

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5457706号  
(P5457706)

(45) 発行日 平成26年4月2日(2014.4.2)

(24) 登録日 平成26年1月17日(2014.1.17)

(51) Int.Cl.

F I

G 1 O L 13/06 (2013.01)

G 1 O L 13/06 1 3 0

請求項の数 8 (全 17 頁)

|           |                               |           |                     |
|-----------|-------------------------------|-----------|---------------------|
| (21) 出願番号 | 特願2009-83563 (P2009-83563)    | (73) 特許権者 | 000003078           |
| (22) 出願日  | 平成21年3月30日 (2009. 3. 30)      |           | 株式会社東芝              |
| (65) 公開番号 | 特開2010-237323 (P2010-237323A) |           | 東京都港区芝浦一丁目1番1号      |
| (43) 公開日  | 平成22年10月21日 (2010. 10. 21)    | (74) 代理人  | 100089118           |
| 審査請求日     | 平成23年3月25日 (2011. 3. 25)      |           | 弁理士 酒井 宏明           |
|           |                               | (74) 代理人  | 100112656           |
|           |                               |           | 弁理士 宮田 英毅           |
|           |                               | (72) 発明者  | ラトレ ハビエル            |
|           |                               |           | 東京都港区芝浦一丁目1番1号 株式会社 |
|           |                               |           | 東芝内                 |
|           |                               | (72) 発明者  | 赤嶺 政巳               |
|           |                               |           | 東京都港区芝浦一丁目1番1号 株式会社 |
|           |                               |           | 東芝内                 |
|           |                               | 審査官       | 五貫 昭一               |
|           |                               |           | 最終頁に続く              |

(54) 【発明の名称】 音声モデル生成装置、音声合成装置、音声モデル生成プログラム、音声合成プログラム、音声モデル生成方法および音声合成方法

(57) 【特許請求の範囲】

【請求項1】

テキスト情報を取得し、前記テキスト情報をテキスト解析することにより、前記テキスト情報に含まれる言語の内容を示す言語情報を生成するテキスト解析部と、

前記テキスト情報に対応する音声信号を取得し、前記音声信号の各フレームから当該フレームのスペクトル形状を表す特徴パラメータを算出するスペクトル分析部と、

前記音声信号の複数フレームを有し、言語レベルを単位とする区間である言語区間の境界位置を示す区切り情報を取得し、前記区切り情報に基づいて、前記音声信号を前記言語区間に分割する分割部と、

対象となる前記言語区間である対象区間に含まれる前記複数フレームそれぞれの前記特徴パラメータに所定の線形変換を行うことにより基本パラメータを算出し、前記対象区間の前記基本パラメータと、前記対象区間の直前の前記言語区間の前記基本パラメータと、前記対象区間の直後の前記言語区間の前記基本パラメータと、に基づいて拡張パラメータを算出することにより、前記基本パラメータと前記拡張パラメータとを含むスペクトルパラメータを取得するパラメータ化部と、

複数の言語区間それぞれに対して算出された複数のスペクトルパラメータを、前記言語情報に基づいて複数のクラスターにクラスタリングするクラスターリング部と、

同一のクラスターに属する複数のスペクトルパラメータから前記複数のスペクトルパラメータの特徴を示すスペクトルモデルを学習するモデル学習部とを備えることを特徴とする音声モデル生成装置。

10

20

## 【請求項 2】

前記モデル学習部は、前記対象区間、前記対象区間の直前および直後の前記言語区間により、前記対象区間を複数のクラスターにクラスタリングすることを特徴とする請求項 1 に記載の音声モデル生成装置。

## 【請求項 3】

音声合成の対象となるテキスト情報を取得し、前記テキスト情報をテキスト解析することにより、前記テキスト情報に含まれる言語の内容を示す言語情報を生成するテキスト解析部と、

音声信号の複数フレームを有し、言語レベルを単位とする区間である言語区間のうち、対象となる前記言語区間である対象区間に含まれる前記複数フレームそれぞれのスペクトル形状を表す特徴パラメータに所定の線形変換を行うことにより基本パラメータを算出し、前記対象区間の前記基本パラメータと、前記対象区間の直前の前記言語区間の前記基本パラメータと、前記対象区間の直後の前記言語区間の前記基本パラメータと、に基づいて拡張パラメータを算出することにより取得された前記基本パラメータと前記拡張パラメータとを含むスペクトルパラメータの特徴を示すスペクトルモデルであって、前記言語区間の前記言語情報により複数のクラスターにクラスタリングされたスペクトルモデルを記憶する記憶部と、

音声合成の対象となるテキスト情報の前記言語区間の前記言語情報に基づいて、前記テキスト情報の前記言語区間が属する前記クラスターの前記スペクトルモデルを前記記憶部から選択する選択部と、

前記選択部により選択された前記スペクトルモデルに基づいて、前記言語区間に対するスペクトルパラメータを生成し、前記スペクトルパラメータを逆変換することにより、特徴パラメータを得る生成部と  
を備えることを特徴とする音声合成装置。

## 【請求項 4】

前記生成部は、前記選択部により選択された前記スペクトルモデルの目的関数を生成し、前記目的関数を最大化することにより、前記言語区間に対するスペクトルパラメータを生成することを特徴とする請求項 3 に記載の音声合成装置。

## 【請求項 5】

音声モデル生成処理をコンピュータに実行させるための音声モデル生成プログラムであって、

前記コンピュータを、

テキスト情報を取得し、前記テキスト情報をテキスト解析することにより、前記テキスト情報に含まれる言語の内容を示す言語情報を生成するテキスト解析部と、

前記テキスト情報に対応する音声信号を取得し、前記音声信号の各フレームから当該フレームのスペクトル形状を表す特徴パラメータを算出するスペクトル分析部と、

前記音声信号の複数フレームを有し、言語レベルを単位とする区間である言語区間の境界位置を示す区切り情報を取得し、前記区切り情報に基づいて、前記音声信号を前記言語区間に分割する分割部と、

対象となる前記言語区間である対象区間に含まれる前記複数フレームそれぞれの前記特徴パラメータに所定の線形変換を行うことにより基本パラメータを算出し、前記対象区間の前記基本パラメータと、前記対象区間の直前の前記言語区間の前記基本パラメータと、前記対象区間の直後の前記言語区間の前記基本パラメータと、に基づいて拡張パラメータを算出することにより、前記基本パラメータと前記拡張パラメータとを含むスペクトルパラメータを取得するパラメータ化部と、

複数の言語区間それぞれに対して算出された複数のスペクトルパラメータを、前記言語情報に基づいて複数のクラスターにクラスタリングするクラスタリング部と、

同一のクラスターに属する複数のスペクトルパラメータから前記複数のスペクトルパラメータの特徴を示すスペクトルモデルを学習するモデル学習部として機能させるためのプログラム。

## 【請求項 6】

音声合成処理をコンピュータに実行させるための音声合成プログラムであって、  
前記コンピュータを、

音声合成の対象となるテキスト情報を取得し、前記テキスト情報をテキスト解析することにより、前記テキスト情報に含まれる言語の内容を示す言語情報を生成するテキスト解析部と、

音声信号の複数フレームを有し、言語レベルを単位とする区間である言語区間のうち、対象となる前記言語区間である対象区間に含まれる前記複数フレームそれぞれのスペクトル形状を表す特徴パラメータに所定の線形変換を行うことにより基本パラメータを算出し、前記対象区間の前記基本パラメータと、前記対象区間の直前の前記言語区間の前記基本パラメータと、前記対象区間の直後の前記言語区間の前記基本パラメータと、に基づいて拡張パラメータを算出することにより取得された前記基本パラメータと前記拡張パラメータとを含むスペクトルパラメータの特徴を示すスペクトルモデルであって、前記言語区間の前記言語情報により複数のクラスターにクラスタリングされたスペクトルモデルを記憶する記憶部と、

10

前記音声合成の対象となるテキスト情報の前記言語区間の前記言語情報に基づいて、前記テキスト情報の前記言語区間が属する前記クラスターの前記スペクトルモデルを前記記憶部から選択する選択部と、

前記選択部により選択された前記スペクトルモデルに基づいて、前記言語区間に対するスペクトルパラメータを生成し、前記スペクトルパラメータを逆変換することにより、特徴パラメータを得る生成部として機能させるためのプログラム。

20

## 【請求項 7】

テキスト解析部が、テキスト情報を取得し、前記テキスト情報をテキスト解析することにより、前記テキスト情報に含まれる言語の内容を示す言語情報を生成するテキスト解析ステップと、

スペクトル分析部が、前記テキスト情報に対応する音声信号を取得し、前記音声信号の各フレームから当該フレームのスペクトル形状を表す特徴パラメータを算出するスペクトル分析ステップと、

分割部が、前記音声信号の複数フレームを有し、言語レベルを単位とする区間である言語区間の境界位置を示す区切り情報を取得し、前記区切り情報に基づいて、前記音声信号を前記言語区間に分割する分割ステップと、

30

パラメータ化部が、対象となる前記言語区間である対象区間に含まれる前記複数フレームそれぞれの前記特徴パラメータに所定の線形変換を行うことにより基本パラメータを算出し、前記対象区間の前記基本パラメータと、前記対象区間の直前の前記言語区間の前記基本パラメータと、前記対象区間の直後の前記言語区間の前記基本パラメータと、に基づいて拡張パラメータを算出することにより、前記基本パラメータと前記拡張パラメータとを含むスペクトルパラメータを取得するパラメータ化ステップと、

クラスタリング部が、複数の言語区間それぞれに対して算出された複数のスペクトルパラメータを、前記言語情報に基づいて複数のクラスターにクラスタリングするクラスタリングステップと、

40

モデル学習部が、同一のクラスターに属する複数のスペクトルパラメータから前記複数のスペクトルパラメータの特徴を示すスペクトルモデルを学習する学習ステップとを有することを特徴とする音声モデル生成方法。

## 【請求項 8】

テキスト解析部が、音声合成の対象となるテキスト情報を取得し、前記テキスト情報をテキスト解析することにより、前記テキスト情報に含まれる言語の内容を示す言語情報を生成するテキスト解析ステップと、

選択部が、音声信号の複数フレームを有し、言語レベルを単位とする区間である言語区間のうち、対象となる前記言語区間である対象区間に含まれる前記複数フレームそれぞれ

50

のスペクトル形状を表す特徴パラメータに所定の線形変換を行うことにより基本パラメータを算出し、前記対象区間の前記基本パラメータと、前記対象区間の直前の前記言語区間の前記基本パラメータと、前記対象区間の直後の前記言語区間の前記基本パラメータと、に基づいて拡張パラメータを算出することにより取得された前記基本パラメータと前記拡張パラメータとを含むスペクトルパラメータの特徴を示すスペクトルモデルであって、前記言語区間の前記言語情報により複数のクラスターにクラスタリングされたスペクトルモデルを記憶する記憶部を参照するステップと、

選択部が、前記音声合成の対象となるテキスト情報の前記言語区間の前記言語情報に基づいて、前記テキスト情報の前記言語区間が属する前記クラスターの前記スペクトルモデルを前記記憶部から選択する選択ステップと、

生成部が、前記選択ステップで選択された前記スペクトルモデルに基づいて、前記言語区間に対するスペクトルパラメータを生成し、前記スペクトルパラメータを逆変換することにより、特徴パラメータを得る生成ステップと  
を有することを特徴とする音声合成方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、音声モデルを生成する音声モデル生成装置、音声モデルを用いて音声を合成する音声合成装置、音声モデル生成プログラム、音声合成プログラム、音声モデル生成方法および音声合成方法に関する。

【背景技術】

【0002】

テキストから音声を生成する音声合成装置は、大別すると、テキスト解析部、韻律生成部及び音声信号生成部の3つの処理部から構成される。テキスト解析部では、言語辞書などを用いて入力されたテキスト（漢字かな混じり文）を解析し、漢字の読みやアクセントの位置、文節（アクセントの句）の区切りなどを定義した言語情報を出力する。韻律生成部では、言語情報に基づいて、声の高さ（基本周波数）の時間変化パターン（ピッチ包絡）と、各音韻の長さなどの音韻・韻律情報を出力する。音声信号生成部は、テキスト解析部からの音韻の系列と韻律生成部からの韻律情報に従って音声波形を生成するものであり、素片接続型合成方式とHMM合成方式の2方式が現在、主流となっている。

【0003】

素片接続型合成方式では、音韻の系列に従って音声素片を選択し、韻律情報に従って音声素片のピッチと継続時間長を変形して接続することで、合成音声を出力する。この方式は録音した音声データの素片を接続して音声波形を作成しているため比較的自然的な音質の合成音を得られる利点がある。しかしながら、素片を蓄積するためのメモリサイズが大きくなるという問題がある。

【0004】

HMM合成方式は、合成フィルタをパルス列または雑音で駆動するボコーダーと呼ばれる合成器に基づいて合成音声を生成するものであり、統計モデルに基づく音声合成方式の一つである。この方式では、合成器のパラメータを統計モデルで表現し、入力された文章に対して統計モデルの尤度が最大となるように合成器のパラメータを生成する。合成器のパラメータは、音声信号のスペクトルを表すLSFやFMCCなど、合成フィルタのパラメータと駆動信号のパラメータであり、それらの時系列は音素毎にHMMとガウス分布により統計的にモデル化される。学習用の音声データが与えられれば、統計モデルは音声データから自動的に学習することができ、メモリサイズも比較的小さくできる利点がある。

【0005】

しかしながら、従来のHMM統計モデルに基づく音声合成方式では、スペクトルが統計的なモデル化により平均化されるため、生成される合成音の音質はメリハリのない籠った音質となるという問題がある。また、音素間でパラメータが不連続になり易く、異音が発生するという問題がある。

10

20

30

40

50

## 【0006】

このようなパラメータの平均化や平滑化による音質の悪化を改善する方法として、文章全体にわたるスペクトルパラメータの分散を学習データから学習し、合成時に学習された分散を制約条件としてパラメータを生成、ダイナミクスを再生する手法が提案されている（非特許文献1）。

## 【先行技術文献】

## 【非特許文献】

## 【0007】

【非特許文献1】Toda, T. and Tokuda K., 2005 “Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis”. Proc. Interspeech 2005, Lisbon, Portugal, pp. 2801-2804

## 【発明の概要】

## 【発明が解決しようとする課題】

## 【0008】

しかしながら、非特許文献1に記載されている方法は、スペクトルのメリハリを回復させる効果があるものの、MFCCパラメータとの組み合わせ以外においては効果が確認されておらず、生成される合成フィルタがしばしば不安定なフィルタとなって異音が発生するという問題がある。

## 【0009】

本発明は、上記に鑑みてなされたものであって、滑らかに変化する自然なスペクトルを生成することのできる音声モデルを生成する音声モデル生成装置、この音声モデルを用いた音声合成装置、プログラムおよび方法を提供することを目的とする。

## 【課題を解決するための手段】

## 【0010】

上述した課題を解決し、目的を達成するために、本発明の一形態は、音声モデル生成装置に係り、テキスト情報を取得し、前記テキスト情報をテキスト解析することにより、前記テキスト情報に含まれる言語の内容を示す言語情報を生成するテキスト解析部と、前記テキスト情報に対応する音声信号を取得し、前記音声信号の各フレームから当該フレームのスペクトル形状を表す特徴パラメータを算出するスペクトル分析部と、前記音声信号の複数フレームを有し、言語レベルを単位とする区間である言語区間の境界位置を示す区切り情報を取得し、前記区切り情報に基づいて、前記音声信号を前記言語区間に分割する分割部と、対象となる前記言語区間である対象区間に含まれる前記複数フレームそれぞれの前記特徴パラメータに所定の線形変換を行うことにより基本パラメータを算出し、前記対象区間の前記基本パラメータと、前記対象区間の直前の前記言語区間の前記基本パラメータと、前記対象区間の直後の前記言語区間の前記基本パラメータと、に基づいて拡張パラメータを算出することにより、前記基本パラメータと前記拡張パラメータとを含むスペクトルパラメータを取得するパラメータ化部と、複数の言語区間それぞれに対して算出された複数のスペクトルパラメータを、前記言語情報に基づいて複数のクラスターにクラスターリングするクラスターリング部と、同一のクラスターに属する複数のスペクトルパラメータから前記複数のスペクトルパラメータの特徴を示すスペクトルモデルを学習するモデル学習部とを備えることを特徴とする。

## 【0011】

また、本発明の他の形態は、音声合成装置に係り、音声合成の対象となるテキスト情報を取得し、前記テキスト情報をテキスト解析することにより、前記テキスト情報に含まれる言語の内容を示す言語情報を生成するテキスト解析部と、音声信号の複数フレームを有し、言語レベルを単位とする区間である言語区間のうち、対象となる前記言語区間である対象区間に含まれる前記複数フレームそれぞれのスペクトル形状を表す特徴パラメータに所定の線形変換を行うことにより基本パラメータを算出し、前記対象区間の前記基本パラ

メータと、前記対象区間の直前の前記言語区間の前記基本パラメータと、前記対象区間の直後の前記言語区間の前記基本パラメータと、に基づいて拡張パラメータを算出することにより取得された前記基本パラメータと前記拡張パラメータとを含むスペクトルパラメータの特徴を示すスペクトルモデルであって、前記言語区間の前記言語情報により複数のクラスターにクラスタリングされたスペクトルモデルを記憶する記憶部と、前記音声合成の対象となるテキスト情報の前記言語区間の前記言語情報に基づいて、前記テキスト情報の前記言語区間が属する前記クラスターの前記スペクトルモデルを前記記憶部から選択する選択部と、前記選択部により選択された前記スペクトルモデルに基づいて、前記言語区間に対するスペクトルパラメータを生成し、前記スペクトルパラメータを逆変換することにより、特徴パラメータを得る生成部とを備えることを特徴とする。

10

【発明の効果】

【0012】

本発明によれば、複数フレームを含む言語区間単位でスペクトルモデルを学習するので、このスペクトルモデルを用いて音声合成を行うことにより、不連続点のない自然なスペクトルを得ることができるという効果を奏する。

【図面の簡単な説明】

【0013】

【図1】本発明の実施の形態にかかる学習モデル生成装置100の構成を示すブロック図である。

【図2】言語区間を説明するための図である。

20

【図3】決定木の一例を示す図である。

【図4】学習モデル生成装置100による学習モデル生成処理を示すフローチャートである。

【図5】パラメータ化部140により得られたスペクトルパラメータを示す図である。

【図6】HMMによりフレーム単位で得られたスペクトルパラメータを示す図である。

【図7】音声合成装置200の構成を示す図である。

【図8】音声合成装置200による音声合成処理を示すフローチャートである。

【図9】学習モデル生成装置100のハードウェア構成を示す図である。

【発明を実施するための形態】

【0014】

30

以下に添付図面を参照して、この発明にかかる音声モデル生成装置、音声合成装置、プログラムおよび方法の最良な実施の形態を詳細に説明する。

【0015】

(第1の実施の形態)

図1は、本発明の実施の形態にかかる学習モデル生成装置100の構成を示すブロック図である。学習モデル生成装置100は、テキスト解析部110と、スペクトル分析部120と、分割部130と、パラメータ化部140と、クラスタリング部150と、モデル学習部160と、モデル記憶部170とを備えている。学習モデル生成装置100は、テキスト情報と、テキスト情報の内容を読み上げた音声信号とを学習データとして取得し、学習データに基づいて、音声合成のための学習モデルを生成する。

40

【0016】

テキスト解析部110は、テキスト情報を取得する。テキスト解析部110は、取得したテキスト情報に対するテキスト解析により言語情報を生成する。ここで、言語情報は、言語レベルを単位とする言語区間の境界位置を示す区間情報、各言語区間の形態素、各言語区間の音素記号、各音素が有声音であるか無声音であるかを示す情報、各音素のアクセントの有無を示す情報、各言語区間の開始時間、終了時間、各言語区間の前後の言語区間の情報、各言語区間と前後の言語区間との言語的な関係を示す情報など言語の内容を示す情報である。言語情報はコンテキストと呼ばれ、クラスタリング部150において、スペクトルパラメータのコンテキストモデル作成に用いられる。なお、言語区間とは、複数フレームを含み、所定の言語レベルを単位とする区間である。言語レベルとしては、音素、

50

音節、単語、句、呼気段階、発声全体などがある。

#### 【0017】

スペクトル分析部120は、音声信号を取得する。音声信号は、テキスト解析部110が取得したテキスト情報の内容を読み上げた発話についての音声の信号である。音声信号は、学習のための音声データを発話単位に分割したものである。

#### 【0018】

スペクトル分析部120は、取得した音声信号に対し、スペクトル分析を行う。すなわち、音声信号を10ms程度のフレームに分割する。そして、フレーム毎に、フレームのスペクトルの形状を表す特徴パラメータとしてのメルケプストラム係数(MFCC)を算出し、各フレームの音声信号とMFCCの組を分割部130に出力する。

10

#### 【0019】

分割部130は、外部から区切り情報を取得する。区切り情報とは、音声信号の言語レベル単位での境界位置、すなわち言語区間の境界位置を示す情報である。区切り情報は、マニュアルまたは自動的なアライメントにより生成される。自動的なアライメントとしては、例えば、HMMで構成される音声認識モデルを用いて、入力された音声信号のフレームを音響モデルの状態に対応付け、この対応付けから言語区間の区切り情報を得る。区切り情報は、学習データとともに与えられるものとする。分割部130は、区切り情報に基づいて、音声信号の言語区間を特定し、スペクトル分析部120から取得したMFCCを言語区間に分割する。

#### 【0020】

20

図2に示すように、例えば[kairo]というテキスト情報に対応するMFCC曲線は、音素単位では、/k/, /a/, /i/, /r/, /o/の4つの音素の言語区間に区切られる。分割部130は、例えば音素、音節、単語、句、呼気段階および発声全体など複数の言語レベルにおいてMFCCを言語区間に分割する。

#### 【0021】

なお、これ以降で説明する処理においても、各言語レベルの言語区間それぞれに対して処理が施されるが、以下の説明においては、一例として、音素を言語レベルとする場合について述べる。

#### 【0022】

パラメータ化部140は、MFCCを分割部130において区切られた単位、すなわち言語区間単位でベクトルとし、そのベクトルからスペクトルパラメータを算出する。なお、スペクトルパラメータは、基本パラメータと拡張パラメータとを有している。

30

#### 【0023】

パラメータ化部140は、言語区間に含まれるフレーム数をkとした場合、複数フレームのMFCCから構成されるk次元ベクトル $MelCep_{i,s}$ に対し、(式1)に示すように、k次のDCTを適用することにより、基本パラメータを算出する。このように、基本パラメータは、対象とする言語区間である対象区間のスペクトルパラメータであり、対象区間の特徴を示すパラメータである。

#### 【数1】

$$X_{i,s} = T_{i,s} \cdot MelCep_{i,s} \cdots \quad (\text{式1})$$

40

なお、 $MelCep_{i,s}$ は、音素sのi次のMFCC係数のk次元ベクトルである。 $T_{i,s}$ は、音素sのフレーム数kに対応するk次のDCTの変換行列である。DCTの次元は言語レベルの単位やフレーム長などに依存する。なお、基本フレームを算出する際には、DCT以外の種々の線形変換を用いてもよい。例えば、逆変換可能な離散コサイン変換、フーリエ変換、ウェーブレット変換、テーラー展開および多項式展開を用いてもよい。

#### 【0024】

パラメータ化部140は、さらに拡張パラメータを算出する。拡張パラメータは、対象区間に隣接する言語区間のMFCCベクトルの傾きで構成される。なお、隣接する区間とは、対象区間の直前の言語区間である直前区間と、対象区間の直後の言語区間である直後

50

区間である。直前区間の拡張パラメータ

【数 2】

$$\Delta \text{MelCep}_{i,s}^{\text{left}}$$

および直後区間の拡張パラメータ

【数 3】

$$\Delta \text{MelCep}_{i,s}^{\text{right}}$$

は、それぞれ（式 2）および（式 3）により算出される。ここで、 $\alpha$  は傾きを計算するための  $W$  次元重みベクトルである。また、カッコ内の負のインデックスはベクトルの最後の要素から数えた場合の要素を示している。

10

【数 4】

$$\Delta \text{MelCep}_{i,s}^{\text{left}} = \sum_{w=0}^W \alpha(w) \cdot \text{MelCep}_{i,s}(w) + \sum_{w=-W}^{-1} \alpha(w) \text{MelCep}_{i,s-1}(-w) \cdots \cdots \quad (\text{式2})$$

【数 5】

$$\Delta \text{MelCep}_{i,s}^{\text{right}} = \sum_{w=-W}^0 \alpha(w) \cdot \text{MelCep}_{i,s}(w) + \sum_{w=1}^W \alpha(w) \cdot \text{MelCep}_{i,s+1}(w) \cdots \cdots \quad (\text{式3})$$

【0 0 2 5】

20

上記の拡張パラメータは、基本パラメータを用いて、それぞれ（式 4）、（式 5）のよう書き換えることができる。すなわち、拡張パラメータを基本パラメータの関数として表すことができる。

【数 6】

$$\Delta \text{MelCep}_{i,s}^{\text{left}} = H_{i,s}^{\text{begin}} \cdot X_{i,s} + H_{i,s-1}^{\text{end}} \cdot X_{i,s-1} \cdots \cdots \quad (\text{式4})$$

【数 7】

$$\Delta \text{MelCep}_{i,s}^{\text{right}} = H_{i,s}^{\text{end}} \cdot X_{i,s} + H_{s+1}^{\text{begin}} \cdot X_{i,s-1} \cdots \cdots \quad (\text{式5})$$

なお、

30

【数 8】

$$H_{i,s}^{\text{begin}}$$

および

【数 9】

$$H_{i,s}^{\text{end}}$$

は、それぞれ、（式 6）、（式 7）で表される。

【数 1 0】

$$H_{i,s}^{\text{begin}} = \sum_{w=0}^W \alpha(w) \cdot T_{i,s}^{-1}(w) \cdots \cdots \quad (\text{式6})$$

40

【数 1 1】

$$H_{i,s}^{\text{end}} = \sum_{w=-W}^0 \alpha(w) \cdot T_{i,s}^{-1}(-w) \cdots \cdots \quad (\text{式7})$$

【0 0 2 6】

パラメータ化部 1 4 0 は、分割部 1 3 0 により算出された基本パラメータおよび拡張パラメータを（式 8）に示すように、1 つのスペクトルパラメータ  $SP_{i,s}$  に統合する。

## 【数 1 2】

$$SP_{i,s} = (X_{i,s}^t, \Delta MelCep_{i,s}^{left}, \Delta MelCep_{i,s}^{right})^t \cdots \cdots \text{(式8)}$$

## 【0 0 2 7】

クラスタリング部 1 5 0 は、パラメータ化部 1 4 0 により得られた各言語区間のスペクトルパラメータを、区切り情報およびテキスト解析部 1 1 0 により生成された言語情報に基づいてクラスタリングする。具体的には、クラスタリング部 1 5 0 は、言語情報、すなわちコンテキスト情報に関する質問を繰り返しながら分岐を繰り返す決定木に基づいて、スペクトルパラメータを複数のクラスターに分割する。例えば、図 3 に示すように、「対象区間は / a / か?」といった質問に対する Yes、No の答えに応じてスペクトルパラメータは Yes の子ノードと No の子ノードに分割される。質問と、回答によるスペクトルパラメータの分割が繰り返されて、図 3 に示すように言語情報に関する条件が等しい複数のスペクトルパラメータが同一クラスターにグループ化される。

10

## 【0 0 2 8】

図 3 に示す例においては、対象区間、直前区間および直後区間の音素が等しい対象区間のスペクトルパラメータが同一のクラスターになるように分類されている。図 3 に示す例においては、対象区間としての音素 / a / であっても、直前の音素と直後の音素との組が異なる [ ( k ) a ( n ) ] と、[ ( k ) a ( m ) ] はそれぞれ異なるクラスターに分類される。

## 【0 0 2 9】

なお、上記において説明したクラスターは一例であり、他の例としては、上述のように、対象区間、直前区間および直後区間の音素のほか、対象区間におけるアクセントの有無、直前区間、直後区間におけるアクセントの有無など、各区間の音素以外の言語情報を用いてより細かいクラスターに分類してもよい。

20

## 【0 0 3 0】

また、クラスタリングは MFCC の全次元の係数ベクトルに対応する基本パラメータと拡張パラメータを統合したスペクトルパラメータに対して行うこととしたが、他の例としては、MFCC の次元ごとに行ってもよい。各次元でクラスタリングする場合は、クラスタリングするスペクトルパラメータの次元が統合したスペクトルパラメータの次元より小さくなる。このため、クラスタリングの精度を向上させることができる。同様に、統合したスペクトルパラメータの次元を PCA ( P r i n c i p a l C o m p o n e n t A n a l y s i s : 主成分分析 ) の手法を用いて次元圧縮した後に行ってもよい。

30

## 【0 0 3 1】

モデル学習部 1 6 0 は、各クラスターに分類された複数のスペクトルパラメータから、これら複数のスペクトルパラメータの分布を近似するガウス分布のパラメータを学習し、コンテキスト依存のスペクトルモデルとして出力する。具体的には、モデル学習部 1 6 0 は、 $SP_{i,s}$ 、平均ベクトル  $m_{i,s}$  および共分散行列  $\Sigma_{i,s}$  の 3 つのパラメータをスペクトルモデルとして出力する。なお、クラスタリングの方法やガウス分布のパラメータ学習法としては、音声認識の分野でよく知られている方法を利用することができる。

## 【0 0 3 2】

モデル記憶部 1 7 0 は、モデル学習部 1 6 0 により出力された学習モデルを、学習モデルに共通する言語情報の条件に対応付けて記憶する。なお、言語情報の条件とは、クラスタリングにおいて質問に用いた言語情報である。

40

## 【0 0 3 3】

図 4 は、学習モデル生成装置 1 0 0 による学習モデル生成処理を示すフローチャートである。学習モデル生成処理においては、まず学習モデル生成装置 1 0 0 は学習データとしてテキスト情報、テキストの区切り位置を示す区切り情報およびテキストに対応する音声信号を取得する (ステップ S 1 0 0)。具体的には、テキスト情報はテキスト解析部 1 1 0、音声信号はスペクトル分析部 1 2 0、区切り情報は、分割部 1 3 0 およびクラスタリング部 1 5 0 に入力される。

50

## 【0034】

次に、テキスト解析部110は、テキスト情報に基づいて、言語情報を生成する（ステップS102）。スペクトル分析部120は、音声信号の各フレームの特徴パラメータMFCCを算出する（ステップS104）。なお、テキスト解析部110による言語情報の生成およびスペクトル分析部120による特徴パラメータ算出の処理は独立に行われるので、両者の処理順番は問わない。

## 【0035】

次に、分割部130は、区切り情報に基づいて、音声信号の言語区間を特定する（ステップS106）。次に、パラメータ化部140は、言語区間に含まれる複数のフレームそれぞれのMFCCから言語区間のスペクトルパラメータを算出する（ステップS108）。パラメータ化部140はより詳しくは、対象区間だけでなく、対象区間の直前区間、直後区間それぞれに含まれる複数のフレームのMFCCに基づいて、基本パラメータおよび拡張パラメータを要素とするスペクトルパラメータ $SP_{i,s}$ を算出する。

10

## 【0036】

次に、クラスタリング部150は、パラメータ化部140によりテキスト情報の各言語区間に対して得られた複数のスペクトルパラメータを、区切り情報および言語情報に基づいてクラスタリングする（ステップS110）。次に、モデル学習部160は、各クラスターに属する複数のスペクトルパラメータから学習モデルとしてのスペクトルモデルを生成する（ステップS112）。次に、モデル学習部160は、スペクトルモデルを、対応するテキスト情報および言語情報（言語情報の条件）に対応付けてモデル記憶部170に記憶する（ステップS114）。以上で、学習モデル生成装置100による学習モデル生成処理が完了する。

20

## 【0037】

図5および図6からわかるように、本実施の形態にかかる学習モデル生成装置100は、HMMによるスペクトルパラメータに比べて、より実際のスペクトルに近いスペクトルパラメータを得ることができる。学習モデル生成装置100は、複数のフレームに対応する言語区間を単位とするスペクトルパラメータからスペクトルモデルを学習するので、より自然なスペクトルモデルを得ることができる。さらに、このスペクトルモデルを利用することにより、より自然なスペクトルパターンを生成することができる。

30

## 【0038】

また、学習モデル生成装置100は、対象区間に対応する基本パラメータだけでなく、直前区間および直後区間に対応する拡張パラメータを考慮することにより、不連続点が生じることなく滑らかに変化するスペクトルモデルを学習することができる。

## 【0039】

さらに、学習モデル生成装置100は、複数の言語レベルそれぞれに対するスペクトルモデルを学習するので、これらのスペクトルモデルを利用して、総合的なスペクトルパターンを生成することができる。

## 【0040】

図7は、音声合成装置200の構成を示す図である。音声合成装置200は、音声合成の対象となるテキスト情報を取得し、学習モデル生成装置100により生成されたスペクトルモデルに基づいて、音声合成を行う。音声合成装置200は、モデル記憶部210と、テキスト解析部220と、モデル選択部230と、継続時間長算出部240と、スペクトルパラメータ生成部250と、F0生成部260と、駆動信号生成部270と、合成フィルタ280とを備えている。

40

## 【0041】

モデル記憶部210は、学習モデル生成装置100において生成された学習モデルを言語情報の条件に対応付けて記憶している。なお、モデル記憶部210は、学習モデル生成装置100のモデル記憶部170と同様である。テキスト解析部220は、外部から音声合成の対象となるテキスト情報を取得する。テキスト解析部220は、テキスト情報に対し、テキスト解析部110と同様の処理を行う。すなわち、取得したテキスト情報に対応

50

する言語情報を生成する。モデル選択部 230 は、言語情報に基づいて、テキスト解析部 220 に入力されたテキスト情報に含まれる複数の言語区間それぞれに対応する、コンテキスト依存のスペクトルモデルをモデル記憶部 210 から選択する。モデル選択部 230 は、テキスト情報に含まれる複数の言語区間それぞれに対して選択されたスペクトルモデルを接続し、これをテキスト情報全体に対応するモデル系列として出力する。

#### 【0042】

継続時間長算出部 240 は、テキスト解析部 220 から言語情報を取得し、言語情報に定義された各言語区間の開始時間と終了時間とに基づいて、各言語区間の継続時間長を算出する。

#### 【0043】

スペクトルパラメータ生成部 250 は、モデル選択部 230 により選択された言語区間のモデル系列と、継続時間長算出部 240 により各言語区間に対して算出された継続時間長を接続した継続時間長系列とを入力とし、入力されたテキスト全体に対応するスペクトルパラメータを算出する。具体的には、モデル系列と継続時間長系列とに基づいて、スペクトルパラメータ  $SP_{i,s}$  の対数尤度（尤度関数）を総目的関数  $F$  とし、目的関数が最大となるようなスペクトルパラメータを算出する。総目的関数  $F$  は、（式 9）で表される。

#### 【数 13】

$$F = \sum_{\forall s} \log(P(SP_{i,s} | s)) \cdots \cdots \text{(式9)}$$

ここで、 $s$  は、単位区間の集合である。スペクトルパラメータはガウス分布でモデル化されているので、その確率は（式 10）に示すように、ガウス分布の確率密度で与えられる。

#### 【数 14】

$$P(SP_{i,s} | s) = N(SP_{i,s}; \mu_{i,s}, \Sigma_{i,s}) \cdots \cdots \text{(式10)}$$

#### 【0044】

スペクトルパラメータを求めるべく、この総目的関数  $F$  を基準となる言語レベル（音素）でのスペクトルパラメータ  $X_{i,s}$  について最大化する。パラメータの最大化は、勾配法などの公知の技術を用いるものとする。このように、目的関数を最大化することにより、適切なスペクトルパラメータを算出することができる。

#### 【0045】

他の例としては、スペクトルパラメータ生成部 250 は、スペクトルのグローバル分散も考慮に入れて目的関数を最大化することとしてもよい。これにより、生成されるスペクトルのパターンが自然音声のスペクトルパターンの変化幅と同様に変化し、より自然な音声を得ることができる。

#### 【0046】

スペクトルパラメータ生成部 250 は、目的関数の最大化で導出されたスペクトルの基本パラメータ  $X_{i,s}$  を逆変換することで、音素に含まれる複数フレームの MFCC 係数を生成する。なお、逆変換は、言語区間に含まれる複数のフレームに渡って行う。

#### 【0047】

F0 生成部 260 は、テキスト解析部 220 から言語情報を取得し、継続時間長算出部 240 から各言語区間の継続時間長を取得する。F0 生成部 260 は、言語情報に含まれるアクセントの有無などの情報および各言語区間の継続時間長に基づいて、ピッチの基本周波数（F0）を生成する。

#### 【0048】

駆動信号生成部 270 は、F0 生成部 260 から基本周波数（F0）を取得し、基本周波数（F0）から駆動信号を生成する。具体的には、対象区間が有声音である場合には、基本周波数（F0）の逆数であるピッチ周期のパルス列を駆動信号として生成する。また

10

20

30

40

50

、対象区間が無声音である場合、白色雑音を駆動信号として生成する。

#### 【0049】

合成フィルタ280は、スペクトルパラメータ生成部250により得られたスペクトルパラメータおよび駆動信号生成部270により生成された駆動信号から合成フィルタを用いて合成音声を生成し出力する。具体的には、まずスペクトルパラメータであるMFCCパラメータをLPCパラメータに変換する。そして、LPCパラメータを有する全極フィルタを適用する。LPCパラメータを $\alpha_i$  ( $i=1, 2, 3 \dots, p$ )とした場合、合成フィルタとしての全極フィルタの伝達関数 $H(z)$ は、(式11)で表される。ここで、 $p$ は合成フィルタの次数である。

#### 【数15】

$$H(Z)=1/\sum_{i=1}^p \alpha_i Z^{-i} \dots \dots \text{(式11)}$$

10

#### 【0050】

また、全極フィルタへの入力信号である駆動信号を $e(n)$ 、全極フィルタの出力を $y(n)$ とした場合、合成フィルタの動作は(式12)の差分方程式で表される。

#### 【数16】

$$y(n)=\sum_{i=1}^p \alpha_i y(n-i)+e(n) \dots \dots \text{(式12)}$$

20

#### 【0051】

図8は、音声合成装置200による音声合成処理を示すフローチャートである。音声合成処理において、まずテキスト解析部220は音声合成の対象となるテキスト情報を取得する(ステップS200)。次に、テキスト解析部220は、取得したテキスト情報に基づいて、言語情報を生成する(ステップS202)。次に、モデル選択部230は、テキスト解析部220が生成した言語情報に基づいて、モデル記憶部210からテキスト情報に含まれる各言語区間に対するスペクトルモデルを選択し、これらを接続したモデル系列を得る(ステップS204)。次に、継続時間長算出部240は、言語情報に含まれる各言語区間の開始時間および終了時間に基づいて、各言語区間の継続時間長を算出する(ステップS206)。なお、モデル選択部230によるモデル選択処理および継続時間長算出部240による継続時間長算出処理は独立した処理であり、これらの処理順番は特に限定されるものではない。

30

#### 【0052】

次に、スペクトルパラメータ生成部250は、モデル系列および継続時間長系列に基づいて、テキスト情報に対応するスペクトルパラメータを算出する(ステップS208)。次に、F0生成部260は、言語情報および継続時間長に基づいて、ピッチの基本周波数(F0)を生成する(ステップS210)。次に、駆動信号生成部270は、駆動信号を生成する(ステップS212)。次に、合成フィルタ280により合成音声信号が生成され外部に出力されて(ステップS214)、音声合成処理が完了する。

#### 【0053】

このように、本実施の形態にかかる音声合成装置200は、学習モデル生成装置100により生成された、DCT係数で表現されたスペクトルモデルを利用して音声合成を行うので、滑らかに変化する自然なスペクトルを生成することができる。

40

#### 【0054】

図9は、学習モデル生成装置100のハードウェア構成を示す図である。学習モデル生成装置100は、CPU(Central Processing Unit)11と、ROM(Read Only Memory)12と、RAM(Random Access Memory)13と、記憶部14と、表示部15と、操作部16と、通信部17とを備え、各部はバス18を介して接続されている。

#### 【0055】

50

CPU 11は、RAM 13を作業領域として、ROM 12又は記憶部 14に記憶されたプログラムとの協働により各種処理を実行し、学習モデル生成装置 100の動作を統括的に制御する。また、CPU 11は、ROM 12又は記憶部 14に記憶されたプログラムとの協働により、上述の各機能部を実現させる。

【0056】

ROM 12は、学習モデル生成装置 100の制御にかかるプログラムや各種設定情報などを書き換え不可能に記憶する。RAM 13は、SDRAMやDDRメモリなどの揮発性メモリであって、CPU 11の作業エリアとして機能する。

【0057】

記憶部 14は、磁氣的又は光学的に記録可能な記憶媒体を有し、学習モデル生成装置 100の制御にかかるプログラムや各種情報を書き換え可能に記憶する。また、記憶部 14は、上述のモデル学習部 160により生成されるスペクトルモデルなどを記憶する。表示部 15は、LCD (Liquid Crystal Display) などの表示デバイスから構成され、CPU 11の制御の下、文字や画像などを表示する。操作部 16は、マウスやキーボードなどの入力デバイスであって、ユーザから操作入力された情報を指示信号として受け付け、CPU 11に出力する。通信部 17は、外部装置との間で通信を行うインターフェイスであって、外部装置から受信した各種情報をCPU 11に出力する。また、通信部 17は、CPU 11の制御の下、各種情報を外部装置に送信する。なお、音声合成装置 200のハードウェア構成は、学習モデル生成装置 100のハードウェア構成と同様である。

【0058】

本実施の形態にかかる学習モデル生成装置 100および音声合成装置 200において実行される学習モデル生成プログラムおよび音声合成プログラムは、ROM等に予め組み込まれて提供される。

【0059】

本実施の形態の学習モデル生成装置 100および音声合成装置 200で実行される学習モデル生成プログラムおよび音声合成プログラムは、インストール可能な形式又は実行可能な形式のファイルでCD-ROM、フレキシブルディスク (FD)、CD-R、DVD (Digital Versatile Disk) 等のコンピュータで読み取り可能な記録媒体に記録して提供するように構成してもよい。

【0060】

さらに、本実施の形態の学習モデル生成装置 100および音声合成装置 200で実行される学習モデル生成プログラムおよび音声合成プログラムを、インターネット等のネットワークに接続されたコンピュータ上に格納し、ネットワーク経由でダウンロードさせることにより提供するように構成しても良い。また、本実施形態の学習モデル生成装置 100および音声合成装置 200で実行される学習モデル生成プログラムおよび音声合成プログラムをインターネット等のネットワーク経由で提供または配布するように構成しても良い。

【0061】

本実施の形態の学習モデル生成装置 100および音声合成装置 200で実行される学習モデル生成プログラムおよび音声合成プログラムは、上述した各部を含むモジュール構成となっており、実際のハードウェアとしてはCPU (プロセッサ) が上記ROMから学習モデル生成プログラムおよび音声合成プログラムを読み出して実行することにより上記各部が主記憶装置上にロードされ、上述した各部が主記憶装置上に生成されるようになっている。

【0062】

なお、本発明は、上記実施の形態そのままに限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で構成要素を変形して具体化することができる。また、上記実施の形態に開示されている複数の構成要素の適宜な組み合わせにより、種々の発明を形成することができる。例えば、実施の形態に示される全構成要素からいくつかの構成要素を削

10

20

30

40

50

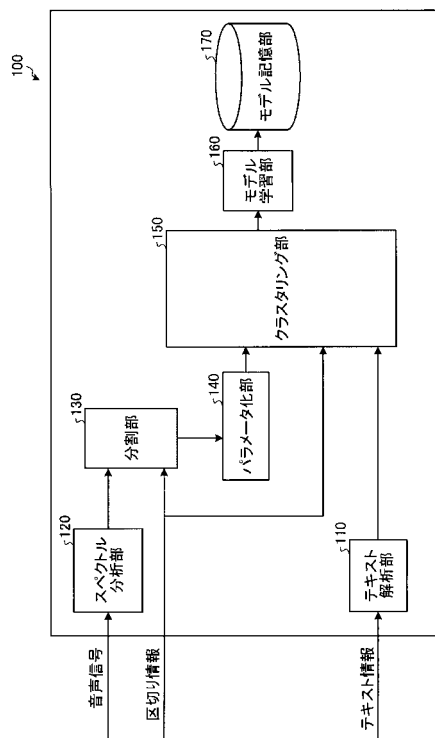
除してもよい。さらに、異なる実施の形態にわたる構成要素を適宜組み合わせても良い。

【符号の説明】

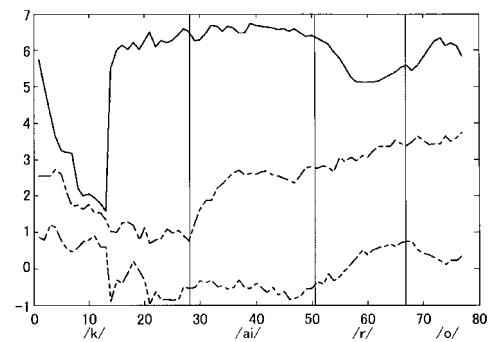
【0063】

- 100 学習モデル生成装置
- 120 スペクトル分析部
- 130 分割部
- 140 パラメータ化部
- 150 クラスタリング部
- 160 モデル学習部

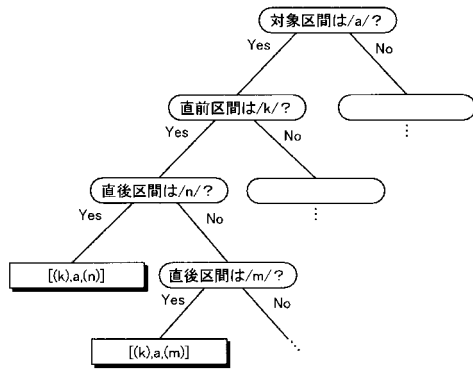
【図1】



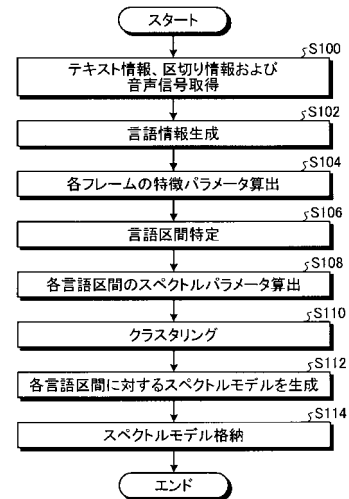
【図2】



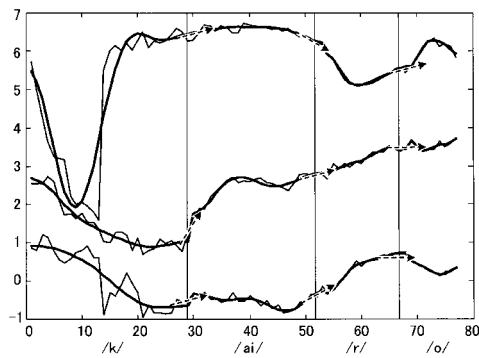
【図 3】



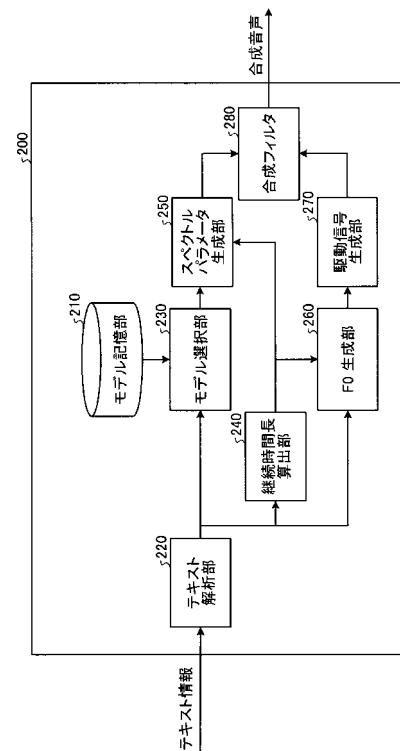
【図 4】



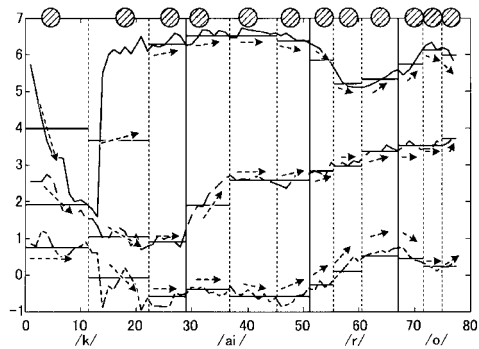
【図 5】



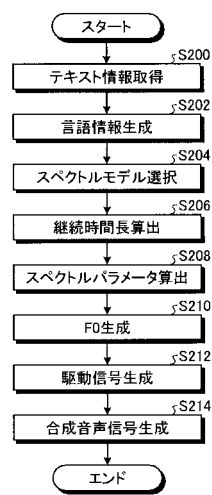
【図 7】



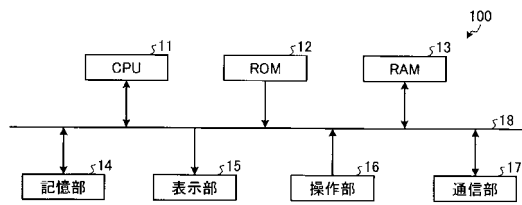
【図 6】



【図 8】



【図 9】



---

フロントページの続き

- (56)参考文献 特開平6-266390 (JP, A)  
特開平8-263520 (JP, A)  
特開2007-163667 (JP, A)  
特開2004-246292 (JP, A)  
特開2010-20166 (JP, A)  
特開2010-224418 (JP, A)

- (58)調査した分野(Int.Cl., DB名)  
G10L 13/06