(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4864783号
(P4864783)
(45)発行日　平成24年2月1日(2012.2.1)　(24)登録日　平成23年11月18日(2011.11.18)

(51)Int.Cl.　F I
G10L 15/20 (2006.01)　G10L 15/20 360Z
G10L 15/02 (2006.01)　G10L 15/02 200Z

請求項の数 7　(全 20 頁)

(21)出願番号　特願2007-76928(P2007-76928)
(22)出願日　平成19年3月23日(2007.3.23)
(65)公開番号　特開2008-233782(P2008-233782A)
(43)公開日　平成20年10月2日(2008.10.2)
審査請求日　平成21年7月10日(2009.7.10)

(73)特許権者　000208891
ＫＤＤＩ株式会社
東京都新宿区西新宿二丁目３番２号
(74)代理人　100106909
弁理士　棚井　澄雄
(74)代理人　100064908
弁理士　志賀　正武
(74)代理人　100089037
弁理士　渡邊　隆
(72)発明者　遠藤　俊樹
埼玉県ふじみ野市大原２丁目１番１５号
株式会社ＫＤＤＩ研究所内
(72)発明者　加藤　恒夫
埼玉県ふじみ野市大原２丁目１番１５号
株式会社ＫＤＤＩ研究所内
最終頁に続く

(54)【発明の名称】パタンマッチング装置、パタンマッチングプログラム、およびパタンマッチング方法

(57)【特許請求の範囲】
【請求項１】
外部より入力された音声データまたは画像データの特徴量を算出する分析手段と、
前記分析手段で算出された前記特徴量を正規化する正規化手段と、
前記正規化手段で正規化された正規化済み特徴量に基づいて、パタンマッチングを行うパタンマッチング手段と、
を備えたパタンマッチング装置において、
前記正規化手段は、
前記音声データの全フレーム数または前記画像データ全体の前記特徴量の平均値である全体平均値を取得する全体平均取得手段と、
前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の平均値である局所平均値を計算する局所平均計算手段と、
前記局所平均値に基づいて、前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の分散値である局所分散値を計算する局所分散計算手段と、
前記全体平均値と複数の前記局所分散値とに基づいて前記特徴量を正規化する正規化処理計算手段と、
を備えたことを特徴とするパタンマッチング装置。
【請求項２】
前記全体平均取得手段は、前記音声データの全フレーム数または前記画像データ全体の前記特徴量から前記全体平均値を計算する

ことを特徴とする、請求項１に記載のパタンマッチング装置。
【請求項３】
前記全体平均取得手段は、予め記憶した所定値を前記全体平均値とする
ことを特徴とする、請求項１に記載のパタンマッチング装置。
【請求項４】
パタンマッチングの対象とする前記特徴量が含まれる範囲を同定する範囲同定手段
を備え、
前記全体平均取得手段は、前記範囲同定手段で同定された範囲に基づく前記音声データの全フレーム数または前記画像データ全体の前記特徴量から前記全体平均値を計算することを特徴とする、請求項２に記載のパタンマッチング装置。
【請求項５】
前記局所平均計算手段は、過去に計算した前記局所平均値を重み付けした値に基づいて、前記局所平均値を計算し、
前記局所分散計算手段は、過去に計算した前記局所分散値を重み付けした値に基づいて、前記局所分散値を計算する
ことを特徴とする請求項１～４に記載のパタンマッチング装置。
【請求項６】
外部より入力された音声データまたは画像データの特徴量を算出する分析手段と、
前記分析手段で算出された前記特徴量を正規化する正規化手段と、
前記正規化手段で正規化された正規化済み特徴量に基づいて、パタンマッチングを行うパタンマッチング手段と、
としてコンピュータを機能させるためのパタンマッチングプログラムにおいて、
前記正規化手段は、
前記音声データの全フレーム数または前記画像データ全体の前記特徴量の平均値である全体平均値を取得する全体平均取得手段と、
前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の平均値である局所平均値を計算する局所平均計算手段と、
前記局所平均値に基づいて、前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の分散値である局所分散値を計算する局所分散計算手段と、
前記全体平均値と複数の前記局所分散値とに基づいて前記特徴量を正規化する正規化処理計算手段と、
としてコンピュータを機能させるためのパタンマッチングプログラム。
【請求項７】
外部より入力された音声データまたは画像データの特徴量を算出する分析ステップと、
前記分析ステップで算出された前記特徴量を正規化する正規化ステップと、
前記正規化ステップで正規化された正規化済み特徴量に基づいて、パタンマッチングを行うパタンマッチングステップと、
を備えたパタンマッチング方法において、
前記正規化ステップは、
前記音声データの全フレーム数または前記画像データ全体の前記特徴量の平均値である全体平均値を取得する全体平均取得ステップと、
前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の平均値である局所平均値を計算する局所平均計算ステップと、
前記局所平均値に基づいて、前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の分散値である局所分散値を計算する局所分散計算ステップと、
前記全体平均値と複数の前記局所分散値とに基づいて前記特徴量を正規化する正規化処理計算ステップと、
を備えたことを特徴とするパタンマッチング方法。
【発明の詳細な説明】
【技術分野】

【０００１】
本発明は、パタンマッチング装置、パタンマッチングプログラム、およびパタンマッチング方法に関する。
【背景技術】
【０００２】
音声認識装置は、入力音声信号から抽出された時系列の音響特徴量を、母音や子音などの音素を単位として、音響特徴量空間における確率密度分布が予め学習された音響モデルと照合することにより認識結果を得る。確率モデルである音響モデルは、音響特徴量の入力に対して、その音素らしさのスコア（音響尤度）を出力する。音声認識装置は文法と単語辞書の制約に従って音素らしさのスコア（音響尤度）を発声全体に渡って累積し、累積スコアが最も高い単語の並びを認識結果として出力する。
【０００３】
音響特徴量は多次元ベクトルの時系列データであり、各次元において各音素に該当するデータの頻度分布を集計すると正規分布に近い形状、もしくは複数の正規分布の和に近い形状になる。こうした音響特徴量の分布を表現するために、音響モデルの確率密度分布は多次元正規分布もしくは複数の多次元正規分布によって表現される。しかし、実際の照合においては、マイク特性のばらつき、話者による違い、背景雑音などにより、入力音響特徴量の分布と音響モデルの確率密度分布との間にミスマッチが生じ、認識率低下の原因となる。入力音響特徴量と音響モデルの照合において、このミスマッチを解消する手法として、ケプストラム平均値正規化（ＣＭＮ: Cepstral Mean Normalization）という手法が広く利用されており、ＣＭＮをさらに発展させた手法として平均値・分散正規化（ＭＶＮ: Mean and Variance Normalization）が提案されている。ＣＭＮは、発声の各時刻の音響特徴量からその発声全体の平均値を減算し、音響特徴量の平均をゼロにすることで、入力音響特徴量の分布と音響モデルの確率密度分布を揃え、ミスマッチを低減する手法である。ＣＭＮ前の各次元の音響特徴量をｘ（ｔ）、ＣＭＮ後の音響特徴量をｘ$_{c}$（ｔ）とすると、ＣＭＮの操作は（１）式、（２）式で表される。Ｔは発声全体のフレーム数を表す。
【０００４】
【数１】

$$x_c(t) = x(t) - E(x) \quad \cdots(1) \qquad E(x) = \frac{1}{T}\sum_{t=1}^{T} x(t) \quad \cdots(2)$$

【０００５】
一方、ＭＶＮとは、発声の各時刻の音響特徴量を、その発声全体の平均値と分散で正規化して、基準系の正規分布Ｎ（平均０、分散１）に揃えることで、マイク特性などによる入力音響特徴量の分布と音響モデルの確率密度分布とのミスマッチを低減する手法である。ＭＶＮ前の各次元の音響特徴量をｘ（ｔ）、ＭＶＮ後の音響特徴量をｘ$_{m}$（ｔ）とすると、ＭＶＮの操作は（３）～（５）式で表される。
【０００６】

【数２】

$$x_m(t)=\frac{x(t)-E(x)}{\sqrt{V(x)}} \quad \cdots(3) \qquad E(x)=\frac{1}{T}\sum_{t=1}^{T}x(t) \quad \cdots(4)$$

$$V(x)=\frac{1}{T-1}\sum_{t=1}^{T}\{x(t)-E(x)\}^2 \quad \cdots(5)$$

【０００７】
また、音声に限らず、静止画および動画についても、ＣＭＮおよびＭＶＮにて正規化が可能である。静止画像の場合、各次元の画像特徴量を$x_{i,j}$、ＣＭＮ後の画像特徴量を$x_{c_{i,j}}$とすると、ＣＭＮの操作は式（６）、（７）で表される。Ｉ、Ｊは静止画の縦軸、横軸のブロック数を表す。
【０００８】
【数３】

$$xc_{i,j}=x_{i,j}-E(x_{i,j}) \quad \cdots(6) \qquad E(x_{i,j})=\frac{1}{I\times J}\sum_{i=1}^{I}\sum_{j=1}^{J}x_{i,j} \quad \cdots(7)$$

【０００９】
一方、ＭＶＮでは、ＭＶＮ前の各次元の画像特徴量を$x_{i,j}$、ＭＶＮ後の画像特徴量を$x_{m_{i,j}}$とすると、ＭＶＮの操作は式（８）～（１０）で表される。
【００１０】
【数４】

$$xm_{i,j}=\frac{x_{i,j}-E(x_{i,j})}{\sqrt{V(x_{i,j})}} \quad \cdots(8) \qquad E(x_{i,j})=\frac{1}{I\times J}\sum_{i=1}^{I}\sum_{j=1}^{J}x_{i,j} \quad \cdots(9)$$

$$V(x_{i,j})=\frac{1}{(I-1)\times(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\{x_{i,j}-E(x_{i,j})\}^2 \quad \cdots(10)$$

【００１１】
動画の場合、各次元の動画特徴量を$x_{i,j,t}$、ＣＭＮ後の動画特徴量を$x_{c_{i,j,t}}$とすると、ＣＭＮの操作は式（１１）、（１２）で表される。Ｉ、Ｊは動画の縦軸、横軸のブロック数、Ｔはフレーム数を表す。
【００１２】

【数５】

$$xc_{i,j,t} = x_{i,j,t} - E\left(x_{i,j,t}\right) \cdots(11) \quad E\left(x_{i,j,t}\right) = \frac{1}{I \times J \times T} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{T=1}^{T} x_{i,j,t} \cdots(12)$$

【００１３】

一方、ＭＶＮでは、ＭＶＮ前の各次元の動画特徴量を$x_{i,j,t}$、ＭＶＮ後の動画特徴量を$x_{mi,j,t}$とすると、ＭＶＮの操作は式（１３）～（１５）で表される。

【００１４】

【数６】

$$xm_{i,j,t} = \frac{x_{i,j,t} - E\left(x_{i,j,t}\right)}{\sqrt{V\left(x_{i,j,t}\right)}} \cdots(13)$$

$$E\left(x_{i,j,t}\right) = \frac{1}{I \times J \times T} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} x_{i,j,t} \cdots(14)$$

$$V\left(x_{i,j,t}\right) = \frac{1}{(I-1) \times (J-1) \times (T-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \{x_{i,j,t} - E(x_{i,j,t})\}^2 \cdots(15)$$

【００１５】

ただし、発声全体の平均値や分散を用いるＣＭＮやＭＶＮは、発声が終わるまで正規化後の音響特徴量が得られないために照合処理の開始が遅れ、発声終了から認識結果出力までの待ち時間を長くしてしまうというデメリットがある。この処理遅れを低減する手法として、発声全体の代わりに数十～数百ミリ秒の局所の区間から平均値や分散を算出して正規化に用いる手法が提案されている。以降、発声全体から計算した平均値を用いて音響特徴量を正規化す手法をバッチＣＭＮ、発声の一部区間から計算した平均値を用いて音響特徴量を正規化する手法をセグメンタルＣＭＮとよぶ。同様に、発声全体から計算した平均値と分散値を用いて音響特徴量を正規化する手法をバッチＭＶＮ、発声の一部区間から計算した平均値と分散値を用いて音響特徴量を正規化する手法をセグメンタルＭＶＮと呼ぶ。

【００１６】

また、特徴量の量子化を仮定しない平均値・分散正規化（ＭＶＮ）において算出した分散の値がゼロもしくはゼロに近い小さな値の場合には分散正規化を行わない手法も知られている（例えば、特許文献１参照）。

【特許文献１】特開２００２－２７８５８６号公報

【発明の開示】

【発明が解決しようとする課題】

【００１７】

しかし、ＣＭＮでは、バッチＣＭＮの方が、セグメンタルＣＭＮよりも長い音声区間から特徴量の平均値を算出するため、精度が高く認識率の改善効果が高いが、入力音響特徴量の分布のばらつきと、参照する音響モデルの確率密度分布のばらつきまで揃えることはできない。

【００１８】
また、ＭＶＮでは、バッチＭＶＮは発声全体の音響特徴量の分布を平均０、分散１に正規化するが、音声認識の単位となる音素ごとの分布に着目すると、分散は正規化されていない。一方、セグメンタルＭＶＮで平均・分散の計算区間を１音素相当の時間長(数十から数百ミリ秒)に設定すれば、音素ごとの分布の分散を正規化するのに近い効果が得られる。ただし、短時間の平均値も０に正規化されるので、すべての音素の分布の平均値が０に近づくため重なりが大きくなり（図2参照）、音素の識別能力の低下を招く。
【００１９】
また、特許文献１では、発声全体の平均値と分散値を用いて正規化するバッチＭＶＮと、局所の平均値と分散値を用いて音響特徴量を正規化するセグメンタルＭＶＮへの適用についてのみ述べられており、前述の音素の識別能力の低下を招くという問題点を解決することができない。
【００２０】
すなわち、ＣＭＮでは分布のばらつき（分散）を正規化することができず、セグメンタルＭＶＮでは、音素ごとの分散の正規化に近い効果があるが、音素間で分布の平均値が近づいてしまい音素の識別能力が低下してしまうという問題がある。
【００２１】
また、上記の課題は、音声に限らず、外部より入力されたデータの特徴量を算出し、算出した特徴量を正規化し、正規化済み特徴量に基づいてパタンマッチングを行うパタンマッチング装置にも当てはまる。
【００２２】
本発明は、上記の課題を解決するためになされたものであり、特徴量の識別能力を低下させること無く特徴量を正規化することが可能なパタンマッチング装置、パタンマッチングプログラム、およびパタンマッチング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】
【００２３】
本発明は、外部より入力された音声データまたは画像データの特徴量を算出する分析手段と、前記分析手段で算出された前記特徴量を正規化する正規化手段と、前記正規化手段で正規化された正規化済み特徴量に基づいて、パタンマッチングを行うパタンマッチング手段と、を備えたパタンマッチング装置において、前記正規化手段は、前記音声データの全フレーム数または前記画像データ全体の前記特徴量の平均値である全体平均値を取得する全体平均取得手段と、前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の平均値である局所平均値を計算する局所平均計算手段と、前記局所平均値に基づいて、前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の分散値である局所分散値を計算する局所分散計算手段と、前記全体平均値と複数の前記局所分散値とに基づいて前記特徴量を正規化する正規化処理計算手段と、を備えたことを特徴とするパタンマッチング装置である。
【００２４】
また、本発明の前記全体平均取得手段は、前記音声データの全フレーム数または前記画像データ全体の前記特徴量から前記全体平均値を計算することを特徴とする。
【００２５】
また、本発明の前記全体平均取得手段は、予め記憶した所定値を前記全体平均値とすることを特徴とする。
【００２６】
また、本発明は、パタンマッチングの対象とする前記特徴量が含まれる範囲を同定する範囲同定手段を備え、前記全体平均取得手段は、前記範囲同定手段で同定された範囲に基づく前記音声データの全フレーム数または前記画像データ全体の前記特徴量から前記全体平均値を計算することを特徴とする、請求項２に記載のパタンマッチング装置である。
【００２７】
また、本発明の前記局所平均計算手段は、過去に計算した前記局所平均値により重み付

けした値に基づいて、前記局所平均値を計算し、前記局所分散計算手段は、過去に計算した前記局所分散値により重み付けした値に基づいて、前記局所分散値を計算することを特徴とする。

【００２８】

また、本発明は、外部より入力された音声データまたは画像データの特徴量を算出する分析手段と、前記分析手段で算出された前記特徴量を正規化する正規化手段と、前記正規化手段で正規化された正規化済み特徴量に基づいて、パタンマッチングを行うパタンマッチング手段と、としてコンピュータを機能させるためのパタンマッチングプログラムにおいて、前記正規化手段は、前記音声データの全フレーム数または前記画像データ全体の前記特徴量の平均値である全体平均値を取得する全体平均取得手段と、前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の平均値である局所平均値を計算する局所平均計算手段と、前記局所平均値に基づいて、前記第２の範囲に含まれる前記特徴量の分散値である局所分散値を計算する局所分散計算手段と、前記全体平均値と複数の前記局所分散値とに基づいて前記特徴量を正規化する正規化処理計算手段と、としてコンピュータを機能させるためのパタンマッチングプログラムである。

【００２９】

また、本発明は、外部より入力された音声データまたは画像データの特徴量を算出する分析ステップと、前記分析ステップで算出された前記特徴量を正規化する正規化ステップと、前記正規化ステップで正規化された正規化済み特徴量に基づいて、パタンマッチングを行うパタンマッチングステップと、を備えたパタンマッチング方法において、前記正規化ステップは、前記音声データの全フレーム数または前記画像データ全体の前記特徴量の平均値である全体平均値を取得する全体平均取得ステップと、前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の平均値である局所平均値を計算する局所平均計算ステップと、前記局所平均値に基づいて、前記音声データの局所のフレーム数または前記画像データの局所範囲の前記特徴量の分散値である局所分散値を計算する局所分散計算ステップと、前記全体平均値と複数の前記局所分散値とに基づいて前記特徴量を正規化する正規化処理計算ステップと、を備えたことを特徴とするパタンマッチング方法である。

【発明の効果】

【００３０】

本発明によれば、特徴量の識別能力を低下させること無く特徴量を正規化することができる。

【発明を実施するための最良の形態】

【００３１】

以下、図面を参照し、本発明の実施形態を説明する。図１は本発明の一実施形態による音声認識装置の構成を示している構成図である。音響分析部１０１は、マイク等より入力された音声データに対して音響分析を行い、音響特徴量を計算する。入力は、プッシュ・ツー・トークで制御することも可能である。また、音響分析部１０１は、計算した音響特徴量を一時的にバッファに記憶させる。正規化処理部１０２は、音響分析部１０１がバッファに記憶させた音響特徴量を、音響特徴量の平均値および分散値を用いて正規化処理を行う。正規化処理については後述する。音響モデル学習部１０３は、学習用音声データに対して、認識対象の音声データと同一の音響分析を音響分析部１０１で行い、正規化処理部１０２で正規化を行って得た、学習用音声データの音響特徴量を音響モデル記憶部１０４に記憶させる。言語モデル記憶部１０５は、単語辞書や文法を記憶する。認識処理部１０６は、認識対象の音声データに対して音響分析部１０１で音響分析を行い、正規化処理部１０２で正規化処理を行って得た、認識対象の音声データの音響特徴量と音響モデル記憶部１０４が記憶している学習用音声データの音響特徴量および言語モデルが記憶している単語辞書や文法を用いてパタンマッチングを行い、認識結果を出力する。

【００３２】

[第１の実施形態]

まず、本発明の第１の実施形態を説明する。図３は、本実施形態による正規化処理部１０２の構成を示している。マイクなどから音声認識装置に入力された１発声全体の音響特徴量は、音響分析部１０１によって図示せぬバッファに格納されている。全体平均計算部３０１は、発声全体に対応したフレーム数Ｔ内の音響特徴量をバッファから読み出し、その平均値を計算する。発声全体の長さとして、単語の長さ、音声の切れ目までの長さ、句読点から句読点までの長さ、入力された音声全体の長さなどを用いることが可能である。発声全体の音響特徴量の平均値Ｅ（ｘ）は計算式（１６）で求める。

【００３３】

【数７】

$$E(x)=\frac{1}{T}\sum_{t=1}^{T}x(t) \qquad \cdots(16)$$

【００３４】

局所平均計算部３０２は、予め設定した局所のフレーム数τ内の発声に対応した音響特徴量をバッファから読み出し、その平均値を計算する。局所のフレーム数τは、音素の長さとして、例えば数十から数百ミリ秒に対応した数である。音素の長さなので発声する単語や人によって変動するが、本実施形態では固定値を使用する。局所のフレーム数τの音響特徴量の平均値E_{τ}（ｘ）は計算式（１７）で求める。

【００３５】

【数８】

$$E_{\tau}(x)=\frac{1}{\tau}\sum_{t=1}^{\tau}x(t) \qquad \cdots(17)$$

【００３６】

局所分散計算部３０３は、予め設定した局所のフレーム数τ内の発声に対応した音響特徴量をバッファから読み出し、その分散値を、局所平均計算部３０２で算出した平均値に基づいて計算する。局所のフレーム数τの音響特徴量の分散値V_{τ}（ｘ）は計算式（１８）で求める。

【００３７】

【数９】

$$V_{\tau}(x)=\frac{1}{\tau-1}\sum_{t=1}^{\tau}\{x(t)-E_{\tau}(x)\}^2 \qquad \cdots(18)$$

【００３８】

正規化処理計算部３０４は、正規化前の音声特徴量から全体平均計算部３０１で算出した発声全体に対しての音響特徴量の平均値を減算し、局所分散計算部３０３で算出した局所のフレーム数τの音響特徴量の分散値で割ることで、正規化後の音響特徴量x_{τ}（ｔ）を求めることができる（計算式（１９）参照）。

【００３９】
【数１０】

$$x_\tau(t)=\frac{x(t)-E(x)}{\sqrt{V_\tau(x)}} \quad \cdots(19)$$

【００４０】
上述したとおり、入力音響特徴量に対して、発声全体の平均値による正規化処理を行うことにより、すべての音素の分布の位置を音響モデルの該当音素の分布に揃え、更に局所の分散値による正規化処理によって、全音素の分布の重なりを抑制しつつ正規分布に近づける効果を持つ（図４参照）。その結果、音素間の識別精度を低減することなく、背景雑音や残響などによる音響モデルと入力された音響特徴量のミスマッチ成分を低減することができ、音声認識精度の劣化を低減することができる。
【００４１】
なお、全体平均計算部３０１で、平均を求めるフレーム数を発声時間に対応する数としたが、代わりにフレーム数τ´を予め設定してもよい。
【００４２】
[第２の実施形態]
次に、本発明の第２の実施形態を説明する。図５は、本実施形態による正規化処理部１０２の構成を示している。本実施形態では、対象とする局所のフレーム数での局所平均値および局所分散値を算出する際に、１つ前の局所のフレーム数の音響特徴量から計算した局所平均値（以下、１つ前の局所平均値と記す。）および１つ前の局所のフレーム数の音響特徴量から計算した局所分散値（以下、１つ前の局所分散値と記す。）を用いることを特徴とする。突発的な雑音が音声認識装置に入力された場合、局所平均値および局所分散値が大きく変わり、入力された音声データを正しく認識することが困難となるが、１つ前の局所平均値および１つ前の局所分散値を用いることで、突発的に音声認識装置に雑音が入力された場合でも局所平均値および局所分散値が大きく変わらず、音声認識制度の劣化を低減することができる。
【００４３】
全体平均計算部５０１は、第１の実施形態と同様に音声認識装置に入力された発声全体対応した音響特徴量をバッファから読み出し、その平均値Ｅ（ｘ）を計算する。局所平均計算部５０２は、予め設定した局所のフレーム数τ内の発声に対応した音響特徴量をバッファから読み出し、その平均値を計算する。その際忘却係数αを予め設定し、1つ前の局所平均値を重み付け加算する。1つ前の局所平均値を重み付け加算した、局所のフレーム数τの音響特徴量の局所平均値E_p（ｔ）は計算式（２０）で求める。
【００４４】
【数１１】

$$E_p(t)=\frac{\alpha}{\tau}\sum_{i=t-\tau}^{t}x(i)+(1-\alpha)E_p(t-1) \quad \cdots(20)$$

【００４５】
局所分散計算部５０３は、予め設定した局所のフレーム数τ内の発声に対応した音響特徴量をバッファから読み出し、その分散値を、局所平均計算部５０２で算出した平均値に

基づいて計算する。その際忘却係数αを予め設定し、1つ前の局所平均値を重み付け加算する。1つ前の局所平均値を重み付け加算した、局所のフレーム数τの音響特徴量の局所分散値V_p（ｔ）は計算式（２１）で求める。

【００４６】

【数１２】

$$V_p(t)=\frac{\alpha}{\tau-1}\sum_{i=t-\tau}^{t}(x(i)-E_p(i))^2+(1-\alpha)V_p(t-1) \quad \cdots(21)$$

【００４７】

正規化処理計算部５０４は、正規化前の音声特徴量から全体平均計算部５０１で算出した発声全体の音響特徴量の平均値を減算し、局所分散計算部５０３で算出した局所のフレーム数τの音響特徴量の分散値で割ることで、正規化後の音響特徴量x_p（ｔ）を求めることができる（計算式（２２）参照）。

【００４８】

【数１３】

$$x_p(t)=\frac{x(t)-E_p(x)}{\sqrt{V_p(x)}} \quad \cdots(22)$$

【００４９】

上述したとおり、入力音響特徴量に対して、発声全体の平均値による正規化処理を行うことにより、すべての音素の分布の位置を音響モデルの該当音素の分布に揃え、更に局所の分散値による正規化処理によって、全音素の分布の重なりを抑制しつつ正規分布に近づける効果を持つ（図４参照）。その結果、音素間の識別精度を低減することなく、背景雑音や残響などによる音響モデルと入力された音響特徴量のミスマッチ成分を低減することができ、音声認識精度の劣化を低減することができる。さらに、突発的な雑音が音声認識装置に入力された場合、入力された音声データを認識することが困難となるが、１つ前の局所平均値および１つ前の局所分散値を用いることで、突発的に音声認識装置に雑音が入力された場合でも平均値が大きく変わらず、音声認識精度の劣化を低減することができる。

【００５０】

[第３の実施形態]

次に、本発明の第３の実施形態を説明する。図６は、本実施形態による正規化処理部１０２の構成を示している。本実施形態では、実施形態１での発声全体の音響特徴量の平均値を算出する代わりに、予め計算した固定の平均値を用いることを特徴とする。これにより、発声全体から音響特徴量の平均値を計算する必要がないため、音響特徴量の正規化が完了するまでの待ち時間が、局所分散の計算に必要な時間となり、リアルタイム処理が可能となる。

【００５１】

固定平均値記憶部６０１は、予め設定した音響特徴量の平均値E_f（ｘ）を記憶する。固定値は、前の発声の平均値を用いる、もしくは過去の莫大な音声データから求めることなどが可能である。

【００５２】

局所平均計算部６０２、局所分散計算部６０３は、第１の実施形態と同様に局所平均値および局所分散値を算出する。正規化処理計算部６０２は、固定平均値記憶部６０１に記憶された固定平均値を用い、正規化前の音声特徴量から固定平均値を減算し、局所分散計算部６０３で算出した局所のフレーム数τの音響特徴量の分散値で割ることで、正規化後の音響特徴量x_f（ｔ）を求めることができる（計算式（２３）参照）。

【００５３】

【数１４】

$$x_f(t)=\frac{x(t)-E_f(x)}{\sqrt{V_\tau(x)}} \quad \cdots(23)$$

【００５４】

上述したとおり、発声全体の音響特徴量の平均値を算出する代わりに、予め計算した固定の平均値を用いることで、発声全体から音響特徴量の平均値をリアルタイムに計算する必要がない。これにより、音響特徴量の正規化が完了するまでの待ち時間が局所分散の計算に必要な時間となり、リアルタイム処理が可能となる。また、入力音響特徴量に対して、発声全体の平均値による正規化処理を行うことにより、すべての音素の分布の位置を音響モデルの該当音素の分布に揃え、更に局所の分散値による正規化処理によって、全音素の分布の重なりを抑制しつつ正規分布に近づける効果を持つ（図４参照）。その結果、音素間の識別精度を低減することなく、背景雑音や残響などによる音響モデルと入力された音響特徴量のミスマッチ成分を低減することができ、音声認識精度の劣化を低減することができる。

【００５５】

[第４の実施形態]

次に、本発明の第４の実施形態を説明する。図７は、本実施形態による正規化処理部１０２の構成を示している。本実施形態では、全体平均計算部７０２の前段に音声検出部を設ける事により音声区間を同定し、音声区間とその前後の数十ミリ秒を加えた時間に対応するフレーム数τ´での平均値を用いて正規化することを特徴とする。これにより、発声終了後に無音区間が長く続いた場合においても、正規化処理までの待ち時間を短くすることが可能となる。

【００５６】

音声検出部７０１は、入力された音響特徴量に音声特有の特徴が含まれていることを検出し、音声区間を同定する。音声特有の特徴としては、音声のパワー、ケプストラム値、周波数などを用いることが可能である。全体平均計算部７０２は、音声検出部７０１で同定した音声区間とその前後の数十ミリ秒を加えた時間に対応するフレーム数τ´での発声に対応した音響特徴量をバッファから読み出し、その平均値を計算する。τ´の音響特徴量の平均値$E_{\tau'}$（ｘ）は計算式（２４）で求める。

【００５７】

【数１５】

$$E_{\tau'}(x)=\frac{1}{\tau'}\sum_{t=1}^{\tau'}x(t) \quad \cdots(24)$$

【００５８】

局所平均計算部７０３は、予め設定した局所のフレーム数τ内の発声に対応した音響特徴量をバッファから読み出し、その平均値を計算する。局所のフレーム数は、音素の長さとして、例えば数十から数百ミリ秒に対応した数である。音素の長さなので発声する単語や人によって変動するが、本実施形態では固定値を使用する。局所のフレーム数τの音響特徴量の平均値$E_{\tau}(x)$は計算式（１７）で求める。

【００５９】

局所分散計算部７０４は、予め設定した局所のフレーム数τ内の発声に対応した音響特徴量をバッファから読み出し、その分散値を、局所平均計算部７０３で算出した平均値に基づいて計算する。局所のフレーム数τの音響特徴量の分散値$V_{\tau}(x)$は計算式（１８）で求める。

【００６０】

正規化処理計算部７０５は、正規化前の音声特徴量から全体平均計算部７０２で算出した発声全体の音響特徴量の平均値を減算し、局所分散計算部７０４で算出した局所のフレーム数τの音響特徴量の分散値で割ることで、正規化後の音響特徴量$x_{\tau'}(t)$を求めることができる（計算式（２５）参照）。

【００６１】

【数１６】

$$x_{\tau'}(t)=\frac{x(t)-E_{\tau'}(x)}{\sqrt{V_{\tau}(x)}} \quad \cdots(25)$$

【００６２】

上述したとおり、全体平均計算部７０２の前段に音声検出部７０１を設ける事により音声区間を同定し、音声区間とその前後の数十ミリ秒を加えた時間に対応するフレーム数τ′での平均値を用いて正規化することにより、発声終了後に無音区間が長く続いた場合においても、正規化処理までの待ち時間を短くすることが可能となる。

【００６３】

[第５の実施形態]

次に、本発明の第５の実施形態を説明する。図８は本実施形態による画像認識装置の構成を示している構成図である。図４において、マイクから入力された音声データの代わりにカメラから入力された画像とし、単語辞書・文法と音響モデルの代わりにオブジェクトモデルとし、音声認識結果の代わりに画像認識結果と置き換えることで、画像認識への適用も可能となる。

【００６４】

画像特徴量分析部８０１は、カメラから入力された画像データに対して画像特徴量分析を行い、画像特徴量を計算する。正規化処理部８０２は、画像特徴量分析部８０１で計算した画像特徴量を画像特徴量の平均値および分散値を用いて正規化処理を行う。正規化処理については後述する。オブジェクトモデル学習部８０３は、学習用画像データに対して、認識対象の画像データと同一の画像特徴量分析を画像特徴量分析部８０１で行い、正規化処理部８０２で正規化を行って得た、学習用画像データの画像特徴量をオブジェクトモデル８０４に記憶させる。認識処理部８０５は、認識対象の画像データに対して画像特徴量分析部８０１で画像特徴量分析を行い、正規化処理部８０２で正規化処理を行って得た、認識対象の画像データの画像特徴量とオブジェクトモデル８０４が記憶している学習用画像データの画像特徴量を用いて認識処理を行い、認識結果を出力する。

【００６５】

図９を参照し本実施形態における画像の正規化処理について説明する。図９は、本実施形態による正規化処理部１０２の構成を示している。全体平均計算部９０１は、カメラ等

から画像認識装置に入力された画像データ全体の画像特徴量をバッファから読み出し、その平均値を計算する。画像データ全体の画像特徴量の平均値Ｅ（$x_{i,j}$）は計算式（２６）で求める。Ｉ、Ｊは静止画の縦軸、横軸のブロック数を表す。

【００６６】

【数１７】

$$E(x_{i,j})=\frac{1}{I\times J}\sum_{i=1}^{I}\sum_{j=1}^{J}x_{i,j} \quad \cdots(26)$$

【００６７】

局所平均計算部９０２は、予め設定した画像データの局所範囲における画像特徴量の平均値を計算する。局所範囲としては、正規化対象画像範囲を含む周囲数ブロックなどを用いることが可能である。局所範囲（ｋ，ｌ）の画像特徴量の平均値Ｅ（$x_{k,l}$）は計算式（２７）で求める。Ｉ、Ｊは静止画の局所範囲での縦軸、横軸のブロック数を表す。

【００６８】

【数１８】

$$E(x_{k,l})=\frac{1}{K\times L}\sum_{k=1}^{K}\sum_{l=1}^{L}x_{k,l} \quad \cdots(27)$$

【００６９】

局所分散計算部９０３は、予め設定した画像データの局所範囲における画像特徴量の分散値を、局所平均計算部９０２で算出した平均値に基づいて計算する。局所範囲（ｋ，ｌ）の画像特徴量の分散値Ｖ（$x_{k,l}$）は計算式（２８）で求める。

【００７０】

【数１９】

$$V(x_{k,l})=\frac{1}{(K-1)\times(L-1)}\sum_{k=1}^{K}\sum_{l=1}^{L}\{x_{k,l}-E(x_{k,l})\}^2 \quad \cdots(28)$$

【００７１】

正規化処理計算部９０４は、正規化前の画像特徴量から全体平均計算部９０１で算出した画像全体の画像特徴量の平均値を減算し、局所分散計算部９０３で算出した予め設定した画像データの範囲における画像特徴量の分散値で割ることで、正規化後の画像特徴量$x_{k,l}$を求めることができる（計算式（２９）参照）。

【００７２】

【数２０】

$$x_{k,l} = \frac{x_{i,j} - E\left(x_{i,j}\right)}{\sqrt{V\left(x_{k,l}\right)}} \qquad \cdots(29)$$

【００７３】
上述したとおり、画像認識においても、画像特徴量に対して画像全体の平均値による正規化処理を行うことにより、すべての画像特徴量の分布の位置をオブジェクトモデルの該当画像特徴量の分布に揃え、更に局所の分散値による正規化処理によって、全画像特徴量の分布の重なりを抑制しつつ正規分布に近づける効果を持つ。その結果、画像特徴量の識別精度を低減することなく、影や輝度などによるオブジェクトモデルと入力された画像特徴量のミスマッチ成分を低減することができ、画像認識精度の劣化を低減することができる。

【００７４】
なお、画像認識については、平面画像だけではなく、３Ｄ画像でも可能である。３Ｄ画像を作成する際にカメラの位置によって、対象物の陰が変わるが、本発明の正規化を用いることで、画像特徴量のミスマッチ成分を低減することができ、画像認識精度の劣化を低減することができる。

【００７５】
[第６の実施形態]
また、画像認識に時間要素を取り入れることで、動画についても動画特徴量のミスマッチ成分を低減することができ、動画認識精度の劣化を低減することができる。

【００７６】
本発明の第６の実施形態を説明する。図１０は本実施形態による動画認識装置の構成を示している構成図である。図４において、マイクから入力された音声データの代わりにカメラから入力された動画とし、単語辞書・文法記憶部と音響モデル記憶部の代わりにオブジェクトモデル記憶部とし、音声認識結果の代わりに動画認識結果と置き換えることで、動画認識への適用も可能となる。

【００７７】
動画特徴量分析部１００１は、カメラから入力された動画データに対して動画特徴量分析を行い、動画特徴量を計算する。正規化処理部１００２は、動画特徴量分析部１００１で計算した動画特徴量を動画特徴量の平均値および分散値を用いて正規化処理を行う。正規化処理については後述する。オブジェクトモデル学習部１００３は、学習用動画データに対して、認識対象の動画データと同一の動画特徴量分析を動画特徴量分析部１００１で行い、正規化処理部１００２で正規化を行って得た、学習用動画データの動画特徴量をオブジェクトモデル１００４に記憶させる。認識処理部１００５は、認識対象の動画データに対して動画特徴量分析部１００１で動画特徴量分析を行い、正規化処理部１００２で正規化処理を行って得た、認識対象の動画データの動画特徴量とオブジェクトモデル１００４が記憶している学習用動画データの動画特徴量を用いて認識処理を行い、認識結果を出力する。

【００７８】
図１１を参照し本実施形態における動画の正規化処理について説明する。図１１は、本実施形態による正規化処理部１０２の構成を示している。全体平均計算部１１０１は、カメラ等から動画認識装置に入力された動画データ全体の動画特徴量をバッファから読み出し、その平均値を計算する。動画データ全体の動画特徴量の平均値Ｅ（$x_{i,j,T}$）は計

算式（３０）で求める。Ｉ、Ｊは動画の縦軸、横軸のブロック数、Ｔはフレーム数を表す。

【００７９】

【数２１】

$$E\left(x_{i,j,t}\right)=\frac{1}{I\times J\times T}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{T=1}^{T}x_{i,j,t} \qquad \cdots(30)$$

【００８０】

局所平均計算部１１０２は、予め設定した動画データの局所範囲における画像特徴量の平均値を計算する。局所範囲としては、正規化対象動画範囲を含む周囲数ブロックおよび局所のフレーム数を用いることが可能である。局所範囲（ｋ，ｌ）および局所のフレーム数τの動画特徴量の平均値Ｅ（$x_{k,l,\tau}$）は計算式（３１）で求める。Ｉ、Ｊは動画の局所範囲での縦軸、横軸の区間のブロック数、τは局所のフレーム数を表す。

【００８１】

【数２２】

$$E\left(x_{k,l,\tau}\right)=\frac{1}{K\times L\times \mathrm{T}}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{\tau=1}^{\mathrm{T}}x_{k,l,\tau} \qquad \cdots(31)$$

【００８２】

局所分散計算部１１０３は、予め設定した動画データの局所範囲における動画特徴量の分散値を、局所平均計算部１１０２で算出した平均値に基づいて計算する。局所範囲（ｋ，ｌ）および局所のフレーム数τの動画特徴量の分散値Ｖ（$x_{k,l,\tau}$）は計算式（３２）で求める。

【００８３】

【数２３】

$$V\left(x_{k,l,\tau}\right)=\frac{1}{(K-1)\times(L-1)\times(\mathrm{T}-1)}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{\tau=1}^{\mathrm{T}}\left\{x_{k,l,\tau}-E(x_{k,l,\tau})\right\}^{2} \qquad \cdots(32)$$

【００８４】

正規化処理計算部１１０４は、正規化前の動画特徴量から全体平均計算部１１０１で算出した動画全体の動画特徴量の平均値を減算し、局所分散計算部１１０３で算出した予め設定した動画データの範囲および局所のフレーム数における動画特徴量の分散値で割ることで、正規化後の動画特徴量$x_{k,l,\tau}$を求めることができる（計算式（３３）参照）。

【００８５】

【数２４】

$$x_{k,l,\tau} = \frac{x_{i,j,t} - E\left(x_{i,j,t}\right)}{\sqrt{V\left(x_{k,l,\tau}\right)}} \qquad \cdots(33)$$

【００８６】

以上、この発明の実施形態について図面を参照して詳述してきたが、具体的な構成はこの実施形態に限られるものではなく、この発明の要旨を逸脱しない範囲の設計等も含まれる。

【００８７】

例えば、音声、画像、および動画について詳述してきたが、音声、画像、および動画に限らず、入力されたデータの特徴量に基づいてパタンマッチングを行う認識装置にも本発明が適用可能である。

【００８８】

また、第２～第４の実施形態については音声認識について説明したが、画像認識および動画認識についても適用可能である。

【００８９】

また、図１などに示す正規化処理部１０２の機能を実現するためのプログラムをコンピュータ読み取り可能な記録媒体に記録して、この記録媒体に記録されたプログラムをコンピュータシステムに読み込ませ、実行することにより、正規化処理を行ってもよい。なお、ここでいう「コンピュータシステム」とは、ＯＳや周辺機器等のハードウェアを含むものであってもよい。

【００９０】

また、「コンピュータシステム」は、WWWシステムを利用している場合であれば、ホームページ提供環境（あるいは表示環境）も含むものとする。また、「コンピュータ読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、フラッシュメモリ等の書き込み可能な不揮発性メモリ、ＣＤ－ＲＯＭ等の可搬媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。

【００９１】

さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムが送信された場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリ（例えばＤＲＡＭ（Dynamic Random Access Memory））のように、一定時間プログラムを保持しているものも含むものとする。

【００９２】

また、上記プログラムは、このプログラムを記憶装置等に格納したコンピュータシステムから、伝送媒体を介して、あるいは、伝送媒体中の伝送波により他のコンピュータシステムに伝送されてもよい。ここで、プログラムを伝送する「伝送媒体」は、インターネット等のネットワーク（通信網）や電話回線等の通信回線（通信線）のように情報を伝送する機能を有する媒体のことをいう。また、上記プログラムは、前述した機能の一部を実現するためのものであっても良い。さらに、前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、いわゆる差分ファイル（差分プログラム）であっても良い。

【図面の簡単な説明】

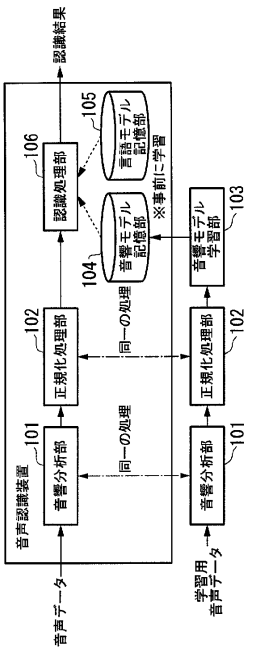
【００９３】

【図１】本発明の一実施形態による音声認識装置の構成を示した構成図である。

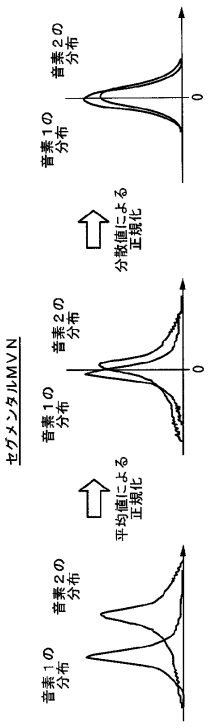
【図２】セグメンタルＭＶＮ手法の正規化処理による分布の変化の様子を示した図である

。
【図３】本発明の第１の実施形態による正規化処理部の構成を示した図である。
【図４】本発明の正規化処理による分布の変化の様子を示した図である。
【図５】本発明の第２の実施形態による正規化処理部の構成を示した図である。
【図６】本発明の第３の実施形態による正規化処理部の構成を示した図である。
【図７】本発明の第４の実施形態による正規化処理部の構成を示した図である。
【図８】本発明の第５の実施形態による画像認識装置の構成を示した構成図である。
【図９】本発明の第５の実施形態による正規化処理部の構成を示した図である。
【図１０】本発明の第６の実施形態による画像認識装置の構成を示した構成図である。
【図１１】本発明の第６の実施形態による正規化処理部の構成を示した図である。
【符号の説明】
【００９４】
１０１・・・音声分析部、１０２,８０２,１００２・・・正規化処理部、１０３・・・音響モデル学習部、１０４・・・音響モデル、１０５・・・言語モデル、１０６,８０５・・・認識処理部、３０１,５０１,７０２,９０１,１１０１・・・全体平均計算部、３０２,５０２,６０２,７０３,９０２,１１０２・・・局所平均計算部、３０３,５０３,６０３,７０４,９０３,１１０３・・・局所分散計算部、３０４,５０４,６０４,７０５,９０４,１１０４・・・正規化処理計算部、６０１・・・固定平均値記憶部、７０１・・・音声検出部、８０１・・・画像特徴量分析部、８０３,１００３・・・オブジェクトモデル学習部、８０４,１００４・・・オブジェクトモデル、１００１・・・動画特徴量分析部

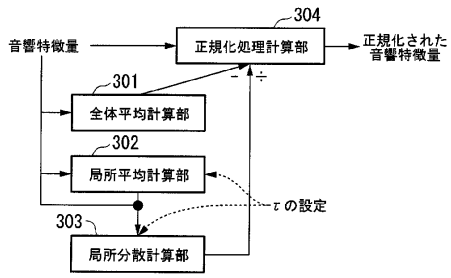
【図１】



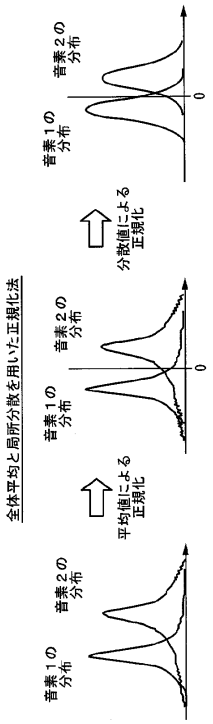
【図２】



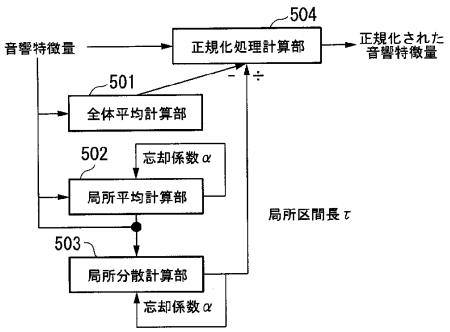
【図３】



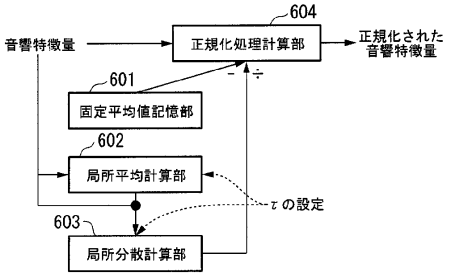
【図４】



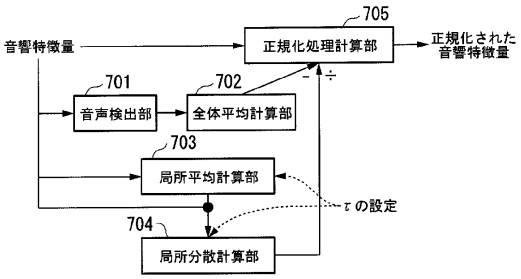
【図５】



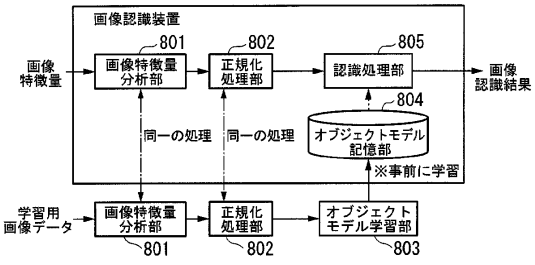
【図６】



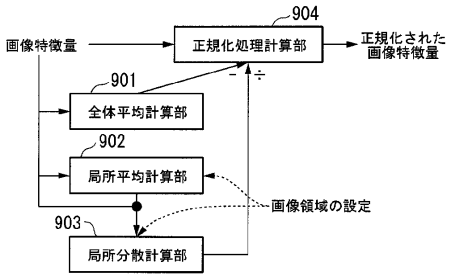
【図７】



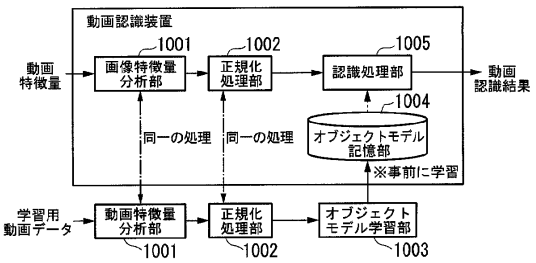
【図８】



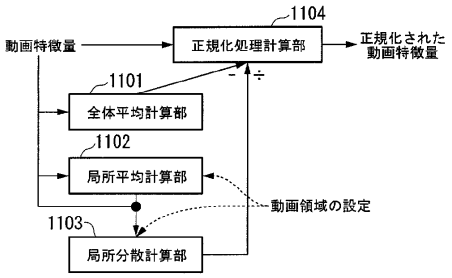
【図９】


904
画像特徴量
正規化処理計算部
正規化された
画像特徴量
901
全体平均計算部
902
局所平均計算部
画像領域の設定
903
局所分散計算部

【図１０】


動画認識装置
1001
1002
1005
動画
特徴量
画像特徴量
分析部
正規化
処理部
認識処理部
動画
認識結果
1004
同一の処理
同一の処理
オブジェクトモデル
記憶部
※事前に学習
学習用
動画データ
動画特徴量
分析部
正規化
処理部
オブジェクト
モデル学習部
1001
1002
1003

【図１１】


1104
動画特徴量
正規化処理計算部
正規化された
動画特徴量
1101
全体平均計算部
1102
局所平均計算部
動画領域の設定
1103
局所分散計算部

フロントページの続き

審査官　前田　祐希

(56)参考文献　特開２００６－０８４６５９（ＪＰ，Ａ）
特表２００７－５３６５６２（ＪＰ，Ａ）
特表２００５－５２１０９１（ＪＰ，Ａ）
特開平１１－０８５２００（ＪＰ，Ａ）
特開２００３－１６７５９９（ＪＰ，Ａ）
原　一眞，柘植　覚，獅子堀　正幹，北　研二，黒岩　眞吾，"実時間分散正規化手法の検討　A Study on Real-time Cepstral Variance Normalizaton"，日本音響学会２００３年秋季研究発表会講演論文集－Ｉ－　THE 2003 AUTUMN MEETING OF THE ACOUSTICAL SOCIETY OF JAPAN，日本，社団法人日本音響学会，２００３年　9月１７日，p151～152

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００　－　１７／００