(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2016-156938
(P2016-156938A)
(43)公開日　平成28年9月1日(2016.9.1)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 21/028	(2013.01)	G10L 21/028	B	
G10L 21/0272	(2013.01)	G10L 21/0272	100Z	
G10L 21/0308	(2013.01)	G10L 21/0308	Z	

審査請求 未請求　請求項の数 13　OL　（全 17 頁）

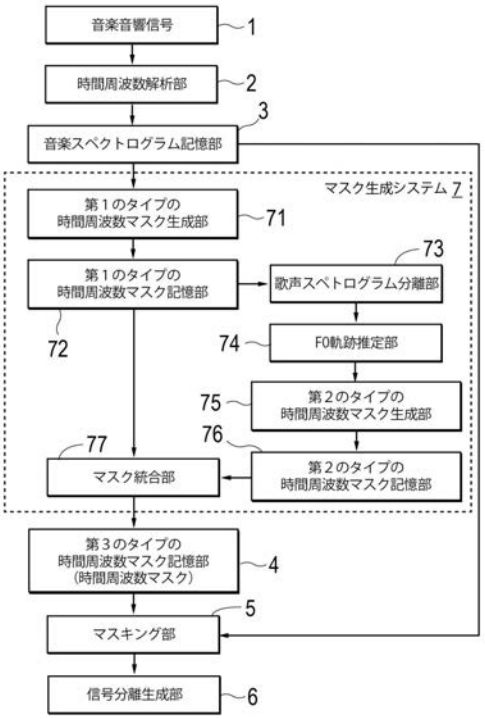
(21)出願番号　特願2015-34339(P2015-34339)
(22)出願日　平成27年2月24日(2015.2.24)

申請有り

(71)出願人　504132272
国立大学法人京都大学
京都府京都市左京区吉田本町３６番地１
(74)代理人　100091443
弁理士　西浦　▲嗣▼晴
(72)発明者　池宮　由楽
京都府京都市左京区吉田本町３６番地１
国立大学法人京都大学内
(72)発明者　吉井　和佳
京都府京都市左京区吉田本町３６番地１
国立大学法人京都大学内
(72)発明者　糸山　克寿
京都府京都市左京区吉田本町３６番地１
国立大学法人京都大学内

(54)【発明の名称】歌声信号分離方法及びシステム

(57)【要約】
【課題】歌声信号と伴奏音信号とを含む音楽音響信号から歌声信号を分離する精度を従来よりも改善することができる歌声信号分離方法及びシステムを提供することにある。
【解決手段】調波構造から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、出現可能性から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビン及び前記出現可能性から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、調波構造から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビンを音楽スペクトログラムからマスキングする機能を有する時間周波数マスクを準備する。次にこの時間周波数マスクを音楽スペクトログラムに適用して分離用歌声スペクトログラムを生成する。そして分離用歌声スペクトログラムに基づいて歌声信号を分離生成する。
【選択図】図１



【特許請求の範囲】
【請求項１】
歌声信号と伴奏音信号とを含む音楽音響信号から前記歌声信号を分離する歌声信号分離方法であって、
前記音楽音響信号を、時間周波数解析を行って音楽スペクトログラムに変換する変換ステップと、
調波構造から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、出現可能性から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビン及び前記出現可能性から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、前記調波構造から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビンを前記音楽スペクトログラムからマスキングする機能を有する時間周波数マスクを準備するマスク準備ステップと、
前記時間周波数マスクを前記音楽スペクトログラムに適用して分離用歌声スペクトログラムを生成するマスキングステップと、
前記分離用歌声スペクトログラムに基づいて前記歌声信号を分離生成する分離生成ステップとからなることを特徴とする歌声信号分離方法。
【請求項２】
前記マスク準備ステップでは、
前記音楽スペクトログラムを、ロバスト主成分分析を用いて低ランク行列とスパース行列とに分解し、
前記低ランク行列と前記スパース行列の比較に基づいて、前記音楽スペクトログラムを歌声が出現している可能性が高いスペクトルビンを含むスパース行列と歌声が出現している可能性が低いスペクトルビンを含む低ランク行列とに分離する機能を有する第１のタイプの時間周波数マスクを生成し、
第１のタイプの時間周波数マスクを前記音楽スペクトログラムに適用して歌声が出現している可能性が高いスペクトルビンを含む歌声スペクトログラムを分離し、
分離された前記歌声スペクトログラムに対して歌声基本周波数Ｆ０を推定して歌声基本周波数Ｆ０軌跡を推定し、
前記歌声基本周波数Ｆ０軌跡に基づいて、前記歌声基本周波数Ｆ０と倍音周辺以外のスペクトルビンをマスキングする機能を有する第２のタイプの時間周波数マスクを生成し、
前記第１のタイプの時間周波数マスクと前記第２のタイプの時間周波数マスクとを統合して、前記第２のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、前記第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビン及び前記第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、前記第２のタイプの時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビンをマスキングする機能を有する第３のタイプの時間周波数マスクを前記時間周波数マスクとして準備することを特徴とする請求項１に記載の歌声信号分離方法。
【請求項３】
前記第１のタイプの時間周波数マスクと前記第２のタイプの時間周波数マスクとの統合とは、前記第１のタイプの時間周波数マスクの選択領域と前記第２のタイプの第２の時間周波数マスクの選択領域との論理積をとることである請求項２に記載の歌声信号分離方法。
【請求項４】
前記第１のタイプの時間周波数マスクと前記第２のタイプの時間周波数マスクとの統合とは、前記第１のタイプの時間周波数マスクの選択領域と前記第２のタイプの時間周波数マスクの選択領域との論理積をとって、仮統合時間周波数マスクを生成し、前記仮統合時間周波数マスクから、歌が無い区間を推定して、推定された時間フレームの全要素を０にすることにより前記第３のタイプの時間周波数マスクとすることである請求項１に記載の歌声信号分離方法。

【請求項５】
前記第１のタイプの時間周波数マスクと前記第２のタイプの時間周波数マスクとの統合とは、前記第１のタイプの時間周波数マスクの選択領域と前記第２のタイプの時間周波数マスクの選択領域との論理積をとって、仮統合時間周波数マスクを生成し、前記仮統合時間周波数マスクから、歌が無い区間を推定して、推定された時間フレームの全要素を０にし、且つ前記第１のタイプの時間周波数マスクから子音を通過させる要素を得て該要素を前記仮統合時間周波数マスクに反映することである請求項１に記載の歌声信号分離方法。
【請求項６】
前記第１のタイプの時間周波数マスク、前記第２のタイプの時間周波数マスク及び第３のタイプの時間周波数マスクは、それぞれバイナリマスクである請求項３，４または５に記載の歌声信号分離方法。
【請求項７】
前記分離生成ステップでは、分離用歌声スペクトログラムを時間領域に逆変換することにより歌声信号を分離生成することを特徴とする請求項１に記載の歌声信号分離方法。
【請求項８】
前記各ステップを１以上のプロセッサで実施することを特徴とする請求項１乃至７に記載の歌声信号分離方法。
【請求項９】
歌声信号と伴奏音信号とを含む音楽音響信号から前記歌声信号を分離する歌声信号分離システムであって、
前記音楽音響信号を、時間周波数解析を行って音楽スペクトログラムに変換する時間周波数解析部と、
調波構造から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、出現可能性から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビン及び前記出現可能性から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、前記調波構造から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビンを前記音楽スペクトログラムからマスキングする機能を有する時間周波数マスクを用いて、前記時間周波数マスクを前記音楽スペクトログラムに適用して分離用歌声スペクトログラムを生成するマスキング部と、
前記分離用歌声スペクトログラムに基づいて前記歌声信号を分離生成する信号分離生成部とからなることを特徴とする歌声信号分離システム。
【請求項１０】
前記音楽スペクトログラムを、ロバスト主成分分析を用いて低ランク行列とスパース行列とに分解し、前記低ランク行列と前記スパース行列の比較に基づいて、前記音楽スペクトログラムを歌声が出現している可能性が高いスペクトルビンを含むスパース行列と歌声が出現している可能性が低いスペクトルビンを含む低ランク行列とに分離する機能を有する第１のタイプの時間周波数マスクを生成する第１のタイプの時間周波数マスク生成部と、
前記第１のタイプの時間周波数マスクを記憶する第１のタイプの時間周波数マスク記憶部と、
前記第１のタイプの時間周波数マスクを前記音楽スペクトログラムに適用して歌声が出現している可能性が高いスペクトルビンを含む歌声スペクトログラムを分離する歌声スペクトログラム分離部と、
分離された前記歌声スペクトログラムに対して歌声基本周波数Ｆ０を推定して歌声基本周波数Ｆ０軌跡を推定するＦ０軌跡推定部と、
前記歌声基本周波数Ｆ０軌跡に基づいて作成されて、前記歌声基本周波数Ｆ０と倍音周辺以外のスペクトルビンをマスキングする機能を有する第２のタイプの時間周波数マスクを生成する第２のタイプの時間周波数マスク生成部と、
前記第２のタイプの時間周波数マスクを記憶する第２のタイプの時間周波数マスク記憶部と、

前記第１のタイプの時間周波数マスクと前記第２のタイプの時間周波数マスクとを統合して作成された、前記第２のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、前記第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビン及び前記第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、前記第２のタイプの時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビンをマスキングする機能を有する第３のタイプの時間周波数マスクを前記時間周波数マスクとして統合するマスク統合部とからなるマスク生成システムによって、前記時間周波数マスクが生成されたものである請求項９に記載の歌語信号分離システム。

【請求項１１】

前記第１のタイプの時間周波数マスク、前記第２のタイプの時間周波数マスク及び第３のタイプの時間周波数マスクは、それぞれバイナリマスクである請求項１０に記載の歌声信号分離システム。

【請求項１２】

前記信号分離生成部は、分離用歌声スペクトログラムを時間領域に逆変換することにより歌声信号を分離生成することを特徴とする請求項９に記載の歌声信号分離システム。

【請求項１３】

上記構成要件は、１以上のプロセッサとメモリによって実現されている請求項９乃至１２のいずれか１項に記載の歌声信号分離システム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、歌声信号と伴奏音信号とを含む音楽音響信号から歌声信号を分離する歌声信号分離方法及びシステムに関するものである。

【背景技術】

【０００２】

非特許文献１［Combining Modeling of Singing Voice and Background Music for Automatic Separation of Musical Mixtures, ISMIR (2013)］には、歌声と伴奏を分離する従来の技術の一例が開示されている。

【０００３】

例えば、非特許文献２［Mixture of Gaussian Process Experts for Predicting Sung Melodic Contour with Expressive Dynamic Fluctuations, ICASSP (2014)］は、歌声のＦ０軌跡を不連続な楽譜成分と微細な変動成分の重ね合わせとして表現する確率モデルを用いて、任意の楽譜から歌声のＦ０軌跡を生成する手法を提案している。同様のモデルは、非特許文献３［混合ガウス過程に基づく歌声音量軌跡の生成過程モデル，情処研報 (2013)］において、歌声の音量軌跡に対しても適用されている。

【先行技術文献】

【非特許文献】

【０００４】

【非特許文献１】Rafii, Z., Germain, F. G., Sun, D. L., and Mysore, G. J.: Combining Modeling of Singing Voice and Background Music for Automatic Separation of Musical Mixtures, ISMIR (2013)

【非特許文献２】Ohishi, Y., Mochihashi, D., Kameoka, H., and Kashino,K.: Mixture of Gaussian Process Experts for Predicting Sung Melodic Contour with Expressive Dynamic Fluctuations, ICASSP (2014)

【非特許文献３】大石康智，持橋大地，亀岡弘和，柏野邦夫：混合ガウス過程に基づく歌声音量軌跡の生成過程モデル，情処研報 (2013)

【発明の概要】

【発明が解決しようとする課題】

【０００５】
混合音中の歌声に対する編集システムを実現するには、高精度な歌声・伴奏音分離と歌声のＦ０推定が必要である。しかしながら従来の技術では、両タスクの相互依存性を考慮して、精度を一挙に改善することができるものはなかった。
【０００６】
本発明の目的は、歌声信号と伴奏音信号とを含む音楽音響信号から歌声信号を分離する精度を従来よりも改善することができる歌声信号分離方法及びシステムを提供することにある。
【０００７】
本発明の他の目的は、高精度な歌声・伴奏音分離と歌声のＦ０推定の相互依存性を考慮して、歌声信号と伴奏音信号とを含む音楽音響信号から歌声信号を分離する精度を一挙に改善することができる歌声信号分離方法及びシステムを提供することにある。
【課題を解決するための手段】
【０００８】
本発明は、歌声信号と伴奏音信号とを含む音楽音響信号から歌声信号を分離する歌声信号分離方法及びシステムを改良の対象とする。本発明の方法では、まず音楽音響信号を、時間周波数解析を行って音楽スペクトログラムに変換する（変換ステップ）。また調波構造から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、出現可能性から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビン及び前記出現可能性から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、調波構造から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビンを音楽スペクトログラムからマスキングする機能を有する時間周波数マスクを準備する（マスク準備ステップ）。
【０００９】
次にこの時間周波数マスクを音楽スペクトログラムに適用して分離用歌声スペクトログラムを生成する（マスキングステップ）。そして分離用歌声スペクトログラムに基づいて歌声信号を分離生成する（分離生成ステップ）。上記のような時間周波数マスクを用いると、歌声信号と伴奏音信号とを含む音楽音響信号から歌声信号を従来よりも精度よく分離することができる。
【００１０】
マスク準備ステップでは、具体的には、次のようにして時間周波数マスクを準備する。まず音楽スペクトログラムを、ロバスト主成分分析を用いて低ランク行列とスパース行列とに分解する。次に低ランク行列とスパース行列の比較に基づいて、音楽スペクトログラムを歌声が出現している可能性が高いスペクトルビンを含むスパース行列と歌声が出現している可能性が低いスペクトルビンを含む低ランク行列とに分離する機能を有する第１のタイプの時間周波数マスクを生成する。次に第１のタイプの時間周波数マスクを音楽スペクトログラムに適用して歌声が出現している可能性が高いスペクトルビンを含む歌声スペクトログラムを分離する。そして分離された歌声スペクトログラムに対して歌声基本周波数Ｆ０を推定して歌声基本周波数Ｆ０軌跡を推定する。次に歌声基本周波数Ｆ０軌跡に基づいて、歌声基本周波数Ｆ０と倍音周辺以外のスペクトルビンをマスキングする機能を有する第２のタイプの時間周波数マスクを生成する。ここで「歌声基本周波数Ｆ０と倍音周辺」とは、歌声基本周波数Ｆ０のピークとその倍音のピークを中心として、予め定めた周波数幅に入る周波数である。この周波数幅は、歌声基本周波数Ｆ０とその倍音のスペクトルの形状から自動的に定めることもできる。
【００１１】
最後に、第１のタイプの時間周波数マスクと第２のタイプの時間周波数マスクとを統合して、第２のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビン及び第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、第２のタイプの

時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビンをマスキングする機能を有する第３のタイプの時間周波数マスクを時間周波数マスクとして準備する。この具体的方法では、基本的にはロバスト主成分分析（Robust Principal Component Analysis; RPCA）を用いてスペクトログラム上で歌声・伴奏音分離を行う。歌声のＦ０情報を用いれば、不要な伴奏音を抑制することができる。一方、混合音に対して歌声のＦ０推定を行うよりも、分離した歌声に対してＦ０推定を行う方がずっと容易である。

【００１２】

第１のタイプの時間周波数マスクと第２のタイプの時間周波数マスクとの統合とは、両マスクの機能を優れた機能を併用可能にすることを意味し、例えば、第１のタイプの時間周波数マスクと第２のタイプの時間周波数マスクとの統合とは、第１のタイプの時間周波数マスクの選択領域と前記第２のタイプの第２の時間周波数マスクの選択領域との論理積をとることにより両マスクを統合することができる。

【００１３】

また統合の他の例では、第１のタイプの時間周波数マスクの選択領域と第２のタイプの時間周波数マスクの選択領域との論理積をとって、仮統合時間周波数マスクを生成し、仮統合時間周波数マスクから、歌が無い区間を推定して、推定された時間フレームの全要素を０にすることにより第３のタイプの時間周波数マスクとすることができる。この時間周波数マスクでは、歌が無い区間を推定して、推定された時間フレームの全要素を０にするため、さらに分離精度を高めることができる。

【００１４】

さらに統合の他の例では、第１のタイプの時間周波数マスクの選択領域と第２のタイプの時間周波数マスクの選択領域との論理積をとって、仮統合時間周波数マスクを生成し、仮統合時間周波数マスクから、歌が無い区間を推定して、推定された時間フレームの全要素を０にし、且つ第１のタイプの時間周波数マスクから子音を通過させる要素を得て該要素を仮統合時間周波数マスクに反映する。このようにするとさらに分離精度を高めることができる。

【００１５】

なお第１のタイプの時間周波数マスク、第２のタイプの時間周波数マスク及び第３のタイプの時間周波数マスクは、それぞれバイナリマスクであるのが好ましい。バイナリマスクを用いると、１と０の組み合わせによりマスクが構成されるため、歌声と伴奏がくっきり分かれ、伴奏音側に歌声が残る可能性はほとんどなくなる。

【００１６】

分離生成ステップでは、分離用歌声スペクトログラムを時間領域に逆変換することにより歌声信号を分離生成することができる。そして各ステップは、１以上のプロセッサで実施することができる。

【００１７】

本発明の方法を実施する本発明の歌声信号分離システムは、時間周波数解析部と、マスキング部と、信号分離生成部とから構成される。

【００１８】

時間周波数解析部は、音楽音響信号を、時間周波数解析を行って音楽スペクトログラムに変換する。マスキング部は、調波構造から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、出現可能性から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビン及び出現可能性から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、調波構造から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビンを音楽スペクトログラムからマスキングする機能を有する時間周波数マスクを用いて、時間周波数マスクを音楽スペクトログラムに適用して分離用歌声スペクトログラムを生成する。そして信号分離生成部は、分離用歌声スペクトログラムに基づいて歌声信号を分離生成する。

【００１９】

時間周波数マスクはマスク生成システムによって生成される。マスク生成システムは、第１のタイプの時間周波数マスク生成部と、第１のタイプの時間周波数マスク記憶部と、歌声スペクトログラム分離部と、Ｆ０軌跡推定部と、第２のタイプの時間周波数マスク生成部と、第２のタイプの時間周波数マスク記憶部と、マスク統合部とから構成される。

【００２０】

第１のタイプの時間周波数マスク生成部は、音楽スペクトログラムを、ロバスト主成分分析を用いて低ランク行列とスパース行列とに分解し、低ランク行列とスパース行列の比較に基づいて、音楽スペクトログラムを歌声が出現している可能性が高いスペクトルビンを含むスパース行列と歌声が出現している可能性が低いスペクトルビンを含む低ランク行列とに分離する機能を有する第１のタイプの時間周波数マスクを生成する。第１のタイプの時間周波数マスク記憶部は、第１のタイプの時間周波数マスクを記憶する。歌声スペクトログラム分離部は、第１のタイプの時間周波数マスクを音楽スペクトログラムに適用して歌声が出現している可能性が高いスペクトルビンを含む歌声スペクトログラムを分離する。Ｆ０軌跡推定部は、分離された歌声スペクトログラムに対して歌声基本周波数Ｆ０を推定して歌声基本周波数Ｆ０軌跡を推定する。第２のタイプの時間周波数マスク生成部は、歌声基本周波数Ｆ０軌跡に基づいて、歌声基本周波数Ｆ０と倍音周辺以外のスペクトルビンをマスキングする機能を有する第２のタイプの時間周波数マスクを生成する。第２のタイプの時間周波数マスク記憶部は、第２のタイプの時間周波数マスクを記憶する。マスク統合部は、第１のタイプの時間周波数マスクと第２のタイプの時間周波数マスクとに基づき、第２のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビン及び第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、第２のタイプの時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビンをマスキングする機能を有する第３のタイプの時間周波数マスクを時間周波数マスクとして統合する。

【００２１】

信号分離生成部は、分離用歌声スペクトログラムを時間領域に逆変換することにより歌声信号を分離生成する。なお上記構成要件は、１以上のプロセッサとメモリによって実現するのが好ましい。

【図面の簡単な説明】

【００２２】

【図１】本発明の歌声信号分離方法を実施する歌声信号分離システムの一例の構成を示すブロック図である。

【図２】図１の実施の形態の歌声信号分離システムをコンピュータ（１以上のプロセッサと１以上のメモリを含む）で実施する際に使用されるソフトウエアのアルゴリズムを示すフローチャートである。

【図３】入力音楽音曲信号を時間周波数解析して得る音楽スペクトログラムの一例を示す図である。

【図４】音楽スペクトログラムからロバスト主成分分析によりスパース行列と低ランク行列とに分析した結果の一例と、両行列の各要素の値を比較して得た第１のタイプの時間周波数分析マスクとしてバイナリマスクの例を示す図である。

【図５】図４の表示内容の理解を高めるために、音楽スペクトログラムの一部を拡大し、またスパース行列と低ランク行列の一部を拡大し、さらに第１のタイプの時間周波数分析マスクとしてバイナリマスクの一部を拡大した図を示している。

【図６】Ｆ０軌跡推定部によって推定された歌声基本周波数Ｆ０軌跡から前記第２のタイプの時間周波数マスク（バイナリマスク）を生成する過程の一例を示す図である。

【図７】第１のタイプの時間周波数マスク（ロバスト主成分分析によるバイナリマスク）と第２のタイプの時間周波数マスク（歌声基本周波数Ｆ０による調波構造のバイナリマスク）を統合する場合の一例を画像で示す図である。

【図８】図７の画像の理解を高めるために、図７に示した複数の画像の一部をそれぞれ拡大して示す図である。
【図９】マスキング部における処理を画像で示すための図である。
【図１０】図９の画像の理解を高めるために、図９に示した複数の画像の一部をそれぞれ拡大して示す図である。
【図１１】（Ａ）乃至（Ｄ）は、マスキング部によるマスキング処理の状況を示す波形図である。
【図１２】歌声信号の再合成を説明するために用いる図である。
【図１３】マスクの統合の他の例を示す概念図である。
【図１４】マスクの統合のさらに他の例を示す概念図である。
【発明を実施するための形態】
【００２３】
以下図面を参照して、本発明の歌声信号分離方法及びシステムの実施の形態の一例を詳細に説明する。図１は、本発明の歌声信号分離方法を実施する歌声信号分離システムの一例の構成を示すブロック図である。図２は、図１の実施の形態の歌声信号分離システムをコンピュータ（１以上のプロセッサと１以上のメモリを含む）で実施する際に使用されるソフトウエアのアルゴリズムを示すフローチャートである。
【００２４】
本発明の方法を実施する本発明の歌声信号分離システムは、時間周波数解析部２と、音楽スペクトログラム記憶部３と、第３のタイプの時間周波数マスク記憶部４と、マスキング部５と、信号分離生成部６とから構成される。図１には、第３のタイプの時間周波数マスク記憶部４に記憶する時間周波数マスクを生成するためのマスク生成システム７も併せて記載してある。説明の都合上、本実施の形態の説明の途中でマスク生成システム７についても説明する。
【００２５】
時間周波数解析部２は、歌声信号と伴奏音信号とを含む音楽音響信号１を、時間周波数解析を行って音楽スペクトログラム（行列）に変換する（ステップＳＴ１）。まず、短時間フーリエ変換（Short-Term Fourier Transform; STFT）あるいは定Ｑ変換を用いて入力音楽音響信号の時間周波数解析を行う。定Ｑ変換については、「Schorkhuber, C. and Klapuri, A.: Constant-Q Transform Toolbox for Music Processing, SMC Conference (2010)」に詳しく記載されている。
【００２６】
実用上、全時間サンプルnにおける対数スペクトルビンを求めるのではなく、例えば１０［msec］などの時間幅で切り出す。以後分かりやすさのため、時間インデクス、周波数インデクスをそれぞれt, fとし、音楽スペクトログラムをX（t, f）と記述する。図３には、入力音楽音曲信号を時間周波数解析して得る音楽スペクトログラムの一例を示している。
【００２７】
第３のタイプの時間周波数マスク記憶部４には、マスク生成システム７で作成した時間周波数マスク（統合マスク）として第３のタイプの時間周波数マスクが記憶される。マスク生成システム７は、第１のタイプの時間周波数マスク生成部７１と、第１のタイプの時間周波数マスク記憶部７２と、歌声スペクトログラム分離部７３と、Ｆ０軌跡推定部７４と、第２のタイプの時間周波数マスク生成部７５と、第２のタイプの時間周波数マスク記憶部７６と、マスク統合部７７とから構成される。第１のタイプの時間周波数マスク生成部７１は、音楽スペクトログラム記憶部３に記憶した音楽スペクトログラム中の音楽スペクトログラムを、ロバスト主成分分析を用いて低ランク行列とスパース行列とに分解し（ステップＳＴ２）、低ランク行列とスパース行列の比較に基づいて、音楽スペクトログラムを歌声が出現している可能性が高いスペクトルビンを含むスパース行列と歌声が出現している可能性が低いスペクトルビンを含む低ランク行列とに分離する機能を有する第１のタイプの時間周波数マスクを生成する（ステップＳＴ３）。そして第１のタイプの時間周

波数マスクは、第１のタイプの時間周波数マスク記憶部７２記憶される。
【００２８】
ロバスト主成分分析は、与えられた行列（２次元配列）を低ランク行列とスパース行列とに分解する手法であり、次式で定式化される。
【００２９】
【数１】

$$\text{minimize } \|L\|_* + \lambda\|S\|_1 \ (\text{subject to } L + S = X)$$

ここで、Ｘ，Ｌ，Ｓはそれぞれ入力行列，低ランク行列およびスパース行列であり、‖・‖*と‖・‖1はそれぞれ核ノルムとＬ1ノルム，λは低ランク性とスパース性のトレードオフパラメータを表す。一般に時間変化するデータ集合などを入力とし、頻出する成分（各フレームで繰り返し現れる成分）が低ランク行列に、それ以外の成分（各フレームに稀にしか現れない成分）がスパース行列に分解される。
【００３０】
音楽スペクトログラムを入力行列Ｘと見なしてロバスト主成分分析を適用すると、繰り返し演奏されるため何度も出現する伴奏音（ドラムやギター）のスペクトルビンは低ランク行列Ｌへ、それ以外の歌声などの時間的な変動が大きいスペクトルビンはスパース行列Ｓへ分解される。本実施の形態では、分析結果から第１のタイプの時間周波数分析マスクとしてバイナリマスクを作成する。
【００３１】
【数２】

$$M_r(t,f) = \begin{cases} 1 & |S(t,f)| > |L(t,f)| \\ 0 & \text{otherwise} \end{cases}$$

このバイナリマスクからなる第１のタイプの時間周波数分析マスクを音楽スペクトログラムX（t，f）へ適用することで歌声スペクトログラムが分離できる。
【００３２】
なお図４には、音楽スペクトログラムからロバスト主成分分析により分析した結果のスパース行列（歌声）と低ランク行列（伴奏）とに分析した結果の一例と、両行列の各要素の値を比較して得た第１のタイプの時間周波数分析マスクとしてバイナリマスクの例を示している。図５は、図４の表示内容の理解を高めるために、音楽スペクトログラムの一部を拡大し、またスパース行列（歌声）と低ランク行列（伴奏）の一部を拡大し、さらに第１のタイプの時間周波数分析マスクとしてバイナリマスクの一部を拡大した図を示している。
【００３３】
歌声スペクトログラム分離部７３は、第１のタイプの時間周波数マスク（バイナリマスク）を音楽スペクトログラムに適用して歌声が出現している可能性が高いスペクトルビンを含む歌声スペクトログラムを分離する（ステップＳＴ４）。Ｆ０軌跡推定部７４は、分離された歌声スペクトログラムに対して歌声基本周波数Ｆ０を推定して歌声基本周波数Ｆ０軌跡を推定する（ステップＳＴ５）。
【００３４】
具体的には、ロバスト主成分分析により分離された歌声スペクトログラムXs^{rpca}（t，f）から、Subharmonic Summation（ＳＨＳ）を用いて歌声のＦ０軌跡を推定する。ＳＨＳについては、「Hermes, D. J.: Measurement of pitch by subharmonic summation, J. Acoust. Soc. Am., Vol. 83, No. 1, pp.257-264 (online), DOI: 10.1121/1.396427 (198

8)」に詳しく説明されている。ＳＨＳは計算コストの低さとノイズへの頑健性を兼ね備えた音高推定法であり、スペクトルビンの各周波数ビンについて、そのビンをＦ０であると仮定したときの倍音に対応する周波数ビンのパワーを重みつきで足し合わせることで、当該ビンにＦ０が存在する尤度を計算する。この音高尤度関数の計算は、対数周波数スケールでは以下で定式化される。

【００３５】

【数３】

$$H(t,s)=\sum_{n=1}^{N}h_n P(t,s+1200\log_2 n),$$

ここで、t，sはそれぞれ時間インデクスと対数周波数［cents］を表し、Ｐ（ｔ，ｓ）は時間フレームt，周波数s［cents］における入力スペクトログラムの振幅である。Nは足し合わせる倍音数，hnは各倍音の重み関数であり、本実施の形態ではそれぞれ１５および０．８６$^{n-1}$とする。人間の聴覚特性の非線形性を考慮するため、ＳＨＳを適用する前に、入力スペクトルビンに対してＡ特性補正をかけるものとする。

【００３６】

ＳＨＳによる音高尤度関数Ｈ（ｔ，ｓ）から歌声音高F(t)は以下の式で計算される。

【００３７】

【数４】

$$F(t)=\underset{c_l^{(t)}\le s\le c_h^{(t)}}{\arg\max}\ H(t,s)$$

ここで、cl$^{(t)}$，ch$^{(t)}$はそれぞれ、時間フレームtにおける音高探索周波数範囲の下限と上限（［cents］）である。

【００３８】

図６は、Ｆ０軌跡推定部７４によって推定された歌声基本周波数Ｆ０軌跡から第２のタイプの時間周波数マスク（バイナリマスク）を生成する過程の一例を示している。第２のタイプの時間周波数マスク生成部７５は、歌声基本周波数Ｆ０軌跡に基づいて、歌声基本周波数Ｆ０と倍音周辺以外のスペクトルビンをマスキングする機能を有する第２のタイプの時間周波数マスクを生成する（ステップＳＴ６）。生成された第２のタイプの時間周波数マスクは第２のタイプの時間周波数マスク記憶部７６に記憶される。

【００３９】

ロバスト主成分分析を用いた従来の歌声分離では、曲の一部しか現れないベースやドラム、メインボーカルと音高をずらして唱和するバックコーラスなども、歌声として分離されてしまう。歌声・伴奏音分離と歌声のＦ０推定は相互依存性をもっている。つまり、歌声のＦ０軌跡が与えられていれば、歌声分離に利用することができる一方、歌声が分離されていれば、そのＦ０軌跡を推定することは比較的容易である。そこでこの相補的な関係を利用した歌声分離のために入力音響信号に対して、統合マスク（第３のタイプの時間周波数マスク）を用いて、精密な歌声分離を行う。そこで、第２のタイプの時間周波数マスク生成部７５は、歌声基本周波数Ｆ０軌跡を利用して、さらに精度の高い歌声分離を行うために、歌声基本周波数Ｆ０軌跡から、基本周波数（Ｆ０）と倍音周辺以外のパワーをマスキングする調波マスクを第２のタイプの時間周波数として生成する。ここで「歌声基本周波数Ｆ０と倍音周辺」とは、歌声基本周波数Ｆ０のピークとその倍音のピークを中心として、予め定めた周波数幅に入る周波数である。この周波数幅幅は、歌声基本周波数Ｆ０とその倍音のスペクトルの形状から自動的に定めることもできる。

【００４０】

【数５】

$$M_{\mathrm{h}}(t,f)=\begin{cases}1 & \left[\begin{array}{l}H_t^h-\frac{\mathrm{w}}{2}<C(f)<H_t^h+\frac{\mathrm{w}}{2}\\ H_t^h=F_t+1200\log_2 h,\ 1\le h\le \mathrm{H}\end{array}\right.\\ 0 & \text{otherwise}\end{cases}$$

ここで、Ftは時間フレームtにおけるＦ０［cents］，C(f)は周波数ビンfに対応する対数周波数［cents］，Hは倍音数，wは各倍音でマスクを取る幅［cents］を示す。ロバスト主成分分析によるバイナリマスクと調波マスクを用いて、最終的な歌声と伴奏のスペクトログラムXs（t，f），Xm（t，f）はそれぞれ以下のように得られる。

【００４１】

【数６】

$$X_s(t,f)=M_r(t,f)M_h(t,f)X(t,f),$$
$$X_m(t,f)=X(t,f)-X_s(t,f)$$

マスク統合部７７は、第１のタイプの時間周波数マスクと第２のタイプの時間周波数マスクとを統合して第３のタイプの時間周波数マスクを時間周波数マスク（統合マスク）として作成する（ステップＳＴ７）。この第３のタイプの時間周波数マスクからなる時間周波数マスク（統合マスク）は、上位概念で言えば、調波構造から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、出現可能性から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビン及び出現可能性から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、調波構造から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビンを音楽スペクトログラムからマスキングする機能を有するものである。より具体的に言えば、第３のタイプの時間周波数マスク（統合マスク）は、第２のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビン及び第１のタイプの時間周波数マスクでは歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、第２のタイプの時間周波数マスクでは歌声を含むスペクトルビンではないと判断されるスペクトルビンをマスキングする機能を有する。第３のタイプの時間周波数マスクを時間周波数マスク（統合マスク）は、第３のタイプの時間周波数マスク記憶部４に記憶される。

【００４２】

図７は、第１のタイプの時間周波数マスク（ロバスト主成分分析によるバイナリマスク）と第２のタイプの時間周波数マスク（歌声基本周波数Ｆ０による調波構造のバイナリマスク）を統合する場合の一例を画像で示している。図８は、図７の画像の理解を高めるために、図７に示した複数の画像の一部をそれぞれ拡大して示す図である。この例では、第１のタイプの時間周波数マスク（ロバスト主成分分析によるバイナリマスク）と第２のタイプの時間周波数マスク（歌声基本周波数Ｆ０によるバイナリマスクまたは調波マスク）との統合を、第１のタイプの時間周波数マスクの選択領域と第２のタイプの時間周波数マスクの選択領域との論理積（ＡＮＤ）をとることにより両マスクを統合して第３のタイプの時間周波数マスク（統合バイナリマスク）を得ている。

【００４３】

マスキング部５は、第３のタイプの時間周波数マスク（統合バイナリマスク）を音楽スペクトログラムに適用して分離用歌声スペクトログラムを生成する（ステップＳＴ８）。このマスキング部５から出力される分離用歌声スペクトログラムを記憶部に記憶しておい

てもよいのは勿論である。図９は、マスキング部５における処理を画像で示すための図である。また図１０は、図９の画像の理解を高めるために、図９に示した複数の画像の一部をそれぞれ拡大して示す図である。図１１（Ａ）乃至（Ｄ）は、マスキング部５によるマスキング処理の状況を示す波形図である。なお図１１（Ａ）乃至（Ｄ）においては、スペクトログラムに含まれる１フレーム分のスペクトルを図示の対象としている。図１１（Ａ）は音楽スペクトログラムに含まれる混合音スペクトルＸ（ｆ）である。そして図１１（Ｂ）は混合音スペクトルＸ（ｆ）に対応する第１のタイプの時間周波数マスク（ロバスト主成分分析マスクに含まれる１フレーム分の周波数マスク（Ｍｂ（ｆ）であり、図１１（Ｃ）は混合音スペクトルＸ（ｆ）に対応する第２のタイプの時間周波数マスクに含まれる１フレーム分の周波数マスク［調波マスクＭｈ（ｆ）］である。そして図１１（Ｄ）は、第１のタイプの時間周波数マスク（ロバスト主成分分析マスクに含まれる１フレーム分の周波数マスク（Ｍｂ（ｆ）と第２のタイプの時間周波数マスクに含まれる１フレーム分の周波数マスク［調波マスクＭｈ（ｆ）］が統合されて生成された第３のタイプの時間周波数マスクに含まれる１フレーム分の周波数マスク［統合マスクＭｂ（ｆ）*Ｍｈ（ｆ）］によってマスキングされて得た分離された歌声スペクトル［Ｘ（ｆ）*Ｍｂ（ｆ）*Ｍｈ（ｆ）］である。図１１（Ｄ）から分かるように、統合マスクを使用してマスキングを行うと分離精度が高くなっているのが分かる。

【００４４】

そして信号分離生成部６は、分離用歌声スペクトログラムに基づいて歌声信号を分離生成する。具体的には、図１２に示すように、信号分離生成部６は、マスキング部５から出力された分離用歌声スペクトログラムを時間領域に逆変換することにより歌声信号を分離生成する（ステップＳＴ９）。

【００４５】

本実施の形態の効果を確認するために、ロバスト主成分分析により歌声信号を分離した場合と、本実施の形態で歌声信号を分離した場合について、目的音源の歪みで分離精度を判定するＮＳＤＲ（Normalized Signal-to-Distortion Ratio[dB]）で、１１０曲の音楽音響信号から歌声信号を分離した結果を比較してみた。その結果、歌声の分離精度に関しては、本実施の形態では５．０６ [dB]、ロバスト主成分分析では２．０９ [dB]、伴奏の分離精度に関しては、本実施の形態では６．２１ [dB]、ロバスト主成分分析では１．７１ [dB]という結果が得られた。歌声分離及び伴奏分離の両方において、本実施の形態のほうが、ＲＰＣＡよりも精度が高いことが確認された。

【００４６】

なお上記各構成要件は、１以上のプロセッサとメモリによって実現するのが好ましい。またマスク生成システム４は、本実施の形態の歌声信号分離システムと一緒に構成する必要はない。すなわち第３のタイプの時間周波数マスク（統合マスク）は、歌声信号分離システムとは別に設けられたマスク生成システムによって事前に生成しておいてもよいのは勿論である。

【００４７】

上記実施の形態では、２つバイナリマスクの統合に論理積（ＡＮＤ）を用いたが、本発明におけるマスクの統合は、上記実施の形態に限定されるものではない。図１３は、マスクの統合の他の例を示す概念図である。この例では、第１のタイプの時間周波数マスク（ＲＰＣＡによるバイナリマスク）の選択領域と第２のタイプの時間周波数マスク（歌声基本周波数Ｆ０による調波構造のバイナリマスク）の選択領域との論理積をとって、仮統合時間周波数マスクを生成する。そしてこの仮統合時間周波数マスクから、歌が無い区間を推定して、推定された時間フレームの全要素を０にすることにより第３のタイプの時間周波数マスク（統合バイナリマスク）とすることができる。この時間周波数マスクでは、歌が無い区間を推定して、推定された時間フレームの全要素を０にするため、さらに分離精度を高めることができる。

【００４８】

図１４は、マスクの統合のさらに他の例を示す概念図である。この例では、第１のタイ

プの時間周波数マスク（ロバスト主成分分析によるバイナリマスク）の選択領域と第２のタイプの時間周波数マスク（歌声基本周波数Ｆ０による調波構造のバイナリマスク）の選択領域との論理積をとって、仮統合時間周波数マスクを生成する。そしてこの仮統合時間周波数マスクから、歌が無い区間を推定して、推定された時間フレームの全要素を０にし、且つ第１のタイプの時間周波数マスク（ロバスト主成分分析によるバイナリマスク）から子音を通過させる要素を得て該要素を仮統合時間周波数マスクに反映して、第３のタイプの時間周波数マスク（統合バイナリマスク）とすることができる。このようにするとさらに分離精度を高めることができる。

【産業上の利用可能性】

【００４９】

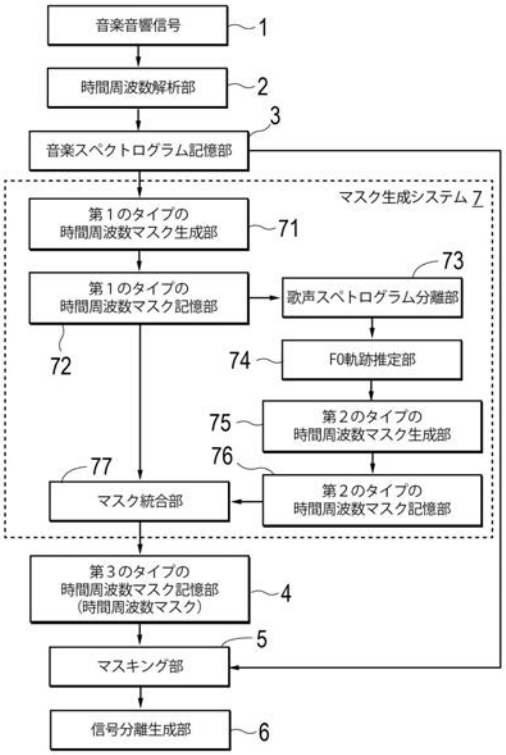
近年、既存楽曲をユーザが自分好みに編集・加工することを可能にする能動的音楽鑑賞システムの研究が盛んである。中でも、混合音中の歌声の編集は最も実現が難しい課題の一つであり、既存の歌声の声質を他の歌唱者の声質に直接変換する技術は提案されているが、歌声がもつ特徴的な音高軌跡、すなわち歌唱表現を編集する技術は実現されていなかったが、本発明によれば、調波構造から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、出現可能性から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビン及び前記出現可能性から判断すると歌声を含むスペクトルビンであると判断されるスペクトルビンであっても、調波構造から判断すると歌声を含むスペクトルビンではないと判断されるスペクトルビンを音楽スペクトログラムからマスキングする機能を有する時間周波数マスクを用いることにより、歌声信号と伴奏音信号とを含む音楽音響信号から歌声信号を従来よりも精度よく分離できる。

【符号の説明】

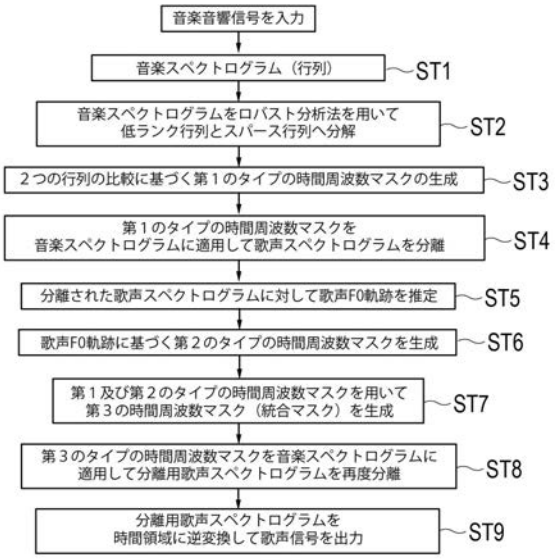
【００５０】

１　音楽音響信号
２　時間周波数解析部
３　音楽スペクトログラム記憶部
４　時間周波数マスク記憶部
５　マスキング部
６　信号分離生成部
７　マスク生成システム
７１　時間周波数マスク生成部
７２　時間周波数マスク記憶部
７３　歌声スペクトログラム分離部
７４　Ｆ０軌跡推定部
７５　時間周波数マスク生成部
７６　時間周波数マスク記憶部
７７　マスク統合部

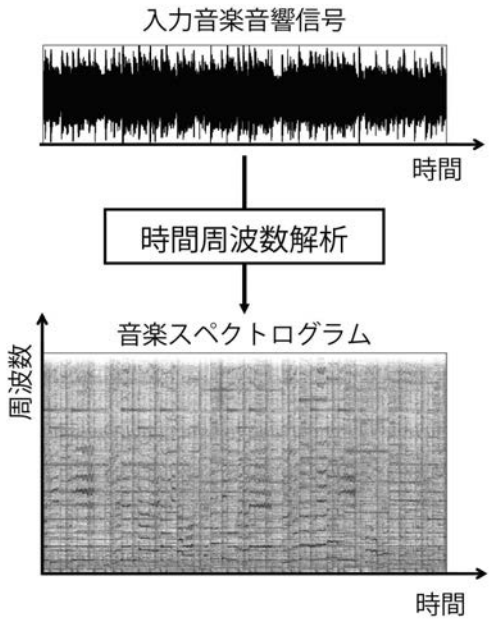
【図１】


音楽音響信号
1
時間周波数解析部
2
音楽スペクトログラム記憶部
3
マスク生成システム 7
第１のタイプの
時間周波数マスク生成部
71
第１のタイプの
時間周波数マスク記憶部
72
歌声スペクトログラム分離部
73
F0軌跡推定部
74
第２のタイプの
時間周波数マスク生成部
75
第２のタイプの
時間周波数マスク記憶部
76
マスク統合部
77
第３のタイプの
時間周波数マスク記憶部
（時間周波数マスク）
4
マスキング部
5
信号分離生成部
6

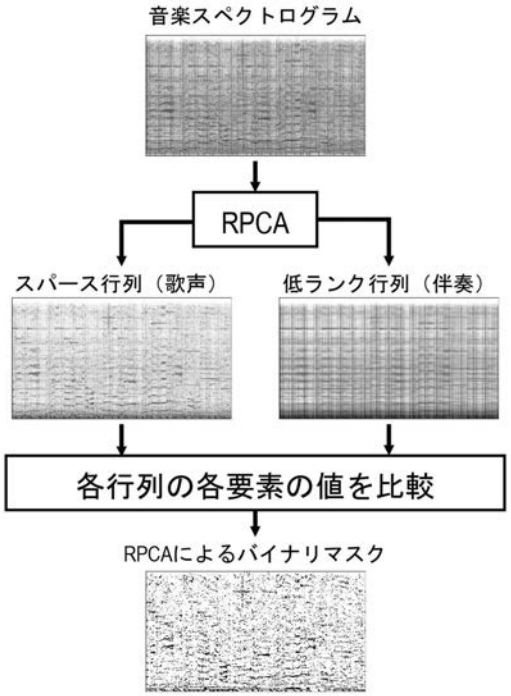
【図２】


音楽音響信号を入力
音楽スペクトログラム（行列）
ST1
音楽スペクトログラムをロバスト分析法を用いて
低ランク行列とスパース行列へ分解
ST2
２つの行列の比較に基づく第１のタイプの時間周波数マスクの生成
ST3
第１のタイプの時間周波数マスクを
音楽スペクトログラムに適用して歌声スペクトログラムを分離
ST4
分離された歌声スペクトログラムに対して歌声F0軌跡を推定
ST5
歌声F0軌跡に基づく第２のタイプの時間周波数マスクを生成
ST6
第１及び第２のタイプの時間周波数マスクを用いて
第３の時間周波数マスク（統合マスク）を生成
ST7
第３のタイプの時間周波数マスクを音楽スペクトログラムに
適用して分離用歌声スペクトログラムを再度分離
ST8
分離用歌声スペクトログラムを
時間領域に逆変換して歌声信号を出力
ST9

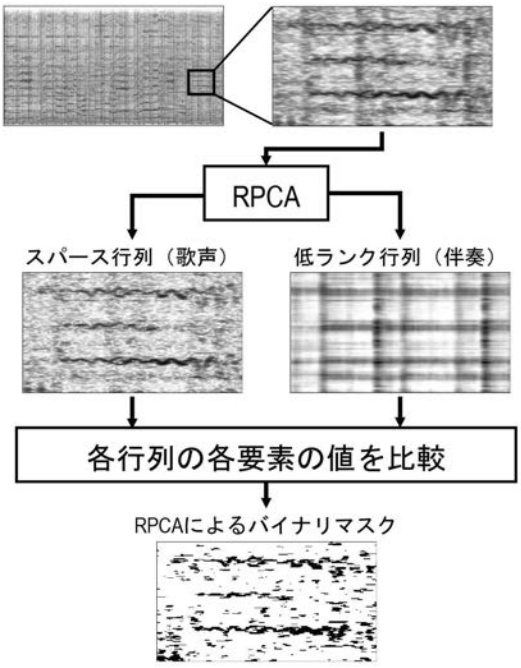
【図３】


入力音楽音響信号
時間
時間周波数解析
音楽スペクトログラム
周波数
時間

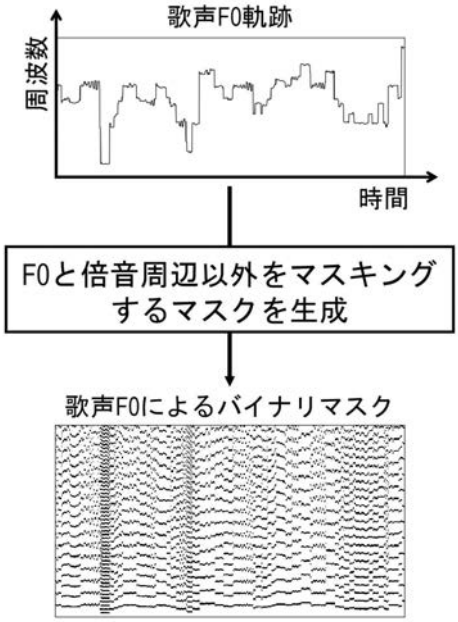
【図４】


音楽スペクトログラム
RPCA
スパース行列（歌声）
低ランク行列（伴奏）
各行列の各要素の値を比較
RPCAIによるバイナリマスク

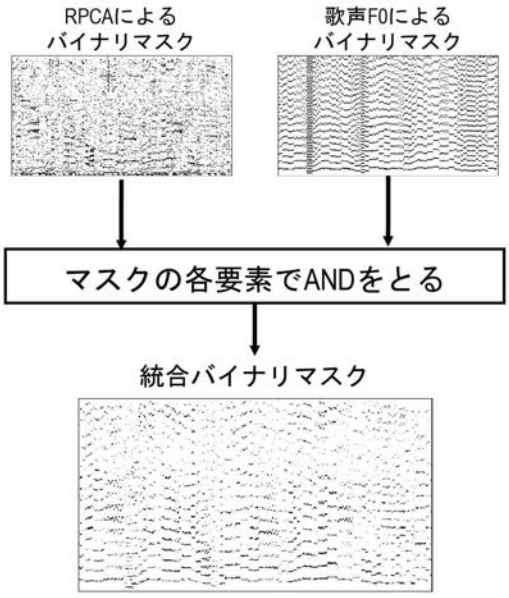
【図５】


RPCA
スパース行列（歌声）
低ランク行列（伴奏）
各行列の各要素の値を比較
RPCAによるバイナリマスク

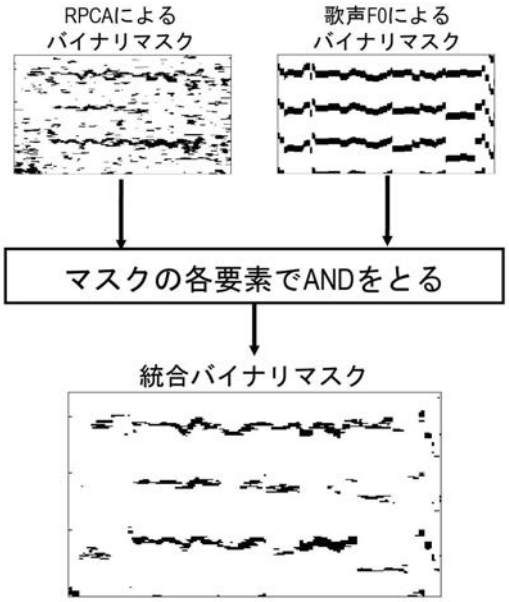
【図６】


歌声F0軌跡
周波数
時間
F0と倍音周辺以外をマスキングするマスクを生成
歌声F0によるバイナリマスク

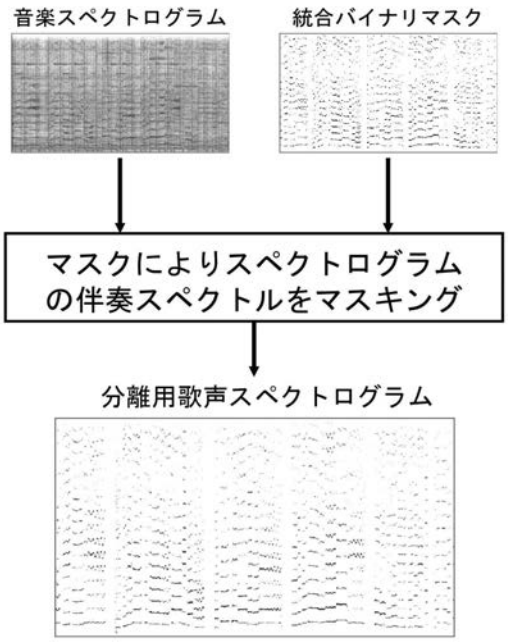
【図７】


RPCAによるバイナリマスク
歌声F0によるバイナリマスク
マスクの各要素でANDをとる
統合バイナリマスク

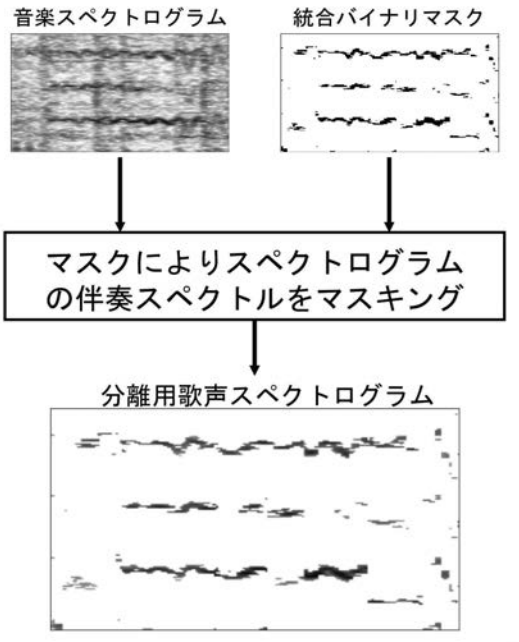
【図８】


RPCAによるバイナリマスク
歌声F0によるバイナリマスク
マスクの各要素でANDをとる
統合バイナリマスク

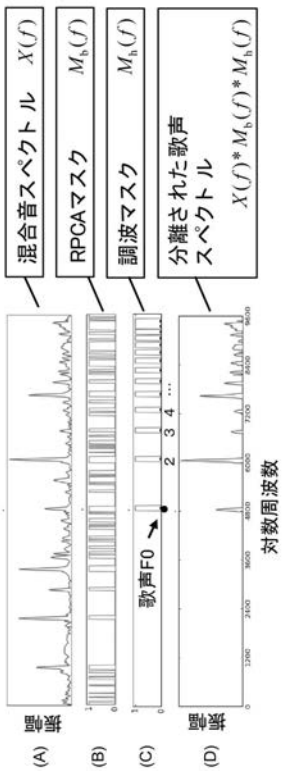
【図９】



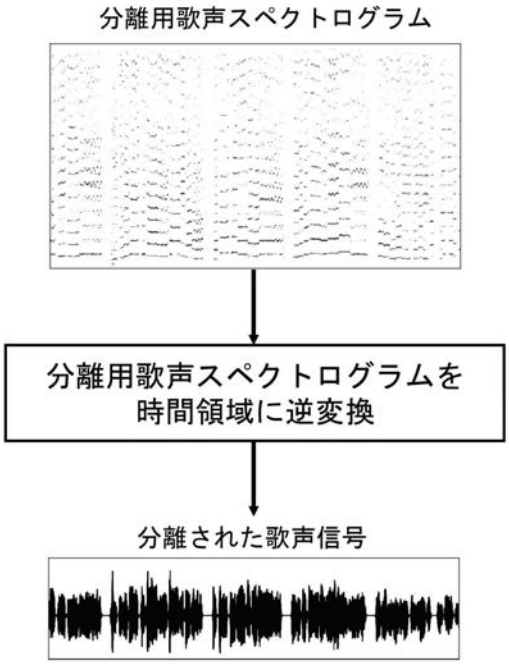
【図１０】



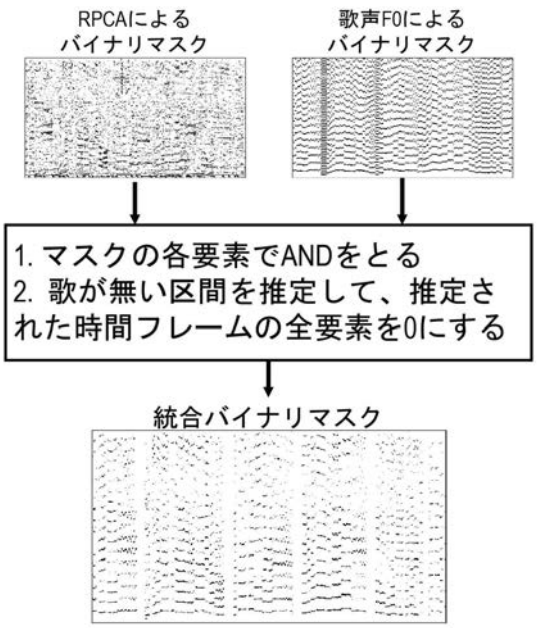
【図１１】



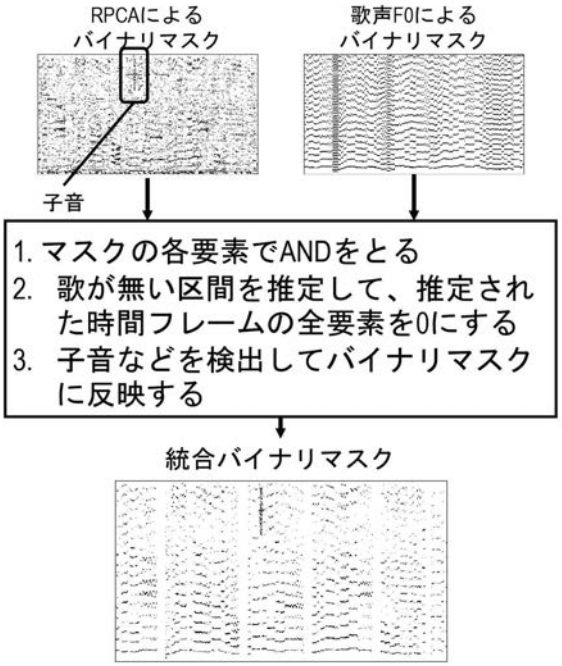
【図１２】



【図１３】


RPCAによる
バイナリマスク
歌声F0による
バイナリマスク
1. マスクの各要素でANDをとる
2. 歌が無い区間を推定して、推定された時間フレームの全要素を0にする
統合バイナリマスク

【図１４】


RPCAによる
バイナリマスク
歌声F0による
バイナリマスク
子音
1. マスクの各要素でANDをとる
2. 歌が無い区間を推定して、推定された時間フレームの全要素を0にする
3. 子音などを検出してバイナリマスクに反映する
統合バイナリマスク