(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2020-38315
(P2020-38315A)
(43)公開日　令和2年3月12日(2020.3.12)

(51)Int.Cl.		F I			テーマコード（参考）
G10L 21/0272	(2013.01)	G10L	21/0272	100B	5D220
G10L 21/028	(2013.01)	G10L	21/028	B	
H04R 3/00	(2006.01)	H04R	3/00	320	
G10L 25/51	(2013.01)	G10L	25/51	400	

審査請求 未請求　請求項の数 15　ＯＬ　（全 16 頁）

(21)出願番号　特願2018-165875 (P2018-165875)
(22)出願日　平成30年9月5日 (2018.9.5)

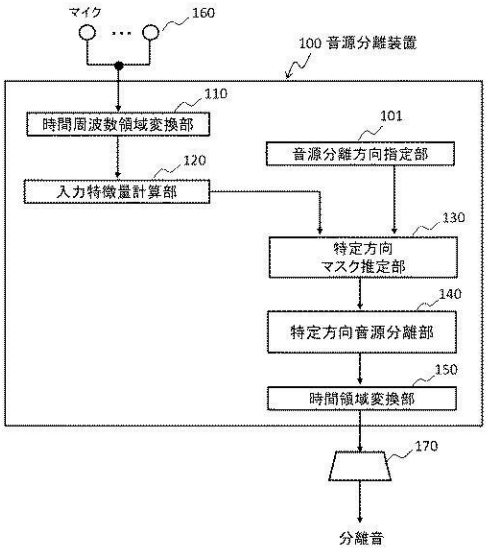
(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　110001689
青稜特許業務法人
(72)発明者　池下　林太郎
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内
(72)発明者　藤岡　拓也
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内
(72)発明者　堀口　翔太
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内

Ｆターム(参考)　5D220 BA06　BC05

(54)【発明の名称】音声情報処理装置および方法

(57)【要約】
【課題】時間周波数領域における音源のスパース性が成立しない場合においても，高性能に音源分離を行うことを目的とする。
【解決手段】
入力装置と、出力装置と、処理部を備える音声情報処理装置である。処理部は、入力装置から入力された音声信号を時間周波数領域に変換する、時間周波数領域変換部と、時間周波数領域から特徴量を計算する、特徴量計算部と、特徴量から特定方向の音を分離するための第１の時間周波数マスクを推定する、第１の時間周波数マスク推定部と、を備える。
【選択図】　図１

図1



【特許請求の範囲】
【請求項１】
入力装置と、出力装置と、処理部を備え、
前記処理部は、
前記入力装置から入力された音声信号を時間周波数領域に変換する、時間周波数領域変換部と、
前記時間周波数領域から特徴量を計算する、特徴量計算部と、
前記特徴量から特定方向の音を分離するための第１の時間周波数マスクを推定する、第１の時間周波数マスク推定部と、
を備える音声情報処理装置。
【請求項２】
前記第１の時間周波数マスク推定部に、１または複数方向を指定する音源分離方向指定部を備える、
請求項１記載の音声情報処理装置。
【請求項３】
時間周波数マスクの類似度、あるいは、時間周波数マスクで抽出された音声信号の類似度を判定する，時間周波数マスクマッチング部を備える，
請求項１記載の音声情報処理装置。
【請求項４】
前記特徴量から特定の音源を分離するための第２の時間周波数マスクを推定する、第２の時間周波数マスク推定部を備える、
請求項３記載の音声情報処理装置。
【請求項５】
前記時間周波数マスクマッチング部は、前記第１の時間周波数マスクと前記第２の時間周波数マスクの類似度を判定する，
請求項４記載の音声情報処理装置。
【請求項６】
前記第１の時間周波数マスク推定部は、異なる方向に対する複数の第１の時間周波数マスクを推定し、
前記第２の時間周波数マスク推定部は、異なる音源に対する複数の第２の時間周波数マスクを推定し、
前記時間周波数マスクマッチング部は、前記複数の前記第１の時間周波数マスクと複数の前記第２の時間周波数マスクの類似度をそれぞれ判定し、指定した方向θにおいて、θ方向に対する第１の時間周波数マスクと、それと最も類似度の高い第２の時間周波数マスクとの類似度を音源存在スコアとして出力する、
請求項５記載の音声情報処理装置。
【請求項７】
前記音源存在スコアが，予め定めた閾値以上であるとき，指定した方向θに音源が存在すると判定し，音源方向推定結果を出力する音源方向推定部を備える、
請求項６記載の音声情報処理装置。
【請求項８】
前記第１の時間周波数マスクを用いて音源分離フィルタを形成し，音源分離を実行する特定方向音源分離部を備え，
前記特定方向音源分離部は，前記音源方向推定部が、指定した方向θに音源が存在すると判定した場合，θ方向に対する第１の時間周波数マスクを用いて音源分離フィルタを形成し，音源分離を実行する，
請求項７記載の音声情報処理装置。
【請求項９】
前記特定方向音源分離部が出力する音源分離結果を，時間領域の分離信号に変換して出力する時間領域変換部を備える，

請求項８記載の音声情報処理装置。
【請求項１０】
前記第１の時間周波数マスク推定部は、異なる方向に対する複数の第１の時間周波数マスクを推定し、
前記第２の時間周波数マスク推定部は、異なる音源に対する複数の第２の時間周波数マスクを推定し、
前記第２の時間周波数マスク推定部が出力する複数の第２の時間周波数マスクを加算し、合成された第２の時間周波数マスクを出力する時間周波数マスク加算部を備え、
前記時間周波数マスクマッチング部は、前記複数の前記第１の時間周波数マスクと前記合成された第２の時間周波数マスクの類似度をそれぞれ判定し、前記類似度に基づいて所定の方向θに複数の音源が存在する可能性を示す複数音源存在スコアを出力する、
請求項５記載の音声情報処理装置。
【請求項１１】
移動機構と、
前記複数音源存在スコアが、予め定めた閾値以上であるとき、所定の方向θに音源が複数存在すると判定し、複数音源同一方向判定結果を出力する、複数音源同一方向判定部を備え、
前記移動機構は、前記複数音源同一方向判定結果が出力された場合に、各音源の方向が別れる位置に移動する、
請求項１０記載の音声情報処理装置。
【請求項１２】
前記第１の時間周波数マスク推定部から出力される、異なる複数の時間における所定の方向θに対する第１の時間周波数マスクを記憶する特定方向マスク貯蔵部を備え、
前記時間周波数マスクマッチング部は、前記特定方向マスク貯蔵部に記憶された前記第１の時間周波数マスク同士の類似度を判定する、
請求項３記載の音声情報処理装置。
【請求項１３】
音声合成機能と、
時間周波数マスクマッチング部が出力する前記類似度が、予め定めた閾値以上であるとき、所定の方向θから同一の発話が２度あったと判定し、同一発話判定結果を出力する同一発話判定部を備え、
前記音声合成機能は、前記同一発話判定結果が出力された場合に、予め定められた内容の発話を行なう、
請求項１２記載の音声情報処理装置。
【請求項１４】
入力装置と、出力装置と、処理部を備える情報処理装置を用いた音声情報処理方法であって、
前記処理部が、前記入力装置から入力された音声信号を時間周波数領域に変換する、時間周波数領域変換処理、
前記処理部が、前記時間周波数領域から特徴量を計算する、特徴量計算処理、
前記処理部が、前記特徴量から特定方向の音を分離するための第１の時間周波数マスクを推定する、第１の時間周波数マスク推定処理、
前記処理部が、前記特徴量から特定の音源を分離するための第２の時間周波数マスクを推定する、第２の時間周波数マスク推定処理、
前記処理部が、前記第１の時間周波数マスクと前記第２の時間周波数マスクのマッチングを行なう周波数マスクマッチング処理、
を行なう音声情報処理方法。
【請求項１５】
前記処理部が、複数の前記第２の時間周波数マスクの合成を行なって、合成時間周波数マスクを出力するマスク加算処理を行ない、

前記周波数マスクマッチング処理では、前記第１の時間周波数マスクと前記合成時間周波数マスクのマッチングを行なう、
請求項１４記載の音声情報処理方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は，音声情報処理技術に関し、音源分離あるいは音源方向推定にかかる技術に関するものである。
【背景技術】
【０００２】
音源分離技術とは，複数音源が混合した観測信号から，混合前の個々の源信号を推定する技術のことである。近年，音声信号の時間周波数スペクトルの特徴を事前に学習しておくことで音源分離を達成する，教師あり機械学習に基づく音源分離手法の研究が盛んに行われている。
【０００３】
非特許文献１には，「マイクロホンで観測した観測信号を時間周波数領域の信号に変換し，観測信号の各時間周波数ビンをニューラルネットワークを用いて高次元空間のベクトルに写像し，写像されたベクトルをクラスタリングすることによって各時間周波数ビンを音源毎のクラスターに分類し，音源分離を達成する」技術が開示されている。
【先行技術文献】
【非特許文献】
【０００４】
【非特許文献１】J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep Clustering: Discriminative embeddings for segmentation and separation," IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 31--35.
【発明の概要】
【発明が解決しようとする課題】
【０００５】
非特許文献１の音源分離方法は，時間周波数ビン（bin）をクラスタリングすることで，音源分離を達成できる。しかしながら，時間周波数ビンのクラスタリングに基づく音源分離方法では，空間に存在する音源の数が多いなどといった，時間周波数領域における音源のスパース性（各時間周波数ビンにおいて原信号のうち１つだけが支配的であること）が成立しない場合において，音源分離性能が低いという問題があった。
【課題を解決するための手段】
【０００６】
本発明の好ましい一側面は、入力装置と、出力装置と、処理部を備える音声情報処理装置である。処理部は、入力装置から入力された音声信号を時間周波数領域に変換する、時間周波数領域変換部と、時間周波数領域から特徴量を計算する、特徴量計算部と、特徴量から特定方向の音を分離するための第１の時間周波数マスクを推定する、第１の時間周波数マスク推定部と、を備える。
【０００７】
好ましい具体例では、時間周波数マスクの類似度、あるいは、時間周波数マスクで抽出された音声信号の類似度を判定する，時間周波数マスクマッチング部を備える。
【０００８】
別の好ましい具体例では、特徴量から特定の音源を分離するための第２の時間周波数マスクを推定する、第２の時間周波数マスク推定部を備える。
【０００９】
さらに好ましい具体例では、時間周波数マスクマッチング部は、第１の時間周波数マスクと第２の時間周波数マスクの類似度を判定する。
【００１０】

他のさらに好ましい具体例では、時間周波数マスクマッチング部は、異なる時間区間における第１の時間周波数マスクの類似度を判定する。

【００１１】

本発明の他の好ましい一側面は、入力装置と、出力装置と、処理部を備える情報処理装置を用いた音声情報処理方法である。処理部が、入力装置から入力された音声信号を時間周波数領域に変換する、時間周波数領域変換処理、時間周波数領域から特徴量を計算する、特徴量計算処理、特徴量から特定方向の音を分離するための第１の時間周波数マスクを推定する、第１の時間周波数マスク推定処理、特徴量から特定の音源を分離するための第２の時間周波数マスクを推定する、第２の時間周波数マスク推定処理、第１の時間周波数マスクと第２の時間周波数マスクのマッチングを行なう周波数マスクマッチング処理、の各処理を行なう。

【００１２】

他の具体的な例では、第１、第２の時間周波数マスク推定部は、教師有り学習で学習されたニューラルネットワークで構成される。

【発明の効果】

【００１３】

本発明によれば、時間周波数領域におけるスパース性が成立しない音源に対しても、高い分離性能を有する音源分離方法を提供できる。

【図面の簡単な説明】

【００１４】

【図１】本発明の第１実施形態に関わるブロック図。
【図２】本発明の第１実施形態に関わるフローチャート。
【図３】本発明の第２実施形態に関わるブロック図。
【図４】本発明の第２実施形態に関わるフローチャート。
【図５】本発明の実施形態に関わる時間周波数マスクの概念図。
【図６】本発明の第３実施形態に関わるブロック図。
【図７】本発明の第３実施形態に関わるフローチャート。
【図８】本発明の第３実施形態に関わる時間周波数マスクの概念図。
【図９】本発明の第４実施形態に関わるブロック図。
【図１０】本発明の第４実施形態に関わるフローチャート。

【発明を実施するための形態】

【００１５】

実施の形態について、図面を用いて詳細に説明する。ただし、本発明は以下に示す実施の形態の記載内容に限定して解釈されるものではない。本発明の思想ないし趣旨から逸脱しない範囲で、その具体的構成を変更し得ることは当業者であれば容易に理解される。

【００１６】

以下に説明する発明の構成において、同一部分又は同様な機能を有する部分には同一の符号を異なる図面間で共通して用い、重複する説明は省略することがある。

【００１７】

同一あるいは同様な機能を有する要素が複数ある場合には、同一の符号に異なる添字を付して説明する場合がある。ただし、複数の要素を区別する必要がない場合には、添字を省略して説明する場合がある。

【００１８】

本明細書等における「第１」、「第２」、「第３」などの表記は、構成要素を識別するために付するものであり、必ずしも、数、順序、もしくはその内容を限定するものではない。また、構成要素の識別のための番号は文脈毎に用いられ、一つの文脈で用いた番号が、他の文脈で必ずしも同一の構成を示すとは限らない。また、ある番号で識別された構成要素が、他の番号で識別された構成要素の機能を兼ねることを妨げるものではない。

【００１９】

図面等において示す各構成の位置、大きさ、形状、範囲などは、発明の理解を容易にす

るため、実際の位置、大きさ、形状、範囲などを表していない場合がある。このため、本発明は、必ずしも、図面等に開示された位置、大きさ、形状、範囲などに限定されない。

【００２０】

本明細書で引用した刊行物、特許および特許出願は、そのまま本明細書の説明の一部を構成する。

【００２１】

本明細書において単数形で表される構成要素は、特段文脈で明らかに示されない限り、複数形を含むものとする。

【００２２】

以下で詳細に説明される実施例の概要を説明する。実施例に関わる代表的な音源分離方法ないし装置は，マイクロホンアレイで観測された複数音源の混合信号から，指定した方向に対する音源分離結果を与える時間周波数マスクを計算するニューラルネットワークを用いる。ここで、時間周波数マスクとは，分離対象とする信号が支配的である時間周波数ビンを推定し，その時間周波数成分をマスキング処理により分離するマスクである（明細書・図面等において、「時間周波数マスク」を単に「マスク」ということがある）。

【００２３】

より具体的な例では、指定した方向に対する音源分離結果を与える第１の時間周波数マスクと、空間の複数の音源（便宜上「全音源」と称することがあるが、必ずしも全ての音源を含む必要はない）に対する音源分離結果を与える第２の時間周波数マスクとを計算し，２つの時間周波数マスクを用いることにより，精度のよい音源分離を行なう。より具体的な例では，計算された２つの時間周波数マスクをマッチングすることによって，所定の方向に音源が存在するか否かを判定する。

【００２４】

例えば第１の時間周波数マスクは，（マイクから見た）方向Ｘ，方向Ｙ，方向Ｚを識別するマスクであり，これは既知の方向Ｘ，方向Ｙ，方向Ｚからの音声を教師データとして，ニューラルネットワークを学習することで生成することができる。また，第２の時間周波数マスクは，音源Ａ，音源Ｂ，音源Ｃを識別するマスクであり，これは既知の音源Ａ，音源Ｂ，音源Ｃからの音声を教師データとして，ニューラルネットワークを学習することで生成することができる。

【実施例１】

【００２５】

図１と図２を用いて，本発明の第一実施形態に関わる音源分離装置１００を説明する。本実施例の音源分離装置１００は、入力装置、出力装置、記憶装置、および処理装置を含む、通常のコンピュータ（例えばサーバ）で構成した。本実施例では計算や制御等の機能は、記憶装置に格納されたプログラムが処理装置によって実行されることで、定められた処理を他のハードウェアと協働して実現される。計算機などが実行するプログラム、その機能、あるいはその機能を実現する手段を、「～部」等と呼ぶ場合がある。また記憶装置と処理装置を総称して「処理部」と呼ぶ場合がある。実施例２以降で説明される音源分離装置も同様である。

【００２６】

また、音源分離装置１００の構成は、単体のコンピュータで構成してもよいし、あるいは、入力装置、出力装置、処理装置、記憶装置の任意の部分が、ネットワークで接続された他のコンピュータで構成されてもよい。また、本実施例中、プログラムで構成した機能と同等の機能は、FPGA（Field Programmable Gate Array）、ASIC（Application Specific Integrated Circuit）などのハードウェアでも実現できる。

【００２７】

図１は第１実施形態に関わる音源分離装置１００のブロック図である。音源分離装置１００は，音源分離方向指定部１０１と，時間周波数領域変換部１１０と，入力特徴量計算部１２０と，特定方向時間周波数マスク推定部１３０と，特定方向音源分離部１４０と，時間領域変換部１５０から構成される。これらは、既述のように、処理装置によって実行

されるプログラムとして実装されている。また、入力装置として例えばマイクロホンアレイ１６０を、分離音の出力装置として例えばスピーカ１７０を備える。

【００２８】

図２も参照しつつ，音源分離装置１００の動作例を説明する。時間周波数領域変換部１１０は，公知の短時間フーリエ変換などにより，マイクロホンアレイ１６０を用いて観測した混合信号（例えば横軸を時間、縦軸を振幅（音圧）で表現される）の時間周波数表現を計算して出力する（ステップS２０１）。公知のように、時間周波数表現としては、例えば横軸を時間、縦軸を周波数とし、周波数の強度をコントラストで示すものがある。本実施例では、時間周波数表現がマイクロホンアレイ１６０のマイクの数得られることになる。

【００２９】

入力特徴量計算部１２０は，時間周波数領域変換部１１０が出力した観測信号の時間周波数表現から音源分離装置が用いる特徴量を計算し，特定方向時間周波数マスク推定部１３０に出力する（ステップS２０２）。特徴量としては，例えば，観測信号の振幅スペクトルとマイクロホン間の位相差を束ねたものを用いることができる。所定の音声について、入力特徴量計算部１２０からはマイクロホンアレイ１６０のマイクの数だけ、観測信号から得た振幅スペクトルと位相差の情報の組からなる情報が出力され、これらを纏めて当該音声の特徴量とする。一般には、マイクロホンアレイ１６０からのアナログ信号をデジタル信号に変換して後続の処理を行なう。この場合、周波数や時間等の量子化の粒度は任意である。

【００３０】

位相差の情報については、マイクロホンアレイの各マイクの位置が異なることにより、各マイクに音波が到達する時刻が異なるので、得られる観測信号に位相差が生じる。位相差の情報から方向情報を得ることができる。位相差の情報は音源の方向を特定するために重要である。

【００３１】

音源分離方向指定部１０１は，音声端末デバイスから見た，音源分離したい方向を指定する。具体的にはマイクロホンアレイ１６０から見た方向である（ステップS２０３）。特定方向時間周波数マスク推定部１３０は，入力特徴量計算部１２０が出力する観測信号の特徴量を用いて，音源分離方向指定部１０１が指定するデバイスから見た音源分離したい方向θに対する時間周波数マスクを推定する（ステップS２０４）。

【００３２】

ステップS２０４における指定した方向θに対する時間周波数マスクの推定には，例えばニューラルネットワークを用いることができる。ニューラルネットワークは，観測信号の特徴量を入力したとき、指定した方向θに対する音源を抽出する時間周波数マスクを出力する。ここで、指定した方向θに対する音源を抽出するとは、物理的な音源そのものを特定するのではなく、指定した方向θから到来する音を特定することを意味する。具体例としては、特定方向時間周波数マスク推定部１３０は学習済みのディープニューラルネットワーク（ＤＮＮ）で構成することができる。ＤＮＮは、方向θが既知の音源からの観測信号を教師データとして用いて、θ方向からの音を選択的に抽出する時間周波数マスクを出力とするように教師有り学習をしておく。

【００３３】

特定方向音源分離部１４０は，特定方向時間周波数マスク推定部１３０が出力する，指定した方向θに対する時間周波数マスクを用いて，指定した方向θに対する音源分離フィルタを形成し，音源分離を実行する（ステップS２０５）。音源分離フィルタは、例えば各時間周波数ビンに対して１または０の重みを付けるものでもよいし、１～０の間で離散的な値を取るものでも良い。

【００３４】

なお、特定方向時間周波数マスク推定部１３０が出力する時間周波数マスクをそのまま用いるのではなく、目的にかなったフィルタとなるように，信号処理技術を用いて再形成

（修正，調整）し，再形成したフィルタを用いて，特定方向からの音を抜き出す処理を行ってもよい。目的にかなったフィルタとは，例えば音声を歪ませないようにマスキングを調整したり、不要な周波数成分を除去するなど、ユーザが任意に設定することができる。

【００３５】

時間領域変換部１５０は，特定方向音源分離部１４０が出力する，指定した方向θに対する音源分離結果を，時間領域の分離信号に変換して出力する。出力は、例えばスピーカ１７０から音声信号として出力される。

【００３６】

上記の特定方向に対する音源分離処理は，音源分離方向指定部１０１が指定する方向として，様々な方向を指定することで，任意方向に対して実行できる。例えば，音源分離方向指定部１０１が指定する方向θとして，予め数式１

【００３７】

【数１】

$$\theta \in \left\{ \frac{2k\pi}{36} \,\middle|\, k = 0, 1, \ldots, 35 \right\}$$

【００３８】

のように全方位を１０度刻みに指定しておくことができる。このために、本実施例では、ＤＮＮに各方向θに対して専用の層を準備しておき、ステップS２０３で選択するように構成される。数式１の例に従うと、ｋごとに３６個の層を準備する。

【００３９】

他の実装方法としては、全方向のθに対して共通の層を準備しておき、補助ネットワークで層の重みパラメータを変化させることにより、指定した方向θに対応した層に変換することも可能である。

【００４０】

上記の実施例では、特定方向時間周波数マスク推定部１３０に入力する特徴量として，観測信号の振幅スペクトルとマイクロホン間の位相差の両方を用いた。他の手法としては，観測信号の振幅スペクトルを用いず、マイクロホン間の位相差のみを特徴量として用いても良い。このようにすると，音の到来方向の情報だけを抽出できるので，音源の種類による影響を小さくすることができる。

【００４１】

時間周波数スペクトルを特徴量に含める場合には，音源の種類が支配的にならないように，様々な音源の種類と、様々な方向の組み合わせを，十分多くの数用意した学習データを用いて，教師有り学習を行なうことが望ましい。この方法により，音源の種類により方向の推定結果が影響されることを緩和することができる。

【００４２】

一方，時間周波数スペクトルを用いる場合の利点として以下が考えられる。例えば，マイクロホンアレイの性質によっては，「右方向の音源からの音がマイク１に小さく入り，マイク２に大きく入る」といったことが起こり得る。位相差の特徴量のみでは、このような音源方向によって各マイクに届く音の大きさが異なるという情報を利用することができない。

【００４３】

実施例１の構成によれば、スパース性が不成立の場合でも、指定した方向θに対しては高い分離性能が得られる。

【実施例２】

【００４４】

図３～図５を用いて，本発明の第二実施形態に関わる音源分離装置３００を説明する。

第二実施形態は，例えば会議室等において，複数の発言者を識別するとともにその方向（あるいは位置）を特定するのに好適である。また，本実施例は２種類の時間周波数マスクをマッチングすることにより，高精度の音源分離、音源方向推定が可能となる。

【００４５】

第二実施例の音源分離装置３００は，図３における全音源時間周波数マスク推定部３１０と，時間周波数マスクマッチング部３２０と，音源方向推定部３３０が加わることを除けば，図１に示した第一実施形態の音源分離装置１００と同じ構成であるので，以下では，全音源時間周波数マスク推定部３１０と，時間周波数マスクマッチング部３２０と，音源方向推定部３３０についてのみ説明し，他の説明を省略する。

【００４６】

図４に示した第二実施形態の処理フローも，ステップS４０１と，ステップS４０２と，ステップS４０３が加わることを除けば，図２に示した第一実施形態の処理フローと同一であるため，以下ではステップS４０１と，ステップS４０２と，ステップS４０３についてのみ説明し，他の説明を省略する。

【００４７】

全音源時間周波数マスク推定部３１０は，入力特徴量計算部１２０が出力する観測信号の特徴量を用いて，空間に存在する全音源に対して，各音源の分離結果に対応する，時間周波数マスクを推定する（ステップS４０１）。ステップS４０１における全音源に対する時間周波数マスクの推定には，例えば，非特許文献１で提案された深層クラスタリングに基づく時間周波数マスクの推定方法を用いることができる。全音源時間周波数マスク推定部３１０は、たとえばＤＮＮで構成することができる。ＤＮＮは、既知の音源からの観測信号を教師データとして、公知の教師有り学習で学習させることができる。

【００４８】

本実施例では，全音源時間周波数マスク推定部３１０に入力する特徴量は，特定方向時間周波数マスク推定部１３０に入力する特徴量と同じ特徴量をとし，観測信号の振幅スペクトルとマイクロホン間の位相差の両方を用いた特徴量とした。ただし，別々の特徴量を用いても良い。また，全音源時間周波数マスク推定部３１０へ入力する特徴量は，マイクロホン間の位相差を省略してもよい。

【００４９】

時間周波数マスクマッチング部３２０は，全音源時間周波数マスク推定部３１０が出力する全音源に対する時間周波数マスクの推定値と，特定方向時間周波数マスク推定部１３０が出力する指定した方向θに対する時間周波数マスクの推定値を入力として受け取り，各音源nの時間周波数マスクの推定値と指定した方向θに対する時間周波数マスクの推定値の類似度を計算する（ステップS４０２）。時間周波数マスクの類似度d_1としては，例えば数式２

【００５０】

【数２】

$$d_1(y_n, y_\theta) = -||y_n - y_\theta||_F^2$$

【００５１】

を用いればよい。ここで，y_nは全音源時間周波数マスク推定部３１０が出力する音源nの時間周波数マスクの推定値，y_θは特定方向時間周波数マスク推定部１３０が出力する方向θに関する時間周波数マスクの推定値，右辺の「$||\cdot||_F$」は行列のフロベニウスノルムである。

【００５２】

ここで類似度d_1は、特定の音源ｎに対する時間周波数マスクと特定の方向θに対する

時間周波数マスクの類似度を示しており、これが類似（あるいは一致）した場合には、特定の音源が特定の方向に存在する可能性が高い。前述のように特定の音源に対する時間周波数マスクは、スパース性が不成立の場合には精度が低くなる可能性はあるが、おおよその音源分離は可能である。一方、特定の方向に対する時間周波数マスクは、定めた方向に対する分離性能は高いが、その方向に音源があるかは不確かである。すなわち、その方向にはノイズ源や壁などの反射物がある可能性もある。よって、両方の時間周波数マスクを利用することで、音源を特定する精度を向上させることが期待できる。

【００５３】

図５に、特定の音源ｎに対する時間周波数マスクと特定の方向θに対する時間周波数マスクの類似度を概念的に示す。単純化するため，マスクは各ビンに対して０（白）と１（黒）の２値を取るものとしている。図５には、音源１～音源４に対する時間周波数マスクと、方向１～方向４に対する時間周波数マスクが概念的に示されている。これらの類似度を計算することにより、例えば音源２のマスクと方向２のマスクがマッチすることが分かる。この場合、音源２が方向２にあり、方向２からの音は音源２の音である可能性が高い。

【００５４】

上記のような理論に基づいて，さらに，時間周波数マスクマッチング部３２０は，指定した方向θに音源が存在するか否かを表す，音源存在スコアを計算し出力する（ステップS４０２）。音源存在スコアとしては，例えば，数式３。

【００５５】

【数３】

$$\ell_1(\theta) = \max_n d_1(y_n, y_\theta)$$

を用いれば良い

【００５６】

数式３の音源存在スコアl_1は、ある方向θにおいて、θ方向に対する時間周波数マスクと、それと最も類似度の高い音源に対する時間周波数マスクとの類似度を示している。音源方向推定部３３０は，時間周波数マスクマッチング部３２０が出力する，指定した方向θに対する音源存在スコアl_1が，予め定めた閾値以上であるとき，指定した方向θに音源が存在すると判定し，音源方向推定結果を出力する（ステップS４０３）。

【００５７】

ステップS４０３で方向θに音源が存在すると判定した場合には、時間領域変換部１５０は，特定方向音源分離部１４０が出力する方向θに対する音源分離結果を，時間領域の分離信号に変換して出力する。出力は、例えばスピーカ１７０から音声信号（分離音）として出力される。このとき、図３のようにあわせて音源方向推定結果を表示することにより、音源の特定がさらに容易となる。例えば、会議室において，音源方向に該当する座席を表示することにより、音源である発言者を特定することができる。

【００５８】

会議室のように、音源が複数想定される場合には、ステップS２０３でθを０～３６０度順次指定し、全方位に渡って同様の処理を行なえばよい。

【００５９】

本実施例では、分離音は，特定方向時間周波数マスク推定部１３０で生成したθ方向に対する時間周波数マスクを用いて得ているが、全音源時間周波数マスク推定部３１０で生成した、音源に対する時間周波数マスクで得たものを用いても良い。あるいは，両者を選択可能としても良い。音源に対する時間周波数マスクとしては、例えばθ方向に対する時間周波数マスクと最も類似度の高い音源に対する時間周波数マスクを用いる。ただし、音

源が多い場合等、スパース性が成立しない場合は、一般に、方向に対する時間周波数マスクを用いたほうが精度が良い。

【００６０】

本実施例では，マスクマッチング部では２つのマスク同士を比較している。他の手法として，マスクを用いて抽出した音源信号同士を比較して類似性を評価しても同様の効果が得られる。ただし，この場合には両方のマスクにより音源分離を行なう必要がある。

【実施例３】

【００６１】

図６～図８を用いて，本発明の第三実施形態に関わる音源分離装置５００を説明する。第三実施形態は，複数の音源が同一方向に存在することを判定可能な例である。本実施例は，例えばロボット等に応用することもできる。

【００６２】

第三実施例の音源分離装置５００は，図６における時間周波数マスク加算部５１０と，時間周波数マスクマッチング部５２０と，複数音源同一方向判定部５３０が加わることを除けば，図３に示した第二実施形態の音源分離装置３００と同じ構成であるので，以下では，時間周波数マスク加算部５１０と，時間周波数マスクマッチング部５２０と，複数音源同一方向判定部５３０についてのみ説明し，他の説明を省略する。

【００６３】

図７に示した第三実施形態の処理フローも，ステップS６０１からステップS６０４が加わることを除けば，図４に示した第二実施形態の処理フローと同一であるため，以下ではステップS６０１からステップS６０４についてのみ説明し，他の説明を省略する。

【００６４】

時間周波数マスク加算部５１０は，全音源時間周波数マスク推定部３１０が出力する全音源に対する時間周波数マスクを入力として受け取り，複数の音源の時間周波数マスクを加算し，時間周波数マスクマッチング部５２０に出力する（ステップS６０１）。例えば，音源数がNで各音源の時間周波数マスクをy_1, …, y_Nとするとき，時間周波数マスク加算部５１０は，数式４

【００６５】

【数４】

$$\mathrm{E} \coloneqq \{ I \subseteq \{1, \ldots, N\} \mid |I| \geq 2 \}$$

【００６６】

と数式５

【００６７】

【数５】

$$\{ \sum_{n \in I} y_n \mid I \in E \}$$

で表される合計2^N - (N + 1)個の加算時間周波数マスクを計算する。

【００６８】

ここで時間周波数マスクの加算とは、例えばマスクＡとマスクＢを加算したマスクＡ＋Ｂがある場合、マスクＡ＋Ｂを使って抽出した音源が、マスクＡを使って抽出した音源と、マスクＢを使って抽出した音源との加算になるようなマスクを合成する処理をいう。

【００６９】
時間周波数マスクマッチング部５２０は，時間周波数マスク加算部５１０が出力する加算時間周波数マスクと，特定方向時間周波数マスク推定部１３０が出力する指定した方向θに対する時間周波数マスクを入力として受け取り，上記２つのマスクの類似度を計算する（ステップS６０２）。時間周波数マスクの類似度d_2としては，例えば数式６
【００７０】
【数６】

$$d_2(\sum_{n\in I} y_n, y_\theta) = -||\sum_{n\in I} y_n - y_\theta||_F^2$$

【００７１】
を用いれば良い。ここで，y_nは全音源時間周波数マスク推定部３１０が出力する音源nの時間周波数マスクの推定値，Iは数式４の要素，y_θは特定方向時間周波数マスク推定部１３０が出力する方向θに対する時間周波数マスクの推定値である。類似度d_2が大きいということは、所定の方向θに複数の音源が存在する可能性を示している。さらに，時間周波数マスクマッチング部５２０は，指定した方向θに複数の音源が存在するか否かを表す，複数音源存在スコアを計算し出力する（ステップS６０２）。複数音源存在スコアl_2としては，例えば，数式７
【００７２】
【数７】

$$\ell_2(\theta) = \max_{I\subseteq E} d_2(\sum_{n\in I} y_n, y_\theta)$$

を用いれば良い。ここで，Eは数式４で定義される。
【００７３】
図８で、所定の方向に複数の音源が存在する場合の判定の原理を概念的に説明する。図８には、音源１と音源２に対する時間周波数マスクから抽出された音源と、各方向からの音源が概念的に示されている。音源１と音源２に対する２つのマスクを加算することにより、音源１＋音源２に対する時間周波数マスクが生成される。加算した時間周波数マスクによる音源と所定方向の時間周波数マスクによる音源の類似度を計算することにより、例えば音源１＋音源２と方向２の音源がマッチすることが分かる。この場合、音源１と音源２が方向２にあり、方向２からの音は音源１および音源２の音である可能性が高い。以上の説明では、理解を容易にするため抽出した音源のマッチングを行なっているが、既に述べたように、音源を抽出するためのマスクのマッチングを行なっても同様の効果がある。
【００７４】
上記の理論を基にして、複数音源同一方向判定部５３０は，時間周波数マスクマッチング部５２０が出力する，指定した方向θに対する複数音源存在スコアl_2が，予め定めた閾値以上であるとき，指定した方向θに音源が複数存在すると判定し，判定結果を出力する（ステップS６０３）。
【００７５】
マイクロホンアレイ１６０を備えた音声アシスタント端末、例えばモーターと操舵機構を備える移動機構を備えるロボットが，音声認識を行なってユーザに応答する用途を想定すると、一方向に複数の音源（複数のユーザ）が存在し、同時に発話する場合、音声認識が困難になる場合が考えられる。そこで、指定した方向θに複数の音源が存在すると判定

されたとき，マイクロホンアレイを備えた音声アシスタント端末は，同一方向にいる複数の音源が別々の方向になるように移動機構５４０を制御し，移動する（ステップS６０４）。

【００７６】

そのためには，ステップS６０３、S２０３，S２０４、S６０２のループ処理を繰り返し，音源が複数存在すると判定された方向θの時間周波数マスクに対する複数音源存在スコアl_2が，閾値を下回るまで移動を行なえばよい。

【実施例４】

【００７７】

図９と図１０を用いて，本発明の第四実施形態に関わる音源分離装置７００を説明する。第四実施形態は，同一方向に存在する音源が複数回発話することを判定可能な例である。本実施例は，例えばロボット等に応用することもできる。

【００７８】

第四実施例の音源分離装置７００は，図９における特定方向時間周波数マスク貯蔵部７１０と，時間周波数マスクマッチング部７２０と，同一発話判定部７３０が加わることを除けば，図１に示した第一実施形態の音源分離装置１００と同じ構成であるので，以下では，特定方向時間周波数マスク貯蔵部７１０と，時間周波数マスクマッチング部７２０と，同一発話判定部７３０についてのみ説明し，他の説明を省略する。

【００７９】

図１０に示した第四実施形態の処理フローも，ステップS８０１からステップS８０３が加わることを除けば，図２に示した第一実施形態の処理フローと同一であるため，以下ではステップS８０１からステップS８０３についてのみ説明し，他の説明を省略する。

【００８０】

特定方向時間周波数マスク貯蔵部７１０は，特定方向時間周波数マスク推定部１３０が毎時刻出力する指定した方向θに関する時間周波数マスクを保存する（ステップS８０１）。毎時刻出力する時間周波数マスクは，それぞれ異なる所定時間区間の観測信号を反映している。

【００８１】

時間周波数マスクマッチング部７２０は，特定方向時間周波数マスク貯蔵部７１０に保存された，指定した方向θに関する２つの時刻の時間周波数マスクを入力として受け取り，類似度を計算する（ステップS８０２）。時間周波数マスクの類似度d_3としては，例えば数式８

【００８２】

【数８】

$$d_3\big(y(t_1,\theta),y(t_2,\theta)\big) = -\,||y(t_1,\theta)-y(t_2,\theta)||_F^2$$

【００８３】

を用いれば良い。ここで，y(t_1，θ)とy(t_2，θ)は，異なる２つの時刻t_1とt_2における指定した方向θに対する時間周波数マスクの推定値である。他にも，時間周波数マスクの類似度として，時間周波数マスクy(t_2，θ)を時間方向に伸縮した結果であるy'(t_2，θ)と，y(t_1，θ)の間の類似度を計算することにしても良い。

【００８４】

同一発話判定部７３０は，時間周波数マスクマッチング部７２０が出力する時間周波数マスクの類似度d_3が，予め定めた閾値以上であるとき，指定した方向θから同一の発話が２度あったと判定し，判定結果を出力する（ステップS８０２）。

【００８５】

マイクロホンアレイ１６０を備えた音声アシスタント端末、例えばロボットが，音声認識を行なってユーザに応答する用途を想定すると、同一のユーザが同じ発話を複数回繰り返す場合、ロボットが音声認識できず応答ができていない場合が考えられる。そこで，指定した方向θから同一の発話が２度あったと判定されたとき，マイクロホンアレイを備えた音声アシスタント端末は，方向θに向きを変え，音声応答を行う（ステップS８０３）。このとき，音声アシスタント端末は必ずしもユーザの発話の音声認識ができている必要はなく，同一方向から同一の発話が２度あったことをトリガーとして，予め定められた応答を行なうことができる。例えば、応答のための音声合成部７４０を備え、同一発話判定部７３０の判定結果により、所定の発話を行なうように音声合成部７４０を制御する。この構成により、ユーザの音声認識ができていなくても、音声合成部７４０が「よく聞こえませんでした。もう一度話してください。」のように発話することにより、円滑な対応が可能となる。

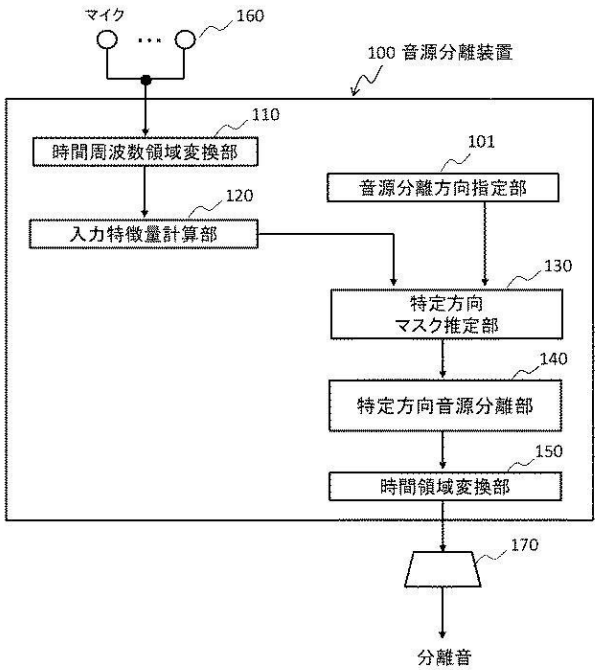
【符号の説明】

【００８６】

音源分離装置１００，音源分離方向指定部１０１，時間周波数領域変換部１１０，入力特徴量計算部１２０，特定方向時間周波数マスク推定部１３０，特定方向音源分離部１４０，時間領域変換部１５０

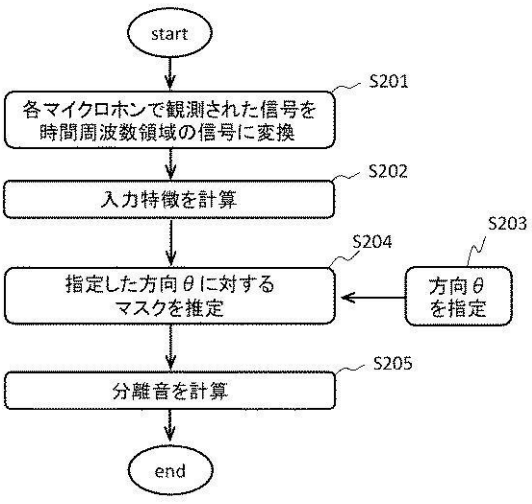
【図１】

図1



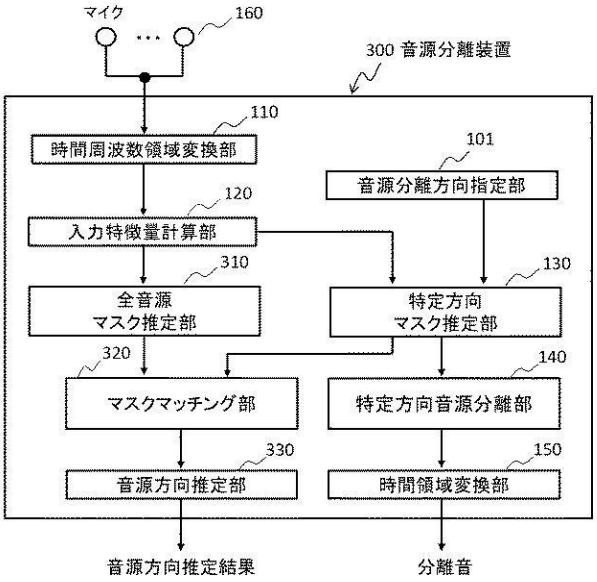
【図２】

図2



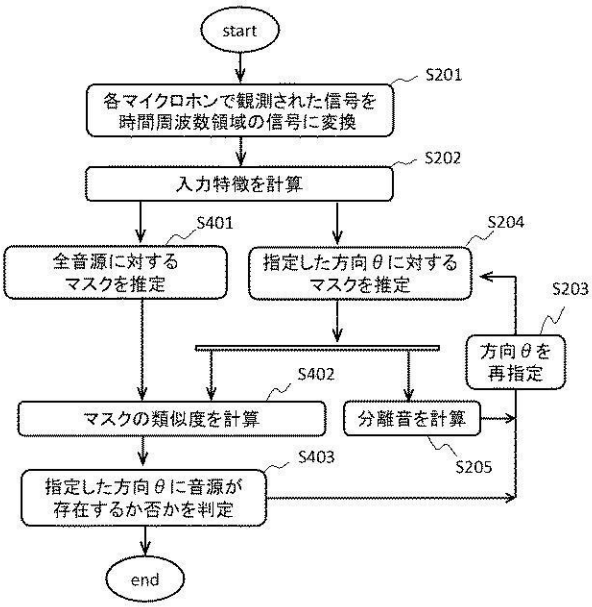
【図３】

図3



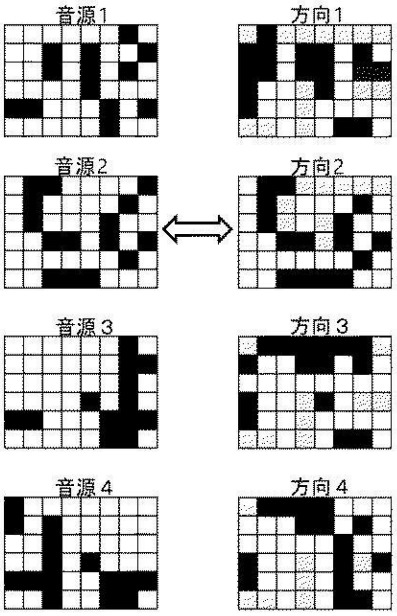
【図４】

図4



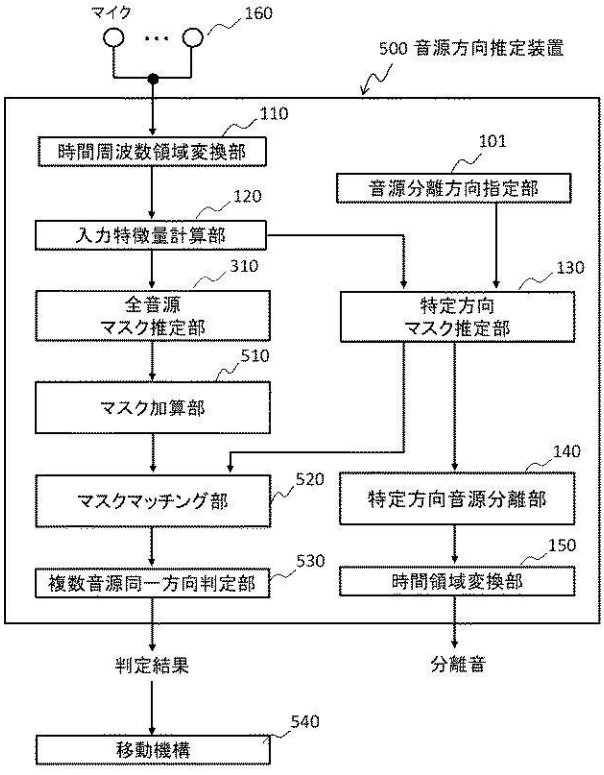
【図５】

図5



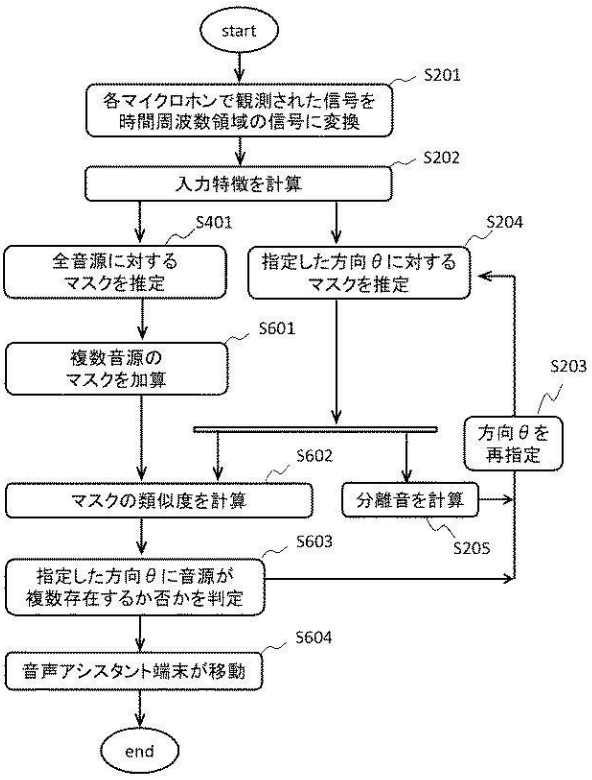
【図６】

図6



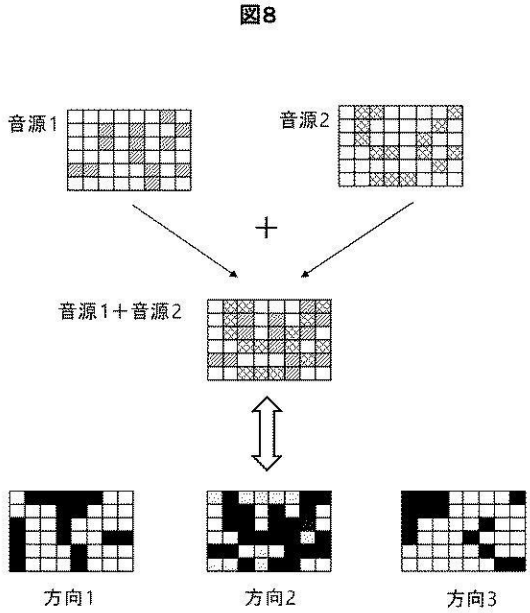
【図７】

図7



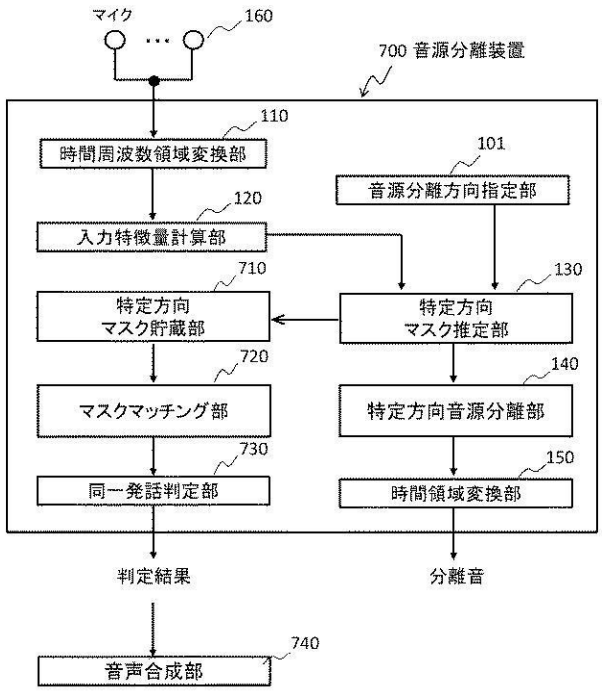
【図８】

図8



【図９】

図9



【図１０】

図10