(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2014-89249
(P2014-89249A)
(43)公開日　平成26年5月15日(2014.5.15)

(51)Int.Cl.		F I			テーマコード（参考）
G10L 21/028	(2013.01)	G10L	21/02	201D	5D018
G10L 21/0308	(2013.01)	G10L	21/02	203Z	5D220
H04R 1/40	(2006.01)	H04R	1/40	320A	
H04R 3/00	(2006.01)	H04R	3/00	320	

審査請求 未請求　請求項の数 5　ＯＬ　（全 13 頁）

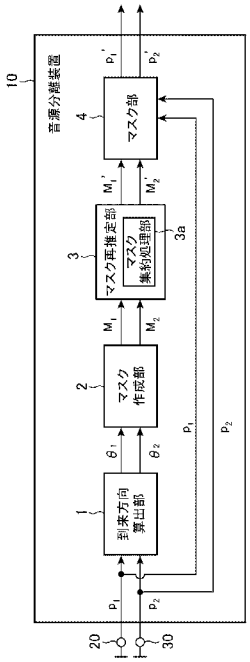
(21)出願番号　特願2012-237835 (P2012-237835)
(22)出願日　平成24年10月29日 (2012.10.29)

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目７番３号
(74)代理人　100123434
弁理士　田澤　英昭
(74)代理人　100101133
弁理士　濱田　初音
(74)代理人　100173934
弁理士　久米　輝代
(74)代理人　100156351
弁理士　河村　秀央
(72)発明者　太刀岡　勇気
東京都千代田区丸の内二丁目７番３号　三菱電機株式会社内

Ｆターム(参考)　5D018 BB21
5D220 BA01　BC05

(54)【発明の名称】音源分離装置

(57)【要約】
【課題】ＴＤＯＡから観測音の到来方向を算出して作成したマスクを音声らしさを用いて修正し、マスクのパーミュテーションを解決する。
【解決手段】各音源から出力された音声の到来方向を算出する到来方向算出部１と、到来方向算出部１が算出した各到来方向の時間周波数平面において、観測信号スペクトルから複数の音源のうち対応する音源から出力された目的音声以外の観測音の観測信号スペクトルをマスキングするマスクを作成するマスク作成部２と、音声の特徴に基づいて、各マスクについて目的音声と当該目的音声以外の観測音との分離性能を検証し、検証結果に基づいてマスクを再推定するマスク再推定部３と、マスク再推定部３が再推定した各マスクを用いて、観測信号スペクトルから目的音声以外の観測音の観測信号スペクトルをマスキングし、目的音声の観測信号スペクトルを取得するマスク部４とを備える。
【選択図】図２



【特許請求の範囲】
【請求項１】
複数の音源からの音声が混合した観測音を時間周波数領域に変換した観測信号スペクトルから、前記各音源から出力された音声の到来方向を算出する到来方向算出部と、
前記到来方向算出部が算出した各到来方向の時間周波数平面において、前記観測信号スペクトルから前記複数の音源のうち対応する音源から出力された目的音声以外の観測音の観測信号スペクトルをマスキングするマスクを作成するマスク作成部と、
前記音声の特徴に基づいて、前記マスク作成部が作成した各マスクについて、前記目的音声と当該目的音声以外の観測音との分離性能を検証し、検証結果に基づいて前記マスクを再推定するマスク再推定部と、
前記マスク再推定部が再推定した各マスクを用いて、前記観測信号スペクトルから前記目的音声以外の観測音の観測信号スペクトルをマスキングし、前記目的音声の観測信号スペクトルを取得するマスク部とを備えた音源分離装置。
【請求項２】
前記マスク部は、前記マスク作成部が作成した各マスクを用いて、前記観測信号スペクトルから前記目的音声以外の観測音の観測信号スペクトルをマスキングし、前記目的音声の観測信号スペクトルを取得し、
前記マスク部において前記マスク再推定部が再推定したマスクを用いて取得した前記目的音声の観測信号スペクトル、および前記マスク部において前記マスク作成部が作成したマスクを用いて取得した前記目的音声の観測信号スペクトルについて、音声モデルに対するそれぞれの尤度を算出する尤度算出部と、
前記尤度算出部が算出した尤度に基づいて、前記マスク再推定部が再推定したマスク、または前記マスク作成部が作成したマスクのいずれか一方を選択し、選択したマスクを用いてマスキングした前記目的音声の観測信号スペクトルを取得するマスク選択部とを備えたことを特徴とする請求項１記載の音源分離装置。
【請求項３】
前記マスク再推定部は、前記音声の時間的連続性、前記音声の倍音構造、または前記音声の話者特性に基づいて、前記マスク作成部が作成した各マスクについて、前記目的音声と当該目的音声以外の観測音との分離性能を検証し、検証結果に基づいて同一の音源から出力された前記目的音声の観測信号スペクトルは同一のマスクのマスキングによって取得されるよう前記マスク作成部が作成した各マスクを再推定するマスク集約処理部を備えることを特徴とする請求項２記載の音源分離装置。
【請求項４】
前記マスク再推定部は、前記音声の時間的連続性、前記音声の倍音構造、または前記音声の話者特性に基づいて、前記マスク作成部が作成した各マスクについて、前記目的音声と当該目的音声以外の観測音との分離性能を検証し、検証結果に基づいて異なる音源から出力された前記目的音声の観測信号スペクトルはそれぞれ異なるマスクのマスキングによって取得されるように、前記マスク作成部が作成したマスクを再推定するマスク分離処理部を備えることを特徴とする請求項２または請求項３記載の音源分離装置。
【請求項５】
前記マスク再推定部は、前記マスク作成部が作成した各マスクの信頼度が低い音声帯域において、前記マスク作成部が作成した複数のマスクを組み合わせてなるマスクの組み合わせを生成するマスク交叉部を備え、
前記マスク部は、前記マスク交叉部が生成したマスクの組み合わせで指定されたマスクを用いて前記観測信号スペクトルから前記目的音声以外の観測音の観測信号スペクトルをマスキングし、前記目的音声の観測信号スペクトルを取得し、
前記尤度算出部は、前記マスク部が取得した各マスクの組み合わせによって取得された前記目的音声の観測信号スペクトルについて、前記音声モデルに対するそれぞれの尤度を算出し、
前記マスク選択部は、前記尤度算出部が算出した尤度に基づいて、前記マスク交叉部が

生成したマスクの組み合わせのうち最も尤度の高いマスクの組み合わせを選択し、選択した組み合わせのマスクを用いてマスキングした前記目的音声の観測信号スペクトルを取得することを特徴とする請求項２記載の音源分離装置。

【発明の詳細な説明】

【技術分野】

【０００１】

この発明は、複数の音源からの音声信号が混在した観測信号から、それぞれの音源に対応する分離信号を得る音源分離装置に関するものである。

【背景技術】

【０００２】

複数人の音声が混ざった音声信号を分離して、各人の音声信号を取り出す技術は音声認識技術の適用範囲拡大に寄与する。音源毎の音声信号の分離方法としては、マイクの死角を対象外の話者に向けるビームフォーミング（以下、ＢＦと称する）による方法や、独立成分分析（ＩＣＡ：independent Component analysis）により混合行列を推定する方法が用いられている。また近年は、時間周波数平面上のスペクトルで音声がスパースなことを利用して、対象話者以外の成分をマスクするバイナリマスクによる分離方法が用いられている。

【０００３】

一方で、ＢＦはノイズの抑圧には優れているが、混成音声の分離にはあまり有効でない。また、ＩＣＡは残響や騒音の影響で性能が低下する。さらに、ＢＦやＩＣＡによる分離方法では、マイクの数が音源数以上でなければならないという制約がある。これに対して、バイナリマスクにはこのような制限がないため、適用先が広く、有望であると言える。

【０００４】

バイナリマスクにもいくつかの手法があるが、ここでは時間・周波数ｂｉｎにおける音声の到来時間差（ＴＤＯＡ）に着目して分類を行う方法について述べる。

２つのマイクで観測された音声信号の短時間フーリエ変換後の時間周波数平面(t,f)におけるスペクトルをp_1，p_2とすると、各スペクトルの位相差αは以下の式（１）で表される。

$$\alpha = \arg\left[p_1(f,t)/p_2(f,t)\right] \quad (-\pi \leq \theta \leq \pi) \quad \cdots (1)$$

【０００５】

さらに、式（１）から各スペクトルの成分の時間差τと音波の到来方向θが、以下の式（２）により求められる。

$$\tau = 1/2\pi f \alpha$$

$$\theta \sin^{-1}(\tau c/l_m) \quad \cdots (2)$$

ｃは音速、l_mはマイク間隔である。音波の到来方向を別手法で推定する、もしくはθをクラスタリングすることにより、音源の方向別に(t,f)領域でのマスクを作成する。

【０００６】

例えば到来角がθ_1の第１の信号に対するマスクがM_1（ｔ，ｆ）であった場合、以下の式（３）のように推定される。

$$M_1(t,f) = 1 \quad for \ |\theta - \theta_1| \leq \theta_t$$

$$M_1(t,f) = \varepsilon \quad for \ |\theta - \theta_1| > \theta_t \quad \cdots (3)$$

θ_tは許容誤差、εは十分小さい数である。

推定されたマスクを用いてマスクされた以下の式（４）で示すスペクトルを、逆フーリエ変換してマスク後の信号を得る。

$$p'_1(f,t) = M_1(t,f)\ p_1(f,t) \quad \cdots (4)$$

【０００７】

ＴＤＯＡによるバイナリマスクを用いた従来の音源分離装置は、例えば上述した式（２）に基づいて到来方向θを算出する手段、上述した式（３）に基づいてマスクを作成する手段、および上述した式（４）に基づいて音声スペクトルをマスクすることにより音声分離スペクトルを得る手段によって構成される。

【０００８】

しかし、バイナリマスクによる分離方法では、バイナリマスクの推定において、マスクを時間および周波数ｂｉｎといった少ない情報から推定するため、推定の精度が誤差の影響を受けやすいという問題があった。特にＴＤＯＡから観測音の到来方向を算出してマスクを作成する方法では、マイクの間隔に比して、波長の長い低周波成分の場合には位相差が付きにくいことから、波長の短い高周波成分の場合には空間的エイリアシングの影響でマスクの推定精度が低下するという問題があった。

【０００９】

そこで、バイナリマスクを用いた音源分離方法において、音声の特徴を生かしてマスクの誤判定を抑制する技術として、例えば特許文献１および特許文献２に開示されているものがある。特許文献１には、ある周波数ｂｉｎに隣接する複数の周波数ｂｉｎのスペクトル成分に対する時間変化を連結する手法が開示されている。特許文献２には、音源分離のためのバイナリマスキングにおいて、パワースペクトルからマスクパターンを生成する手法が開示されている。

【００１０】

また、バイナリマスクの妥当性を、音声モデルを用いて検証する技術として、例えば特許文献３から特許文献５に開示されているものがある。

特許文献３には、ブラインド音声分離にＥＭアルゴリズムを適用し、最大尤度を与える音源方向と、各時間周波数成分への各音源の寄与率をＥＭアルゴリズムによって推定する手法が開示されている。特許文献４には、信号分離において、事後確率の類似度を指標として観測信号のクラスタリングを行う手法が開示されている。特許文献５には、音源分離装置において、確率モデルのモデルパラメタと各音源の存在確率を用いて有効音源を抽出する手法が開示されている。

【先行技術文献】

【特許文献】

【００１１】

【特許文献１】特開２００８－０２６６２５号公報
【特許文献２】特開２０１０－２３９４２４号公報
【特許文献３】特開２００８－１４５６１０号公報
【特許文献４】特開２００９－０５３３４９号公報
【特許文献５】特開２０１１－１６４４６７号公報

【発明の概要】

【発明が解決しようとする課題】

【００１２】

しかしながら、上述した特許文献１および特許文献２に開示された技術では、マスクのスパース性を利用していないため、滑らかではあるが分離性能の低い非合理的なマスクを生じるという課題があった。また特許文献３に開示された技術では、マスクがスパースになるような基準が設けられておらず、分離性能の低い非合理的なマスクを生じるという課題があった。また、特許文献４および特許文献５に開示された技術では、音声らしさを基準として用いていないため、分離音に聴感上や音声認識にとって悪影響を及ぼすひずみが入りやすいという課題があった。

【００１３】

この発明は、上記のような課題を解決するためになされたもので、ＴＤＯＡから観測音の到来方向を算出して作成したマスクを音声らしさを用いて修正し、マスクのパーミュテーションを解決する音源分離装置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】

【００１４】
この発明に係る音源分離装置は、複数の音源からの音声が混合した観測音を時間周波数領域に変換した観測信号スペクトルから、各音源から出力された音声の到来方向を算出する到来方向算出部と、到来方向算出部が算出した各到来方向の時間周波数平面において、観測信号スペクトルから複数の音源のうち対応する音源から出力された目的音声以外の観測音の観測信号スペクトルをマスキングするマスクを作成するマスク作成部と、音声の特徴に基づいて、マスク作成部が作成した各マスクについて、目的音声と当該目的音声以外の観測音との分離性能を検証し、検証結果に基づいてマスクを再推定するマスク再推定部と、マスク再推定部が再推定した各マスクを用いて、観測信号スペクトルから目的音声以外の観測音の観測信号スペクトルをマスキングし、目的音声の観測信号スペクトルを取得するマスク部とを備えるものである。
【発明の効果】
【００１５】
この発明によれば、分離性能の高いマスクを作成することができ、明瞭な目的音声を取得することができる。
【図面の簡単な説明】
【００１６】
【図１】実施の形態１による音源分離装置のマスク再推定処理を示す説明図である。
【図２】実施の形態１による音源分離装置の構成を示すブロック図である。
【図３】実施の形態２による音源分離装置の構成を示すブロック図である。
【図４】実施の形態３による音源分離装置の構成を示すブロック図である。
【図５】実施の形態４による音源分離装置の構成を示すブロック図である。
【図６】１６ｋＨｚサンプリングでの波形とスペクトログラムを示す図である。
【図７】実施の形態５による音源分離装置の構成を示すブロック図である。
【発明を実施するための形態】
【００１７】
実施の形態１.
混合前の音声を用いて、それぞれの音源に対応する分離信号を得るためのマスク（理想マスク）を作成して観察すると、$M_1(t,f)=1$となる(t,f)は、ある程度まとまっている傾向にある。すなわち時間・周波数方向にはスペクトルは局所的にはある程度の連続性がある。ところがＴＤＯＡにより作成したマスクは孤立点が多い。これはＴＤＯＡの推定誤差の影響で、１つの音源からの音が異なるマスクに分類されてしまうためである。そこで、この実施の形態１では、$M_1(t,f)=1$となる(t,f)を近い範囲にまとめることで、より分離性能の高いマスクを作成する。
【００１８】
なお以下では、説明の簡単化のため上述した式（３）においてεを０とする。また、２つの音源から出力されて混合された音声信号を分離する場合を例に説明する。なお、本発明の構成は、３つ以上の音源から出力されて混合された音声信号を分離する場合にも適用可能である。
【００１９】
ここである(t,f)において$M_1(t,f)=1, M_2(t,f)=0$であったとする。この時、時間周波数平面でのスペクトルの局所的な連続性を考慮すると、$M_2(t,f)$の周囲の点に１が多く、$M_1(t,f)$の周囲の点に０が多い場合は、推定誤りであって実は$M_1(t,f)=0, M_2(t,f)=1$である可能性が高い。ここでは例えば、密集度の指標として以下の式（５）を用いることとする。

$$\sigma_i(t,f) = \sum_{(|t-t'|\leq \Delta t,\ |f-f'|\leq \Delta f)} M_i(t', f') \quad \cdots (5)$$

Δｔ，Δｆはそれぞれ時間・周波数領域での近接範囲を示す。$\sigma_1(t,f)<\sigma_2(t,f)$であった場合には$M_1(t,f)=0, M_2(t,f)=1$とする。この操作を時間周波数平面に対して行う。さらにそれを繰り返すことで、２つのマスクのうち正しいマスクに集約させることができる

。言い換えると、マスクの密集度を高め凝縮させることができる。

【００２０】

時間周波数平面上でマスクが分散している場合、スパース性が低くなり音声らしさが失われる。そのため、近接範囲のマスクの状況を参考にして、対象マスクの｛０，１｝を切り替えることによりできるだけ小さい範囲にマスクをまとめることができる。この処理を繰り返すことにより、時間周波数平面上でマスクが局所的に分布するようになり、マスクの密集度を高める前と比較してマスクのスパース性を向上させることができる。

【００２１】

次に、マスクの密集度を高める処理を具体的に説明する。

図１は、この発明の実施の形態１による音源分離装置のマスク再推定処理を示す説明図である。図１（ａ）はマスクの初期状態を示し、図１（ｂ）はマスク再推定処理を１回行った状態を示し、図１（ｃ）はマスク再推定処理を２回行った状態を示している。

図１で示す表の列方向は時間を変化させた領域であり、行方向は周波数を変化させた領域である。

図１（ａ）の初期状態において、マスクM_1の領域Ａは$M_1(t,f)=1$であり、領域Ｂは$M_1(t+1,f)=1$である。一方、マスクM_2の領域Ａ´は$M_2(t,f)=0$であり、領域Ｂ´は$M_2(t+1,f)=0$である。領域Ａ，Ａ´では密集度が$\sigma_1 < \sigma_2$であることから、$M_1(t,f)=0$，$M_2(t,f)=1$と再推定される。一方、領域Ｂ，Ｂ´では密集度が$\sigma_1 > \sigma_2$であることから、$M_1(t,f)=1$，$M_2(t,f)=0$と再推定される。

【００２２】

図１（ｂ）の再推定処理１回目の状態において、マスクM_1の領域Ａは$M_1(t,f)=0$であり、領域Ｂは$M_1(t+1,f)=1$である。一方、マスクM_2の領域Ａ´は$M_2(t,f)=1$であり、領域Ｂ´は$M_2(t+1,f)=0$である。領域Ａ，Ａ´では密集度が$\sigma_1 < \sigma_2$であることから、$M_1(t,f)=0$，$M_2(t,f)=1$と再推定される。同様に、領域Ｂ，Ｂ´においても密集度が$\sigma_1 < \sigma_2$であることから、$M_1(t+1,f)=0$，$M_2(t+1,f)=1$と再推定される。

【００２３】

図１（ｃ）の２回目の状態において、マスクM_1の領域Ａは$M_1(t,f)=0$であり、領域Ｂは$M_1(t+1,f)=0$である。一方、マスクM_2の領域Ａ´は$M_2(t,f)=1$であり、領域Ｂ´は$M_2(t+1,f)=1$である。領域Ａ，Ａ´および領域Ｂ，Ｂ´では共に密集度が$\sigma 1 < \sigma 2$であり、上述した再推定結果である$M_1(t,f)=0$，$M_2(t,f)=1$と変化はない。

【００２４】

図１（ａ）で示した初期状態と、図１（ｃ）で示した２回の再推定処理を行った状態との密度σにより、マスクM_1，M_2の密集度が高まったと判定される。図１の例では、再推定処理を２回行う構成を示したが、再推定処理を行う回数はあらかじめ設定しておいてもよいし、密集度σの変化が閾値以下になった時に処理を終了するように構成してもよい。

また、図１の例では領域Ａ，Ａ´，Ｂ，Ｂ´に対して再推定処理を行う構成を示したが、図１で示したその他全ての領域を構成する全ての要素に対して近接要素の影響を勘案して再推定処理を行う。

【００２５】

図２は、この発明の実施の形態１による音源分離装置の構成を示すブロック図である。

音源分離装置１０は、到来方向算出部１、マスク作成部２、マスク再推定部３およびマスク部４で構成されている。

到来方向算出部１は、第１のマイク２０および第２のマイク３０でそれぞれ観測された２つの混合音声信号の時間周波数領域（ｔ，ｆ）におけるスペクトルから、第１のマイク２０および第２のマイク３０からの音波の到来角θ_1、θ_2を算出する。到来方向θの算出は、上述した式（１）および式（２）を用いて行われる。

【００２６】

マスク作成部２は、到来方向算出部１が算出した音波の到来方向θをクラスタリングする、または音波の到来方向を異なる手法で推定した結果を取得することにより、第１のマ

イク２０の時間周波数領域(t,f)のマスクM_1および第２のマイク３０の時間周波数領域(t,f)のマスクM_2を作成する。例えば、到来角θ_1の第１のマイク２０からの第１の音声信号に対するマスクM_1の時間周波数領域M_1（ｔ，ｆ）は、上述した式（３）で示したように推定される。

【００２７】

マスク再推定部３のマスク集約処理部３ａは、上述した式（５）の基準に従って、マスク間で推定誤りを解消するため、例えばM_1（ｔ，ｆ）=1となる時間周波数(ｔ，ｆ)領域を所定の範囲内にまとめるようにマスクM_1，M_2の再推定処理を行い、マスクの集約を行う。マスクM_1，M_2の再推定処理として、図１で示した再推定処理を適用する。再推定されたマスクM_1'，M_2'は、マスク部４に出力される。

【００２８】

マスク部４は、マスク再推定部３から入力されたマスクM_1，M_2を用いて上述した式（４）に基づいて、第１のマイク２０および第２のマイク３０で観測された音声スペクトルp_1，p_2をマスクし、分離音声スペクトルp_1'，p_2'を得る。

【００２９】

以上のように、この実施の形態１によれば、マスク間での推定誤りを解消するために時間周波数平面でのスペクトルの局所的な連続性を考慮してマスクの密集度を高めて適切なマスクへの集約を行うマスク集約処理部３ａを備えたマスク再推定部３を備えるように構成したので、１つの音源に寄与すべきマスクが分散していた場合に、１つのマスクの集約することができ、分離音声に他の音源の音声が漏れ出すことが抑制され、対象話者の音声を明瞭化することができる。

【００３０】

実施の形態２．

この実施の形態２では、複数の音源に寄与するマスクが、１つの分離音声スペクトルに寄与する１つのマスクにまとめられている場合に、適切なマスクに分離する構成を示す。

図３は、この発明の実施の形態２による音源分離装置の構成を示すブロック図である。なお、以下では、実施の形態１による音源分離装置１０の構成要素と同一または相当する部分には実施の形態１で使用した符号と同一の符号を付して説明を省略または簡略化する。

【００３１】

この実施の形態２では、マスク再推定部３をマスク分離処理部３ｂで構成している。マスク分離処理部３ｂには、例えば異なる周波数ｂｉｎでのマスクの共起確率を用いる方法を適用する。ある周波数f_1とf_2［Ｈｚ］の共起する確率を、学習データによりあらかじめ求めておく。共起確率p_sとしては、$p_s(f_1,f_2)=N(f_1,f_2)/N_t$を用いることができる。ここでＮｔは総フレーム数、Ｎ（f_1，f_2）は$M(t,f_1)=M(t,f_2)=1$であったフレーム数である。これにより、共起しやすい周波数の組み合わせを知ることができる。例えば、音声は倍音構造を持つのでｆとｎ＊ｆは共起しやすい。ｆは任意の周波数［Ｈｚ］、ｎは自然数である。ここで、ｆとf_1が共起する確率は低いとする（f_1はｆでない任意の周波数［Ｈｚ］）。ここでマスクM_1において$M_1(t,f)=M_1(t,f_1)=1$であり、マスクM_2において$M_2(t,f_1)=0$且つ$M_2(t,n*f_1)=1$であった場合$M_1(t,f_1)=0$ $M_2(t,f_1)=1$である確率が高いと考えられる。このように倍音構造の利用では、Ｆ０に対してその倍音成分のマスクを確認して音声らしさを推定することによりマスクを再度推定する。

【００３２】

以上のようにこの実施の形態２によれば、マスク分離処理部３ｂが共起しやすい周波数の組み合わせに注目して音声らしさを推定するように構成したので、複数の音源に寄与するマスクを適切なマスクに分離することができる。これにより、本来複数の音源に寄与すべきマスクが１つにまとめられていた場合に、適切なマスクに分離されるため、分離音声の雑音が抑制され、対象話者の音声を明瞭化することができる。

【００３３】

実施の形態３．

この実施の形態３では、実施の形態２で再推定されたマスクが、再推定前のマスクよりも妥当であるか否か音声モデルを用いて検証する構成を示す。

図４は、この発明の実施の形態３による音源分離装置の構成を示すブロック図である。なお、以下では、実施の形態２による音源分離装置１０の構成要素と同一または相当する部分には実施の形態１で使用した符号と同一の符号を付して説明を省略または簡略化する。

【００３４】

実施の形態２で示した音源分離装置１０に対して、マスク部４の後段に尤度算出部５、音声モデル記憶部６およびマスク選択部７を追加して設けている。

上述した実施の形態２の処理を行うことにより、マスク作成部２が作成した元のマスクM_1，M_2と、マスク再推定部３のマスク分離処理部３ｂにより再推定されたマスクM_1'，M_2'の２通りのマスクが得られる。

マスク部４は、マスク再推定部３から入力されたマスクM_1'，M_2'を用いて上述した式（４）に基づいて、第１のマイク２０および第２のマイク３０で観測された音声スペクトルp_1，p_2をマスキングし、分離音声スペクトルp_1'，p_2'を得る。さらにマスク部４は、マスク作成部２が作成した元のマスクM_1，M_2を用いて、第１のマイク２０および第２のマイク３０で観測された音声スペクトルp_1，p_2をマスキングし、音声スペクトルp_1''，p_2''を得る。

【００３５】

これら全ての分離音声スペクトルp_1'，p_2'，p_1''，p_2''について、尤度算出部５が音声モデル記憶部６に記憶された音声モデルに対する尤度をフレーム単位で計算する。通常、複数の話者のそれぞれの発話内容は異なると仮定できるので、異なる話者の音声が混在した場合には、異なる話者の音声が混在したスペクトルよりも、単一の話者による音声のスペクトルの方が音声モデルに対する尤度が高くなり音声らしいと判断されることになる。例えば、以下の式（６）のＧＭＭ(Gaussian Mixture Model)によりモデル化された音声のモデル中の最大尤度を求めることで、音声らしさを判断することができる。

$$p(x)=\sum_{k=1}^{K}\pi_k \mathrm{N}(\mu_k,\Sigma_k) \quad \cdots \quad (6)$$

式（６）において、Ｎは平均μ_k、共分散Σ_k、混合率π_kの正規分布である。

【００３６】

マスク選択部７は、尤度算出部５が算出した尤度を参照し、分離音声スペクトルp_1'，p_2'，p_1''，p_2''のうち最も音声らしい組み合わせを選択し、選択した分離音声スペクトルに対応したマスクを選択する。これにより、音声らしさの高い分離音声を出力するマスクを選択することができる。マスク選択部７が選択したマスクを用いて、再度音声スペクトルp_1，p_2をマスキングしてもよいが、マスク部４の処理により分離音声スペクトルp_1'，p_2'，p_1''，p_2''が既に得られているので、対応するマスクの分離音声スペクトルを選択して最終的な分離音声スペクトルを得ることができる。

尤度算出部５およびマスク選択部７を備えたことにより、元のマスクM_1，M_2と再推定されたマスクM_1'，M_2'のうち音声らしさの高いマスクを選択することができる。

【００３７】

以上のように、この実施の形態３によれば、マスク分離処理部３ｂが再推定したマスクおよびマスク作成部２が作成したマスクを用いてマスキングされた分離音声スペクトルp_1'，p_2'，p_1''，p_2''の尤度を計算する尤度算出部５と、算出された尤度に基づいて音声らしい組み合わせとなる分離音声スペクトルに対応したマスクを選択するマスク選択部７とを備えるように構成したので、マスク分離処理部３ｂによるマスクの分離の誤りを検出し、適切なマスクの選択を行うことができる。

【００３８】

実施の形態４．

上述したように、ＴＤＯＡによって作られたマスクは低周波数域と高周波数域において性能が低い。低周波数域では位相の変化が小さいため、誤差が生じるためである。また高周波数域でもマイクの間隔よりも短い波長の音波が到来した場合には、位相が２π回転したものと区別がつかない空間的エイリアシングの影響で推定精度が低くなる。この実施の形態４では、信頼性の低い周波数域の分離結果を、音声モデルを用いて補正する構成を示す。

【００３９】

図５は、この発明の実施の形態４による音源分離装置の構成を示すブロック図である。なお、以下では、実施の形態３による音源分離装置１０の構成要素と同一または相当する部分には実施の形態３で使用した符号と同一の符号を付して説明を省略または簡略化する。

この実施の形態４では、マスク再推定部３をマスク交叉部３ｃで構成している。マスク交叉部３ｃは、マスクの性能が低い領域（以下、低信頼領域と称する）においてパーミュテーションが起こっているものとして、マスク作成部２が作成したマスクをそれぞれ交叉させて得られるマスクの組み合わせを生成する。マスク部４は、マスク交叉部３ｃが生成したマスクの組み合わせを用いて、上述した式（４）に基づいて、第１のマイク２０および第２のマイク３０で観測された音声スペクトルp_1，p_2をマスキングし、分離音声スペクトルを得る。

【００４０】

次に、具体例を挙げながら実施の形態４の音源分離装置１０の処理内容を説明する。以下では、高周波数域での場合を例に説明を行うが、低周波数域であっても同様に適用することができる。

マイクアレイを用いたＴＤＯＡの精度は、音波の半波長がマイク間隔以下になると低下する。例えば６ｃｍの間隔のアレイを用いた場合、３ｋＨｚ以上で空間的エイリアシングが起こることになる。図６は、１６ｋＨｚサンプリングでの波形とスペクトログラムを示す図である。図６において０Ｈｚ～４ｋＨｚを高信頼領域、４ｋＨｚ～８ｋＨｚを低信頼領域とする。まず低信頼領域をいくつかの帯域に分割する。例えば、低信頼領域である４ｋＨｚ～８ｋＨｚを、４ｋＨｚ～６ｋＨｚと６ｋＨｚ～８ｋＨｚの２つの領域に分割した場合を想定する。

【００４１】

ここで、マスク作成部２はマスクM_1，M_2を作成する。低信頼領域ではパーミュテーションが起こっているものとして、マスク交叉部３ｃは以下に示す（ａ）から（ｄ）に示す４通りのマスクのかけ方の組み合わせを生成する。

（ａ）（$M_1-M_1-M_1$，$M_2-M_2-M_2$）
（ｂ）（$M_1-M_1-M_2$，$M_2-M_2-M_1$）
（ｃ）（$M_1-M_2-M_1$，$M_2-M_1-M_2$）
（ｄ）（$M_1-M_2-M_2$，$M_2-M_1-M_1$）

【００４２】

マスク部４は、上述の（ａ）～（ｄ）の４通りのマスクのかけ方を用いて８つの分離音声スペクトルを作成する。

そこで、この４通りのマスクのかけ方を用いて８つの分離音声スペクトルを作成する。例えば（ａ）のマスクのかけ方により２つの分離音声スペクトルが生成されるので、それぞれp_{a-1}，p_{a-2}と呼ぶ。上述した実施の形態３と比較して分離音声スペクトルの生成数が増加する。これは、低信頼領域を２つの領域に分割しているためである。

【００４３】

尤度算出部５は、マスク部４が作成した８つの分離音声スペクトルに対して、音声モデル記憶部６に記憶された音声モデルに対する尤度をフレーム単位で計算する。実施の形態３と同様に、異なる話者の音声が混在した分離音声スペクトルよりも、単一の話者による音声スペクトルの方が音声らしいと判断されることになる。

音声モデルとしては、例えばモノフォンやトライフォンといった単位での音声のＧＭＭが考えられる。上述した式（６）で示したＧＭＭによりモデル化された音声のモデル中の最大尤度を求めることで、音声らしさを判断することができる。

【００４４】

その他にもＬＰＣ係数などスペクトル包絡の滑らかさなどを基準とすることもできる。分離音声スペクトルp_{a-1}に対するモデル中の最大尤度と、分離音声スペクトルp_{a-2}に対するモデル中の最大尤度を加算したものを$L_{(1)}$とする。同様に（ｂ）から（ｄ）のマスクに対する分離音声スペクトルに対する$L_{(2)}$，$L_{(3)}$，$L_{(4)}$を算出し、最も尤度の高いマスクの組み合わせを選択することで高信頼領域と接続のよい分離音声信号を選び出すことができる。

【００４５】

マスク選択部７は、尤度算出部５が算出した尤度を参照し、最も音声らしい組み合わせの分離音声スペクトルを選択し、選択した分離音声スペクトルに対応したマスクを選択する。これにより、高信頼領域の情報を活用すると共に、低信頼領域のパーミュテーションを解決することができる。

【００４６】

以上のように、この実施の形態４によれば、低信頼領域において、マスク作成部２が作成したマスクをそれぞれ交叉させて得られるマスクの組み合わせを生成するマスク交叉部３ｃと、マスク交叉部３ｃが生成したマスクの組み合わせに基づいて音声スペクトルをマスキングして分離音声スペクトルを生成するマスク部４と、生成された全ての分離音声スペクトルの尤度を計算する尤度算出部５と、算出された尤度に基づいて最も音声らしい組み合わせとなる分離音声スペクトルに対応したマスクを選択するマスク選択部７とを備えるように構成したので、音声らしさの高いマスクを選択することができる。また、高信頼領域と接続のよい分離音声信号を選び出すことができると共に、低信頼領域のパーミュテーションを解決することができる。さらにマスク作成部２が作成したマスクの分離の誤りを検証することができ、適切なマスクを選択することができる。

【００４７】

実施の形態５．

マスクの再推定処理には、様々な方法が適用可能であり、パラメータの調整の余地もある。また、音素や話者によって分離性能の高いマスク推定方法が異なる場合も存在する。そこで、この実施の形態５では、マスク再推定部３がマスクの再推定を行う複数の構成を備える例を示す。

【００４８】

図７は、この発明の実施の形態６の音源分離装置の構成を示すブロック図である。図７の例では、実施の形態１で示したマスク集約処理部３ａおよび実施の形態２で示したマスク分離処理部３ｂを用いてマスク再推定部３を構成している。マスク部４は、マスク集約処理部３ａおよびマスク分離処理部３ｂにより再推定されたマスクを用いて、音声スペクトルp_1、p_2をマスキングする。なお、尤度算出部５、音声モデル記憶部６およびマスク選択部７の動作は上述した実施の形態３および実施の形態４と同一であるため、説明を省略する。

【００４９】

以上のように、この実施の形態５によれば、マスク作成部２が作成したマスクの再推定処理を行うマスク集約処理部３ａおよびマスク分離処理部３ｂを備えるように構成したので、複数のマスク再推定方法により、密集度が高められたマスク、あるいは適切な分離が行われたマスクから、最も音声らしい分離スペクトルが得られるマスクを選択することができ、最適なマスクの再推定処理を選択することができる。

【００５０】

なお、上述した実施の形態１から実施の形態５では、ＴＤＯＡに着目した構成を示したが、ＴＤＯＡ以外のその他の手法に関しても、マスクを作成可能であれば、本願発明の構成を適用することができる。

【００５１】

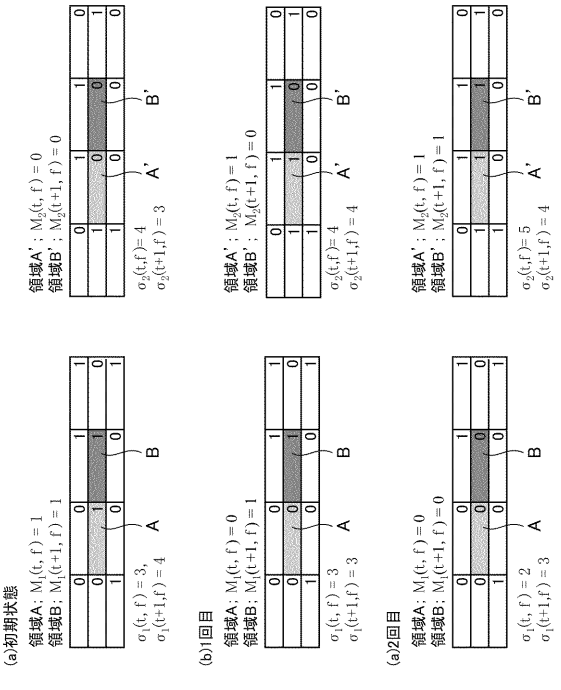
なお、本願発明はその発明の範囲内において、各実施の形態の自由な組み合わせ、あるいは各実施の形態の任意の構成要素の変形、もしくは各実施の形態において任意の構成要素の省略が可能である。

【符号の説明】

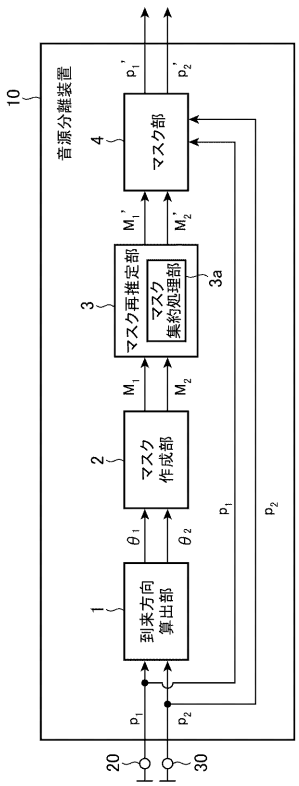
【００５２】

１　到来方向算出部、２　マスク作成部、３　マスク再推定部、３ａ　マスク集約処理部、３ｂ　マスク分離処理部、３ｃ　マスク交叉部、４　マスク部、５　尤度算出部、６　音声モデル記憶部、７　マスク選択部、１０　音源分離装置、２０　第１のマイク、３０　第２のマイク。

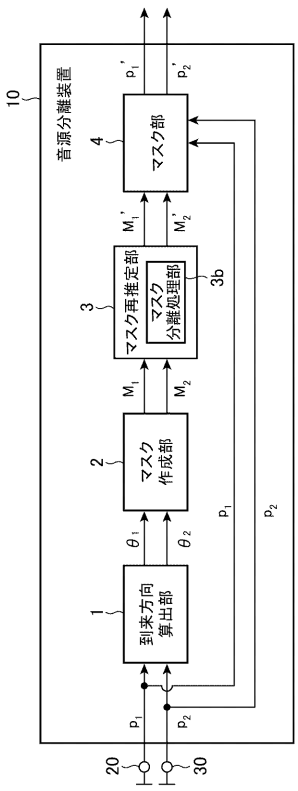
【図１】



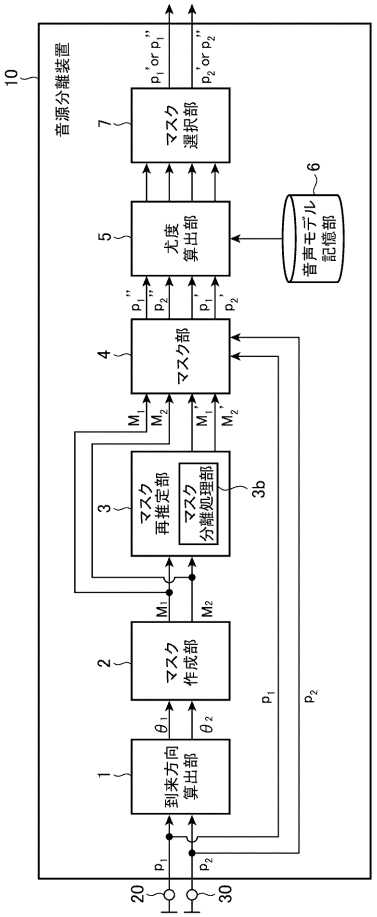
【図２】



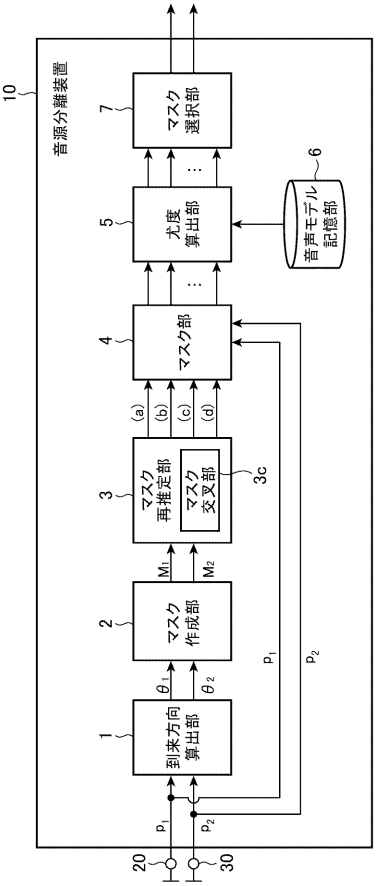
【図３】


音源分離装置
10
到来方向算出部
1
マスク作成部
2
マスク再推定部
マスク分離処理部
3
3b
マスク部
4
20
30

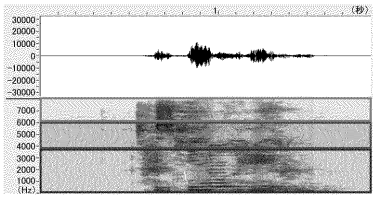
【図４】


音源分離装置
10
到来方向算出部
1
マスク作成部
2
マスク再推定部
マスク分離処理部
3
3b
マスク部
4
尤度算出部
5
音声モデル記憶部
6
マスク選択部
7
20
30

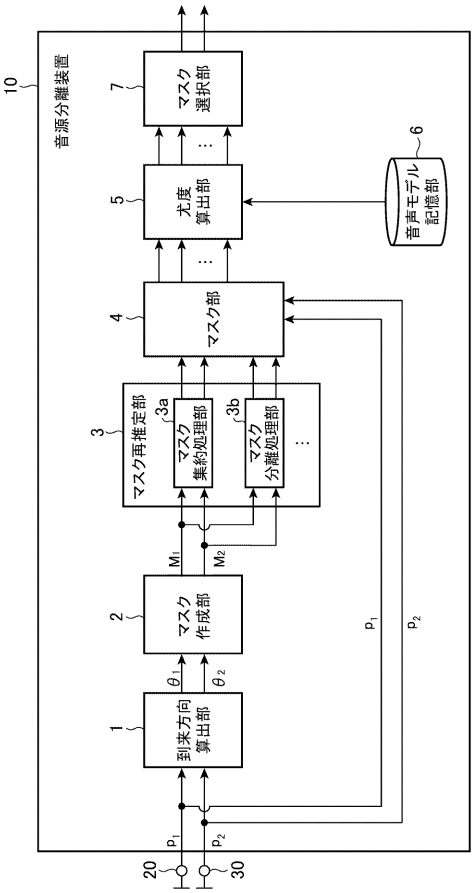
【図５】


音源分離装置
10
到来方向算出部
1
マスク作成部
2
マスク再推定部
マスク交叉部
3
3c
マスク部
4
尤度算出部
5
音声モデル記憶部
6
マスク選択部
7
20
30

【図６】



【図７】


10
音源分離装置
1
到来方向算出部
2
マスク作成部
3
マスク再推定部
3a
マスク集約処理部
3b
マスク分離処理部
4
マスク部
5
尤度算出部
6
音声モデル記憶部
7
マスク選択部
20
30
p_1
p_2
θ_1
θ_2
M_1
M_2