

(19) 日本国特許庁(JP)

(12) 公 開 特 許 公 報(A)

(11) 特許出願公開番号  
特開2012-163918  
(P2012-163918A)

(43) 公開日 平成24年8月30日 (2012. 8. 30)

|                         |                       |             |
|-------------------------|-----------------------|-------------|
| (51) Int.Cl.            | F I                   | テーマコード (参考) |
| G 1 O L 15/02 (2006.01) | G 1 O L 15/02 3 O O J | 5 D O 1 5   |
| G 1 O L 15/20 (2006.01) | G 1 O L 15/20 2 O O Z |             |
| G 1 O L 11/00 (2006.01) | G 1 O L 11/00 4 O 2 L |             |
| G O 6 N 3/00 (2006.01)  | G O 6 N 3/00 5 6 O G  |             |

審査請求 未請求 請求項の数 11 O L (全 18 頁)

|           |                            |          |                |
|-----------|----------------------------|----------|----------------|
| (21) 出願番号 | 特願2011-26240 (P2011-26240) | (71) 出願人 | 000002185      |
| (22) 出願日  | 平成23年2月9日 (2011. 2. 9)     |          | ソニー株式会社        |
|           |                            |          | 東京都港区港南1丁目7番1号 |
|           |                            | (74) 代理人 | 100093241      |
|           |                            |          | 弁理士 宮田 正昭      |
|           |                            | (74) 代理人 | 100101801      |
|           |                            |          | 弁理士 山田 英治      |
|           |                            | (74) 代理人 | 100086531      |
|           |                            |          | 弁理士 澤田 俊夫      |
|           |                            | (74) 代理人 | 100095496      |
|           |                            |          | 弁理士 佐々木 榮二     |
|           |                            | (74) 代理人 | 110000763      |
|           |                            |          | 特許業務法人大同特許事務所  |

最終頁に続く

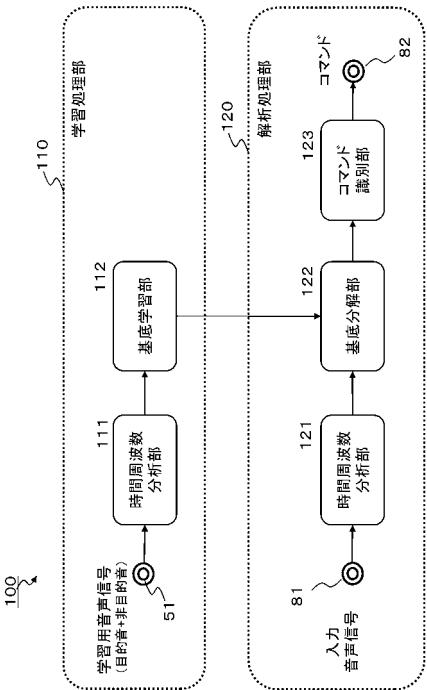
(54) 【発明の名称】 音声信号処理装置、および音声信号処理方法、並びにプログラム

(57) 【要約】

【課題】複数の音が混在した音声信号から、特定音源のコマンドを高精度に分離する装置、方法を提供する。

【解決手段】複数音源からの音を含む学習用音声信号に基づいて各音源対応の基底周波数Bからなる学習データを生成し、この各音源対応の基底周波数Bを結合した全基底周波数B a l lを生成する。さらに、入力音声信号に対して、時間周波数分析を実行して時間周波数分析結果を生成する。この入力音声信号に対する時間周波数分析結果に対して、全基底周波数B a l lを適用した基底分解を行い、入力音声信号に対する基底アクティビティHを生成する。最後に生成した基底アクティビティHの識別処理を実行してコマンド識別を行う。これらの学習データに基づく音源分離処理により、高精度なコマンド識別が実現される。

【選択図】 図 1



## 【特許請求の範囲】

## 【請求項 1】

入力音声信号の時間周波数分析を実行する時間周波数分析部と、

複数音源からの音を含む学習用音声信号に基づいて予め生成した学習データであり、各音源対応の基底周波数 B からなる学習データを入力し、前記各音源対応の基底周波数 B を結合した全基底周波数 B a l l を適用して、前記時間周波数分析部から入力する入力音声信号に対する時間周波数分析結果の基底分解を行い、入力音声信号に対する基底アクティビティ H を生成する基底分解部と、

前記基底分解部から、前記基底アクティビティ H を入力し、入力した基底アクティビティ H の識別処理を実行してコマンド識別を行うコマンド識別部を有する音声信号処理装置

10

## 【請求項 2】

前記学習データは、

前記コマンドとして識別する音に対応する基底周波数を持つ目的音と、識別対象としない非目的音とを含む学習用音声信号に基づいて生成される学習データであり、

前記基底分解部は、

前記目的音対応の基底周波数と、前記非目的音対応の基底周波数を結合した全基底周波数 B a l l を適用して、前記時間周波数分析部から入力する入力音声信号に対する時間周波数分析結果の基底分解を行い、入力音声信号に対する基底アクティビティ H を生成する請求項 1 に記載の音声信号処理装置。

20

## 【請求項 3】

前記時間周波数分析部は、

前記入力音声信号の時間周波数分析を行い、時間周波数スペクトルを生成し、さらに前記時間周波数スペクトルに基づくパワースペクトルを算出し、該パワースペクトルを前記時間周波数分析結果として、前記基底分解部に提供する請求項 1 に記載の音声信号処理装置。

## 【請求項 4】

前記基底分解部は、

前記入力音声信号に基づいて生成されたパワースペクトルを、前記時間周波数分析部から入力し、入力パワースペクトルに対して、前記全基底周波数 B a l l を適用して基底分解を行い、入力音声信号に対する基底アクティビティ H を生成する請求項 3 に記載の音声信号処理装置。

30

## 【請求項 5】

前記コマンド識別部は、

前記基底分解部から、前記基底アクティビティ H を入力し、入力した基底アクティビティ H と、予め設定した閾値との比較処理を行ってコマンドと非コマンドを判定する処理を実行する請求項 1 に記載の音声信号処理装置。

## 【請求項 6】

前記音声信号処理装置は、

複数音源からの音を含む学習用音声信号に基づいて、各音源対応の基底周波数 B からなる学習データを生成する学習処理部を有し、

40

前記基底分解部は、

前記学習処理部の生成した学習データを適用して、前記入力音声信号の基底アクティビティ H を生成する請求項 1 に記載の音声信号処理装置。

## 【請求項 7】

音声コマンドの正負判定に必要な特徴量を予め算出する学習処理部と、

前記学習処理部において学習された特徴量を用いて音源分離処理を行う解析処理部を備えた音声信号処理装置。

## 【請求項 8】

前記学習処理部において算出する音声コマンドの正負判定に必要な特徴量は、前記解析

50

処理部における音声コマンド認識処理において実行する音声コマンド対応の目的音と音声コマンド非対応の非目的音との判別処理である正負判定処理に必要な特徴量である請求項 7 に記載の音声信号処理装置。

【請求項 9】

音声信号処理装置において、入力音声信号からのコマンド識別処理を行う音声信号処理方法であり、

時間周波数分析部が、入力音声信号の時間周波数分析を実行する時間周波数分析ステップと、

基底分解部が、複数音源からの音を含む学習用音声信号に基づいて予め生成した学習データであり、各音源対応の基底周波数 B からなる学習データを入力し、前記各音源対応の基底周波数 B を結合した全基底周波数 B a l l を適用して、前記時間周波数分析部から入力する入力音声信号に対する時間周波数分析結果の基底分解を行い、入力音声信号に対する基底アクティビティ H を生成する基底分解ステップと、

コマンド識別部が、前記基底分解ステップにおいて生成した前記基底アクティビティ H を入力し、入力した基底アクティビティ H の識別処理を実行してコマンド識別を行うコマンド識別ステップを実行する音声信号処理方法。

【請求項 10】

音声信号処理装置において、入力音声信号からのコマンド識別処理を行う音声信号処理方法であり、

学習処理部が、音声コマンドの正負判定に必要な特徴量を予め算出する学習処理ステップと、

解析処理部が、前記学習処理ステップにおいて学習された特徴量を用いて音源分離処理を行う解析処理ステップを実行する音声信号処理方法。

【請求項 11】

音声信号処理装置において、入力音声信号からのコマンド識別処理を実行させるプログラムであり、

時間周波数分析部に。入力音声信号の時間周波数分析を実行させる時間周波数分析ステップと、

基底分解部に、複数音源からの音を含む学習用音声信号に基づいて予め生成した学習データであり、各音源対応の基底周波数 B からなる学習データを入力し、前記各音源対応の基底周波数 B を結合した全基底周波数 B a l l を適用して、前記時間周波数分析部から入力する入力音声信号に対する時間周波数分析結果の基底分解を行い、入力音声信号に対する基底アクティビティ H を生成させる基底分解ステップと、

コマンド識別部に、前記基底分解ステップにおいて生成した前記基底アクティビティ H を入力し、入力した基底アクティビティ H の識別処理を実行してコマンド識別を行わせるコマンド識別ステップを実行させるプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、音声信号処理装置、および音声信号処理方法、並びにプログラムに関する。さらに、詳細には複数の信号が混合された信号を例えば音源単位で分離する処理を実行する音声信号処理装置、および音声信号処理方法、並びにプログラムに関する。

【0002】

本発明は、声や雑音などの様々な音源からの音が混合して入力される環境において、例えばユーザの声に対応する音声コマンドなど、特定の音源からの音を選択分離する信号処理装置、および信号処理方法、並びにプログラムに関する。

【背景技術】

【0003】

昨今の情報処理機器や家電機器などの装置では、音声入力部としてのマイクを備え、マイクから入力するユーザの声を認識して、認識結果に基づいて様々な動作を実行するもの

10

20

30

40

50

がある。すなわちユーザの発する言葉を解析して音声コマンドとして解釈し、そのコマンドに従って処理を実行するものである。

#### 【0004】

音声コマンドによる処理を実行する装置においては、正確な音声認識を行うことが要請されるが、様々な雑音やノイズの発生している環境では、音声入力部としてのマイクを介して入力する音信号に、ユーザの声の他、様々な音源からのノイズが混在した信号となってしまう。

#### 【0005】

このような混合信号からユーザの声を抽出するため、多くの装置ではマイクを介した入力信号を音源分離処理を実行する信号処理部に入力し、ユーザの声を分離する処理を行う。その後、分離抽出されたユーザの声に基づいてコマンド解釈が行われる。

10

#### 【0006】

なお、音源分離処理について開示した従来技術としては、例えば特許文献1（特開2006-238409号公報）や、特許文献2（特開2008-134298号公報）がある。これらの特許文献は、独立成分分析（ICA：Independent Component Analysis）に基づく音源分離処理を開示している。

#### 【0007】

しかし、音源分離処理における問題点として、簡易な構成では分離処理機能が不十分であり、一方、高い分離機能を実現するためには処理負荷や処理時間が増大し、装置としてのコストも大きくなるという問題点がある。一般的な家電機器等に備えるためには、処理負荷やコストを低く抑えることが要求される。また、従来の音源分離処理は、前段の分離処理と後段の認識処理が別のモジュールとして独立していたため、認識に必要な特徴量の情報を用いて分離処理を行うなどの全体最適化を行うことが困難であるという問題があった。

20

#### 【先行技術文献】

#### 【特許文献】

#### 【0008】

【特許文献1】特開2006-238409号公報

【特許文献2】特開2008-134289号公報

#### 【発明の概要】

30

#### 【発明が解決しようとする課題】

#### 【0009】

本発明は、このような状況に鑑みてなされたものであり、簡易な構成で実現でき、かつ全体最適化が行われ、より精度の高い音源分離を実現する音声信号処理装置、および音声信号処理方法、並びにプログラム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】

本発明の第1の側面は、

入力音声信号の時間周波数分析を実行する時間周波数分析部と、

40

複数音源からの音を含む学習用音声信号に基づいて予め生成した学習データであり、各音源対応の基底周波数Bからなる学習データを入力し、前記各音源対応の基底周波数Bを結合した全基底周波数B<sub>all</sub>を適用して、前記時間周波数分析部から入力する入力音声信号に対する時間周波数分析結果の基底分解を行い、入力音声信号に対する基底アクティビティHを生成する基底分解部と、

前記基底分解部から、前記基底アクティビティHを入力し、入力した基底アクティビティHの識別処理を実行してコマンド識別を行うコマンド識別部を有する音声信号処理装置にある。

#### 【0011】

さらに、本発明の音声信号処理装置の一実施態様において、前記学習データは、前記コマンドとして識別する音に対応する基底周波数を持つ目的音と、識別対象としない非目的

50

音とを含む学習用音声信号に基づいて生成される学習データであり、前記基底分解部は、前記目的音対応の基底周波数と、前記非目的音対応の基底周波数を結合した全基底周波数  $B_{all}$  を適用して、前記時間周波数分析部から入力する入力音声信号に対する時間周波数分析結果の基底分解を行い、入力音声信号に対する基底アクティビティ  $H$  を生成する。

【0012】

さらに、本発明の音声信号処理装置の一実施態様において、前記時間周波数分析部は、前記入力音声信号の時間周波数分析を行い、時間周波数スペクトルを生成し、さらに前記時間周波数スペクトルに基づくパワースペクトルを算出し、該パワースペクトルを前記時間周波数分析結果として、前記基底分解部に提供する。

【0013】

さらに、本発明の音声信号処理装置の一実施態様において、前記基底分解部は、前記入力音声信号に基づいて生成されたパワースペクトルを、前記時間周波数分析部から入力し、入力パワースペクトルに対して、前記全基底周波数  $B_{all}$  を適用して基底分解を行い、入力音声信号に対する基底アクティビティ  $H$  を生成する。

【0014】

さらに、本発明の音声信号処理装置の一実施態様において、前記コマンド識別部は、前記基底分解部から、前記基底アクティビティ  $H$  を入力し、入力した基底アクティビティ  $H$  と、予め設定した閾値との比較処理を行ってコマンドと非コマンドを判定する処理を実行する。

【0015】

さらに、本発明の音声信号処理装置の一実施態様において、前記音声信号処理装置は、複数音源からの音を含む学習用音声信号に基づいて、各音源対応の基底周波数  $B$  からなる学習データを生成する学習処理部を有し、前記基底分解部は、前記学習処理部の生成した学習データを適用して、前記入力音声信号の基底アクティビティ  $H$  を生成する。

【0016】

さらに、本発明の第2の側面は、  
音声コマンドの正負判定に必要な特徴量を予め算出する学習処理部と、  
前記学習処理部において学習された特徴量を用いて音源分離処理を行う解析処理部を備えた音声信号処理装置にある。

【0017】

さらに、本発明の音声信号処理装置の一実施態様において、前記学習処理部において算出する音声コマンドの正負判定に必要な特徴量は、前記解析処理部における音声コマンド認識処理において実行する音声コマンド対応の目的音と音声コマンド非対応の非目的音との判別処理である正負判定処理に必要な特徴量である。

【0018】

さらに、本発明の第3の側面は、  
音声信号処理装置において、入力音声信号からのコマンド識別処理を行う音声信号処理方法であり、

時間周波数分析部が、入力音声信号の時間周波数分析を実行する時間周波数分析ステップと、

基底分解部が、複数音源からの音を含む学習用音声信号に基づいて予め生成した学習データであり、各音源対応の基底周波数  $B$  からなる学習データを入力し、前記各音源対応の基底周波数  $B$  を結合した全基底周波数  $B_{all}$  を適用して、前記時間周波数分析部から入力する入力音声信号に対する時間周波数分析結果の基底分解を行い、入力音声信号に対する基底アクティビティ  $H$  を生成する基底分解ステップと、

コマンド識別部が、前記基底分解ステップにおいて生成した前記基底アクティビティ  $H$  を入力し、入力した基底アクティビティ  $H$  の識別処理を実行してコマンド識別を行うコマンド識別ステップを実行する音声信号処理方法にある。

【0019】

さらに、本発明の第4の側面は、

10

20

30

40

50

音声信号処理装置において、入力音声信号からのコマンド識別処理を行う音声信号処理方法であり、

学習処理部が、音声コマンドの正負判定に必要な特徴量を予め算出する学習処理ステップと、

解析処理部が、前記学習処理ステップにおいて学習された特徴量を用いて音源分離処理を行う解析処理ステップを実行する音声信号処理方法にある。

#### 【0020】

さらに、本発明の第5の側面は、

音声信号処理装置において、入力音声信号からのコマンド識別処理を実行させるプログラムであり、

時間周波数分析部に、入力音声信号の時間周波数分析を実行させる時間周波数分析ステップと、

基底分解部に、複数音源からの音を含む学習用音声信号に基づいて予め生成した学習データであり、各音源対応の基底周波数Bからなる学習データを入力し、前記各音源対応の基底周波数Bを結合した全基底周波数B a l lを適用して、前記時間周波数分析部から入力する入力音声信号に対する時間周波数分析結果の基底分解を行い、入力音声信号に対する基底アクティビティHを生成させる基底分解ステップと、

コマンド識別部に、前記基底分解ステップにおいて生成した前記基底アクティビティHを入力し、入力した基底アクティビティHの識別処理を実行してコマンド識別を行わせるコマンド識別ステップを実行させるプログラムにある。

#### 【0021】

なお、本発明のプログラムは、例えば、様々なプログラム・コードを実行可能な画像処理装置やコンピュータ・システムに対して、コンピュータ可読な形式で提供する記憶媒体、通信媒体によって提供可能なプログラムである。このようなプログラムをコンピュータ可読な形式で提供することにより、情報処理装置やコンピュータ・システム上でプログラムに応じた処理が実現される。

#### 【0022】

本発明のさらに他の目的、特徴や利点は、後述する本発明の実施例や添付する図面に基づくより詳細な説明によって明らかになるであろう。なお、本明細書においてシステムとは、複数の装置の論理的集合構成であり、各構成の装置が同一筐体内にあるものには限らない。

#### 【発明の効果】

#### 【0023】

本発明の一実施例の構成によれば、複数の音が混在した音声信号から、特定音源のコマンドを高精度に分離する装置、方法が実現される。具体的には、例えば、複数音源からの音を含む学習用音声信号に基づいて各音源対応の基底周波数Bからなる学習データを生成し、この各音源対応の基底周波数Bを結合した全基底周波数B a l lを生成する。さらに、入力音声信号に対して、時間周波数分析を実行して時間周波数分析結果を生成する。この入力音声信号に対する時間周波数分析結果に対して、全基底周波数B a l lを適用した基底分解を行い、入力音声信号に対する基底アクティビティHを生成する。最後に生成した基底アクティビティHの識別処理を実行してコマンド識別を行う。

これらの学習データに基づく音源分離処理により、高精度なコマンド識別が実現される。

#### 【図面の簡単な説明】

#### 【0024】

【図1】音声信号処理装置の構成例について説明する図である。

【図2】時間周波数分析部において実行する時間周波数分析処理について説明する図である。

【図3】1つの行列を2つの行列に分解する処理例について説明する図である。

【図4】図1に示す上段の学習処理部110において基底を学習した後、下段の解析処理

10

20

30

40

50

部 1 2 0 において学習された基底を結合して使用する例について説明する図である。

【発明を実施するための形態】

【0 0 2 5】

以下、図面を参照しながら本発明の音声信号処理装置、および音声信号処理方法、並びにプログラムの詳細について説明する。説明は以下の項目に従って行う。

1. 音声信号処理装置の全体構成について
2. 音声信号処理装置の各構成部の処理について
  2. 1. 時間周波数分析部について
  2. 2. 基底学習部について
  2. 3. 基底分解部について
  2. 4. コマンド識別部について

10

【0 0 2 6】

[ 1. 音声信号処理装置の全体構成について ]

まず、図 1 を参照して、この発明の実施の形態に係る音声信号処理装置の全体構成について説明する。

図 1 は、この発明の実施の形態に係る音声信号処理装置 1 0 0 の一例を示す図である。図 1 に示す音声信号処理装置 1 0 0 は、ユーザの言葉を入力して、ユーザの言葉から装置に対する要求である音声コマンドの認識処理を実行する装置である。

【0 0 2 7】

図 1 に示す音声信号処理装置は、音声コマンドの正負判定に必要な特徴量を予め算出する学習処理部 1 1 0 と、学習処理部 1 1 0 において学習された特徴量を用いて音源分離処理を行う解析処理部 1 2 0 を備えた構成である。学習処理部 1 1 0 において算出する音声コマンドの正負判定に必要な特徴量は、例えば解析処理部 1 2 0 における音声コマンド認識処理において実行する音声コマンド対応の目的音と音声コマンド非対応の非目的音との判別処理である正負判定処理に必要な特徴量である。

20

【0 0 2 8】

図 1 に示すように、音声信号処理装置 1 0 0 は、上段の学習処理部 1 1 0 と、下段の解析処理部 1 2 0 を有する。

上段の学習処理部 1 1 0 は、予め目的音声と非目的音声の特徴量空間における基底学習を行い、その学習結果を学習データとして解析処理部 1 2 0 に提供する。

30

【0 0 2 9】

解析処理部 1 2 0 は、学習処理部 1 1 0 から提供される目的音と非目的音の特徴量空間における基底学習結果を利用して、実際に解析対象となるユーザの声を含む音を入力し、入力音から目的とするユーザの声を分離し、分離結果に基づくコマンド識別処理を行う。

【0 0 3 0】

図 1 に示すように、学習処理部 1 1 0 は、時間周波数分析部 1 1 1 と、基底学習部 1 1 2 を有する。

また解析処理部 1 2 0 は、時間周波数分析部 1 2 1 と、基底分解部 1 2 2 と、コマンド識別部 1 2 3 を有する。

【0 0 3 1】

学習処理部 1 1 0 の処理と、解析処理部 1 2 0 の処理の概要について説明する。

学習処理部 1 1 0 は、目的音と非目的音からなる学習用音声信号 5 1 を入力し、この学習用音声信号 5 1 に対して、時間周波数分析部 1 1 1 において時間周波数分析を行う。さらに、基底学習部 1 1 2 が、時間周波数分析結果を用いて学習処理を実行し、学習結果として、目的音の基底周波数行列  $W_1$  の要素である基底周波数  $B_1(k, p)$  と、目的外音の基底周波数行列  $W_2$  の要素である基底周波数  $B_2(k, p)$  を生成する。これを学習データとして、解析処理部 1 2 0 に提供する。

【0 0 3 2】

解析処理部 1 2 0 は、抽出対象とするコマンドを含むユーザの声（＝目的音）やノイズ（＝非目的音）を含む入力音声信号 8 1 を入力する。時間周波数分析部 1 2 1 は、入力音

40

声信号 8 1 に対する時間周波数分析を実行し、分析結果を基底分解部 1 2 2 に提供する。

【0033】

基底分解部 1 2 2 は、時間周波数分析部 1 2 1 から入力する時間周波数分析結果と、学習処理部 1 1 0 の基底学習部 1 1 2 から入力する学習データ、すなわち、目的音と非目的音に対応する基底周波数データを適用して、基底分解を行い、基底アクティビティ  $H(p, 1)$  を得る。

【0034】

さらに、コマンド識別部 1 2 3 が、基底分解部 1 2 2 から供給された基底アクティビティ  $H(p, 1)$  に対する識別処理を行い、コマンド 8 2 を取得する。このコマンドが識別結果としてのコマンド 8 2 は、コマンドに基づくデータ処理を実行する次段のデータ処理部に提供されることになる。

10

以下、各構成部の処理の詳細について説明する。

【0035】

[2. 音声信号処理装置の各構成部の処理について]

(2. 1. 時間周波数分析部について)

図 1 に示すように、時間周波数分析部は、学習処理部 1 1 0 と解析処理部 1 2 0 の双方の処理部に設定されている。

図 1 に示す学習処理部 1 1 0 の時間周波数分析部 1 1 1 は、目的音と非目的音からなる学習用音声信号 5 1 を入力し、この学習用音声信号 5 1 に対して時間周波数分析を行う。

また、解析処理部 1 2 0 の時間周波数分析部 1 2 1 は、抽出対象とするコマンドを含むユーザの声 (= 目的音) や、コマンド抽出対象とならないユーザの声以外のノイズ (= 非目的音) を含む入力音声信号 8 1 に対して時間周波数分析を行う。

20

【0036】

なお、学習処理部 1 1 0 において学習対象として入力する学習用音声信号 5 1 は、解析処理部 1 2 0 の入力する音声信号と同様のユーザの声 (= 目的音) と、ユーザの声以外のノイズ (= 非目的音) を含む音声信号に設定することが好ましい。

【0037】

学習処理部 1 1 0 の時間周波数分析部 1 1 1 と、解析処理部 1 2 0 の時間周波数分析部 1 2 1 において実行する時間周波数分析処理について、図 2 を参照して説明する。

時間周波数分析部 1 1 1 と、時間周波数分析部 1 2 1 は、入力する音信号の時間周波数情報を分析する。

30

【0038】

マイク等を介して入力する入力信号を  $x$  とする。図 2 の最上段に入力信号  $x$  の例を示す。横軸が時間 (またはサンプル番号)、縦軸が振幅である。

入力信号  $x$  は、様々な音源からの音が混合した信号である。

学習処理部 1 1 0 の時間周波数分析部 1 1 1 に対する入力信号  $x$  は、目的音と非目的音からなる学習用音声信号 5 1 である。

また、解析処理部 1 2 0 の時間周波数分析部 1 2 1 に対する入力信号  $x$  は、抽出対象とするコマンドを含むユーザの声 (= 目的音) やノイズ (= 非目的音) を含む入力音声信号 8 1 である。

40

【0039】

まず、入力信号  $x$  から固定サイズのフレーム分割を行い、入力フレーム信号  $x(n, 1)$  を得る。

図 2 のステップ S 101 の処理である。

図 2 に示す例では、フレーム分割のサイズを  $N$  とし、各フレームのシフト量 ( $sf$ ) をフレームの大きさ  $N$  の 50% とし、各フレームを重複させた設定としている。

【0040】

さらに、入力フレーム信号  $x(n, 1)$  に対して、所定の窓関数:  $w$  を乗算して窓関数適用信号  $wx(n, 1)$  を得る。窓関数は例えばハミング窓が適用可能である。

窓関数適用信号  $wx(n, 1)$  は、以下の (式 1) によって示される。

50

【 0 0 4 1 】

【 数 1 】

$$wx(n, l) = w(n) * x(n, l) \quad \left( \begin{array}{l} x: \text{入力信号} \\ n: \text{時間インデックス} \\ l: \text{フレーム番号} \\ w: \text{窓関数} \\ wx: \text{窓関数適用信号} \end{array} \right)$$

10

$$w(n) = 0.54 - 0.46 * \cos\left(2\pi \frac{n}{N}\right) \quad (N: \text{フレームの大きさ})$$

．．．．．（式 1）

【 0 0 4 2 】

上記式（式 1）において、

x：入力信号、

n：時間インデックス、 $n = 0, \dots, N-1$ 、 $l = 0, \dots, L-1$   
(N はフレームの大きさ)l：フレーム番号、 $l = 0, \dots, L-1$   
(L は全フレーム数)

w：窓関数、

w x：窓関数適用信号、

である。

【 0 0 4 3 】

なお、窓関数：w は、ハミング窓の他、ハニング窓やブラックマンハリス窓などのその他の窓関数も使用可能である。

フレームの大きさ N は、例えば、0.02 sec 相当のサンプル数（ $N = \text{サンプリング周波数 } f_s * 0.02$ ）である。ただし、それ以外の大きさでも構わない。

また、図 2 に示す例では、フレームのシフト量（s f）はフレームの大きさ（N）の 50% とし、各フレームを重複した設定としているが、それ以外のシフト量でも構わない。

【 0 0 4 4 】

上記（式 1）に従って得られた窓関数適用信号 w x（n，l）に対して、以下に示す（式 2）に従って時間周波数分析を行い時間周波数スペクトル X（k，l）を得る。

【 0 0 4 5 】

20

30

【数 2】

$$X(k, l) = \sum_{n=0}^{M-1} wx(n, l) * \exp\left(-j2\pi \frac{k * n}{M}\right)$$

$\left( \begin{array}{l} wx: \text{窓関数適用信号} \\ j: \text{純虚数} \\ M: \text{DFTのポイント数} \\ k: \text{周波数インデックス} \\ X: \text{時間周波数スペクトル} \end{array} \right)$

10

$$wx(n, l) = \begin{cases} wx(n, l) & n = 0, \dots, N-1 \\ 0 & n = N, \dots, M-1 \end{cases}$$

・・・ (式 2)

【0046】

上記式 (式 2) において、

w x : 窓関数適用信号、

j : 純虚数、

M : D F T ( 離散フーリエ変換 ) のポイント数、

k : 周波数インデックス、

X : 時間周波数スペクトル、

である。

20

【0047】

窓関数適用信号 w x ( n , l ) に対する時間周波数分析処理としては、例えば、D F T ( 離散フーリエ変換 ) による周波数分析が適用される。その他、D C T ( 離散コサイン変換 ) や M D C T ( 修正離散コサイン変換 ) などの他の周波数分析を用いてもよい。また、必要であれば、D F T ( 離散フーリエ変換 ) のポイント数 M に合わせて適切に零詰めを行ってもよい。D F T のポイント数 M は、N 以上の 2 のべき乗の値としているが、それ以外

30

【0048】

次に、上記式 (式 2) に従って得られた時間周波数スペクトル X ( k , l ) から、以下に示す (式 3) に従って、パワースペクトル P X ( k , l ) を得る。

【0049】

【数 3】

$$PX(k, l) = X(k, l) * conj(X(k, l))$$

$\left( \begin{array}{l} X: \text{時間周波数スペクトル} \\ conj: \text{複素共役} \\ PX: \text{パワースペクトル} \end{array} \right)$

40

・・・ (式 3)

【0050】

上記式 (式 3) において、

X : 時間周波数スペクトル、

c o n j : 複素共役、

50

PX：パワースペクトル、  
である。

#### 【0051】

この処理は、図2に示すステップS102の処理に対応する。

図1に示す学習処理部110の時間周波数分析部111に対する入力信号xは、目的音と非目的音からなる学習用音声信号51である。学習処理部110の時間周波数分析部111は、この目的音と非目的音からなる学習用音声信号51に対する時間周波数分析結果として得られるパワースペクトル $PX(k, 1)$ を基底学習部112へ供給する。

#### 【0052】

また、解析処理部120の時間周波数分析部121に対する入力信号xは、抽出対象とするコマンドを含むユーザの声（＝目的音）やノイズ（＝非目的音）を含む入力音声信号81である。解析処理部120の時間周波数分析部121は、この抽出対象とするコマンドを含むユーザの声（＝目的音）やノイズ（＝非目的音）を含む入力音声信号81に対する時間周波数分析結果として得られるパワースペクトル $PX(k, 1)$ を基底分解部122へ供給する。

10

#### 【0053】

図2に示すステップS103には、フレーム毎に算出したパワースペクトル $PX(k, 1)$ を行列として表現した場合の行列の要素を示している。

周波数（周波数ビン）を行、

時間（フレーム）を列、

としたM行L列の行列として、行列の各要素を示したものである。

20

#### 【0054】

（2.2. 基底学習部について）

上述したように、図1に示す学習処理部110の時間周波数分析部111に対する入力信号xは、目的音と非目的音からなる学習用音声信号51である。学習処理部110の時間周波数分析部111は、この目的音と非目的音からなる学習用音声信号51に対する時間周波数分析結果として得られるパワースペクトル $PX(k, 1)$ を学習データとして基底学習部112へ供給する。

#### 【0055】

基底学習部112では、時間周波数分析部111より供給されたパワースペクトル $PX(k, 1)$ を、M行L列の行列として捉え、それを新たな二つの行列に分解する。

30

#### 【0056】

M行L列の行列とは、図2に示すステップS103に示す行列である。

基底学習部112は、このM行L列の行列形式のパワースペクトル $PX(k, 1)$ を新たな二つの行列に分解する。

この行列分解には、例えば、NMF（非負値行列分解）を適用する。

分解数をPとして、基底数P個の基底周波数 $B(k, p)$ と、その各々に対応する基底数P個の基底アクティビティ $H(p, 1)$ を得る。

なお、pは基底インデックスを表し、 $p = 0, \dots, P-1$ である。

#### 【0057】

本実施例の場合、

基底周波数 $B(k, p)$ は、入力信号の時間周波数情報を表すパワースペクトル $PX(k, 1)$ の周波数方向の性質を表し、

基底アクティビティ $H(p, 1)$ は時間方向の性質を表している。

40

#### 【0058】

入力信号xに対する分解数をPとし、以下の（式4）によって定義される誤差関数Eを最小化することで、基底周波数 $B(k, p)$ と基底アクティビティ $H(p, 1)$ を得る。

#### 【0059】

【数 4】

$$PX(k,l) = X(k,l) * conj(X(k,l)) \quad \left( \begin{array}{l} X: \text{時間周波数スペクトル} \\ conj: \text{複素共役} \\ PX: \text{パワースペクトル} \end{array} \right)$$

・・・ (式 4)

【0060】

10

上記式 (式 4) において、

E: 誤差関数

V: パワースペクトル行列、

W: 基底周波数行列、

H: 基底アクティビティ行列、

である。

【0061】

なお、パワースペクトル  $PX(k, l)$  は、図 2 の (S103) に示すような K 行 L 列の行列 V に対応する。

基底周波数  $B(k, p)$  は K 行 P 列の行列 W、

20

基底アクティビティ  $H(p, l)$  は P 行 L 列の行列 H、

で表す。

【0062】

1 つの行列を 2 つの行列に分解する処理について図 3 を参照して説明する。

図 3 に示す例は、

パワースペクトル  $PX(k, l)$  を示す M 行 L 列の 1 つの行列  $V_{201}$  を、

基底周波数  $B(k, p)$  を示す M 行 P 列の行列  $W_{202}$  と、

基底アクティビティ  $H(p, l)$  を P 行 L 列の行列  $H_{203}$ 、

これらの 2 つの行列に分解した例を示している。

【0063】

30

上記 (式 4) で示される誤差関数 E を勾配法により最小化することにより以下の (式 5) に示す更新式を得る。

【0064】

【数 5】

$$W_{kp} \leftarrow W_{kp} \frac{(V * H^T)_{kp}}{(W * H * H^T)_{kp}}$$

$$H_{pl} \leftarrow H_{pl} \frac{(W^T * V)_{pl}}{(W^T * W * H)_{pl}}$$

40

・・・ (式 5)

50

## 【0065】

なお、上記（式4）で示される誤差関数Eを勾配法により最小化する場合、例えば予測結果と観測結果との差の算出にユークリッド距離を用いる。その他、KLダイバージェンスやその他の距離なども利用可能である。

## 【0066】

基底学習部112は、上記処理によって得られた基底周波数行列Wの要素である基底周波数B(k, p)を、解析処理部120の基底分解部122へ供給する。

## 【0067】

すなわち、図1に示す学習処理部110では、まず、時間周波数分析部111が、目的音と非目的音からなる学習用音声信号51に対する時間周波数分析を実行し、時間周波数分析結果としてパワースペクトルPX(k, l)を生成する。

次に、基底学習部112が、目的音と非目的音からなる学習用音声信号51に対する時間周波数分析結果であるパワースペクトルPX(k, l)に基づいて、上記（式5）に示す更新式によって基底周波数行列Wの要素である基底周波数B(k, p)を算出し、算出した基底周波数B(k, p)を解析処理部120の基底分解部122へ供給する。

## 【0068】

なお、基底学習部112が算出する基底周波数B(k, p)は、

（1）目的音の基底周波数行列W1の要素である基底周波数B1(k, p)

（2）非目的音の基底周波数行列W2の要素である基底周波数B2(k, p)

である。

## 【0069】

このように、図1に示す学習処理部110は、目的音の基底周波数行列W1の要素である基底周波数B1(k, p)と、非目的音の基底周波数行列W2の要素である基底周波数B2(k, p)を、学習用音声信号51に基づく学習データとして生成し、解析処理部120に提供する。

## 【0070】

ただし、基底数Pの値は、それぞれの音源で同じである必要はなく、適宜変化させてよい。図4は、図1に示す上段の学習処理部110において基底を学習した後、下段の解析処理部120において学習された基底が結合されて使用される概念を説明する図である。

## 【0071】

図4に示す例は、

（1）目的音について、

パワースペクトルPXを示す1つの行列V<sub>1, 311</sub>を、

基底周波数B(k, p)を示す行列W<sub>1, 312</sub>と、

基底アクティビティH(p, l)を示す行列H<sub>1, 313</sub>、

これらの2つの行列への分解例。

## 【0072】

（2）非目的音について、

パワースペクトルPXを示す1つの行列V<sub>2, 321</sub>を、

基底周波数B(k, p)を示す行列W<sub>2, 322</sub>と、

基底アクティビティH(p, l)を示す行列H<sub>2, 323</sub>、

これらの2つの行列への分解例。

## 【0073】

（3）目的音と非目的音の混合信号について、

パワースペクトルPXを示す1つの行列V<sub>3, 331</sub>を、

基底周波数B(k, p)を示す行列W<sub>3, 332</sub>と、

基底アクティビティH(p, l)を示す行列H<sub>3, 333</sub>、

これらの2つの行列への分解例。

これらを示している。

## 【0074】

図 1 に示す上段の学習処理部 110 における基底学習により、図 4 (1), (2) のデータが生成される。

下段の解析処理部 120 の基底分解部 122 は、これら図 4 (1), (2) のデータを適用して、図 4 (3) に示す目的音と非目的音の混合信号から得られるパワースペクトル  $PX$  を示す 1 つの行列  $V_{3, 331}$  から得られる基底周波数  $B(k, p)$  を示す行列  $W_{3, 332}$  と、基底アクティビティ  $H(p, l)$  を示す行列  $H_{3, 333}$  について、目的音対応の行列 (a) と、非目的音対応の行列 (b) とに分離することを行うものである。

【0075】

(2. 3. 基底分解部について)

10

次に、図 1 に示す解析処理部 120 の基底分解部 122 の処理について説明する。

基底分解部 122 は、前段の時間周波数分析部 121 において、入力信号 81 に対する時間周波数分析によって生成されたパワースペクトル  $PX(k, l)$  を入力する。

さらに、基底分解部 122 は、学習処理部 110 の基底学習部 112 から、学習済みの様々な音源の基底周波数  $B(k, p)$  を入力する。

【0076】

基底分解部 122 は、学習処理部 110 の基底学習部 112 から、学習済みの様々な音源の個別の基底周波数  $B(k, p)$  に基づいて、これらを結合させた全基底周波数  $B_{all}(k, p)$  を生成する。

この処理は、図 4 に示す (3) の処理に相当する。

20

基底分解部 122 は、個別の基底周波数  $B(k, p)$  を結合させた全基底周波数  $B_{all}(k, p)$  を用いて基底分解を行い、基底アクティビティ  $H(p, l)$  を得る。ただし、 $p = 0, \dots, P_{all} - 1$  である。 $P_{all}$  は様々な音源毎に定められた基底数  $P$  の数の総和をとったものである。

【0077】

なお、パワースペクトル  $PX(k, l)$  を  $K$  行  $L$  列の行列  $V$  で表し、全基底周波数  $B_{all}(k, p)$  を  $K$  行  $P_{all}$  列の行列  $W_{all}$  で表し、基底アクティビティ  $H(p, l)$  を  $P_{all}$  行  $L$  列の行列  $H$  で表す。

図 4 に示すように、全基底周波数  $B_{all}(k, p)$  は学習処理部 110 において既に学習されているため、勾配法による更新は行わず、基底アクティビティ  $H(p, l)$  のみを更新する。

30

基底アクティビティ  $H(p, l)$  の更新処理は、以下の (式 6) に従って行われる。

【0078】

【数 6】

$$H_{pl} \leftarrow H_{pl} \frac{(W_{all}^T * V)_{pl}}{(W_{all}^T * W_{all}^T * H)_{pl}}$$

40

・・・・・・ (式 6)

【0079】

基底分解部 122 の算出した基底アクティビティ  $H(p, l)$  は、基底分解部 122 からコマンド識別部 123 へ供給される。

【0080】

(2. 4. コマンド識別部について)

次に、図 1 に示す解析処理部 120 のコマンド識別部 123 の処理について説明する。

コマンド識別部 123 では、基底分解部 122 より供給された基底アクティビティ  $H(p, l)$  に対する識別処理を行い、コマンド結果を得る。例えば、以下の (式 7) に従っ

50

て、閾値比較を実行してコマンド結果を得る。

【0081】

【数7】

$$\prod_{p=0}^{P_{all}-1} (H(p,l) \geq Thre(p,l)) = \begin{cases} 1, \text{コマンド} \\ 0, \text{非コマンド} \end{cases} \quad (Thre: \text{基底毎の閾値})$$

・・・・・・（式7）

10

【0082】

なお、上記の（式7）は、予め設定した閾値との比較処理を行ってコマンドと非コマンドを判定する処理を行っているが、この方法に限らず、例えば一般化線形判別など活性化関数を伴う非線形な識別を行ってもよい。また、上記（式7）では閾値処理の結果をAND演算しているがOR演算などその他の論理演算を適用してもよい。

コマンド識別部123は、上記の（式7）に従った判定処理により得られたコマンド情報を、図1に示すコマンド出力82とする。

【0083】

このコマンド出力82が、例えばコマンドに応じたデータ処理を実行するデータ処理部に入力され、コマンドに従って様々な処理が実行される。

20

【0084】

なお、上記の実施例では、図1に示す音声信号処理装置100として、学習処理部110と、解析処理部120の2つの処理部を有する構成例として説明したが、学習処理部110の学習結果として得られる学習データを予め記憶部に格納する構成としてもよい。すなわち、記憶部に格納された学習データを解析処理部120が必要に応じて取得して入力信号に対する処理を行う構成としてもよい。この構成の場合は、学習処理部を省略した解析処理部と学習結果としての学習データを格納した記憶部によって音声信号処理装置を構成することか可能である。

【0085】

以上、特定の実施例を参照しながら、本発明について詳解してきた。しかしながら、本発明の要旨を逸脱しない範囲で当業者が実施例の修正や代用を成し得ることは自明である。すなわち、例示という形態で本発明を開示してきたのであり、限定的に解釈されるべきではない。本発明の要旨を判断するためには、特許請求の範囲の欄を参酌すべきである。

30

【0086】

また、明細書中において説明した一連の処理はハードウェア、またはソフトウェア、あるいは両者の複合構成によって実行することが可能である。ソフトウェアによる処理を実行する場合は、処理シーケンスを記録したプログラムを、専用のハードウェアに組み込まれたコンピュータ内のメモリにインストールして実行させるか、あるいは、各種処理が実行可能な汎用コンピュータにプログラムをインストールして実行させることが可能である。例えば、プログラムは記録媒体に予め記録しておくことができる。記録媒体からコンピュータにインストールする他、LAN（Local Area Network）、インターネットといったネットワークを介してプログラムを受信し、内蔵するハードディスク等の記録媒体にインストールすることができる。

40

【0087】

なお、明細書に記載された各種の処理は、記載に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されてもよい。また、本明細書においてシステムとは、複数の装置の論理的集合構成であり、各構成の装置が同一筐体内にあるものには限らない。

【産業上の利用可能性】

【0088】

50

以上、説明したように、本発明の一実施例の構成によれば、複数の音が混在した音声信号から、特定音源のコマンドを高精度に分離する装置、方法が実現される。具体的には、例えば、複数音源からの音を含む学習用音声信号に基づいて各音源対応の基底周波数  $B$  からなる学習データを生成し、この各音源対応の基底周波数  $B$  を結合した全基底周波数  $B_{all}$  を生成する。さらに、入力音声信号に対して、時間周波数分析を実行して時間周波数分析結果を生成する。この入力音声信号に対する時間周波数分析結果に対して、全基底周波数  $B_{all}$  を適用した基底分解を行い、入力音声信号に対する基底アクティビティ  $H$  を生成する。最後に生成した基底アクティビティ  $H$  の識別処理を実行してコマンド識別を行う。

これらの学習データに基づく音源分離処理により、高精度なコマンド識別が実現される

10

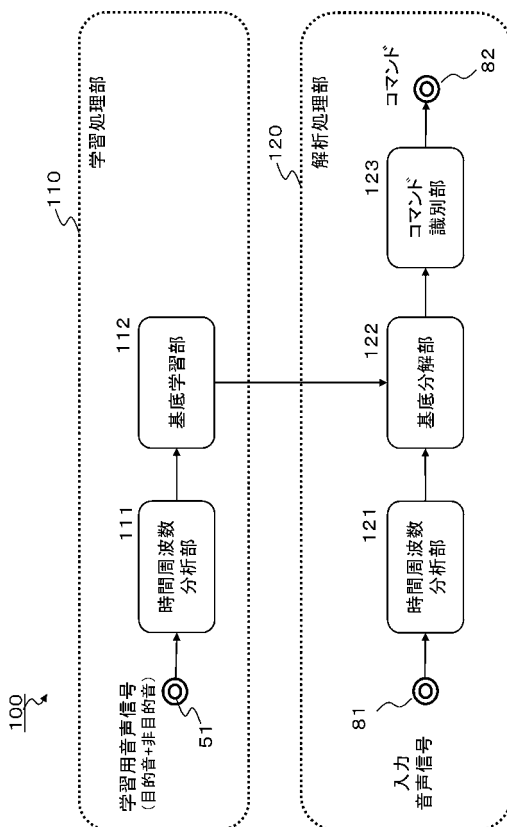
# 【符号の説明】

## 【0089】

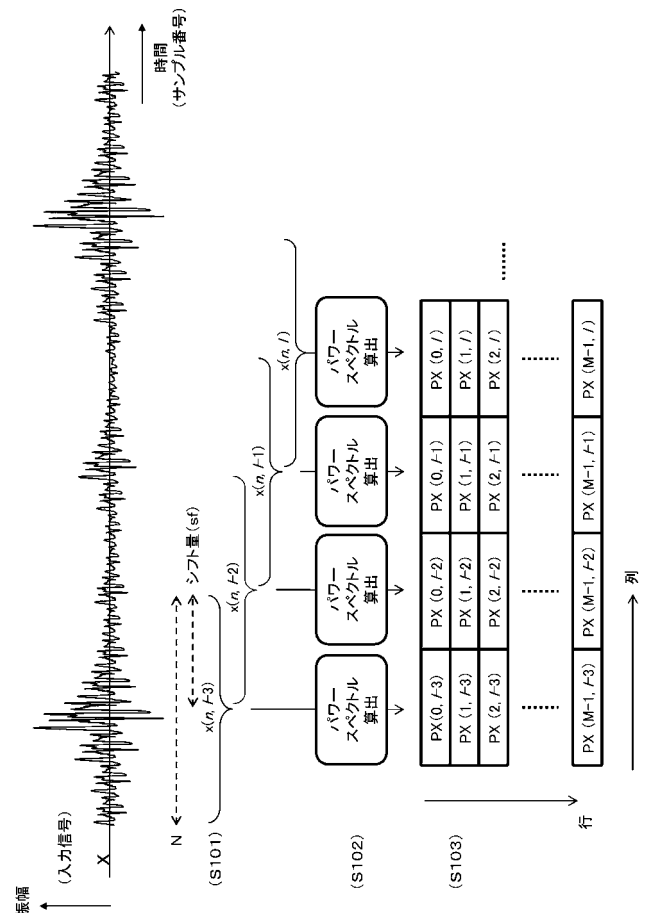
- 51 学習用音声信号
- 81 入力音声信号
- 82 出力音声信号
- 100 音声信号処理装置
- 110 学習処理部
- 111 時間周波数分析部
- 112 基底学習部
- 120 解析処理部
- 121 時間周波数分析部
- 122 基底分解部
- 123 コマンド識別部

20

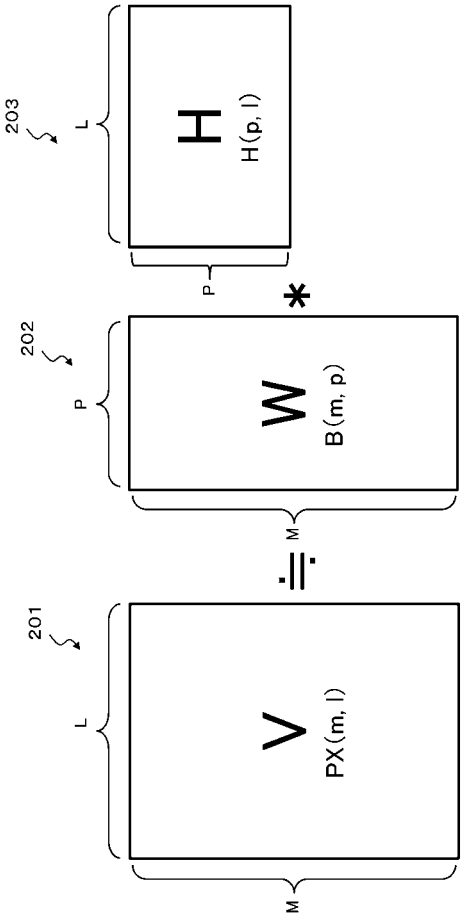
【図 1】



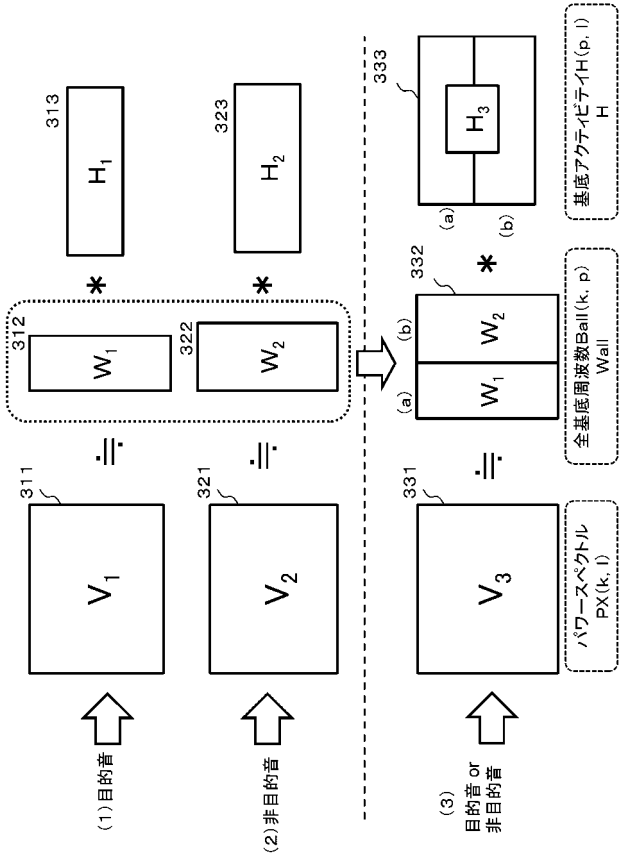
【図 2】



【図 3】



【図 4】



---

フロントページの続き

- (72)発明者 光藤 祐基  
東京都港区港南1丁目7番1号 ソニー株式会社内
- (72)発明者 西口 正之  
東京都港区港南1丁目7番1号 ソニー株式会社内
- Fターム(参考) 5D015 CC00 JJ00