(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7443823号
(P7443823)

(45)発行日　令和6年3月6日(2024.3.6)

(24)登録日　令和6年2月27日(2024.2.27)

(51)国際特許分類

		FI		
G10L 21/0272(2013.01)		G10L	21/0272	100Z
G10L 25/30 (2013.01)		G10L	25/30	

請求項の数　1　（全24頁）

(21)出願番号　特願2020-33347(P2020-33347)
(22)出願日　令和2年2月28日(2020.2.28)
(65)公開番号　特開2021-135446(P2021-135446A)
(43)公開日　令和3年9月13日(2021.9.13)
審査請求日　令和4年12月20日(2022.12.20)

(73)特許権者　000004075
ヤマハ株式会社
静岡県浜松市中央区中沢町10番1号
(74)代理人　110003177
弁理士法人旺知国際特許事務所
(72)発明者　北村 大地
香川県三豊市詫間町香田551 香川高等専門学校内
(72)発明者　渡辺 瑠伊
香川県三豊市詫間町香田551 香川高等専門学校内
審査官　大野 弘

最終頁に続く

(54)【発明の名称】　音響処理方法

(57)【特許請求の範囲】

【請求項１】

第１音源に対応する第１音のうち第１周波数帯域の成分を表す第１入力データと、前記第１音源とは異なる第２音源に対応する第２音のうち前記第１周波数帯域の成分を表す第２入力データと、前記第１音と前記第２音との混合音のうち前記第１周波数帯域とは異なる第２周波数帯域を含む周波数帯域の成分を含む音を表す混合音データと、を含む入力データを取得し、

学習済の推定モデルに前記入力データを入力することで、前記第１音のうち前記第２周波数帯域を含む周波数帯域の成分を表す第１出力データと、前記第２音のうち前記第２周波数帯域を含む周波数帯域の成分を表す第２出力データとの少なくとも一方を生成する

コンピュータにより実現される音響処理方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本開示は、音響処理に関する。

【背景技術】

【０００２】

相異なる音源が発生した複数の音の混合音を音源毎に分離する音源分離技術が従来から提案されている。例えば非特許文献１には、信号の独立性と音源の低ランク性とを同時に考慮することで高精度な音源分離を実現する独立低ランク行列分析（ILRMA：Independ

ent Low- Rank Matrix Analysis）が開示されている。また、非特許文献２には、振幅スペクトログラムをニューラルネットワークに入力することで、音源分離のための時間-周波数領域マスクを生成する技術が開示されている。

【先行技術文献】

【非特許文献】

【０００３】

【文献】Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 9, pp. 1626?1641, September 2016

【０００４】

【文献】Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, Tillman Weyde, "Singing Voice Separation with Deep U-Net Convolutional Networks," Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), 2017

【発明の概要】

【発明が解決しようとする課題】

【０００５】

しかし、非特許文献１および非特許文献２に開示された技術においては、音源分離のための処理負荷が過大であるという問題がある。以上の事情を考慮して、本開示のひとつの態様は、音源分離のための処理負荷を軽減することを目的とする。

【課題を解決するための手段】

【０００６】

以上の課題を解決するために、本開示のひとつの態様に係る音響処理方法は、第１音源に対応する第１音のうち第１周波数帯域の成分を表す第１入力データと、前記第１音源とは異なる第２音源に対応する第２音のうち前記第１周波数帯域の成分を表す第２入力データと、前記第１音と前記第２音との混合音のうち前記第１周波数帯域とは異なる第２周波数帯域を含む周波数帯域の成分を含む音を表す混合音データと、を含む入力データを取得し、学習済の推定モデルに前記入力データを入力することで、前記第１音のうち前記第２周波数帯域を含む周波数帯域の成分を表す第１出力データと、前記第２音のうち前記第２周波数帯域を含む周波数帯域の成分を表す第２出力データとの少なくとも一方を生成する。

【図面の簡単な説明】

【０００７】

【図１】音響処理システムの構成を例示するブロック図である。

【図２】音響処理システムの機能的な構成を例示するブロック図である。

【図３】入力データおよび出力データの説明図である。

【図４】推定モデルの構成を例示するブロック図である。

【図５】音響処理の具体的な手順を例示するフローチャートである。

【図６】訓練データの説明図である。

【図７】学習処理の具体的な手順を例示するフローチャートである。

【図８】第２実施形態における入力データおよび出力データの説明図である。

【図９】第３実施形態における入力データの模式図である。

【図１０】第３実施形態における音響処理システムの機能的な構成を例示するブロック図である。

【図１１】第１実施形態および第３実施形態による効果の説明図である。

【図１２】第１実施形態から第３実施形態に関する観測結果の図表である。

【図１３】第５実施形態における入力データおよび出力データの説明図である。

【図１４】第５実施形態における訓練データの説明図である。

【図１５】第５実施形態に係る音響処理システムの機能的な構成を例示するブロック図で

ある。
【発明を実施するための形態】
【０００８】
Ａ：第１実施形態
図１は、本開示の第１実施形態に係る音響処理システム１００の構成を例示するブロック図である。音響処理システム１００は、制御装置１１と記憶装置１２と放音装置１３とを具備するコンピュータシステムである。音響処理システム１００は、例えばスマートフォン，タブレット端末またはパーソナルコンピュータ等の情報端末により実現される。なお、音響処理システム１００は、単体の装置で実現されるほか、相互に別体で構成された複数の装置（例えばクライアントサーバシステム）でも実現される。
【０００９】
記憶装置１２は、制御装置１１が実行するプログラムと制御装置１１が使用する各種のデータとを記憶する単数または複数のメモリである。記憶装置１２は、例えば磁気記録媒体もしくは半導体記録媒体等の公知の記録媒体、または、複数種の記録媒体の組合せで構成される。なお、音響処理システム１００とは別体の記憶装置１２（例えばクラウドストレージ）を用意し、例えば移動体通信網またはインターネット等の通信網を介して、制御装置１１が記憶装置１２に対する書込および読出を実行してもよい。すなわち、記憶装置１２は音響処理システム１００から省略されてもよい。
【００１０】
記憶装置１２は、音波形を表す時間領域の音響信号Ｓｘを記憶する。音響信号Ｓｘは、第１音源から発音される音（以下「第１音」という）と第２音源から発音される音（以下「第２音」という）とが混合された音（以下「混合音」という）を表す。第１音源と第２音源とは別個の音源である。第１音源および第２音源の各々は、歌唱者または楽器等の発音源である。例えば、第１音は、歌唱者（第１音源）が発音する歌唱音声であり、第２音は、打楽器等の楽器（第２音源）が発音する楽器音である。音響信号Ｓｘは、第１音源と第２音源とが並列に発音する環境において例えばマイクロホンアレイ等の収音装置を利用して収録される。ただし、公知の合成技術により合成された信号が音響信号Ｓｘとして利用されてもよい。すなわち、第１音源および第２音源の各々は仮想的な音源でもよい。
【００１１】
なお、単体の音源のほか複数の音源の集合を第１音源または第２音源として把握してもよい。また、第１音源と第２音源とは基本的には別種の音源であり、第１音と第２音とは音響特性が相違する。ただし、第１音源と第２音源とが相異なる位置に設置された場合のように、各音源の位置を利用して第１音と第２音とを分離可能であれば、第１音源と第２音源とは同種の音源でもよい。すなわち、第１音の音響特性と第２音の音響特性とは、相互に近似または一致してもよい。
【００１２】
制御装置１１は、音響処理システム１００の各要素を制御する単数または複数のプロセッサである。具体的には、例えばＣＰＵ（Central Processing Unit）、ＳＰＵ（Sound Processing Unit）、ＤＳＰ（Digital Signal Processor）、ＦＰＧＡ（Field Programmable Gate Array）、またはＡＳＩＣ（Application Specific Integrated Circuit）等の１種類以上のプロセッサにより、制御装置１１が構成される。制御装置１１は、記憶装置１２に記憶された音響信号Ｓｘから音響信号Ｓｚを生成する。音響信号Ｓｚは、第１音および第２音の一方が他方に対して強調された音を表す時間領域の信号である。すなわち、音響処理システム１００は、音響信号Ｓｘを音源毎に分離する音源分離を実行する。
【００１３】
放音装置１３は、制御装置１１が生成した音響信号Ｓｚが表す音を放音する。放音装置１３は、例えばスピーカまたはヘッドホンである。なお、音響信号Ｓｚをデジタルからアナログに変換するＤ/Ａ変換器と、音響信号Ｓｚを増幅する増幅器とは、便宜的に図示が省略されている。また、図１においては、放音装置１３を音響処理システム１００に搭載した構成を例示したが、音響処理システム１００とは別体の放音装置１３が有線または無線によ

り音響処理システム１００に接続されてもよい。
【００１４】
［１］音響処理部２０
図２は、音響処理システム１００の機能的な構成を例示するブロック図である。図２に例示される通り、制御装置１１は、記憶装置１２に記憶された音響処理プログラムＰ1を実行することで音響処理部２０として機能する。音響処理部２０は、音響信号Ｓxから音響信号Ｓzを生成する。音響処理部２０は、周波数解析部２１と音源分離部２２と帯域拡張部２３と波形合成部２４と音量調整部２５とを具備する。
【００１５】
周波数解析部２１は、音響信号Ｓxの強度スペクトルＸ(m)を時間軸上の単位期間（フレーム）毎に順次に生成する。記号ｍは、時間軸上の１個の単位期間を意味する。強度スペクトルＸ(m)は、例えば振幅スペクトルまたはパワースペクトルである。強度スペクトルＸ(m)の生成には、例えば短時間フーリエ変換またはウェーブレット変換等の公知の周波数分析が任意に採用される。なお、音響信号Ｓxから算定される複素スペクトルが強度スペクトルＸ(m)とされてもよい。
【００１６】
図３には、音響信号Ｓxから生成される強度スペクトルＸ(m)の時系列（…，Ｘ(m-1)，Ｘ(m)，Ｘ(m+1)，…）が例示されている。強度スペクトルＸ(m)は、周波数軸上の所定の周波数帯域（以下「全帯域」という）ＢＦ内に分布する。全帯域ＢＦは、例えば０ｋＨｚから８ｋＨｚまでの範囲である。
【００１７】
音響信号Ｓxが表す混合音は、周波数帯域ＢＬの成分と周波数帯域ＢＨの成分とを含む。周波数帯域ＢＬおよび周波数帯域ＢＨは、全帯域ＢＦ内の相異なる周波数帯域である。周波数帯域ＢＬは周波数帯域ＢＨよりも低域側に位置する。具体的には、周波数帯域ＢＬは、全帯域ＢＦのうち周波数軸上の所定の周波数を下回る帯域であり、周波数帯域ＢＨは、全帯域ＢＦのうち当該周波数を上回る帯域である。したがって、周波数帯域ＢＬと周波数帯域ＢＨとは相互に重複しない。例えば、周波数帯域ＢＬは０ｋＨｚから４ｋＨｚまでの範囲であり、周波数帯域ＢＨは４ｋＨｚから８ｋＨｚまでの範囲である。なお、周波数帯域ＢＬの帯域幅と周波数帯域ＢＨの帯域幅との異同は不問である。混合音を構成する第１音および第２音の各々は、周波数帯域ＢＬの成分と周波数帯域ＢＨの成分との双方を含む。なお、周波数帯域ＢＬは「第１周波数帯域」の一例であり、周波数帯域ＢＨは「第２周波数帯域」の一例である。
【００１８】
図２の音源分離部２２は、強度スペクトルＸ(m)に対する音源分離を実行する。具体的には、音源分離部２２は、全帯域ＢＦにわたる強度スペクトルＸ(m)のうち周波数帯域ＢＬの成分を対象として音源分離を実行する。すなわち、強度スペクトルＸ(m)のうち周波数帯域ＢＨの成分については音源分離の処理対象から除外される。
【００１９】
音源分離部２２による強度スペクトルＸ(m)の処理には、公知の音源分離が任意に採用される。例えば、独立成分分析（ICA：Independent Component Analysis），独立ベクトル分析（IVA：Independent Vector Analysis），非負行列因子分解（NMF：Non-negative Matrix Factorization），多チャンネル非負行列因子分解（MNMF：Multichannel NMF），独立低ランク行列分析（ILRMA：Independent Low-Rank Matrix Analysis），独立低ランクテンソル分析（ILRTA：Independent Low-Rank Tensor Analysis），または独立深層学習行列分析（IDLMA：Independent Deeply-Learned Matrix Analysis）等の技術が、音源分離部２２による音源分離に利用される。なお、以上の説明では周波数領域における音源分離を例示したが、音源分離部２２は、時間領域における音源分離を音響信号Ｓxに対して実行してもよい。
【００２０】
音源分離部２２は、強度スペクトルＸ(m)のうち周波数帯域ＢＬの成分に対する音源分離

により強度スペクトルＹ1(m)と強度スペクトルＹ2(m)とを生成する。図３に例示される通り、強度スペクトルＹ1(m)は、混合音に含まれる第１音のうち周波数帯域ＢＬ内の成分（以下「第１成分」という）のスペクトルを意味する。すなわち、強度スペクトルＹ1(m)は、混合音のうち周波数帯域ＢＬ内の成分に含まれる第１音を第２音に対して強調した結果（理想的には第２音を除去した結果）を表すスペクトルである。他方、強度スペクトルＹ2(m)は、混合音に含まれる第２音のうち周波数帯域ＢＬ内の成分（以下「第２成分」という）のスペクトルを意味する。すなわち、強度スペクトルＹ2(m)は、混合音のうち周波数帯域ＢＬ内の成分に含まれる第２音を第１音に対して強調した結果（理想的には第１音を除去した結果）を表すスペクトルである。以上の説明から理解される通り、混合音のうち周波数帯域ＢＨの成分は、強度スペクトルＹ1(m)および強度スペクトルＹ2(m)には含まれない。

【００２１】

以上の通り、第１実施形態においては、音響信号Ｓxが表す混合音のうち周波数帯域ＢＨの成分が音源分離の対象から除外される。したがって、周波数帯域ＢＬおよび周波数帯域ＢＨの双方を含む全帯域ＢＦを対象として混合音の音源分離を実行する構成と比較して、音源分離部２２による処理負荷が軽減される。

【００２２】

図２の帯域拡張部２３は、混合音の強度スペクトルＸ(m)と第１成分の強度スペクトルＹ1(m)と第２成分の強度スペクトルＹ2(m)とを利用して出力データＯ(m)を生成する。出力データＯ(m)は、第１出力データＯ1(m)と第２出力データＯ2(m)とで構成される。第１出力データＯ1(m)は、強度スペクトルＺ1(m)を表すデータであり、第２出力データＯ2(m)は、強度スペクトルＺ2(m)を表すデータである。

【００２３】

第１出力データＯ1(m)が表す強度スペクトルＺ1(m)は、図３に例示される通り、周波数帯域ＢＬと周波数帯域ＢＨとを含む全帯域ＢＦにわたる第１音のスペクトルである。すなわち、音源分離において周波数帯域ＢＬに制限された第１音の強度スペクトルＹ1(m)が、帯域拡張部２３の処理により、全帯域ＢＦにわたる強度スペクトルＺ1(m)に変換される。他方、第２出力データＯ2(m)が表す強度スペクトルＺ2(m)は、全帯域ＢＦにわたる第２音のスペクトルである。すなわち、音源分離において周波数帯域ＢＬに制限された第２音の強度スペクトルＹ2(m)が、帯域拡張部２３の処理により、全帯域ＢＦにわたる強度スペクトルＺ2(m)に変換される。以上の説明から理解される通り、帯域拡張部２３は、第１音および第２音の各々の周波数帯域を、周波数帯域ＢＬから全帯域ＢＦ（周波数帯域ＢＬおよび周波数帯域ＢＨ）に拡張する。

【００２４】

図２に例示される通り、帯域拡張部２３は、取得部２３１と生成部２３２とを具備する。取得部２３１は、単位期間毎に入力データＤ(m)を生成する。入力データＤ(m)は、混合音の強度スペクトルＸ(m)と第１成分の強度スペクトルＹ1(m)と第２成分の強度スペクトルＹ2(m)とに応じたベクトルを表すデータである。

【００２５】

図３に例示される通り、入力データＤ(m)は、混合音データＤx(m)と第１入力データＤ1(m)と第２入力データＤ2(m)とを含む。混合音データＤx(m)は、混合音の強度スペクトルＸ(m)を表すデータである。具体的には、任意の１個の単位期間（以下「目標期間」という）について生成される混合音データＤx(m)は、当該目標期間の強度スペクトルＸ(m)と、目標期間の周囲に位置する他の単位期間の強度スペクトルＸ（Ｘ(m-4)，Ｘ(m-2)，Ｘ(m+2)，Ｘ(m+4)）とを含む。具体的には、混合音データＤx(m)は、目標期間の強度スペクトルＸ(m)と、目標期間の２個前の単位期間の強度スペクトルＸ(m-2)と、目標期間の４個前の単位期間の強度スペクトルＸ(m-4)と、目標期間の２個後の単位期間の強度スペクトルＸ(m+2)と、目標期間の４個後の単位期間の強度スペクトルＸ(m+4)とを含む。

【００２６】

第１入力データＤ1(m)は、第１音の強度スペクトルＹ1(m)を表すデータである。具体

的には、任意の１個の目標期間について生成される第１入力データＤ1(m)は、当該目標期間の強度スペクトルＹ1(m)と、目標期間の周囲に位置する他の単位期間の強度スペクトルＹ1（Ｙ1(m-4)，Ｙ1(m-2)，Ｙ1(m+2)，Ｙ1(m+4)）とを含む。具体的には、第１入力データＤ1(m)は、目標期間の強度スペクトルＹ1(m)と、目標期間の２個前の単位期間の強度スペクトルＹ1(m-2)と、目標期間の４個前の単位期間の強度スペクトルＹ1(m-4)と、目標期間の２個後の単位期間の強度スペクトルＹ1(m+2)と、目標期間の４個後の単位期間の強度スペクトルＹ1(m+4)とを含む。以上の説明から理解される通り、第１入力データＤ1(m)は、第１音のうち周波数帯域ＢＬ内の第１成分を表すデータである。

【００２７】

第２入力データＤ2(m)は、第２音の強度スペクトルＹ2(m)を表すデータである。具体的には、任意の１個の目標期間について生成される第２入力データＤ2(m)は、当該目標期間の強度スペクトルＹ2(m)と、目標期間の周囲に位置する他の単位期間の強度スペクトルＹ2（Ｙ2(m-4)，Ｙ2(m-2)，Ｙ2(m+2)，Ｙ2(m+4)）とを含む。具体的には、第２入力データＤ2(m)は、目標期間の強度スペクトルＹ2(m)と、目標期間の２個前の単位期間の強度スペクトルＹ2(m-2)と、目標期間の４個前の単位期間の強度スペクトルＹ2(m-4)と、目標期間の２個後の単位期間の強度スペクトルＹ2(m+2)と、目標期間の４個後の単位期間の強度スペクトルＹ2(m+4)とを含む。以上の説明から理解される通り、第２入力データＤ2(m)は、第２音のうち周波数帯域ＢＬ内の第２成分を表すデータである。

【００２８】

入力データＤ(m)の全体で表現されるベクトルＶの各要素は、当該ベクトルＶの大きさが１（すなわち単位ベクトル）となるように正規化される。例えば、正規化前の入力データＤ(m)において、第１入力データＤ1(m)と第２入力データＤ2(m)と混合音データＤx(m)とにより、Ｎ個の要素ｅ1～ｅNが配列されたＮ次元のベクトルＶが構成されると想定する。正規化後の入力データＤ(m)を構成するＮ個の要素Ｅ1～ＥNの各々は、以下の数式(1)で表現される（ｎ＝１～Ｎ）。

【数１】

$$E_n = \frac{e_n}{\|V\|_2} \quad (1)$$

【００２９】

数式(1)の記号‖‖2は、以下の数式(2)で表現されるＬ2ノルムを意味し、ベクトルＶの大きさを表す指標（以下「強度指標α」という）に相当する。

【数２】

$$\|V\|_2 = \alpha = \left(\sum_{n=1}^{N} |e_n|^2 \right)^{\frac{1}{2}} \quad (2)$$

【００３０】

図２の生成部２３２は、入力データＤ(m)から出力データＯ(m)を生成する。出力データＯ(m)は、単位期間毎に順次に生成される。具体的には、生成部２３２は、各単位期間の

入力データＤ(m)から当該単位期間の出力データＯ(m)を生成する。出力データＯ(m)の生成には推定モデルＭが利用される。推定モデルＭは、入力データＤ(m)を入力として出力データＯ(m)を出力する統計的モデルである。すなわち、推定モデルＭは、入力データＤ(m)と出力データＯ(m)との関係を学習した学習済モデルである。

【００３１】

推定モデルＭは、例えばニューラルネットワークで構成される。図４は、推定モデルＭの構造を例示するブロック図である。推定モデルＭは、例えば、入力層Ｌinと出力層Ｌoutとの間の隠れ層Ｌhに４層の全結合層Ｌaを含む深層ニューラルネットワークである。活性化関数は、例えばReLU（Rectified Linear Unit）である。入力データＤ(m)は、隠れ層Ｌhの第１層において出力層Ｌoutと同等の次元数に圧縮される。なお、推定モデルＭの構造は以上の例示に限定されない。例えば、再帰型ニューラルネットワーク（RNN：Recurrent Neural Network）、または畳込ニューラルネットワーク（CNN：Convolutional Neural Network）等の任意の形式のニューラルネットワークが推定モデルＭとして利用される。複数種のニューラルネットワークの組合せが推定モデルＭとして利用されてもよい。また、長短期記憶（LSTM：Long Short-Term Memory）等の付加的な要素が推定モデルＭに搭載されてもよい。

【００３２】

推定モデルＭは、入力データＤ(m)から出力データＯ(m)を生成する演算を制御装置１１に実行させる推定プログラムと、当該演算に適用される複数の変数Ｋ（具体的には加重値およびバイアス）との組合せで実現される。推定プログラムと複数の変数Ｋとは記憶装置１２に記憶される。複数の変数Ｋの各々の数値は、機械学習により事前に設定される。

【００３３】

図２の波形合成部２４は、帯域拡張部２３が順次に生成する出力データＯ(m)の時系列から音響信号Ｓz0を生成する。具体的には、波形合成部２４は、第１出力データＯ1(m)および第２出力データＯ2(m)の何れかの時系列から音響信号Ｓz0を生成する。例えば、第１音の強調が利用者から指示された場合、波形合成部２４は、第１出力データＯ1(m)（強度スペクトルＺ1(m)）の時系列から音響信号Ｓz0を生成する。すなわち、第１音が強調された音響信号Ｓz0が生成される。他方、第２音の強調が利用者から指示された場合、波形合成部２４は、第２出力データＯ2(m)（強度スペクトルＺ2(m)）の時系列から音響信号Ｓz0を生成する。すなわち、第２音が強調された音響信号Ｓz0が生成される。音響信号Ｓz0の生成には、例えば短時間逆フーリエ変換が利用される。

【００３４】

前述の通り、入力データＤ(m)を構成する各要素Ｅnは、強度指標αを利用して正規化された数値である。したがって、音響信号Ｓz0の音量は、音響信号Ｓxとは相違する可能性がある。音量調整部２５は、音響信号Ｓz0の音量を音響信号Ｓxと同等の音量に調整すること（すなわちスケーリング）で音響信号Ｓzを生成する。音響信号Ｓzが放音装置１３に供給されることで音波として放射される。具体的には、音量調整部２５は、音響信号Ｓxの音量と音響信号Ｓz0の音量との相違に応じた調整値Ｇを音響信号Ｓz0に乗算することで音響信号Ｓzを生成する。調整値Ｇは、音響信号Ｓxと音響信号Ｓzとの音量差が最小化されるように設定される。

【００３５】

図５は、制御装置１１が音響信号Ｓxから音響信号Ｓzを生成する処理（以下「音響処理Ｓa」という）の具体的な手順を例示するフローチャートである。例えば音響処理システム１００に対する利用者からの指示を契機として音響処理Ｓaが開始される。

【００３６】

音響処理Ｓaが開始されると、制御装置１１（周波数解析部２１）は、複数の単位期間の各々について音響信号Ｓxの強度スペクトルＸ(m)を生成する（Ｓa1）。制御装置１１（音源分離部２２）は、強度スペクトルＸ(m)のうち周波数帯域ＢＬ内の成分に対する音源分離により各単位期間の強度スペクトルＹ1(m)と強度スペクトルＹ2(m)とを生成する（Ｓa2）。

【００３７】
制御装置１１（取得部２３１）は、強度スペクトルＸ(m)と強度スペクトルＹ1(m)と強度スペクトルＹ2(m)とから各単位期間の入力データＤ(m)を生成する（Ｓa3）。制御装置１１（生成部２３２）は、入力データＤ(m)を推定モデルＭに入力することで各単位期間の出力データＯ(m)を生成する（Ｓa4）。制御装置１１（波形合成部２４）は、第１出力データＯ1(m)または第２出力データＯ2(m)の時系列から音響信号Ｓz0を生成する（Ｓa5）。制御装置１１（音量調整部２５）は、音響信号Ｓz0に調整値Ｇを乗算することで音響信号Ｓzを生成する（Ｓa6）。
【００３８】
以上に説明した通り、第１実施形態においては、周波数帯域ＢＬの成分を表す第１入力データＤ1(m)および第２入力データＤ2(m)を含む入力データＤ(m)から、周波数帯域ＢＬを含む全帯域ＢＦの音を表す出力データＯ(m)が生成される。すなわち、音響信号Ｓxが表す混合音のうち周波数帯域ＢＬについてのみ限定的に音源分離を実行する構成にも関わらず、全帯域ＢＦの成分を含む出力データＯ(m)が生成される。したがって、音源分離のための処理負荷を軽減できる。
【００３９】
［２］学習処理部３０
図２に例示される通り、制御装置１１は、記憶装置１２に記憶された機械学習プログラムＰ2を実行することで学習処理部３０として機能する。学習処理部３０は、音響処理Ｓaに利用される推定モデルＭを機械学習により確立する。学習処理部３０は、取得部３１と訓練部３２とを具備する。
【００４０】
記憶装置１２には、推定モデルＭの機械学習に利用される複数の訓練データＴが記憶される。図６は、訓練データＴの説明図である。複数の訓練データＴの各々は、訓練用の入力データＤt(m)と訓練用の出力データＯt(m)との組合せで構成される。図３の入力データＤ(m)と同様に、訓練用の入力データＤt(m)は、混合音データＤx(m)と第１入力データＤ1(m)と第２入力データＤ2(m)とを含む。
【００４１】
図６には、参照信号Ｓrと第１信号Ｓr1と第２信号Ｓr2とが図示されている。参照信号Ｓrは、第１音源から発音される第１音と第２音源から発音される第２音との混合音を表す時間領域の信号である。参照信号Ｓrが表す混合音は、周波数帯域ＢＬと周波数帯域ＢＨとを含む全帯域ＢＦにわたる。参照信号Ｓrは、例えば、第１音源と第２音源とが並列に発音する環境において収音装置を利用して収録される。また、第１信号Ｓr1は、第１音を表す時間領域の信号であり、第２信号Ｓr2は、第２音を表す時間領域の信号である。第１音および第２音の各々は、周波数帯域ＢＬと周波数帯域ＢＨとを含む全帯域ＢＦにわたる。第１信号Ｓr1は、第１音源のみが発音する環境において収録され、第２信号Ｓr2は、第２音源のみが発音する環境において収録される。なお、相互に個別に収録された第１信号Ｓr1と第２信号Ｓr2とを混合することで参照信号Ｓrが生成されてもよい。
【００４２】
図６には、参照信号Ｓrの強度スペクトルＸ(m)の時系列（…，Ｘ(m-1)，Ｘ(m)，Ｘ(m+1)，…）と、第１信号Ｓr1の強度スペクトルＲ1(m)の時系列（…，Ｒ1(m-1)，Ｒ1(m)，Ｒ1(m+1)，…）と、第２信号Ｓr2の強度スペクトルＲ2(m)の時系列（…，Ｒ2(m-1)，Ｒ2(m)，Ｒ2(m+1)，…）とが図示されている。訓練用の入力データＤt(m)のうちの混合音データＤx(m)は、参照信号Ｓrの強度スペクトルＸ(m)から生成される。具体的には、任意の１個の目標期間の混合音データＤx(m)は、図３の例示と同様に、当該目標期間の強度スペクトルＸ(m)と、目標期間の周囲に位置する他の単位期間の強度スペクトルＸ（Ｘ(m-4)，Ｘ(m-2)，Ｘ(m+2)，Ｘ(m+4)）とを含む。
【００４３】
第１信号Ｓr1は、周波数帯域ＢＬの成分と周波数帯域ＢＨの成分とを含む。第１信号Ｓr1の強度スペクトルＲ1(m)は、周波数帯域ＢＬ内の強度スペクトルＹ1(m)と周波数帯域Ｂ

H内の強度スペクトルＨ1(m)とで構成される。訓練用の入力データＤt(m)の第１入力データＤ1(m)は、周波数帯域ＢＬの強度スペクトルＹ1(m)を表すデータである。具体的には、目標期間の第１入力データＤ1(m)は、当該目標期間の強度スペクトルＹ1(m)と、当該目標期間の周囲に位置する他の単位期間の強度スペクトルＹ1（Ｙ1(m-4)，Ｙ1(m-2)，Ｙ1(m+2)，Ｙ1(m+4)）とを含む。

【００４４】

第１信号Ｓr1と同様に、第２信号Ｓr2は、周波数帯域ＢＬの成分と周波数帯域ＢＨの成分とを含む。第２信号Ｓr2の強度スペクトルＲ2(m)は、周波数帯域ＢＬ内の強度スペクトルＹ2(m)と周波数帯域ＢＨ内の強度スペクトルＨ2(m)とで構成される。訓練用の入力データＤt(m)の第２入力データＤt2(m)は、周波数帯域ＢＬの強度スペクトルＹ2(m)を表すデータである。具体的には、目標期間の第２入力データＤt2(m)は、当該目標期間の強度スペクトルＹ2(m)と、目標期間の周囲に位置する他の単位期間の強度スペクトルＹ2（Ｙ2(m-4)，Ｙ2(m-2)，Ｙ2(m+2)，Ｙ2(m+4)）とを含む。

【００４５】

他方、各訓練データＴを構成する訓練用の出力データＯt(m)は、第１出力データＯt1(m)と第２出力データＯt2(m)とで構成される正解データである。第１出力データＯt1(m)は、第１信号Ｓr1の強度スペクトルＲ1(m)を表す。すなわち、第１出力データＯt1(m)は、参照信号Ｓrが表す混合音のうち全帯域ＢＦにわたる第１音のスペクトルである。第２出力データＯt2(m)は、第２信号Ｓr2の強度スペクトルＲ2(m)を表す。すなわち、第２出力データＯt2(m)は、参照信号Ｓrが表す混合音のうち全帯域ＢＦにわたる第２音のスペクトルである。

【００４６】

訓練用の入力データＤt(m)の全体で表現されるベクトルＶの各要素は、前述の入力データＤt(m)と同様に、当該ベクトルＶの大きさが１となるように正規化される。同様に、訓練用の出力データＯt(m)の全体で表現されるベクトルＶの各要素は、当該ベクトルＶの大きさが１となるように正規化される。

【００４７】

図２の取得部３１は、複数の訓練データＴの各々を記憶装置１２から取得する。なお、参照信号Ｓrと第１信号Ｓr1と第２信号Ｓr2とが記憶装置１２に記憶された構成においては、取得部３１が参照信号Ｓrと第１信号Ｓr1と第２信号Ｓr2とから複数の訓練データＴを生成する。すなわち、取得部３１による「取得」は、事前に用意された訓練データＴを記憶装置１２から読出する処理のほか、当該取得部３１自身が訓練データＴを生成する処理も包含する。

【００４８】

訓練部３２は、複数の訓練データＴを利用した処理（以下「学習処理Ｓb」という）により推定モデルＭを確立する。学習処理Ｓbは、複数の訓練データＴを利用した教師あり機械学習である。具体的には、訓練部３２は、各訓練データＴの入力データＤt(m)を入力した場合に暫定的な推定モデルＭが生成する出力データＯ(m)と、当該訓練データＴに含まれる出力データＯt(m)との誤差を表す損失関数Ｌが低減（理想的には最小化）されるように、推定モデルＭを規定する複数の変数Ｋを反復的に更新する。したがって、推定モデルＭは、複数の訓練データＴにおける入力データＤt(m)と出力データＯt(m)との間に潜在する関係を学習する。すなわち、訓練部３２による訓練後の推定モデルＭは、未知の入力データＤ(m)に対して当該関係のもとで統計的に妥当な出力データＯ(m)を出力する。

【００４９】

損失関数Ｌは、例えば以下の数式(3)で表現される。

【数３】

$$L = \varepsilon[O_1(m), O_{t1}(m)] + \varepsilon[O_2(m), O_{t2}(m)] \ldots (3)$$

数式(3)の記号ε[a,b]は、要素ａと要素ｂとの誤差（例えば平均二乗誤差またはクロスエントロピー関数）である。

【００５０】

図７は、学習処理Ｓｂの具体的な手順を例示するフローチャートである。例えば音響処理システム１００に対する利用者からの指示を契機として学習処理Ｓｂが開始される。

【００５１】

制御装置１１（取得部３１）は、訓練データＴを記憶装置１２から取得する（Ｓｂ1）。制御装置１１（訓練部３２）は、当該訓練データＴを利用した機械学習を実行する（Ｓｂ2）。すなわち、訓練データＴの入力データＤt(m)から推定モデルＭが生成する出力データＯ(m)と、当該訓練データＴの出力データＯt(m)（すなわち正解値）との間の損失関数Ｌが低減されるように、推定モデルＭの複数の変数Ｋを反復的に更新する。損失関数Ｌに応じた複数の変数Ｋの更新には、例えば誤差逆伝播法が利用される。

【００５２】

制御装置１１は、学習処理Ｓｂに関する終了条件が成立したか否かを判定する（Ｓｂ3）。終了条件は、例えば、損失関数Ｌが所定の閾値を下回ること、または、損失関数Ｌの変化量が所定の閾値を下回ることである。終了条件が成立しない場合（Ｓｂ3：ＮＯ）、制御装置１１（取得部３１）は、未取得の訓練データＴを記憶装置１２から取得する（Ｓｂ1）。すなわち、終了条件の成立まで、訓練データＴの取得（Ｓｂ1）と当該訓練データＴを利用した複数の変数Ｋの更新（Ｓｂ2）とが反復される。終了条件が成立した場合（Ｓｂ3：ＹＥＳ）、制御装置１１は学習処理Ｓｂを終了する。

【００５３】

以上に説明した通り、第１実施形態においては、周波数帯域ＢＬの成分を表す第１入力データＤ1(m)および第２入力データＤ2(m)を含む入力データＤ(m)から、周波数帯域ＢＬおよび周波数帯域ＢＨの音を表す出力データＯ(m)が生成されるように、推定モデルＭが確立される。すなわち、音響信号Ｓxが表す混合音のうち周波数帯域ＢＬについてのみ限定的に音源分離を実行する構成でも、推定モデルＭを利用することで、周波数帯域ＢＨの成分を含む出力データＯ(m)が生成される。したがって、音源分離のための処理負荷を軽減できる。

【００５４】

Ｂ：第２実施形態

第２実施形態について以下に説明する。なお、以下に例示する各形態において機能が第１実施形態と同様である要素については、第１実施形態の説明で使用した符号を流用して各々の詳細な説明を適宜に省略する。

【００５５】

第１実施形態においては、混合音データＤx(m)が周波数帯域ＢＬの成分と周波数帯域ＢＨの成分とを双方を含む構成を例示した。しかし、第１音のうち周波数帯域ＢＬ内の成分は第１入力データＤ1(m)に含まれ、第２音のうち周波数帯域ＢＨ内の成分は第２入力データＤ2(m)に含まれるから、混合音データＤx(m)が周波数帯域ＢＬの成分を含む構成は必須ではない。以上の事情を考慮して、第２実施形態においては、混合音データＤx(m)が混合音のうち周波数帯域ＢＬの成分を含まない。

【００５６】

図８は、第２実施形態における入力データＤ(m)の模式図である。音響信号Ｓxの強度スペクトルＸ(m)は、周波数帯域ＢＬ内の強度スペクトルＸＬ(m)と周波数帯域ＢＨ内の強度スペクトルＸＨ(m)とに分割される。入力データＤ(m)の混合音データＤx(m)は、周波数帯域ＢＨの強度スペクトルＸＨ(m)を表すデータである。具体的には、１個の目標期間について生成される混合音データＤx(m)は、当該目標期間の強度スペクトルＸＨ(m)と、当該目標期間の周囲に位置する他の単位期間の強度スペクトルＸＨ（ＸＨ(m-4)，ＸＨ(m-2)，ＸＨ(m+2)，ＸＨ(m+4)）とを含む。すなわち、第２実施形態の混合音データＤx(m)は、混合音のうち周波数帯域ＢＬの成分（強度スペクトルＸＬ(m)）を含まない。なお、音源分離部２２が強度スペクトルＸ(m)のうち周波数帯域ＢＬの成分を対象として音源分離を実行する点

は第１実施形態と同様である。

【００５７】

以上の説明においては、音響処理Ｓａに利用される入力データＤ(m)を例示したが、学習処理Ｓｂに利用される訓練用の入力データＤｔ(m)についても同様に、参照信号Ｓｒが表す混合音のうち周波数帯域ＢＨの成分を表す混合音データＤｘ(m)が含まれる。すなわち、訓練用の混合音データＤｘ(m)は、参照信号Ｓｒの強度スペクトルＸ(m)のうち周波数帯域ＢＨ内の強度スペクトルＸＨ(m)を表し、周波数帯域ＢＬ内の強度スペクトルＸＬ(m)は混合音データＤｘ(m)に反映されない。

【００５８】

第２実施形態においても第１実施形態と同様の効果が実現される。また、第２実施形態においては、混合音データＤｘ(m)が混合音のうち周波数帯域ＢＬの成分を含まない。したがって、混合音データＤｘ(m)が全帯域ＢＦの成分を含む構成と比較して、学習処理Ｓｂの処理負荷および推定モデルＭの規模が低減されるという利点がある。

【００５９】

第１実施形態においては、全帯域ＢＦにわたる混合音を表す混合音データＤｘ(m)を例示した。第２実施形態においては、混合音のうち周波数帯域ＢＨの成分を表す混合音データＤｘ(m)を例示した。以上の例示から理解される通り、混合音データＤｘ(m)は、混合音のうち周波数帯域ＢＨを含む周波数帯域の成分を表すデータとして包括的に表現される。

【００６０】

Ｃ：第３実施形態

図９は、第３実施形態における入力データＤ(m)の模式図である。第３実施形態の入力データＤ(m)は、混合音データＤｘ(m)と第１入力データＤ1(m)と第２入力データＤ2(m)とに加えて強度指標αを含む。強度指標αは、前述の通り、入力データＤ(m)の全体で表現されるベクトルＶの大きさ（例えばＬ2ノルム）を表す指標であり、前掲の数式(2)で算定される。学習処理Ｓｂに利用される訓練用の入力データＤｔ(m)についても同様に、混合音データＤｘ(m)と第１入力データＤ1(m)と第２入力データＤ2(m)とに加えて、当該入力データＤｔ(m)で表現されるベクトルＶの大きさに応じた強度指標αが含まれる。なお、混合音データＤｘ(m)と第１入力データＤ1(m)と第２入力データＤ2(m)とは、第１実施形態または第２実施形態と同様である。

【００６１】

図１０は、第３実施形態に係る音響処理システム１００の機能的な構成を例示するブロック図である。第３実施形態の入力データＤ(m)には強度指標αが含まれるから、当該強度指標αが反映された出力データＯ(t)が推定モデルＭから出力される。具体的には、波形合成部２４が出力データＯ(t)から生成する音響信号Ｓｚは、音響信号Ｓｘと同等の音量となる。したがって、第１実施形態において例示した音量調整部２５（図５のステップＳａ6）が第３実施形態においては省略される。すなわち、波形合成部２４による出力信号（第１実施形態における音響信号Ｓｚ0）が最終的な音響信号Ｓｚとして出力される。

【００６２】

第３実施形態においても第１実施形態と同様の効果が実現される。また、第３実施形態においては、強度指標αが入力データＤ(m)に含まれるから、混合音に対応する音量の音を表す出力データＯ(m)が生成される。したがって、第１出力データＯ1(m)および第２出力データＯ2(m)が表す音の強度を調整する処理（音量調整部２５）が不要であるという利点がある。

【００６３】

図１１は、第１実施形態および第３実施形態による効果の説明図である。図１１の結果Ａは、第１実施形態により生成された音響信号Ｓｚの振幅スペクトログラムであり、図１１の結果Ｂは、第３実施形態により生成された音響信号Ｓｚの振幅スペクトログラムである。結果Ａおよび結果Ｂにおいては、打楽器音（第１音）と歌唱音声（第２音）との混合音を表す音響信号Ｓｘに対して音響処理Ｓａを実行することで、打楽器音を表す音響信号Ｓｚを生成した場合が想定されている。図１１の正解Ｃは、単独で発音された打楽器音の振幅ス

ペクトログラムである。

【００６４】

図１１の結果Ａからは、第１実施形態により、正解Ｃに近い音響信号Ｓｚを生成できることが確認できる。また、図１１の結果Ｂからは、入力データＤ(m)が強度指標αを含む第３実施形態により、第１実施形態と比較しても正解Ｃに充分に近い音響信号Ｓｚを生成できることが確認される。

【００６５】

図１２は、第１実施形態から第３実施形態に関する観測結果の図表である。図１２においては、打楽器音（第１音）と歌唱音声（第２音）との混合音を表す音響信号Ｓｘに対して音響処理Ｓａを実行することで、打楽器音（Drums）を表す音響信号Ｓｚと、歌唱音声（Vocals）を表す音響信号Ｓｚとを生成した場合が想定されている。図１２には、評価指標として有効なＳＡＲ（信号対非線形歪比：Sources to Artifacts Ratio）およびＳＡＲ改善量が、第１実施形態から第３実施形態の各々について図示されている。ＳＡＲ改善量は、比較例を基準としたＳＡＲの改善量である。比較例については、音響信号Ｓｚのうち周波数帯域ＢＨの成分を一律にゼロとした場合のＳＡＲが基準として例示されている。

【００６６】

第１実施形態および第２実施形態においてもＳＡＲが改善することが図１２から確認できる。また、第３実施形態によれば、打楽器音および歌唱音声の何れについても、第１実施形態および第２実施形態と比較して非常に高精度な音源分離が実現されることが図１２から確認できる。

【００６７】

Ｄ：第４実施形態

第４実施形態の学習処理Ｓｂにおいては、前掲の数式(3)で表現される損失関数Ｌが、以下の数式(4)で表現される損失関数Ｌに置換される。

【数４】

$$L = \varepsilon[O_1(m), O_{t1}(m)] + \varepsilon[O_2(m), O_{t2}(m)] + \varepsilon[X_H(m), O_{1H}(m) + O_{2H}(m)] \quad (4)$$

【００６８】

数式(4)における記号Ｏ1Ｈ(m)は、第１出力データＯ1(m)が表す強度スペクトルＺ1(m)のうち周波数帯域ＢＨ内の強度スペクトルであり、記号Ｏ2Ｈ(m)は、第２出力データＯ2(m)が表す強度スペクトルＺ2(m)のうち周波数帯域ＢＨ内の強度スペクトルである。すなわち、数式(4)の右辺における第３項は、参照信号Ｓｒの強度スペクトルＸ(m)のうち周波数帯域ＢＨ内の強度スペクトルＸＨ(m)と、強度スペクトルＨ1(m)および強度スペクトルＨ2(m)の合計（Ｈ1(m)＋Ｈ2(m)）との誤差を意味する。以上の説明から理解される通り、第４実施形態の訓練部３２は、強度スペクトルＺ1(m)のうち周波数帯域ＢＨ内の成分と、強度スペクトルＺ2(m)のうち周波数帯域ＢＨ内の成分とを混合した結果が、混合音の強度スペクトルＸ(m)のうち周波数帯域ＢＨの成分（強度スペクトルＸＨ(m)）に近似または一致するという条件（以下「追加条件」という）のもとで、推定モデルＭを訓練する。

【００６９】

第４実施形態においても第１実施形態と同様の効果が実現される。また、第４実施形態によれば、追加条件なしで訓練された推定モデルＭを利用する構成と比較して、第１音のうち周波数帯域ＢＨの成分（第１出力データＯ1(m)）と第２音のうち周波数帯域ＢＨの成分（第２出力データＯ2(m)）とを高精度に推定できる。なお、第４実施形態の構成は、第２実施形態および第３実施形態にも同様に適用される。

【００７０】
Ｅ：第５実施形態
図１３は、第５実施形態における入力データＤ(m)および出力データＯ(m)の模式図である。第１実施形態の出力データＯ(m)における第１出力データＯ1(m)は、全帯域ＢＦにわたる強度スペクトルＺ1(m)を表し、第２出力データＯ2(m)は、全帯域ＢＦにわたる強度スペクトルＺ2(m)を表す。第５実施形態における第１出力データＯ1(m)は、第１音のうち周波数帯域ＢＨの成分を表す。すなわち、第１出力データＯ1(m)は、第１音の強度スペクトルＺ1(m)のうち周波数帯域ＢＨ内の強度スペクトルＨ1(m)を表し、周波数帯域ＢＬ内の強度スペクトルを含まない。同様に、第５実施形態における第２出力データＯ2(m)は、第２音のうち周波数帯域ＢＨの成分を表す。すなわち、第２出力データＯ2(m)は、第２音の強度スペクトルＺ2(m)のうち周波数帯域ＢＨ内の強度スペクトルＨ2(m)を表し、周波数帯域ＢＬ内の強度スペクトルを含まない。
【００７１】
図１４は、第５実施形態における訓練用の入力データＤt(m)および出力データＯt(m)の模式図である。第１実施形態において、訓練用の出力データＯt(m)における第１出力データＯt1(m)は、全帯域ＢＦにわたる第１音の強度スペクトルＲ1(m)を表し、第２出力データＯt2(m)は、全帯域ＢＦにわたる第２音の強度スペクトルＲ2(m)を表す。第５実施形態における第１出力データＯt1(m)は、第１音のうち周波数帯域ＢＨの成分を表す。すなわち、第１出力データＯt1(m)は、第１音の強度スペクトルＲ1(m)のうち周波数帯域ＢＨ内の強度スペクトルＨ1(m)を表し、周波数帯域ＢＬ内の強度スペクトルＹ1(m)を含まない。同様に、第５実施形態における第２出力データＯt2(m)は、第２音のうち周波数帯域ＢＨの成分を表す。すなわち、第２出力データＯt2(m)は、第２音の強度スペクトルＲ2(m)のうち周波数帯域ＢＨ内の強度スペクトルＨ2(m)を表し、周波数帯域ＢＬ内の強度スペクトルＹ2(m)を含まない。
【００７２】
図１５は、第５実施形態における音響処理部２０の部分的な構成を例示するブロック図である。第５実施形態の波形合成部２４には、第１音のうち周波数帯域ＢＨ内の強度スペクトルＨ1(m)を表す第１出力データＯ1(m)が音響処理部２０から供給されるほか、第１音のうち周波数帯域ＢＬ内の強度スペクトルＹ1(m)が音源分離部２２から供給される。第１音の強調が利用者から指示された場合、波形合成部２４は、強度スペクトルＨ1(m)と強度スペクトルＹ1(m)とを合成することで全帯域ＢＦにわたる強度スペクトルＺ1(m)を生成し、強度スペクトルＺ1(m)の時系列から音響信号Ｓz0を生成する。
【００７３】
また、第５実施形態の波形合成部２４には、第２音のうち周波数帯域ＢＨ内の強度スペクトルＨ2(m)を表す第２出力データＯ2(m)が音響処理部２０から供給されるほか、第２音のうち周波数帯域ＢＬ内の強度スペクトルＹ2(m)が音源分離部２２から供給される。第２音の強調が利用者から指示された場合、波形合成部２４は、強度スペクトルＨ2(m)と強度スペクトルＹ2(m)とを合成することで全帯域ＢＦにわたる強度スペクトルＺ2(m)を生成し、強度スペクトルＺ2(m)の時系列から音響信号Ｓz0を生成する。
【００７４】
第５実施形態においても第１実施形態と同様の効果が実現される。また、第５実施形態においては、出力データＯ(m)が周波数帯域ＢＬの成分を含まない。したがって、出力データＯ(m)が全帯域ＢＦの成分を含む構成（例えば第１実施形態）と比較して、学習処理Ｓbの処理負荷および推定モデルＭの規模が低減されるという利点がある。他方、出力データＯ(m)が全帯域ＢＦの成分を含む第１実施形態によれば、第５実施形態と比較して、全帯域ＢＦにわたる音響を簡便に生成できるという利点がある。
【００７５】
第１実施形態においては、第１音のうち周波数帯域ＢＬと周波数帯域ＢＨとを含む全帯域ＢＦの成分を表す第１出力データＯ1(m)を例示した。第５実施形態においては、第１音のうち周波数帯域ＢＨの成分を表す第１出力データＯ1(m)を例示した。以上の例示から理解

される通り、第１出力データＯ1(m)は、第１音のうち周波数帯域ＢＨを含む周波数帯域の成分を表すデータとして包括的に表現される。同様に、第２出力データＯ2(m)は、第２音のうち周波数帯域ＢＨを含む周波数帯域の成分を表すデータとして包括的に表現される。

【００７６】

Ｆ：変形例

以上に例示した各態様に付加される具体的な変形の態様を以下に例示する。以下の例示から任意に選択された２以上の態様を、相互に矛盾しない範囲で適宜に併合してもよい。

【００７７】

（１）前述の各形態においては、目標期間の強度スペクトルＸ(m)と他の単位期間の強度スペクトルＸとを含む混合音データＤx(m)を例示したが、混合音データＤx(m)の内容は以上の例示に限定されない。例えば、目標期間の混合音データＤx(m)が当該目標期間の強度スペクトルＸ(m)のみを含む構成が想定される。目標期間の混合音データＤx(m)が、当該目標期間に対して過去および未来の一方の単位期間の強度スペクトルＸを含んでもよい。また、前述の各形態においては、目標期間の混合音データＤx(m)が、当該目標期間に間隔をあけて前後する他の単位期間の強度スペクトルＸ（Ｘ(m-4)，Ｘ(m-2)，Ｘ(m+2)，Ｘ(m+4)）を含む構成を例示したが、目標期間の直前の単位期間の強度スペクトルＸ(m-1)または直後の単位期間の強度スペクトルＸ(m+1)を混合音データＤx(m)が含んでもよい。

【００７８】

以上の説明においては混合音データＤx(m)に着目したが、第１入力データＤ1(m)および第２入力データＤ2(m)についても同様である。例えば、目標期間の第１入力データＤ1(m)は、当該目標期間の強度スペクトルＹ1(m)のみで構成されてもよいし、当該目標期間の過去および未来の一方の単位期間の強度スペクトルＹ1を含んでもよい。また、目標期間の第１入力データＤ1(m)が、当該目標期間の直前の単位期間の強度スペクトルＹ2(m-1)、または直後の単位期間の強度スペクトルＹ1(m+1)を含んでもよい。第２入力データＤ2(m)についても同様である。

【００７９】

（２）前述の各形態においては、所定の周波数を下回る周波数帯域ＢＬと当該周波数を上回る周波数帯域ＢＨとに着目したが、周波数帯域ＢＬと周波数帯域ＢＨとの関係は以上の例示に限定されない。例えば、周波数帯域ＢＬが所定の周波数を上回り、周波数帯域ＢＨが当該周波数を下回る構成も想定される。また、周波数帯域ＢＬおよび周波数帯域ＢＨの各々は、周波数軸上で連続する周波数帯域に限定されない。例えば、周波数軸を区分した複数の周波数帯域のうち奇数番目および偶数番目の一方に属する２以上の周波数帯域の集合が周波数帯域ＢＬとされ、奇数番目および偶数番目の他方に属する２以上の周波数帯域の集合が周波数帯域ＢＨとされてもよい。

【００８０】

（３）前述の各形態においては、事前に用意された音響信号Ｓxを処理する場合を例示したが、音響処理部２０は、音響信号Ｓxの収録に並行して実時間的に、音響信号Ｓxに対する音響処理Ｓaを実行してもよい。なお、前述の各形態における例示のように混合音データＤx(m)が目標期間の後方の強度スペクトルＸ(m+4)を含む構成では、単位期間の４個分に相当する時間長の遅延が発生する。

【００８１】

（４）前述の各形態においては、第１音が強調された強度スペクトルＺ1(m)を表す第１出力データＯ1(m)と第２音が強調された強度スペクトルＺ2(m)を表す第２出力データＯ2(m)との双方を帯域拡張部２３が生成したが、第１出力データＯ1(m)および第２出力データＯ2(m)の一方のみを出力データＯ(m)として帯域拡張部２３が生成してもよい。例えば、歌唱音声（第１音）と楽器音（第２音）との混合音に対する音響処理Ｓaで歌唱音声を抑制するという用途に使用される音響処理システム１００においては、第２音が強調された強度スペクトルＺ2(m)を表す出力データＯ(m)（第２出力データＯ2(m)）を帯域拡張部２３が生成すれば充分である。すなわち、第１音が強調された強度スペクトルＺ1(m)の生成は省略される。以上の説明から理解される通り、生成部２３２は、第１出力データＯ1(m)

および第２出力データＯ2(m)の少なくとも一方を生成する要素として表現される。
【００８２】
（５）前述の各形態においては、第１音および第２音の一方が強調された音響信号Ｓzを生成したが、音響処理部２０による処理の内容は以上の例示に限定されない。例えば、第１出力データＯ1(m)の時系列から生成される第１音響信号と第２出力データＯ2(m)の時系列から生成される第２音響信号との加重和を、音響処理部２０が音響信号Ｓzとして出力してもよい。第１音響信号は第１音が強調された信号であり、第２音響信号は第２音が強調された信号である。また、第１音響信号および第２音響信号の各々に対して、例えば効果付与等の音響処理を相互に独立に実行し、処理後の第１音響信号と第２音響信号とを加算することで、音響処理部２０が音響信号Ｓzを生成してもよい。
【００８３】
（６）携帯電話機またはスマートフォン等の端末装置との間で通信するサーバ装置により音響処理システム１００が実現されてもよい。例えば、音響処理システム１００は、端末装置から受信した音響信号Ｓxに対する音響処理Ｓaにより音響信号Ｓzを生成し、当該音響信号Ｓzを端末装置に送信する。端末装置に搭載された周波数解析部２１が生成した強度スペクトルＸ(m)を音響処理システム１００が受信する構成においては、音響処理システム１００から周波数解析部２１が省略される。また、波形合成部２４（および音量調整部２５）が端末装置に搭載された構成においては、帯域拡張部２３が生成した出力データＯ(m)が音響処理システム１００から端末装置に送信される。したがって、波形合成部２４および音量調整部２５は音響処理システム１００から省略される。
【００８４】
　また、周波数解析部２１および音源分離部２２は端末装置に搭載されてもよい。音響処理システム１００は、周波数解析部２１が生成した強度スペクトルＸ(m)と、音源分離部２２が生成した強度スペクトルＹ1(m)および強度スペクトルＹ2(m)とを、端末装置から受信する。以上の説明から理解される通り、音響処理システム１００から音源分離部２２が省略されてもよい。音響処理システム１００が音源分離部２２を具備しない構成でも、端末装置等の外部装置において実行される音源分離の処理負荷を軽減できる、という所期の効果は実現される。
【００８５】
（７）前述の各形態においては、音響処理部２０と学習処理部３０とを具備する音響処理システム１００を例示したが、音響処理部２０および学習処理部３０の一方が省略されてもよい。学習処理部３０を具備するコンピュータシステムは、推定モデル訓練システム（機械学習システム）とも換言される。推定モデル訓練システムにおける音響処理部２０の有無は不問である。
【００８６】
（８）以上に例示した音響処理システム１００の機能は、前述の通り、制御装置１１を構成する単数または複数のプロセッサと、記憶装置１２に記憶されたプログラム（Ｐ1，Ｐ2）との協働により実現される。本開示に係るプログラムは、コンピュータが読取可能な記録媒体に格納された形態で提供されてコンピュータにインストールされ得る。記録媒体は、例えば非一過性（non-transitory）の記録媒体であり、ＣＤ-ＲＯＭ等の光学式記録媒体（光ディスク）が好例であるが、半導体記録媒体または磁気記録媒体等の公知の任意の形式の記録媒体も包含される。なお、非一過性の記録媒体とは、一過性の伝搬信号（transitory, propagating signal）を除く任意の記録媒体を含み、揮発性の記録媒体も除外されない。また、配信装置が通信網を介してプログラムを配信する構成では、当該配信装置においてプログラムを記憶する記憶装置１２が、前述の非一過性の記録媒体に相当する。
【００８７】
Ｇ：付記
　以上に例示した形態から、例えば以下の構成が把握される。
【００８８】
　本開示のひとつの態様（態様１）に係る音響処理方法は、第１音源に対応する第１音の

うち第１周波数帯域の成分を表す第１入力データと、前記第１音源とは異なる第２音源に対応する第２音のうち前記第１周波数帯域の成分を表す第２入力データと、前記第１音と前記第２音との混合音のうち前記第１周波数帯域とは異なる第２周波数帯域を含む周波数帯域の成分を含む音を表す混合音データと、を含む入力データを取得し、学習済の推定モデルに前記入力データを入力することで、前記第１音のうち前記第２周波数帯域を含む周波数帯域の成分を表す第１出力データと、前記第２音のうち前記第２周波数帯域を含む周波数帯域の成分を表す第２出力データとの少なくとも一方を生成する。

【００８９】

以上の構成によれば、第１音のうち第１周波数帯域の成分を表す第１入力データと、第２音のうち第１周波数帯域の成分を表す第２入力データとを含む入力データから、第１音のうち第２周波数帯域を含む周波数帯域の成分を表す第１出力データと、第２音のうち第２周波数帯域を含む周波数帯域の成分を表す第２出力データとの少なくとも一方が生成される。すなわち、第１入力データが表す音は第１音のうち第１周波数帯域の成分であれば足り、第２入力データが表す音は第２音のうち第１周波数帯域の成分であれば足りる。以上の構成によれば、第１音源に対応する第１音と第２音源に対応する第２音との混合音を第１音と第２音とに分離する音源分離を、第１周波数帯域についてのみ限定的に実行すれば足りる。したがって、音源分離のための処理負荷が軽減される。

【００９０】

「第１音源に対応する第１音」は、第１音源から発音される音を優勢に含む音を意味する。すなわち、第１音源から発音される音単独のほか、例えば第１音源から発音される第１音に加えて第２音源からの第２音（例えば音源分離により完全には除去されなかった第２音）が僅かに含まれる音も、「第１音源に対応する第１音」の概念には包含される。同様に、「第２音源に対応する第２音」は、第２音源から発音される音を優勢に含む音を意味する。すなわち、第２音源から発音される音単独のほか、例えば第２音源から発音される第２音に加えて第１音源からの第１音（例えば音源分離により完全には除去されなかった第１音）が僅かに含まれる音も、「第２音源に対応する第２音」の概念には包含される。

【００９１】

混合音データが表す音は、混合音のうち第１周波数帯域および第２周波数帯域の双方の成分を含む音（例えば全帯域にわたる混合音）と、混合音のうち第１周波数帯域の成分を含まない音とを包含する。

【００９２】

第１周波数帯域および第２周波数帯域は、周波数軸上の相異なる周波数帯域である。典型的には、第１周波数帯域と第２周波数帯域とは相互に重複しない。ただし、第１周波数帯域と第２周波数帯域とが部分的に重複してもよい。第１周波数帯域の周波数軸上の位置と第２周波数帯域の周波数軸上の位置との関係は任意である。また、第１周波数帯域の帯域幅と第２周波数帯域の帯域幅との異同は不問である。

【００９３】

第１出力データは、第１音のうち第２周波数帯域の成分のみを表すデータ、または、第１音のうち第１周波数帯域および第２周波数帯域を含む周波数帯域の成分を表すデータである。同様に、第２出力データは、第２音のうち第２周波数帯域の成分のみを表すデータ、または、第２音のうち第１周波数帯域および第２周波数帯域を含む周波数帯域の成分を表すデータである。

【００９４】

推定モデルは、入力データと出力データ（第１出力データおよび第２出力データ）との関係を学習した統計的モデルである。推定モデルの典型例はニューラルネットワークであるが、推定モデルの種類は以上の例示に限定されない。

【００９５】

態様１の具体例（態様２）において、前記混合音は、前記第１周波数帯域の成分と前記第２周波数帯域の成分とを含み、前記混合音データは、前記混合音のうち前記第１周波数帯域の成分を含まない音を表す。以上の構成によれば、混合音データが表す音が第１周波

数帯域の成分を含まないから、混合音データが表す音が第１周波数帯域の成分と第２周波数帯域の成分とを含む構成と比較して、推定モデルの機械学習に必要な処理負荷および当該推定モデルの規模が低減されるという利点がある。

【００９６】

態様１または態様２の具体例（態様３）において、前記第１入力データは、前記第１音のうち前記第１周波数帯域の成分の強度スペクトルを表し、前記第２入力データは、前記第２音のうち前記第１周波数帯域の成分の強度スペクトルを表し、前記混合音データは、前記混合音のうち前記第２周波数帯域を含む周波数帯域の成分の強度スペクトルを表し、前記入力データは、前記第１入力データと前記第２入力データと前記混合音データとで構成される正規化されたベクトルと、当該ベクトルの大きさを表す強度指標とを含む。以上の構成によれば、強度指標が入力データに含まれるから、混合音に対応する音量の音を表す第１出力データおよび第２出力データが生成される。したがって、第１出力データおよび第２出力データが表す音の強度を調整する処理（スケーリング）が不要であるという利点がある。

【００９７】

態様１から態様３の何れかの具体例（態様４）において、前記推定モデルは、前記第１出力データが表す音のうち前記第２周波数帯域の成分と、前記第２出力データが表す音のうち前記第２周波数帯域の成分とを混合した結果が、前記混合音のうち前記第２周波数帯域の成分に近似するように訓練されたモデルである。以上の構成によれば、第１出力データが表す音のうち第２周波数帯域の成分と、第２出力データが表す音のうち第２周波数帯域の成分とを混合した結果が、混合音のうち第２周波数帯域の成分に近似するように、推定モデルが訓練される。したがって、以上の条件を加味せずに訓練された推定モデルを利用する構成と比較して、第１音のうち第２周波数帯域の成分（第１出力データ）と第２音のうち第２周波数帯域の成分（第２出力データ）とを高精度に推定できる。

【００９８】

態様１から態様４の何れかの具体例（態様５）において、さらに、前記混合音のうち前記第１周波数帯域の成分に対する音源分離により、前記第１音のうち第１周波数帯域の第１成分と、前記第２音のうち前記第１周波数帯域の第２成分とを生成し、前記入力データの取得においては、前記第１成分を表す前記第１入力データと、前記第２成分を表す前記第２入力データとを取得する。以上の構成によれば、混合音のうち第１周波数帯域の成分に対して音源分離が実行されるから、混合音の全帯域を対象として音源分離を実行する構成と比較して、音源分離のための処理負荷が軽減される。

【００９９】

態様１から態様５の何れかの具体例（態様６）において、前記第１出力データは、前記第１音のうち前記第１周波数帯域の成分と前記第２周波数帯域の成分とを表し、前記第２出力データは、前記第２音のうち前記第１周波数帯域の成分と前記第２周波数帯域の成分とを表す。以上の構成によれば、第１周波数帯域および第２周波数帯域の双方の成分を含む第１出力データおよび第２出力データが生成される。したがって、第１出力データが第１音のうち第２周波数帯域の成分のみを表すデータであり、第２出力データが第２音のうち第２周波数帯域の成分のみを表すデータである構成と比較して、第１周波数帯域および第２周波数帯域の双方にわたる音響を簡便に生成できる。

【０１００】

本開示のひとつの態様（態様７）に係る推定モデルの訓練方法は、入力データと出力データとを各々が含む複数の訓練データを取得し、前記複数の訓練データを利用した機械学習により、前記入力データと前記出力データとを関係を学習した推定モデルを確立し、前記入力データは、第１音源に対応する第１音のうち第１周波数帯域の成分を表す第１入力データと、前記第１音源とは異なる第２音源に対応する第２音のうち前記第１周波数帯域の成分を表す第２入力データと、前記第１音と前記第２音との混合音のうち前記第１周波数帯域とは異なる第２周波数帯域を含む周波数帯域の成分を含む音を表す混合音データとを含み、前記出力データは、前記第１音のうち前記第２周波数帯域を含む周波数帯域の成

分を表す第１出力データと、前記第２音のうち前記第２周波数帯域を含む周波数帯域の成分を表す第２出力データとを含む。

【０１０１】

以上の構成によれば、第１音のうち第１周波数帯域の成分を表す第１入力データと、第２音のうち第１周波数帯域の成分を表す第２入力データとを含む入力データから、第１音のうち第２周波数帯域を含む周波数帯域の成分を表す第１出力データと、第２音のうち第２周波数帯域を含む周波数帯域の成分を表す第２出力データとの少なくとも一方を生成する推定モデルが確立される。以上の構成によれば、第１音源に対応する第１音と第２音源に対応する第２音との混合音を第１音と第２音とに分離する音源分離を、第１周波数帯域についてのみ限定的に実行すれば足りる。したがって、音源分離のための処理負荷が軽減される。

【０１０２】

なお、本開示は、以上に例示した各態様（態様１から態様６）に係る音響処理方法を実現する音響処理システム、または、当該音響処理方法をコンピュータに実行させるプログラム、としても実現される。また、本開示は、前述の態様７に係る訓練方法を実現する推定モデル訓練システム、または、当該訓練方法をコンピュータに実行させるプログラム、としても実現される。

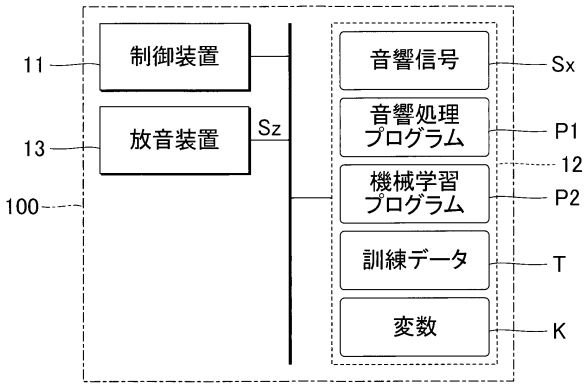
【符号の説明】

【０１０３】

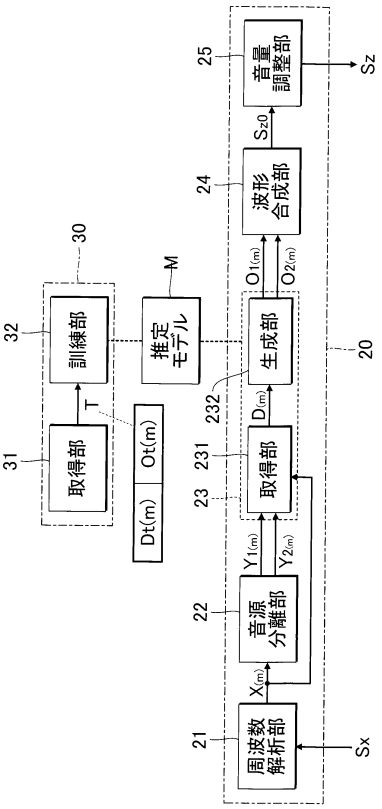
１００…音響処理システム、１１…制御装置、１２…記憶装置、１３…放音装置、２０…音響処理部、２１…周波数解析部、２２…音源分離部、２３…帯域拡張部、２３１…取得部、２３２…生成部、２４…波形合成部、２５…音量調整部、３０…学習処理部、３１…取得部、３２…訓練部、Ｍ…推定モデル。

【図面】

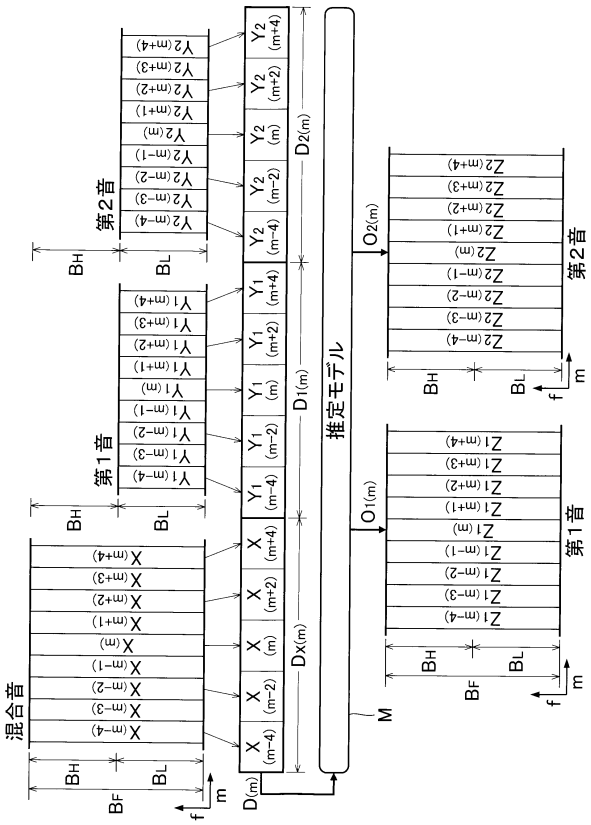
【図１】


11
制御装置
13
放音装置
Sz
100
音響信号
Sx
音響処理プログラム
P1
12
機械学習プログラム
P2
訓練データ
T
変数
K

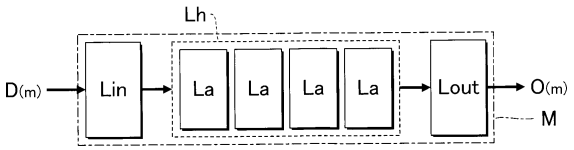
【図２】


25
音量調整部
Sz
Sz≥0
24
波形合成部
O1(m)
O2(m)
30
M
32
訓練部
推定モデル
T
20
232
Dt(m)
31
取得部
Ot(m)
Dt(m)
231
取得部
23
Y1(m)
Y2(m)
22
音源分離部
X(m)
21
周波数解析部
Sx

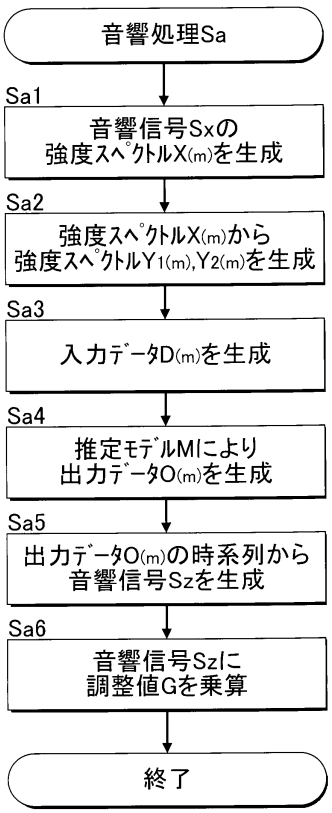
【図３】


第２音
第１音
混合音
推定モデル
BH
BL
BF
D1(m)
D2(m)
Dx(m)
O1(m)
O2(m)
M
D(m)
f
m

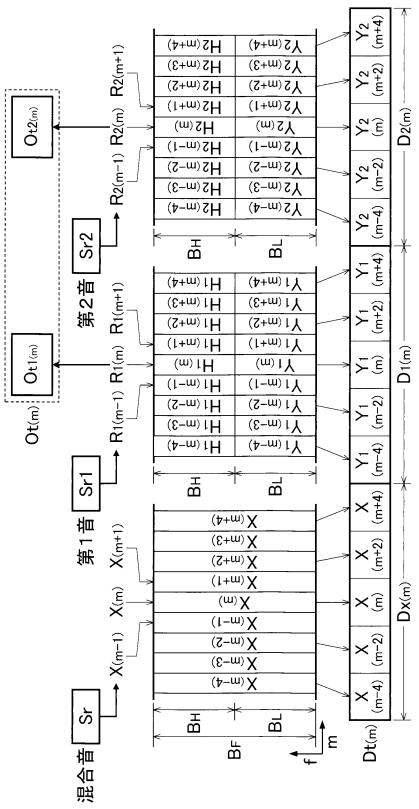
【図４】


Lh
D(m)
Lin
La
La
La
La
Lout
O(m)
M

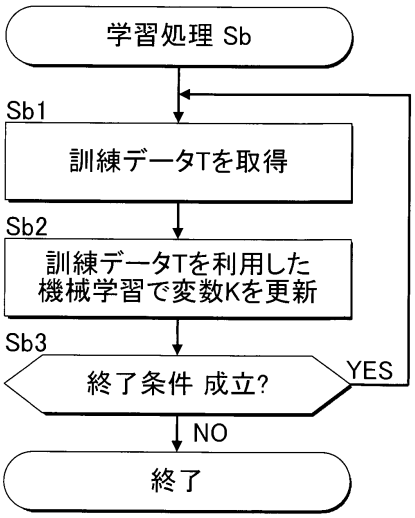
【図５】


音響処理Sa
Sa1
音響信号Sxの
強度ｽﾍﾟｸﾄﾙX(m)を生成
Sa2
強度ｽﾍﾟｸﾄﾙX(m)から
強度ｽﾍﾟｸﾄﾙY1(m),Y2(m)を生成
Sa3
入力ﾃﾞｰﾀD(m)を生成
Sa4
推定ﾓﾃﾞﾙMにより
出力ﾃﾞｰﾀO(m)を生成
Sa5
出力ﾃﾞｰﾀO(m)の時系列から
音響信号Szを生成
Sa6
音響信号Szに
調整値Gを乗算
終了

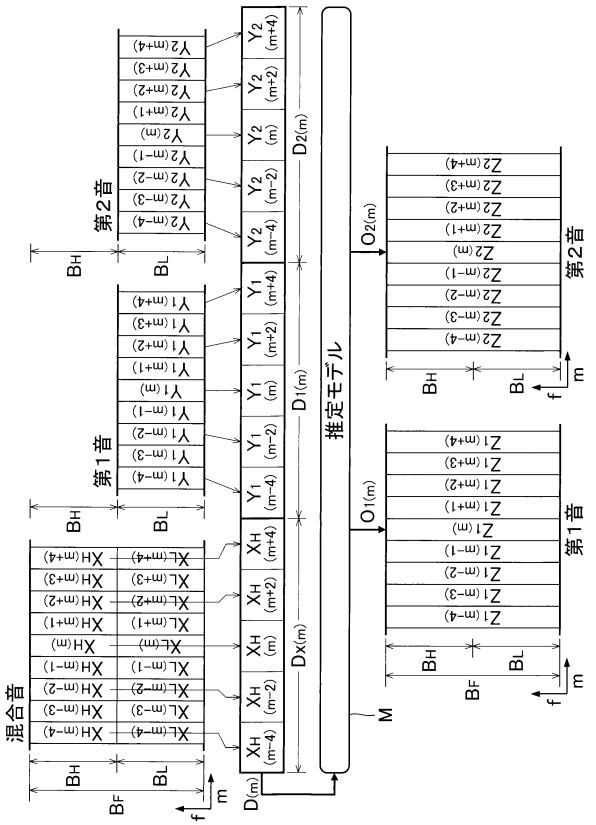
【図６】


第２音
第１音
混合音
Sr2
Sr1
Sr
Ot2(m)
Ot1(m)
Ot(m)
BH
BL
BF
Dt(m)
D1(m)
D2(m)
Dx(m)
f
m

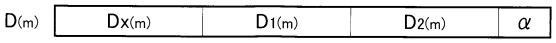
【図７】


学習処理 Sb
Sb1
訓練データTを取得
Sb2
訓練データTを利用した
機械学習で変数Kを更新
Sb3
終了条件 成立?
YES
NO
終了

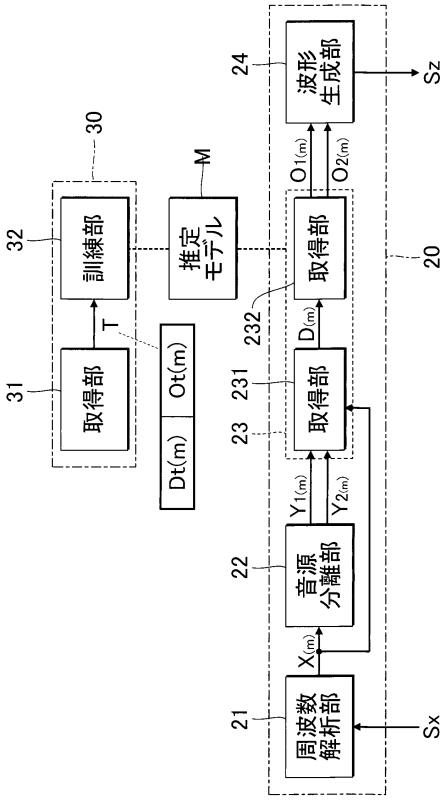
【図８】


混合音
第1音
第2音
推定モデル
第1音
第2音
BH
BL
BF
m
f
M

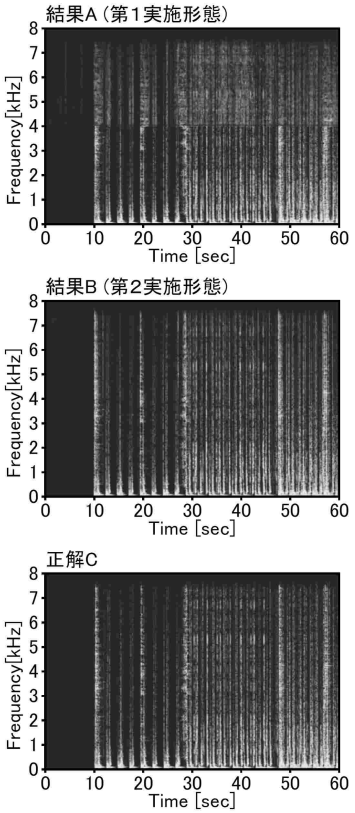
【図９】


D(m)
Dx(m)
D1(m)
D2(m)
α

【図１０】


波形生成部
訓練部
推定モデル
取得部
取得部
取得部
音源分離部
周波数解析部
Sz
Sx
T
Ot(m)
Dt(m)
D(m)
X(m)
Y1(m)
Y2(m)
O1(m)
O2(m)
M
20
21
22
23
24
30
31
32
231
232

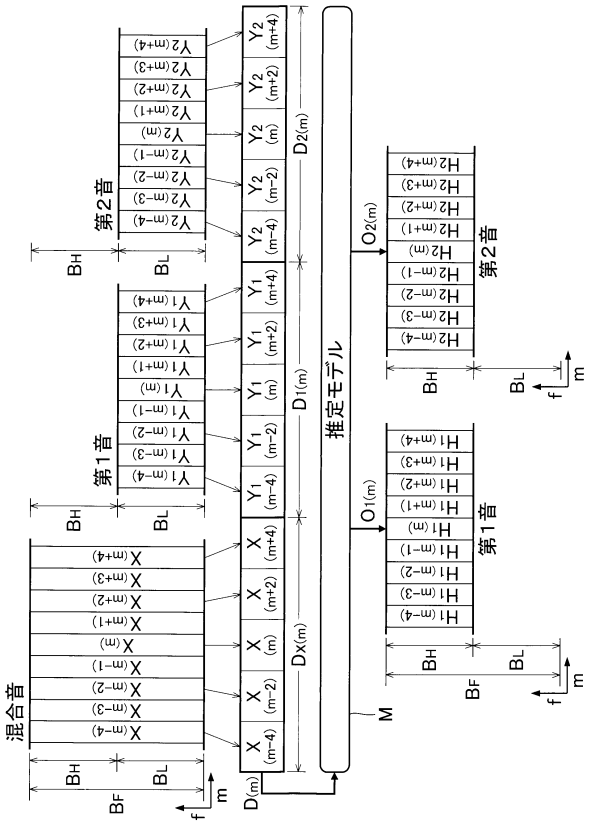
【図１１】



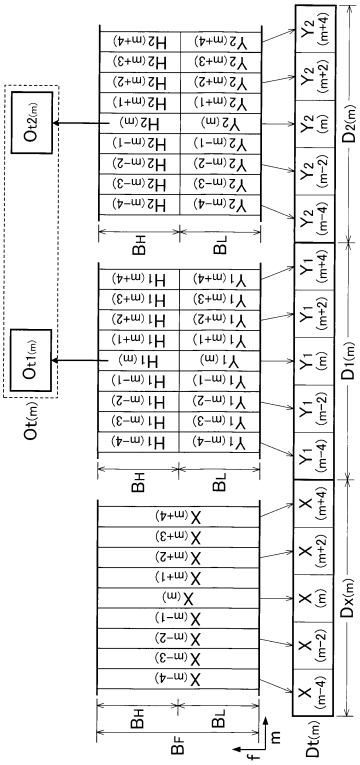
【図１２】

		SAR [dB]	改善量 [dB]
Drums	比較例	21.96	—
	第1実施形態	22.51	0.55
	第2実施形態	22.50	0.54
	第3実施形態	27.60	5.64
Vocals	比較例	25.00	—
	第1実施形態	25.62	0.62
	第2実施形態	25.62	0.62
	第3実施形態	29.28	4.29

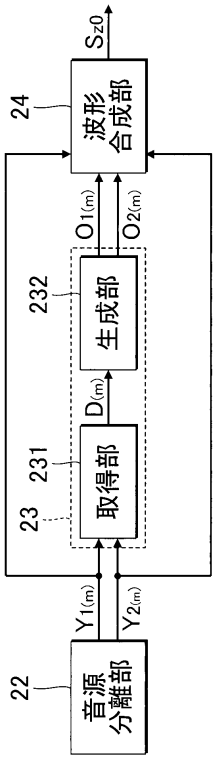
【図１３】



【図１４】



【図１５】



フロントページの続き

(56)参考文献　　特開２０１２－２２１２０（ＪＰ，Ａ）
特開２００８－２７８４０６（ＪＰ，Ａ）

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　２１／０２７２
Ｇ１０Ｌ　２５／３０