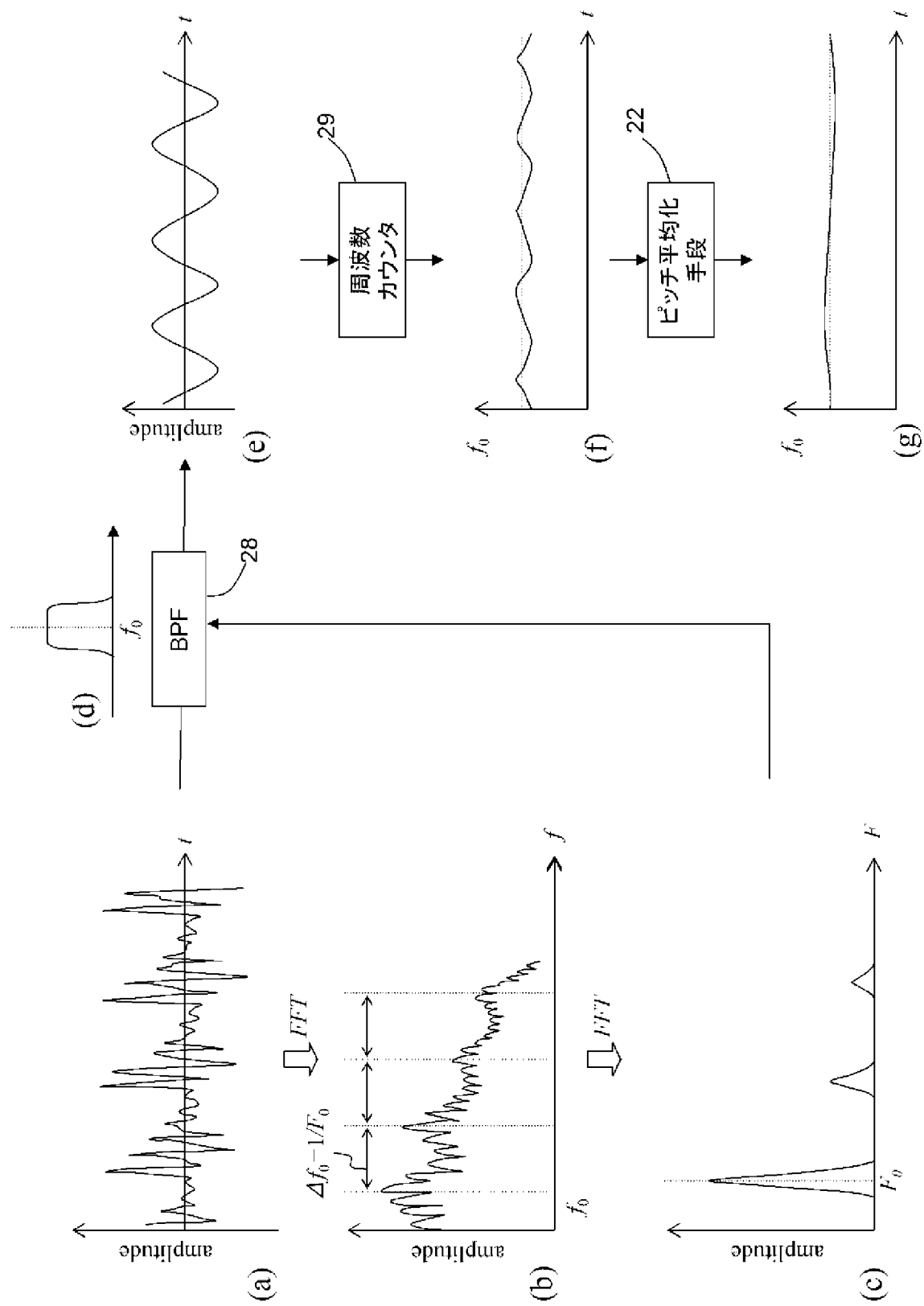
[図1]



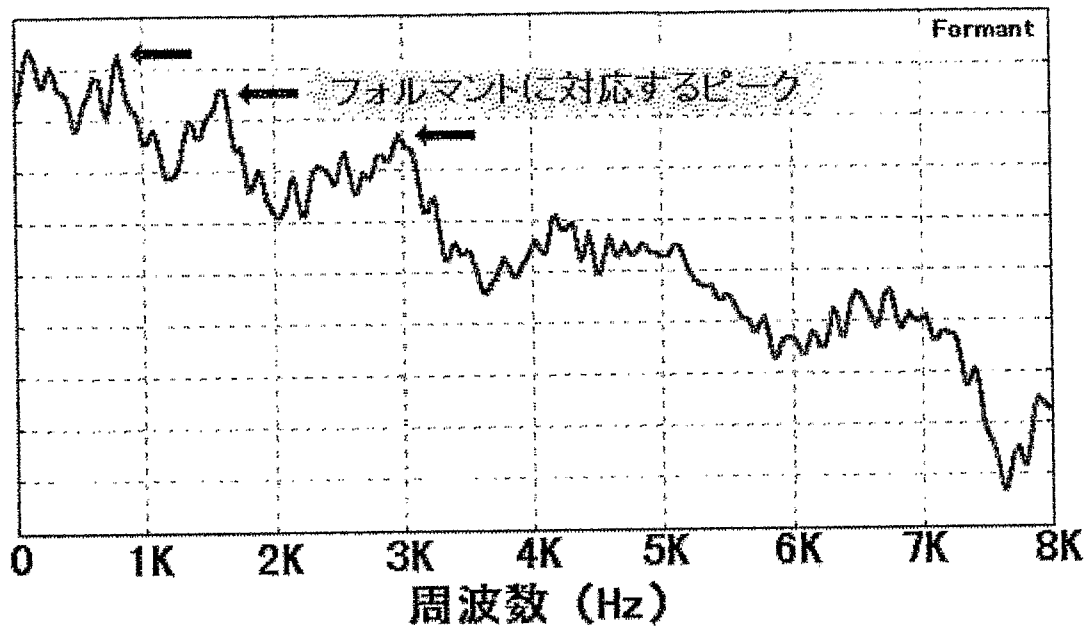
[図3]



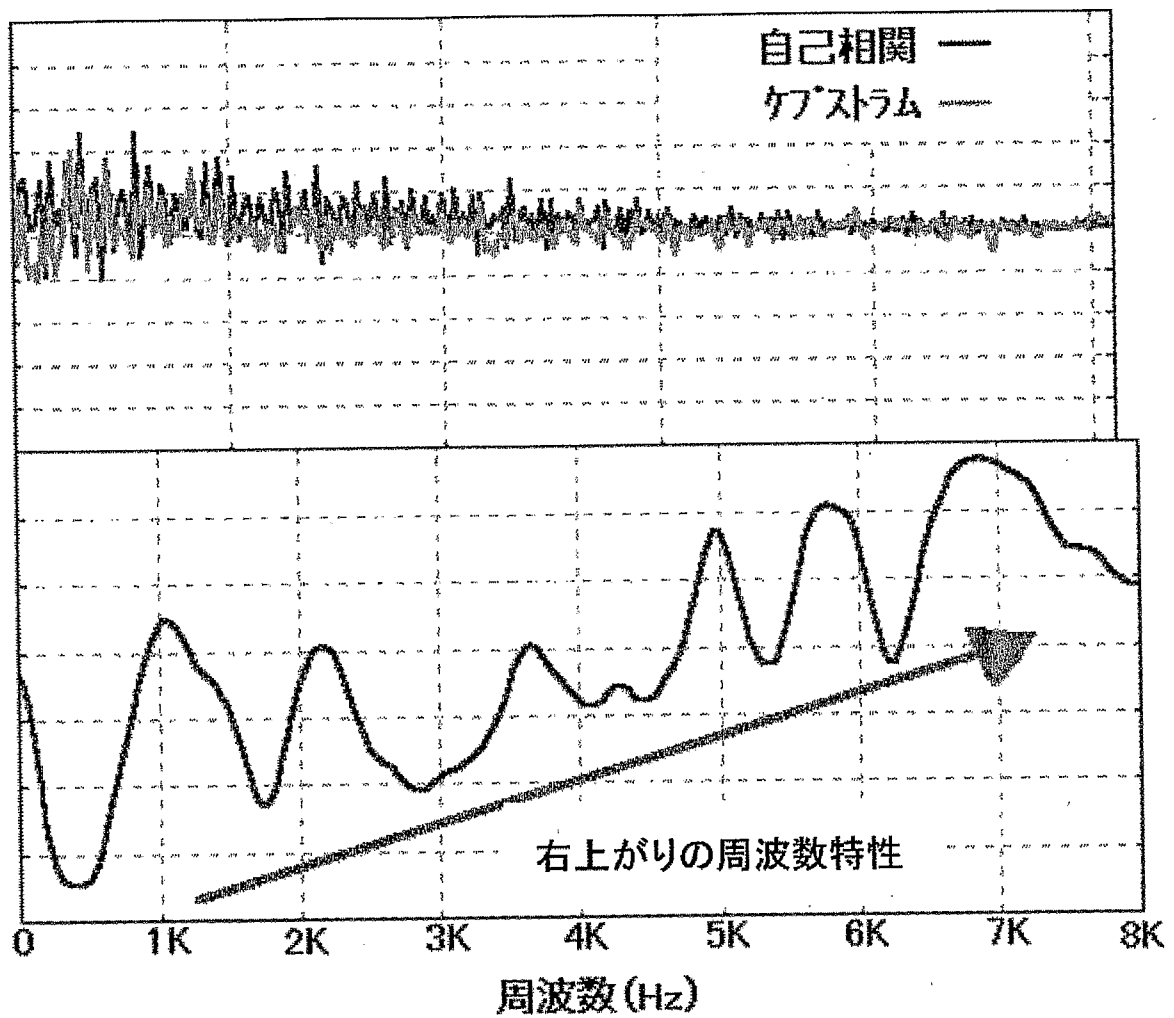
[図4]



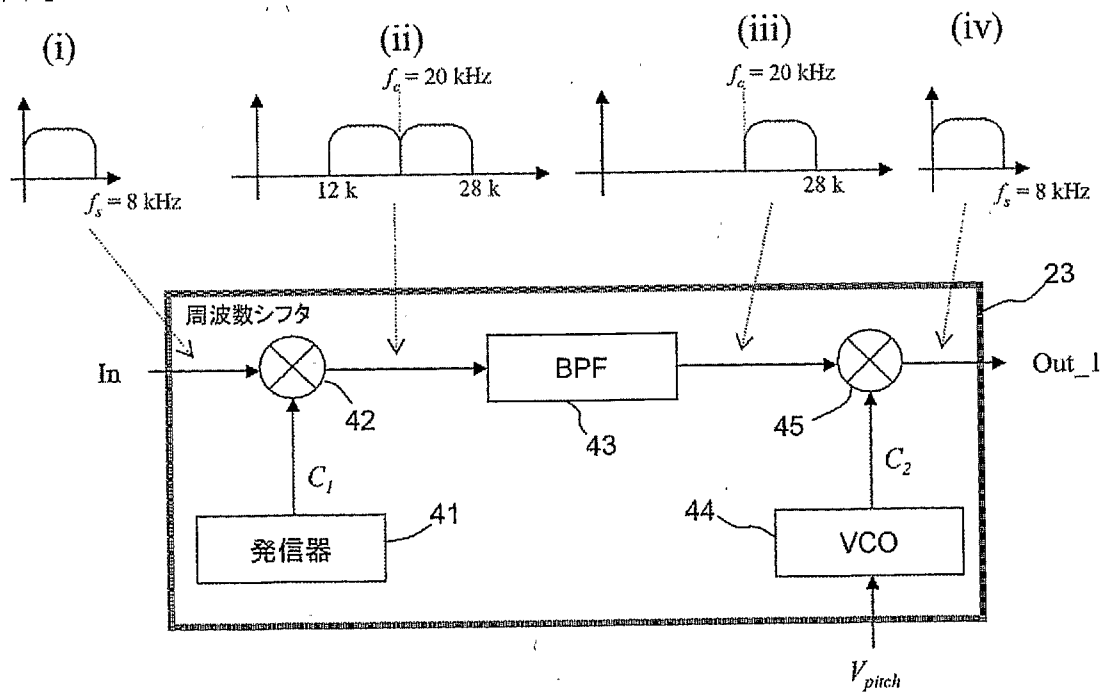
[図5]



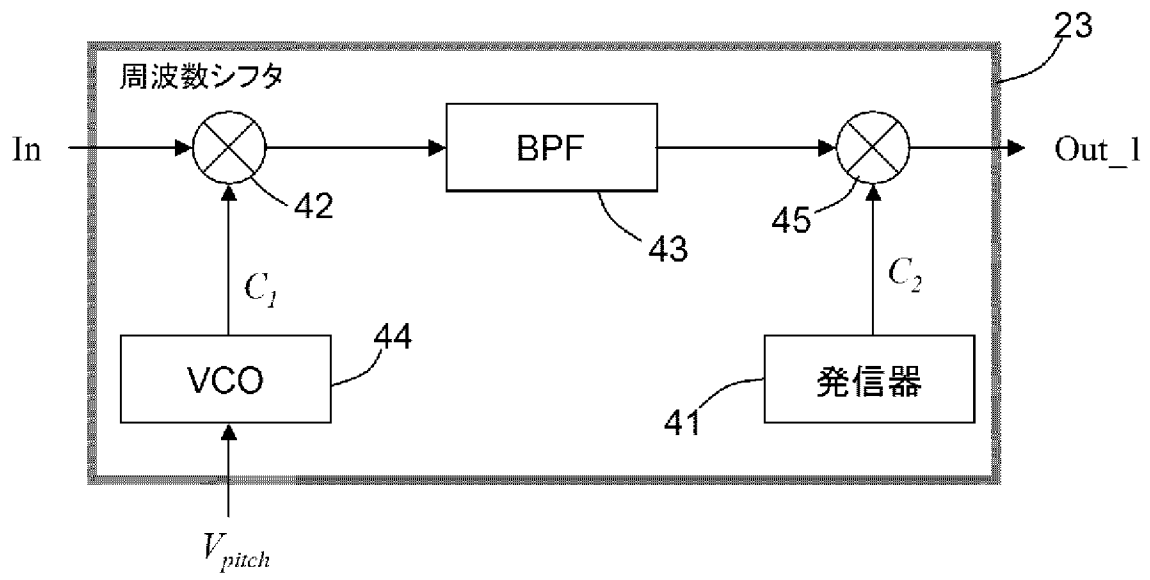
[図6]



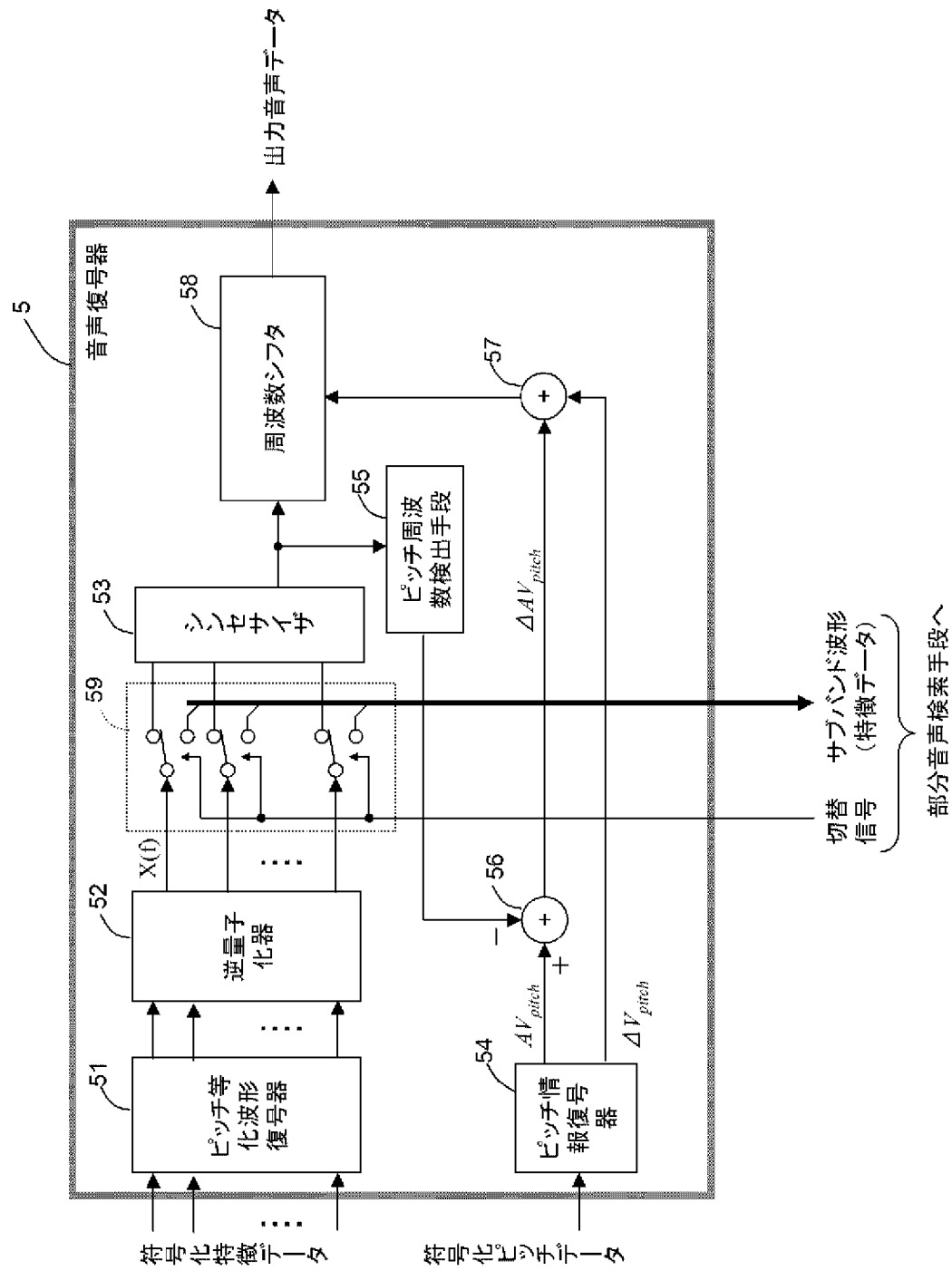
[図7]



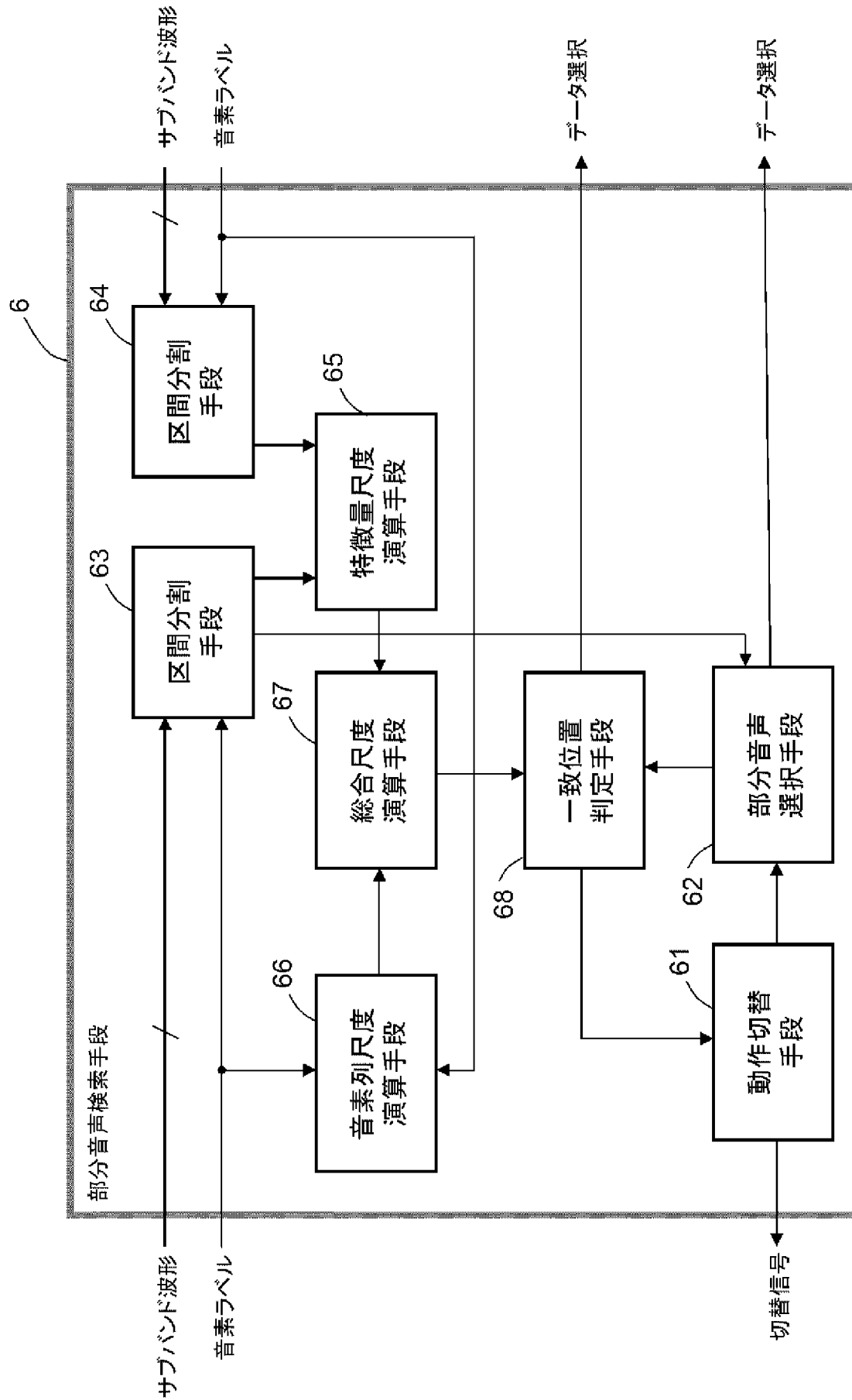
[図8]



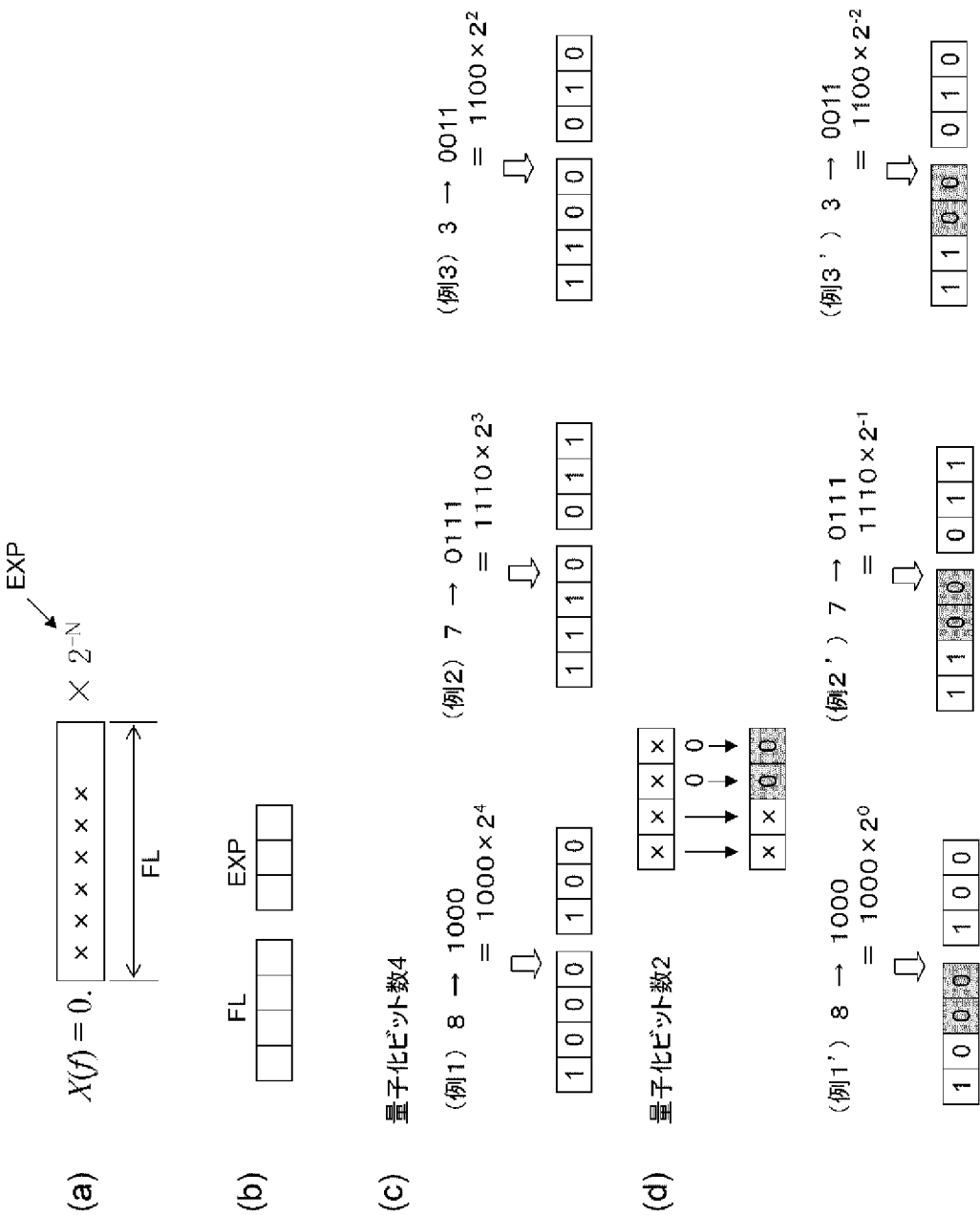
[図9]



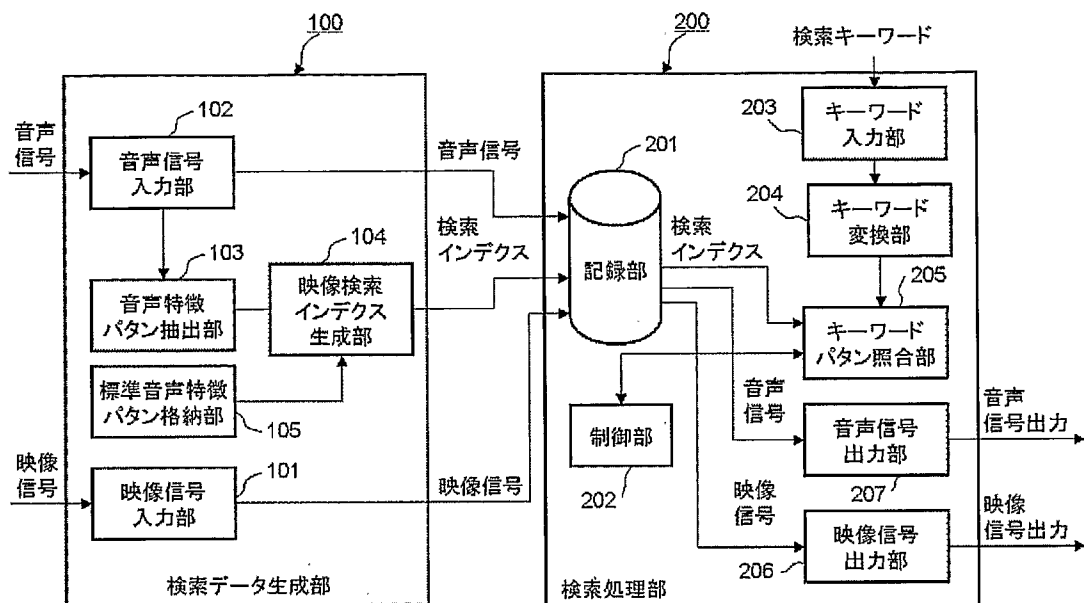
[図10]



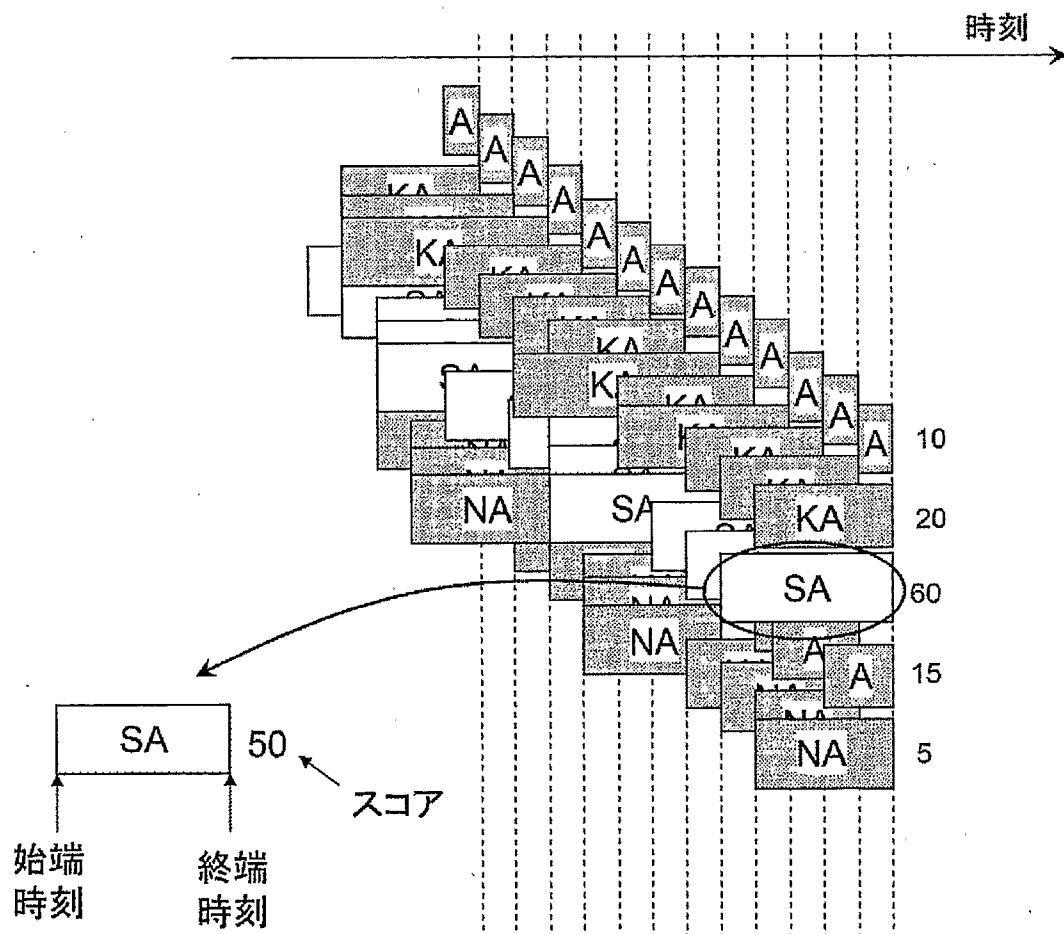
[図11]



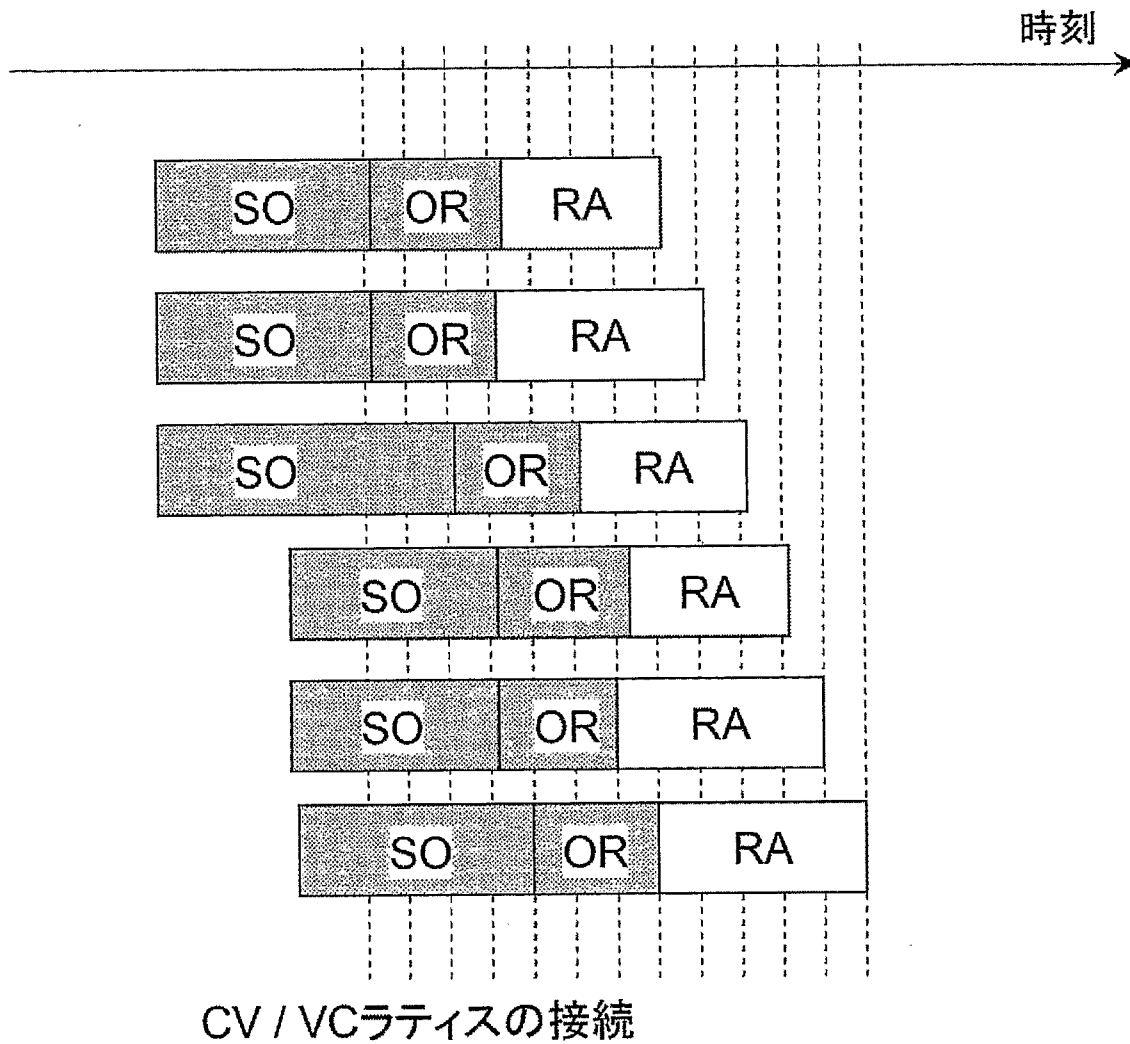
[図12]



[図13]



[図14]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2006/315228

A. CLASSIFICATION OF SUBJECT MATTER

G10L15/00(2006.01)i, G06F17/30(2006.01)i, G10L11/00(2006.01)i, G10L11/04(2006.01)i, G10L15/02(2006.01)i, G10L15/10(2006.01)i, G10L15/20(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L15/00, G06F17/30, G10L11/00, G10L11/04, G10L15/02, G10L15/10, G10L15/20

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2006
Kokai Jitsuyo Shinan Koho	1971-2006	Toroku Jitsuyo Shinan Koho	1994-2006

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

JSTPlus (JDream2), IEEE

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	H. Singer et al., "Pitch dependent phone modelling for HMM-based speech recognition", The Journal of the Acoustical Society of Japan (E), 1994.03, Vol.15, No.2, p.77-86	1-19
A	JP 2834471 B2 (Nippon Telegraph And Telephone Corp.), 02 October, 1998 (02.10.98), Full text; all drawings (Family: none)	1-19
A	JP 3252282 B2 (Matsushita Electric Industrial Co., Ltd.), 22 November, 2001 (22.11.01), Full text; all drawings & US 6611803 B1 & EP 1058453 B1	1-19



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

16 August, 2006 (16.08.06)

Date of mailing of the international search report

29 August, 2006 (29.08.06)

Name and mailing address of the ISA/

Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2006/315228

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 59-99500 A (Nippon Telegraph & Telephone Public Corp.), 08 June, 1984 (08.06.84), Full text; all drawings (Family: none)	9, 18

A. 発明の属する分野の分類（国際特許分類（I P C）） Int.Cl. G10L15/00(2006.01)i, G06F17/30(2006.01)i, G10L11/00(2006.01)i, G10L11/04(2006.01)i, G10L15/02(2006.01)i, G10L15/10(2006.01)i, G10L15/20(2006.01)i			
B. 調査を行った分野 調査を行った最小限資料（国際特許分類（I P C）） Int.Cl. G10L15/00, G06F17/30, G10L11/00, G10L11/04, G10L15/02, G10L15/10, G10L15/20			
最小限資料以外の資料で調査を行った分野に含まれるもの 日本国実用新案公報 日本国公開実用新案公報 日本国実用新案登録公報 日本国登録実用新案公報 1 9 2 2 - 1 9 9 6 年 1 9 7 1 - 2 0 0 6 年 1 9 9 6 - 2 0 0 6 年 1 9 9 4 - 2 0 0 6 年			
国際調査で使用した電子データベース（データベースの名称、調査に使用した用語） JSTPlus(JDream2) IEEE			
C. 関連すると認められる文献			
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号	
A	H. Singer et al., "Pitch dependent phone modelling for HMM-based speech recognition", The Journal of the Acoustical Society of Japan (E), 1994.03, Vol.15, No.2, p.77-86	1 - 1 9	
A	J P 2 8 3 4 4 7 1 B 2（日本電信電話株式会社）1998.10.02, 全文, 全図（ファミリーなし）	1 - 1 9	
☒ C欄の続きにも文献が列挙されている。 ☐ パテントファミリーに関する別紙を参照。			
* 引用文献のカテゴリー 「A」特に関連のある文献ではなく、一般的技術水準を示すもの 「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの 「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す） 「O」口頭による開示、使用、展示等に言及する文献 「P」国際出願日前で、かつ優先権の主張の基礎となる出願日の後に公表された文献 「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの 「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの 「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの 「&」同一パテントファミリー文献			
国際調査を完了した日 1 6 . 0 8 . 2 0 0 6		国際調査報告の発送日 2 9 . 0 8 . 2 0 0 6	
国際調査機関の名称及びあて先 日本国特許庁（I S A / J P） 郵便番号100-8915 東京都千代田区霞が関三丁目4番3号		特許庁審査官（権限のある職員） 榎本 剛 電話番号 03-3581-1101 内線 3541	5 Z 9 3 7 9

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	JP 3252282 B2 (松下電器産業株式会社) 2001.11.22, 全文, 全図 & US 6611803 B1 & EP 1058453 B1	1 - 19
A	JP 59-99500 A (日本電信電話公社) 1984.06.08, 全文, 全図 (ファミリーなし)	9, 18