(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2025年1月23日(23.01.2025)

WIPO | PCT

(10) 国際公開番号
WO 2025/017883 A1

(51) 国際特許分類:
G10L 21/0272 (2013.01) *G10L 25/30* (2013.01)
G10L 21/0208 (2013.01)

(21) 国際出願番号： PCT/JP2023/026455

(22) 国際出願日： 2023年7月19日(19.07.2023)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 日本電信電話株式会社 (**NIPPON TELEGRAPH AND TELEPHONE CORPORATION**) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目５番１号 Tokyo (JP).

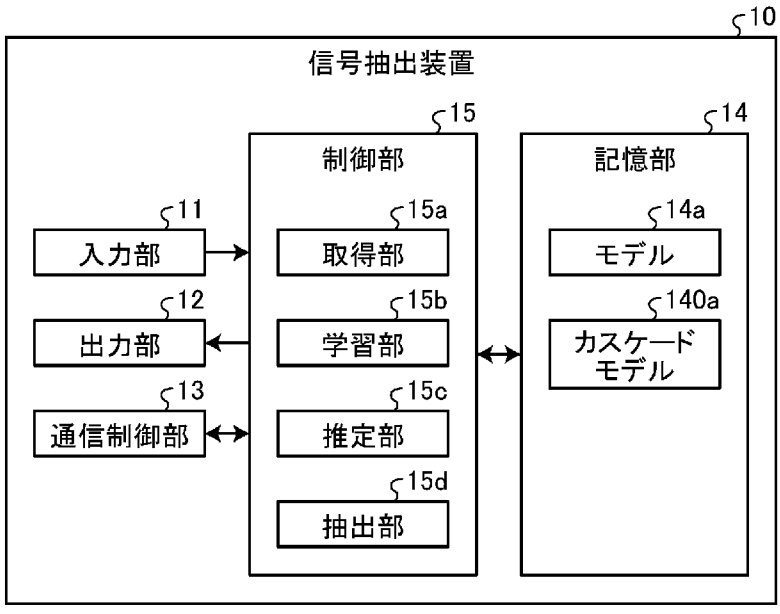
(72) 発明者: 加茂 直之(**KAMO, Naoyuki**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ 知的財産センタ内 Tokyo (JP). デルクロア マーク(**DELCROIX, Marc**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ 知的財産センタ内 Tokyo (JP). 中谷 智広(**NAKATANI, Tomohiro**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ 知的財産センタ内 Tokyo (JP).

(74) 代理人: 弁理士法人酒井国際特許事務所 (**SAKAI INTERNATIONAL PATENT OFFICE**); 〒1000013 東京都千代田区霞が関３丁目８番１号 虎ノ門ダイビルイースト Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK,

(54) **Title:** SIGNAL EXTRACTION DEVICE AND SIGNAL EXTRACTION METHOD

(54) 発明の名称： 信号抽出装置、および信号抽出方法



10 Signal extraction device
11 Input unit
12 Output unit
13 Communication control unit
14 Storage unit
14a Model
15 Control unit
15a Acquisition unit
15b Learning unit
15c Estimation unit
15d Extraction unit
140a Cascade model

(57) **Abstract:** In a signal extraction device (10), an acquisition unit (15a) acquires a mixed signal of a plurality of sound sources and clue information for a target sound source. An estimation unit (15c) inputs, into a trained model (14a), data from a sampled step and the acquired mixed signal and clue information of the target sound source, and estimates a score corresponding to said data. An extraction unit (15d) uses the estimated score to extract a signal of the target sound source from the mixed signal of the plurality of sound sources.

WO 2025/017883 A1

［続葉有］

SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

一 国際調査報告(条約第21条(3))

(57) 要約: 信号抽出装置(10)において、取得部(15a)が、複数音源の混合信号と、目的音源の手がかり情報とを取得する。推定部(15c)が、取得された混合信号と目的音源の手がかり情報と、サンプリングされたステップにおけるデータとを学習済みモデル(14a)に入力し、該データに対応するスコアを推定する。抽出部(15d)が、推定されたスコアを用いて、複数音源の混合信号から目的音源の信号を抽出する。

明　細　書

発明の名称 ： 信号抽出装置、および信号抽出方法

技術分野

[0001]　本発明は、信号抽出装置、および信号抽出方法に関する。

背景技術

[0002]　近年、音声認識・話者認識での高性能化や会話アプリ等で音声を聞き取りやすくするため等に、混合音声信号から特定の話者の音声を抽出する技術が検討されている。例えば、観測信号と、事前に録音した対象話者の声等の目的音を特定できる何らかの手掛かりとを、ＤＮＮ（Deep Neural Network）モデルに入力し、目的音源信号を出力する決定論的（Discriminative）な目的音源抽出システムが提案されている（非特許文献１参照）。

[0003]　また、生成対象のデータからガウシアンノイズまで変化させるマルコフ過程を逆方向に変化させることにより、データを生成するDiffusion modelを用いて、観測信号から雑音除去を行うが開示されている（非特許文献２参照）。

先行技術文献

非特許文献

[0004] 非特許文献1：Katerina Zmolikova, “SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures”, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 13, NO. 4, 2019年8月

非特許文献2：Julius Richter, “Speech Enhancement and Dereverberation with Diffusion-based Generative Models”, IEEE/ACM Transactions on Audio, Speech and Language Processing, 2023年

発明の概要

発明が解決しようとする課題

[0005]　しかしながら、従来技術では、混合信号から目的音源信号を精度高く抽出

することが困難であった。例えば、抽出された目的音源信号の聴感上の品質が不十分であった。また、目的音源信号に対象信号と異なる信号の抽出等の失敗が含まれていた場合でも改善が難しかった。

[0006] 本発明は、上記に鑑みてなされたものであって、従来技術では、混合信号から目的音源信号を精度高く抽出することを目的とする。

課題を解決するための手段

[0007] 上述した課題を解決し、目的を達成するために、本発明に係る信号抽出装置は、複数音源の混合信号と、目的音源の手がかり情報とを取得する取得部と、取得された前記混合信号と前記目的音源の手がかりと、サンプリングされたステップにおけるデータとを学習済みモデルに入力し、該データに対応するスコアを推定する推定部と、推定された前記スコアを用いて、前記複数音源の混合信号から前記目的音源の信号を抽出する抽出部と、を有することを特徴とする。

発明の効果

[0008] 本発明によれば、混合信号から目的音源信号を精度高く抽出することが可能となる。

図面の簡単な説明

[0009] [図1]図１は、本実施形態の信号抽出装置の概要を説明するための図である。

[図2]図２は、本実施形態の信号抽出装置の概要を説明するための図である。

[図3]図３は、本実施形態の信号抽出装置の概要を説明するための図である。

[図4]図４は、本実施形態の信号抽出装置の概要を説明するための図である。

[図5]図５は、本実施形態の信号抽出装置の概要を説明するための図である。

[図6]図６は、本実施形態の信号抽出装置の概略構成を例示する模式図である。

[図7]図７は、信号抽出処理を説明するための図である。

[図8]図８は、信号抽出処理を説明するための図である。

[図9]図９は、信号抽出処理手順を示すフローチャートである。

[図10]図１０は、実施例を説明するための図である。

[図11]図１１は、実施例を説明するための図である。

[図12]図１２は、信号抽出プログラムを実行するコンピュータの一例を示す図である。

発明を実施するための形態

[0010] 以下、図面を参照して、本発明の一実施形態を詳細に説明する。なお、この実施形態により本発明が限定されるものではない。また、図面の記載において、同一部分には同一の符号を付して示している。

[0011] ［信号抽出装置の概要］

図１～図５は、本実施形態の信号抽出装置の概要を説明するための図である。まず、図１には、従来の決定論的な目的音源抽出の処理が例示されている。図１に示すように、決定論的な目的音源抽出では、観測信号ｙと、事前に録音した目的話者の音声等の目的音声を特定できる何らかの手がかりｃを、音声抽出モデルに入力し、目的音源信号が出力される。この音声抽出モデルは、ＤＮＮを用いて、正解信号と出力信号との差を最小にするように学習される。この音声抽出モデルは確率分布に従うデータ生成モデルではないため、タイミングに関わらず同じ結果を出力する決定論的な音声の生成しかできない。そのため、目的と異なる信号等の失敗が含まれていた場合に改善する仕組みがない。

[0012] 一方、図２には、従来のDiffusion modelの処理が例示されている。図２に示すように、Diffusion modelでは生成対象のデータx_0からガウシアンノイズx_Tまで変化させるマルコフ過程（forward process）に対し、x_0からx_Tへ逆方向に変化させるマルコフ過程（Reverse process）を用いて、データx_0を生成する。この場合のＤＮＮは、Reverse processの１ステップｐ（$x_T \mid x_{T-1}$）を実現するように学習が行われる。

[0013] また、図３には、図２に示したDiffusion modelに、Scoreと呼ばれる値を用いるScore Based Diffusion modelの処理が例示されている。具体的には、図３（ａ）に示すように、Score Based Diffusion modelでは、連続的な状態遷移を記述する確率微分方程式（ＳＤＥ、Stochastic Differentia

l Equations）が用いられる。この場合のＤＮＮ（Score model）は、図３（ｂ）に例示するように、生成するデータに対応する正解Scoreの値を教師データとして用いて学習される。そして、図３（ｃ）に例示するように、学習されたScore modelを用いて、１ステップの推移（ステップｔでのデータx_t）の推論を繰り返すことにより、データが生成される。

[0014] また、図４には、図３に示したScore Based Diffusion modelを用いて、音声信号の雑音除去を行う処理（非特許文献２参照）が例示されている。図４（ａ）（ｂ）に示すように、この技術では、図３と同様のＤＮＮ（Score model）を用いて、観測信号ｙの条件の下に、雑音除去されたクリーン音声が生成される。このＤＮＮは、雑音除去モデルであり、目的音源の抽出はできない。また、失敗が含まれていた場合にも改善する仕組みがない。

[0015] そこで、本実施形態の信号抽出装置は、図３に示したScore Based Diffusion model（Score model）を用いて、観測信号ｙおよび目的音源の手がかりｃの条件の下に、目的音源の信号を生成する。

[0016] ここで、図５には、本実施形態の信号抽出処理が例示されている。具体的には、信号抽出装置のScore modelは、図５（ａ）に示すように、生成対象のデータx_0からガウシアンノイズx_Tまで変化させる次式（１）に示すforward processに対し、x_0からx_Tへ逆方向に変化させる次式（２）に示すReverse processを用いる。そして、ステップｔでのデータx_tの推論を繰り返すことにより、データx_0を生成する。

[0017] [数1]

$$\mathbf{dx} = \mathbf{f}(\mathbf{x}_t, \mathbf{y})\mathrm{d}t + g(t)\mathrm{d}\mathbf{w} \quad \cdots (1)$$

[数2]

$$\mathbf{dx}_t = [-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{y}, \mathbf{c})]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \quad \cdots (2)$$

[0018] このように、信号抽出装置は、生成モデルであるDiffusion modelを用いて目的音源の音声信号を生成する。これにより、ＰＥＳＱ(Perceptual Evaluation of Speech Quality)の性能向上等、聴覚上の品質向上が可能となる。

[0019] また、信号抽出装置では、図５（ｂ）（ｃ）に例示する、推論のステップ番号ｔ等のサンプリングの際の乱数のシードによって確率的に推論結果が変わる。そこで、信号抽出装置は、推論結果のアンサンブルを行う。これにより、推論結果の精度向上が可能となる。

[0020] ［信号抽出装置の構成］

図６は、本実施形態の信号抽出装置の概略構成を例示する模式図である。図６に例示するように、本実施形態の信号抽出装置１０は、パソコン等の汎用コンピュータで実現され、入力部１１、出力部１２、通信制御部１３、記憶部１４、および制御部１５を備える。

[0021] 入力部１１は、マイクやキーボードやマウス等の入力デバイスを用いて実現され、操作者による入力操作に対応して、制御部１５に対して処理開始などの各種指示情報を入力する。出力部１２は、液晶ディスプレイなどの表示装置、プリンター等の印刷装置等によって実現される。例えば、出力部１２には、後述する信号抽出処理の結果が表示される。

[0022] 通信制御部１３は、ＮＩＣ（Network Interface Card）等で実現され、ＬＡＮ（Local Area Network）やインターネットなどの電気通信回線を介した外部の装置と制御部１５との通信を制御する。例えば、通信制御部１３は、各種情報を管理する管理装置等と制御部１５との通信を制御する。

[0023] 記憶部１４は、ＲＡＭ（Random Access Memory）、フラッシュメモリ（Flash Memory）等の半導体メモリ素子、または、ハードディスク、光ディスク等の記憶装置によって実現される。記憶部１４には、信号抽出装置１０を動作させる処理プログラムや、処理プログラムの実行中に使用されるデータなどが予め記憶され、あるいは処理の都度一時的に記憶される。なお、記憶部１４は、通信制御部１３を介して制御部１５と通信する構成でもよい。

[0024] 本実施形態において、記憶部１４は、後述する信号抽出処理で使用されるモデル１４ａまたはカスケードモデル１４０ａ等を記憶する。

[0025] 制御部１５は、ＣＰＵ（Central Processing Unit）やＮＰ（Network Processor）やＦＰＧＡ（Field Programmable Gate Array）等を用いて実

現され、メモリに記憶された処理プログラムを実行する。これにより、制御部１５は、図６に例示するように、取得部１５ａ、学習部１５ｂ、推定部１５ｃ、および抽出部１５ｄとして機能して、信号抽出処理を実行する。なお、これらの機能部は、それぞれ、あるいは一部が異なるハードウェアに実装されてもよい。また、制御部１５は、その他の機能部を備えてもよい。

[0026] 取得部１５ａは、複数音源の混合信号と、目的音源の手がかり情報とを取得する。例えば、取得部１５ａは、入力部１１あるいは通信制御部１３を介して、マイクや各種情報を管理する管理装置等から、複数音源の混合信号である観測信号ｙと、事前に録音した目的話者の音声等の目的音源を特定できる何らかの手がかりｃとを取得する。取得部１５ａは、取得した観測信号ｙと、目的音源の手がかり情報ｃとを記憶部１４に記憶させてもよい。

[0027] 学習部１５ｂは、取得された混合信号ｙと目的音源の手がかりｃと、サンプリングされたｔ番目のステップにおけるデータx_tとを入力すると、該データx_tに対応するスコアを推定するモデル１４ａを学習により生成する。具体的には、図５（ｂ）に例示したように、ステップｔでのデータx_tと観測信号ｙと手がかりｃとを入力すると、x_tに対応するスコアを出力するモデル（Score model）１４ａを学習により生成する。

[0028] このモデル１４ａは、ガウシアンノイズx_Tから生成対象のデータx_0にステップごとに変化させる逆過程（Reverse Process）により該データx_0を生成するDiffusion modelである。具体的には、モデル１４ａは、図５（ａ）に例示したように、スコアを用いたステップｔでのデータx_tの推論を繰り返すことにより、データx_0を生成するScore Baced diffusion modelである。

[0029] 推定部１５ｃは、取得された混合信号ｙと目的音源の手がかりｃと、サンプリングされたｔ番目のステップにおけるデータx_tとを学習済みのモデル１４ａに入力し、該データに対応するスコアを推定する。具体的には、図５（ｃ）に例示したように、モデル１４ａに混合信号ｙと目的音源の手がかりｃとx_tとを入力してスコアの出力を得る。

[0030] 抽出部１５ｄは、推定されたスコアを用いて、複数音源の混合信号ｙから目的音源の信号を抽出する。具体的には、抽出部１５ｄは、図５（ｃ）に例示したように、推定部１５ｃにより出力されたスコアを用いて、$x_{T-\Delta t}$を推定する処理を、最終的な出力として推論結果x_0を得るまで繰り返すことにより、目的音源の信号x_0を抽出する。

[0031] ここで、サンプリングの際の乱数のシードを変えると、推論結果は確率的に変わる。そこで、抽出部１５ｄは、所定回数の各処理において抽出された目的音源の信号の平均値を抽出結果としてもよい。つまり、抽出部１５ｄは、乱数のｓｅｅｄを変えて所定回数の推論を繰り返し、推論結果の平均値を求めるアンサンブルを行ってもよい。これにより、推論の精度を向上させることが可能となる。

[0032] ここで、図７および図８は、信号抽出処理を説明するための図である。まず、図７には、図５に示したモデル１４ａ（ＤＮＮ、Score model）の処理が例示されている。

[0033] 具体的には、学習時（図５（ｂ）参照）には、学習部１５ｂは、まず、観測信号として混合音声ｙ、目的話者音声x_0、目的音源の手がかりとしてランダムに選んだ目的話者の音声ｃを用いて、データ（ｙ，x_0，ｃ）を作成する。

[0034] また、学習部１５ｂは、０≦ｔ≦Ｔの一様分布からステップ番号ｔのサンプリングを行うとともに、次式（３）～（５）に従い、x_0を用いてx_tのサンプリングを行う。

[0035] [数3]

$$p_t(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y}) = \mathcal{N}_c(\boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},t),\sigma(t)^2\mathbf{I}), \cdots (3)$$

[数4]

$$\boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},t) = e^{-\gamma t}\mathbf{x}_0 + (1-e^{-\gamma t})\mathbf{y}, \cdots (4)$$

[数5]

$$\sigma(t)^2 = \frac{\sigma_0^2\left((\sigma_1/\sigma_0)^{2t}-e^{-2\gamma t}\right)\log(\sigma_1/\sigma_0)}{\gamma+\log(\sigma_1/\sigma_0)} \cdots (5)$$

[0036] そして、学習部１５ｂは、図７のモデル１４ａにデータ（ｙ，x_t，ｃ）を入力してスコア（score）の出力を得る。また、学習部１５ｂは、正解スコアと出力スコアとの差を最小化するようにｌｏｓｓ関数を計算し、ｂａｃｋｗａｒｄ計算を行う。

[0037] また、推論時（図５（ｃ）参照）には、推定部１５ｃが、上記式（３）～（５）に従い、x_Tのサンプリングを行う。また、［０，Ｔ］をＮ分割して$\Delta t = T/N$とし、（x_T，$x_{T-\Delta t}$、…，x_0）において、reverse processによる不連続な状態遷移を行う。具体的には、推定部１５ｃは、モデル１４ａにデータ（ｙ，x_t，ｃ）を入力してスコアの出力を得る。また、抽出部１５ｄは、出力されたスコアを用いて、$x_{T-\Delta t}$を推定する処理を、最終的な出力として推論結果x_0を得るまで繰り返す。

[0038] なお、サンプリングの際の乱数のシードを変えると、推論結果は確率的に変わる。そこで、抽出部１５ｄは、乱数のｓｅｅｄを変えて所定回数の推論を繰り返し、推論結果の平均値を求めるアンサンブルを行ってもよい。これにより、推論の精度を向上させることが可能となる。

[0039] また、図８（ａ）には、図１に例示した決定論的な目的音源抽出処理のモデル（Discriminative ＴＳＥ（Target Source Extraction））１４１ａの処理が例示されている。また、図８（ｂ）には、図８（ａ）の決定論的な目的音源抽出処理のモデル１４１ａと、図７に示した信号抽出処理のモデル１４ａとを組み合わせたモデル（カスケードモデル）１４０ａの処理が例示されている。

[0040] カスケードモデル１４０ａでは、学習部１５ｂは、スコアのｌｏｓｓ関数と目的話者音声ｘのｌｏｓｓ関数との２つのｌｏｓｓ関数を用いてマルチタスク学習を行う。また、カスケードモデル１４０ａは、スコアを用いた推論を行うDiffusion modelとして、またｘの直接推定値を用いる決定論的なモデルとして、２つの使い方が可能である。

[0041] 具体的には、カスケードモデル１４０ａの学習時には、学習部１５ｂは、図７のモデル１４ａの学習時と同様に、観測信号として混合音声ｙ、目的話

者音声x_0、目的音源の手がかりとしてランダムに選んだ目的話者の音声ｃを用いて、データ（ｙ，x_0，ｃ）を作成する。

[0042] また、学習部１５ｂは、０≦ｔ≦Ｔの一様分布からステップ番号ｔのサンプリングを行うとともに、上記式（１）～（３）に従い、x_0を用いてx_tのサンプリングを行う。

[0043] そして、学習部１５ｂは、図８（ｂ）のモデル１４０ａのうち、モデル１４１ａにデータ（ｙ，ｃ）を入力して$\hat{x}$の出力を得る。また、学習部１５ｂは、モデル１４２ａにデータ（x_t，$\hat{x}$）を入力してスコア（score）の出力を得る。そして、学習部１５ｂは、正解スコアと出力スコアとの差のｌｏｓｓ関数と、$\hat{x}$とx_0との差のｌｏｓｓ関数とをそれぞれ最小化するように計算し、ｂａｃｋｗａｒｄ計算を行うことにより学習する。

[0044] カスケードモデル１４０ａの推論時には、推定部１５ｃが、図７のモデル１４ａの推論時と同様に、上記式（３）～（５）に従い、x_Tのサンプリングを行う。また、推定部１５ｃは、［０，Ｔ］をＮ分割して$\Delta t = T/N$とし、（x_T，$x_{T-\Delta t}$、…，x_0）において、reverse processによる不連続な状態遷移を行う。具体的には、推定部１５ｃは、カスケードモデル１４０ａにデータ（ｙ，x_t，ｃ）を入力してスコアの出力を得る。また、抽出部１５ｄが、出力されたスコアを用いて、$x_{T-\Delta t}$を推定する処理を、最終的な出力として推論結果x_0を得るまで繰り返す。

[0045] なお、図７に示した例と同様に、サンプリングの際の乱数のｓｅｅｄを変えると、推論結果は確率的に変わる。そこで、抽出部１５ｄは、乱数のｓｅｅｄを変えて所定回数の推論を繰り返し、推論結果の平均値を求めるアンサンブルを行ってもよい。これにより、推論の精度を向上させることが可能となる。

[0046] ［信号抽出処理］

次に、図９を参照して、本実施形態に係る信号抽出装置１０による信号抽出処理について説明する。図９は、信号抽出処理手順を示すフローチャートである。図９のフローチャートは、例えば、ユーザが開始を指示する操作入

力を行ったタイミングで開始される。

[0047] まず、取得部１５ａが、複数音源の混合信号である観測信号ｙと、目的音源の手がかりｃとを取得する（ステップＳ１）。

[0048] 次に、推定部１５ｃが、取得された観測信号ｙと目的音源の手がかりｃと、サンプリングされたｔ番目のステップにおけるデータx_tとを学習済みのモデル１４ａに入力し、該データに対応するスコアを推定する（ステップＳ２）。具体的には、学習済のモデル１４ａに混合信号ｙと目的音源の手がかりｃとx_tとを入力してスコアの出力を得る。

[0049] そして、抽出部１５ｄが、推定されたスコアを用いて、複数音源の混合信号ｙから目的音源の信号を抽出する（ステップＳ３）。具体的には、抽出部１５ｄは、推定部１５ｃにより出力されたスコアを用いて、$x_{T-\Delta t}$を推定する処理を、最終的な出力として推論結果x_0を得るまで繰り返すことにより、目的音源の信号x_0を抽出する。これにより、一連の信号抽出処理が終了する。

[0050] ［効果］

以上、説明したように、本実施形態の信号抽出装置１０において、取得部１５ａが、複数音源の混合信号と、目的音源の手がかり情報とを取得する。推定部１５ｃが、取得された混合信号と目的音源の手がかり情報と、サンプリングされたｔ番目のステップにおけるデータとを学習済みモデル１４ａに入力し、該データに対応するスコアを推定する。抽出部１５ｄが、推定されたスコアを用いて、複数音源の混合信号から目的音源の信号を抽出する。

[0051] 具体的には、学習済みモデル１４ａは、ガウシアンノイズから生成対象のデータにステップごとに変化させる逆過程により該データを生成するDiffusion modelである。

[0052] このように、確率分布に基づく生成モデルを用いることにより、ＰＥＳＱの性能向上等、聴感上の品櫃向上が可能となる。したがって、混合信号から目的音源信号を精度高く抽出することが可能となる。

[0053] また、抽出部１５ｄは、所定回数の各処理において抽出された目的音源の

信号の平均値を抽出結果とする。このように、乱数のｓｅｅｄを変えると推論結果が確率的に変わることを用いて、推論の精度を向上させることが可能となる。

[0054] ［実施例］

図１０および図１１は、実施例を説明するための図である。まず、図１０に示す実施例では、上記実施形態のモデル１４ａを用いて、乱数のシードを変えて１０回推論処理を繰り返した。図１０には、各推論結果と、推論結果を時間周波数領域で単純平均したアンサンブル結果とが例示されている。図１０（ａ）に例示するように、「・」で示す各推論結果（Extracted samples）のＳＤＲ（Source to Distortion Ratio）より、「×」で示すアンサンブル結果（Ensemble）のＳＤＲの方が良いことが確認された。

[0055] なお、図１０（ｂ）には、アンサンブルの一例が示されている。図に太枠で囲んだ部分でわかるように、混合信号（Mixture）に対する目的音源（target）の推論結果を、アンサンブル結果（Ensemble）がよりよく表すことが確認された。

[0056] また、図１１には、上記の各手法の評価結果として、音声品質の知覚評価指標（ＰＥＳＱ、Perceptual Evaluation of Speech Quality）が例示されている。図１１に示すように、図７に例示した本発明のモデル１４ａでは、アンサンブルにより、図１、図８（ａ）に例示した従来法のDiscriminativeＴＳＥより性能が高いことが確認された。また、図８（ｂ）に例示した本発明のカスケードモデル１４０ａでは、アンサンブルの有無にかかわらず、従来法のDiscriminativeＴＳＥより性能が高いことが確認された。

[0057] ［プログラム］

上記実施形態に係る信号抽出装置１０が実行する処理をコンピュータが実行可能な言語で記述したプログラムを作成することもできる。一実施形態として、信号抽出装置１０は、パッケージソフトウェアやオンラインソフトウェアとして上記の信号抽出処理を実行する信号抽出プログラムを所望のコンピュータにインストールさせることによって実装できる。例えば、上記の信

号抽出プログラムを情報処理装置に実行させることにより、情報処理装置を信号抽出装置１０として機能させることができる。ここで言う情報処理装置には、デスクトップ型またはノート型のパーソナルコンピュータが含まれる。また、その他にも、情報処理装置にはスマートフォン、携帯電話機やＰＨＳ（Personal Handyphone System）などの移動体通信端末、さらには、ＰＤＡ（Personal Digital Assistant）などのスレート端末などがその範疇に含まれる。また、信号抽出装置１０の機能を、クラウドサーバに実装してもよい。

[0058] 図１２は、信号抽出プログラムを実行するコンピュータの一例を示す図である。コンピュータ１０００は、例えば、メモリ１０１０と、ＣＰＵ１０２０と、ハードディスクドライブインタフェース１０３０と、ディスクドライブインタフェース１０４０と、シリアルポートインタフェース１０５０と、ビデオアダプタ１０６０と、ネットワークインタフェース１０７０とを有する。これらの各部は、バス１０８０によって接続される。

[0059] メモリ１０１０は、ＲＯＭ（Read Only Memory）１０１１およびＲＡＭ１０１２を含む。ＲＯＭ１０１１は、例えば、ＢＩＯＳ（Basic Input Output System）等のブートプログラムを記憶する。ハードディスクドライブインタフェース１０３０は、ハードディスクドライブ１０３１に接続される。ディスクドライブインタフェース１０４０は、ディスクドライブ１０４１に接続される。ディスクドライブ１０４１には、例えば、磁気ディスクや光ディスク等の着脱可能な記憶媒体が挿入される。シリアルポートインタフェース１０５０には、例えば、マウス１０５１およびキーボード１０５２が接続される。ビデオアダプタ１０６０には、例えば、ディスプレイ１０６１が接続される。

[0060] ここで、ハードディスクドライブ１０３１は、例えば、ＯＳ（Operating System）１０９１、アプリケーションプログラム１０９２、プログラムモジュール１０９３およびプログラムデータ１０９４を記憶する。上記実施形態で説明した各情報は、例えばハードディスクドライブ１０３１やメモリ１０

１０に記憶される。

[0061] また、信号抽出プログラムは、例えば、コンピュータ１０００によって実行される指令が記述されたプログラムモジュール１０９３として、ハードディスクドライブ１０３１に記憶される。具体的には、上記実施形態で説明した信号抽出装置１０が実行する各処理が記述されたプログラムモジュール１０９３が、ハードディスクドライブ１０３１に記憶される。

[0062] また、信号抽出プログラムによる情報処理に用いられるデータは、プログラムデータ１０９４として、例えば、ハードディスクドライブ１０３１に記憶される。そして、ＣＰＵ１０２０が、ハードディスクドライブ１０３１に記憶されたプログラムモジュール１０９３やプログラムデータ１０９４を必要に応じてＲＡＭ１０１２に読み出して、上述した各手順を実行する。

[0063] なお、信号抽出プログラムに係るプログラムモジュール１０９３やプログラムデータ１０９４は、ハードディスクドライブ１０３１に記憶される場合に限られず、例えば、着脱可能な記憶媒体に記憶されて、ディスクドライブ１０４１等を介してＣＰＵ１０２０によって読み出されてもよい。あるいは、信号抽出プログラムに係るプログラムモジュール１０９３やプログラムデータ１０９４は、ＬＡＮやＷＡＮ（Wide Area Network）等のネットワークを介して接続された他のコンピュータに記憶され、ネットワークインタフェース１０７０を介してＣＰＵ１０２０によって読み出されてもよい。

[0064] 以上、本発明者によってなされた発明を適用した実施形態について説明したが、本実施形態による本発明の開示の一部をなす記述および図面により本発明は限定されることはない。すなわち、本実施形態に基づいて当業者等によりなされる他の実施形態、実施例および運用技術等は全て本発明の範疇に含まれる。

符号の説明

[0065] １０ 信号抽出装置

１１ 入力部

１２ 出力部

１３　通信制御部

１４　記憶部

１４ａ、１４０ａ、１４１ａ、１４２ａ　モデル

１５　制御部

１５ａ　取得部

１５ｂ　学習部

１５ｃ　推定部

１５ｄ　抽出部

請求の範囲

[請求項1] 複数音源の混合信号と、目的音源の手がかり情報とを取得する取得部と、

取得された前記混合信号と前記目的音源の手がかり情報と、サンプリングされたステップにおけるデータとを学習済みモデルに入力し、該データに対応するスコアを推定する推定部と、

推定された前記スコアを用いて、前記複数音源の混合信号から前記目的音源の信号を抽出する抽出部と、

を有することを特徴とする信号抽出装置。

[請求項2] 前記学習済みモデルは、ガウシアンノイズから生成対象のデータにステップごとに変化させる逆過程により該データを生成するDiffusion modelであることを特徴とする請求項１に記載の信号抽出装置。

[請求項3] 前記抽出部は、所定回数の各処理において抽出された前記目的音源の信号の平均値を抽出結果とすることを特徴とする請求項１に記載の信号抽出装置。

[請求項4] 信号抽出装置が実行する信号抽出方法であって、

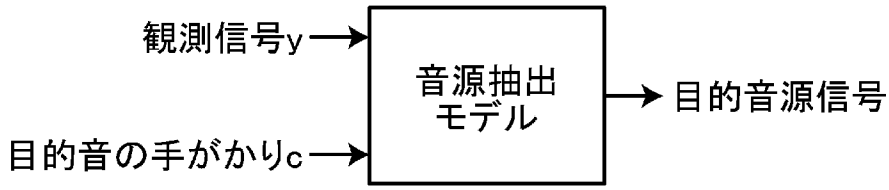
複数音源の混合信号と、目的音源の手がかり情報とを取得する取得工程と、

取得された前記混合信号と前記目的音源の手がかり情報と、サンプリングされたステップにおけるデータとを学習済みモデルに入力し、該データに対応するスコアを推定する推定工程と、

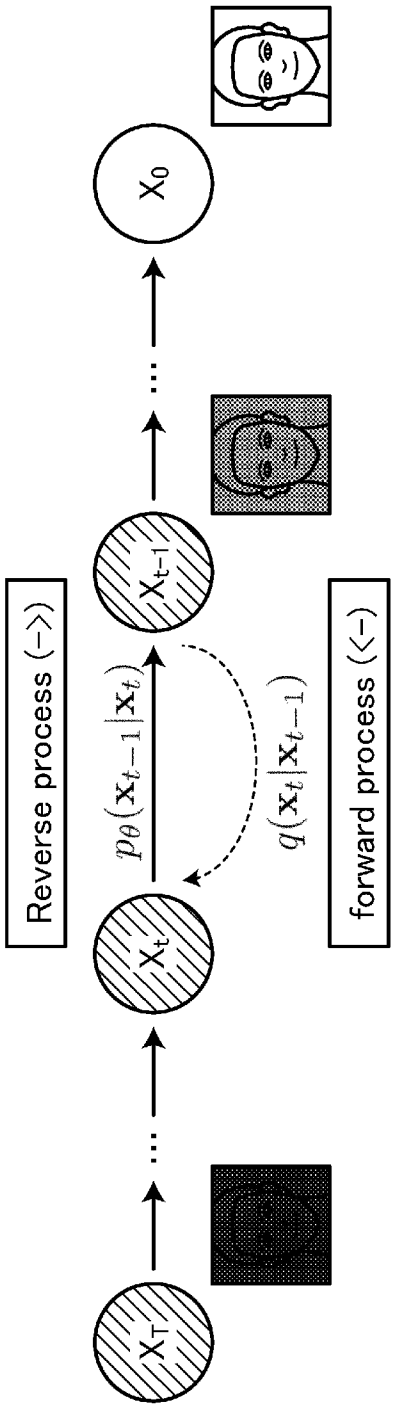
推定された前記スコアを用いて、前記複数音源の混合信号から前記目的音源の信号を抽出する抽出工程と、

を含んだことを特徴とする信号抽出方法。

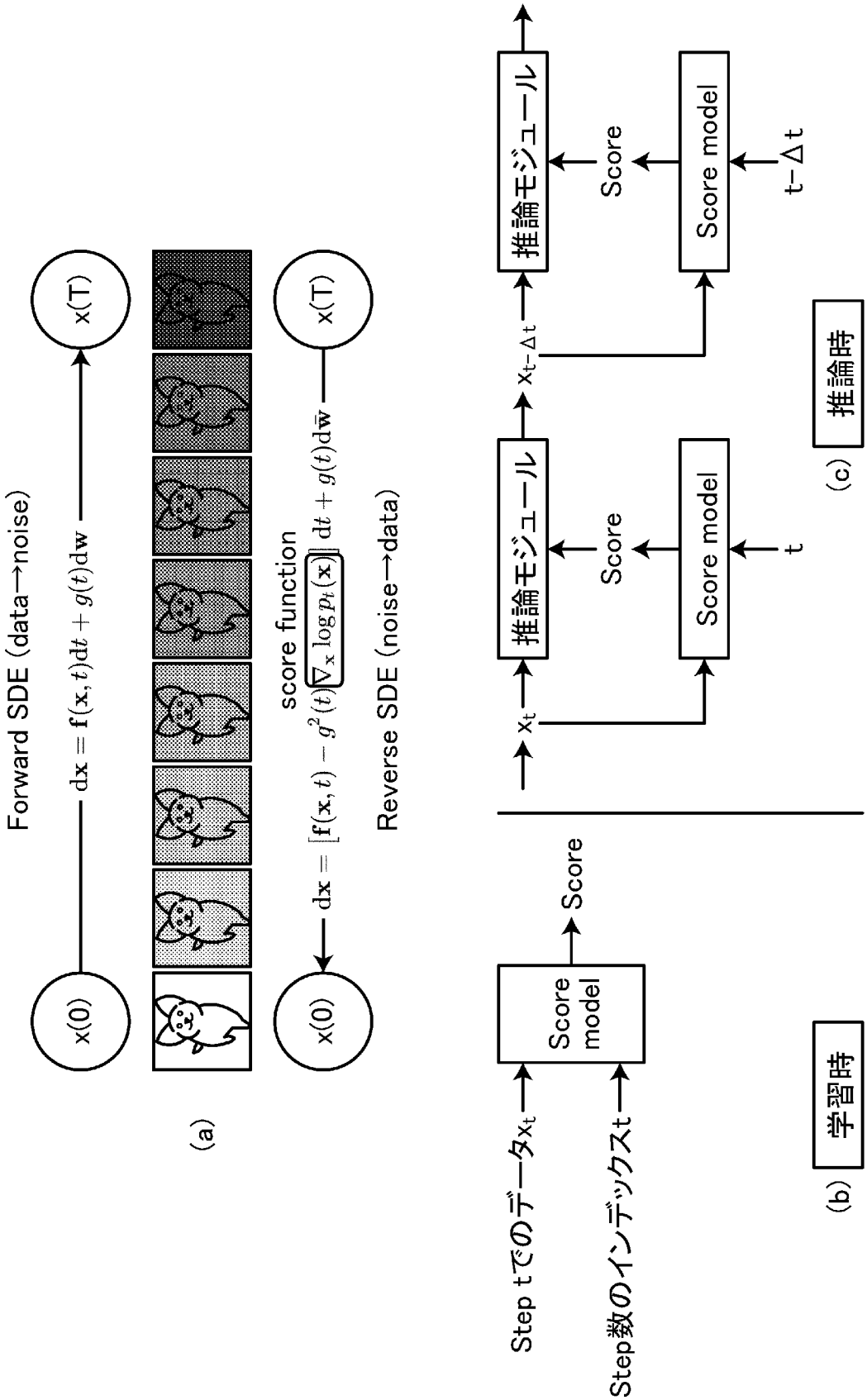
[図1]



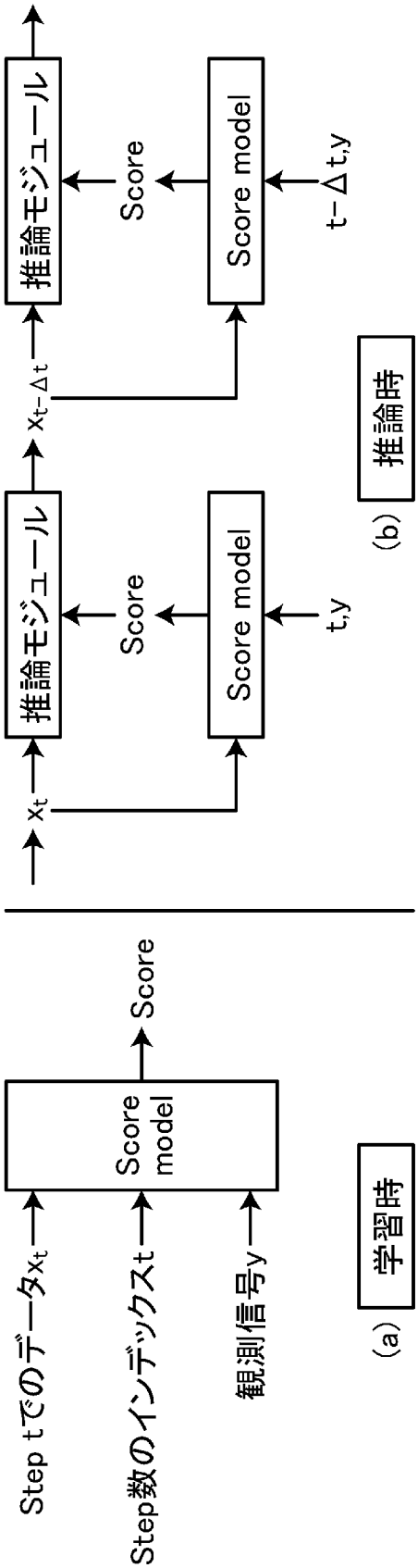
[図2]


x_0
x_{t-1}
x_t
x_T
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
Reverse process (->)
forward process (<-)

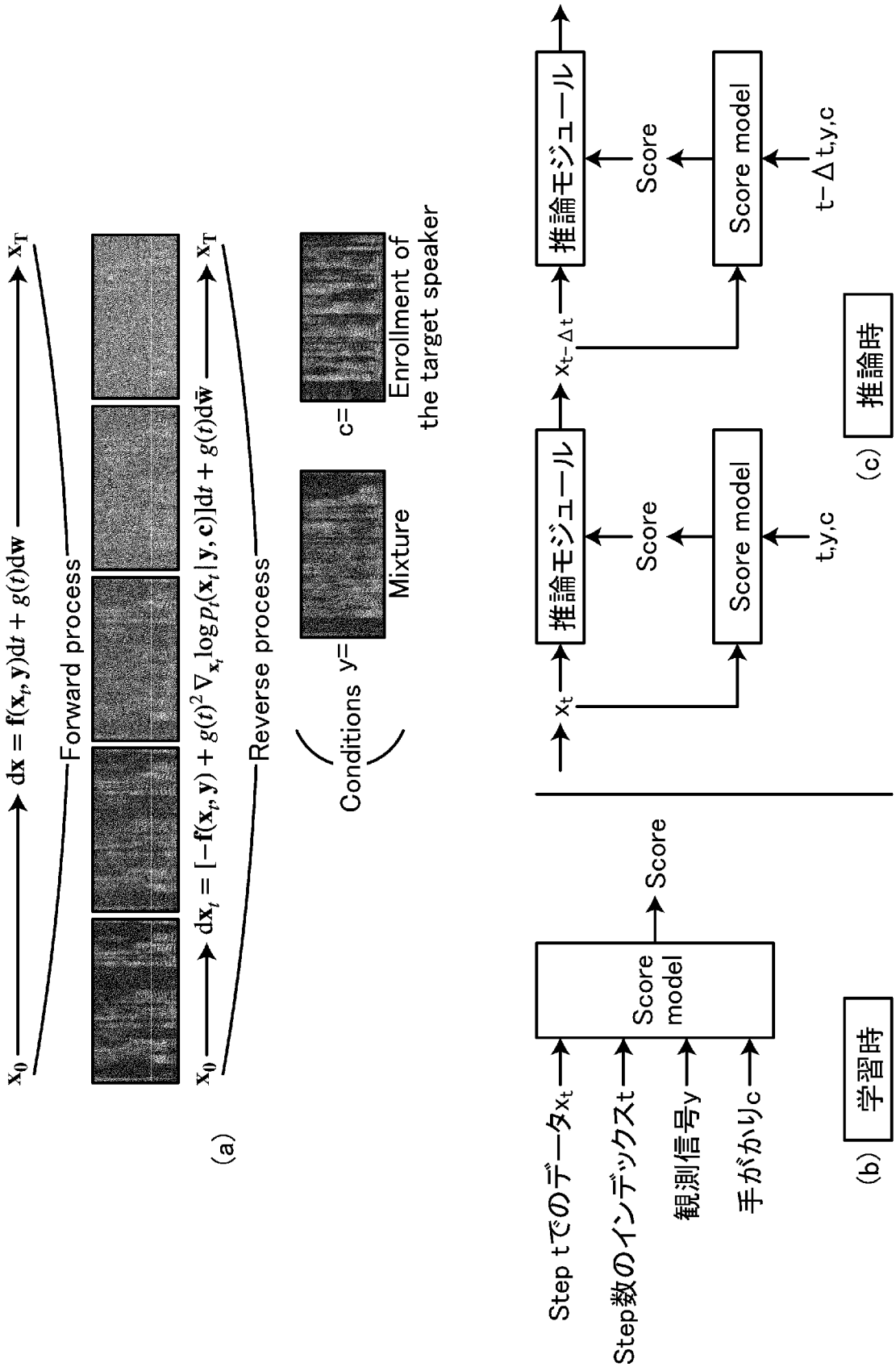
[図3]


Forward SDE (data→noise)
$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$
x(0)
x(T)
(a)
score function
$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]\,\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$
Reverse SDE (noise→data)
x(T)
x(0)
Step tでのデータx_t
Step数のインデックスt
Score model
Score
(b) 学習時
x_t
推論モジュール
Score
Score model
t
$x_{t-\Delta t}$
推論モジュール
Score
Score model
t−Δt
(c) 推論時

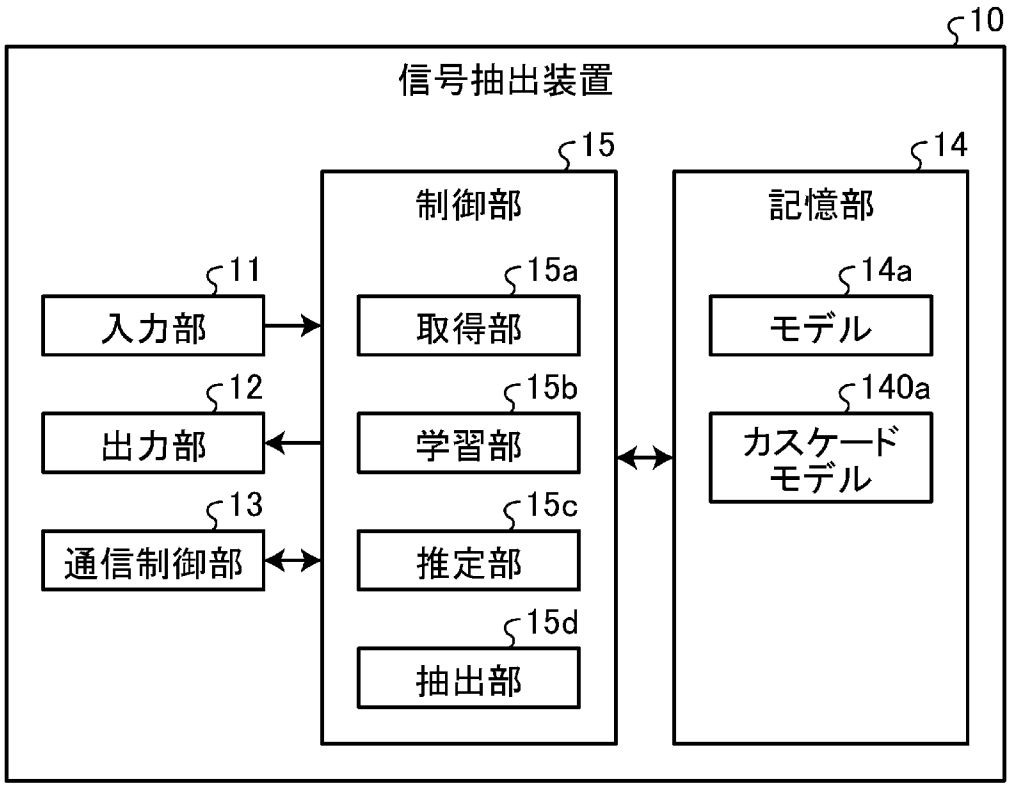
[図4]


Step tでのデータx_t
Step数のインデックスt
観測信号y
Score model
Score
(a) 学習時
x_t
$x_{t-\Delta t}$
推論モジュール
Score
Score model
t,y
$t-\Delta t,y$
(b) 推論時

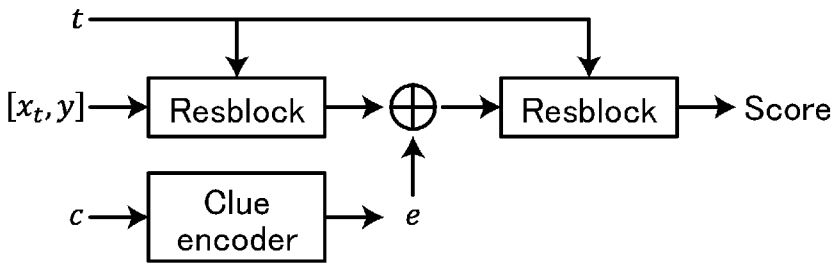
[図5]


x_0
$dx = f(x_t, y)dt + g(t)dw$
x_T
Forward process
x_0
$dx_t = [-f(x_t, y) + g(t)^2 \nabla_{x_t} \log p_t(x_t|y, c)]dt + g(t)d\bar{w}$
x_T
Reverse process
(a)
Conditions y=
Mixture
c=
Enrollment of the target speaker
Step tでのデータ x_t
Step数のインデックスt
観測信号y
手がかりc
Score model
Score
(b) 学習時
x_t
推論モジュール
Score
Score model
t,y,c
$x_{t-\Delta t}$
推論モジュール
Score
Score model
$t-\Delta t$,y,c
(c) 推論時

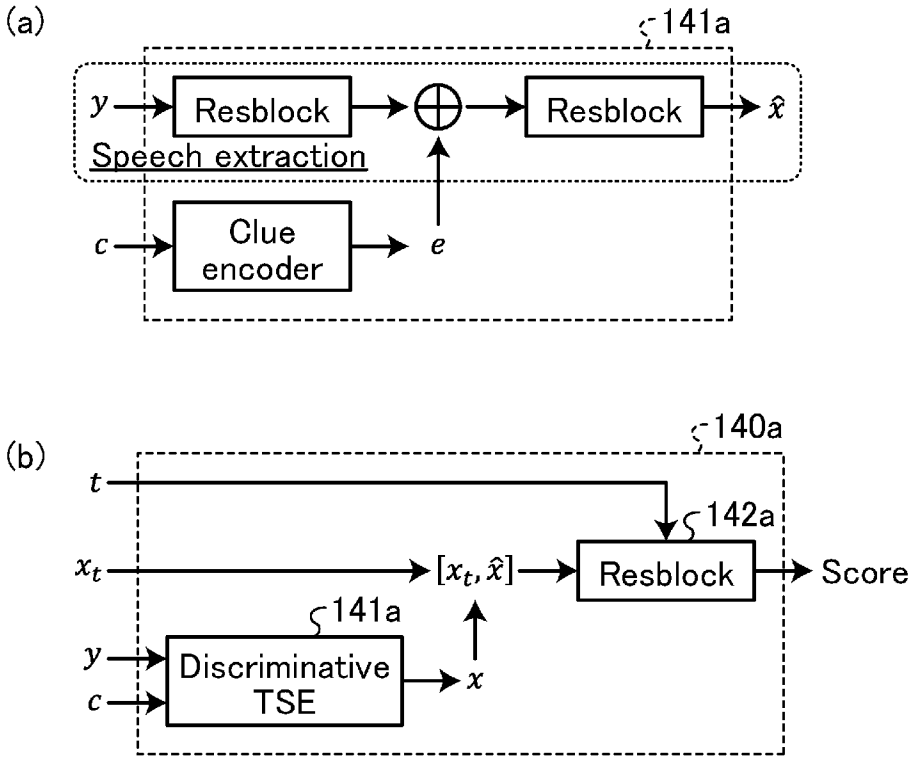
[図6]


10
信号抽出装置
15
制御部
14
記憶部
11
入力部
15a
取得部
14a
モデル
12
出力部
15b
学習部
140a
カスケード
モデル
13
通信制御部
15c
推定部
15d
抽出部

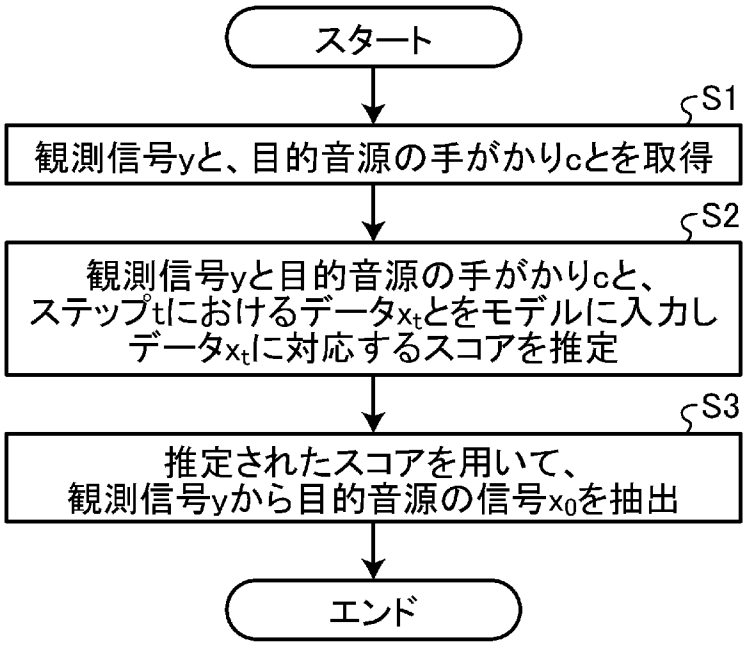
[図7]


t
[x_t, y]
Resblock
Resblock
Score
c
Clue encoder
e

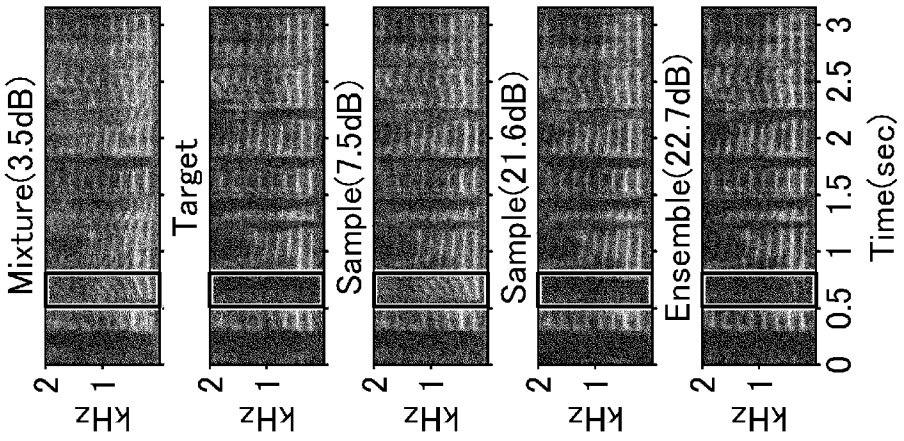
[図8]


(a)
141a
y
Resblock
Resblock
x̂
Speech extraction
c
Clue encoder
e
(b)
140a
t
142a
x_t
[x_t, x̂]
Resblock
Score
141a
y
c
Discriminative TSE
x

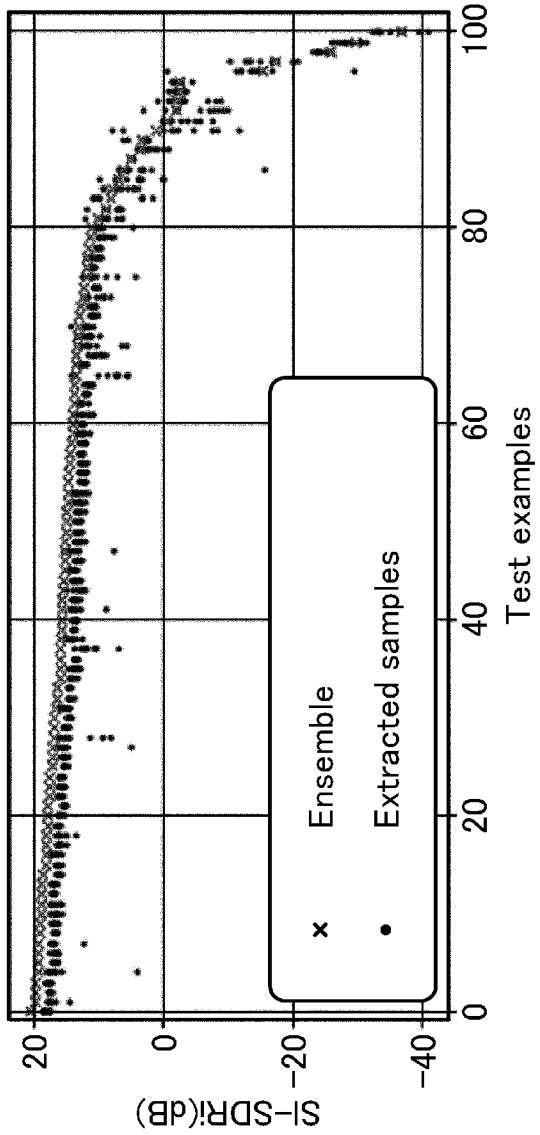
[図9]


スタート
S1
観測信号yと、目的音源の手がかりcとを取得
S2
観測信号yと目的音源の手がかりcと、
ステップtにおけるデータx_tとをモデルに入力し
データx_tに対応するスコアを推定
S3
推定されたスコアを用いて、
観測信号yから目的音源の信号x_0を抽出
エンド

[図10]


Mixture(3.5dB)
Target
Sample(7.5dB)
Sample(21.6dB)
Ensemble(22.7dB)
kHz
2
1
0 0.5 1 1.5 2 2.5 3
Time(sec)

(b)

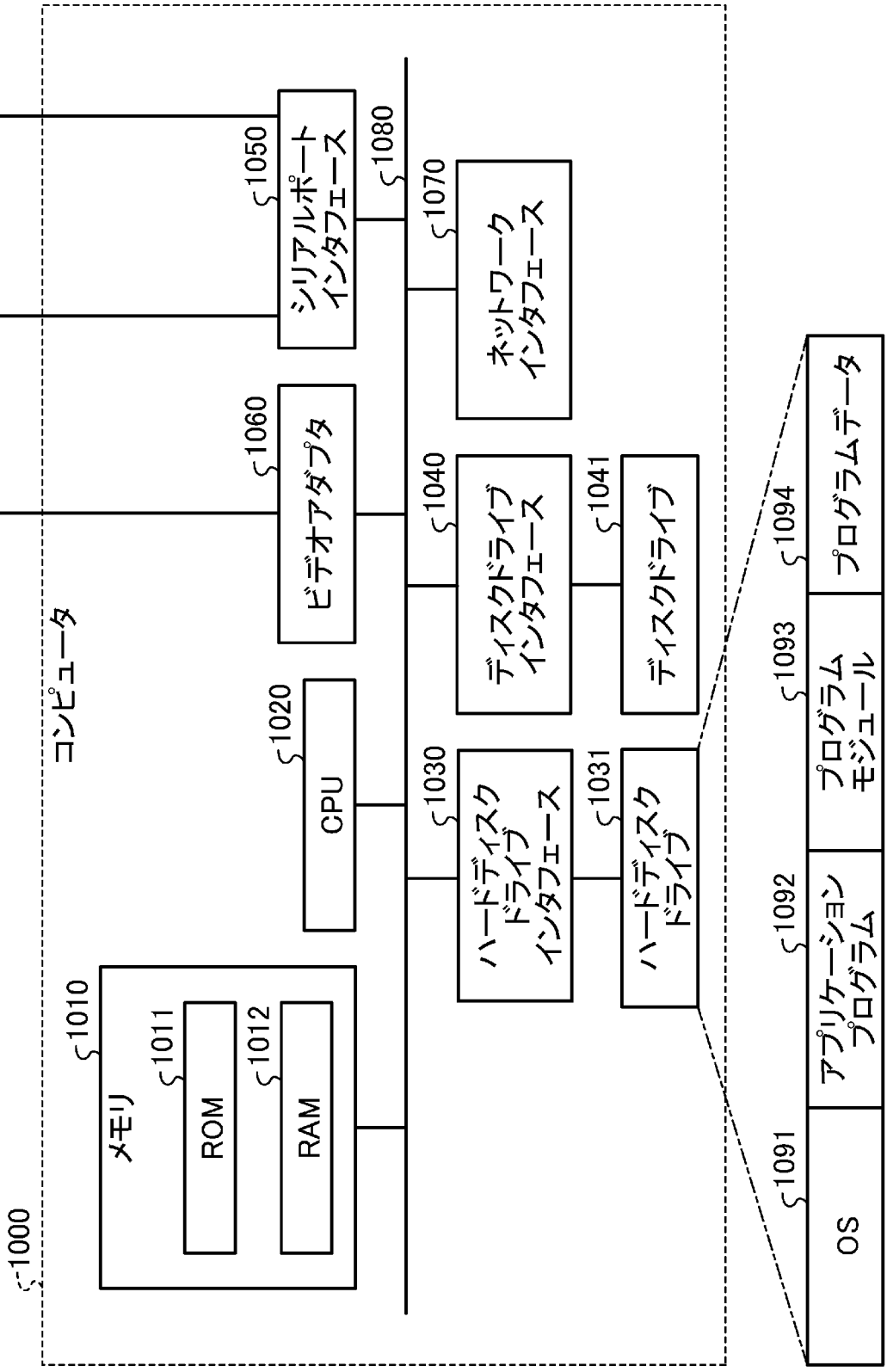

SI-SDRi(dB)
20
0
−20
−40
0 20 40 60 80 100
Test examples
Ensemble
Extracted samples

(a)

[図11]

	PESQ
従来法(Discriminative TSE)	2.58
本発明のモデル	2.56
本発明のモデルのアンサンブル	2.90
本発明のカスケードモデル	2.79
本発明のカスケードモデルのアンサンブル	**3.08**

[図12]


1000
コンピュータ
1010
メモリ
1011
ROM
1012
RAM
1020
CPU
1030
ハードディスクドライブインタフェース
1031
ハードディスクドライブ
1040
ディスクドライブインタフェース
1041
ディスクドライブ
1050
シリアルポートインタフェース
1052
キーボード
1051
マウス
1060
ビデオアダプタ
1061
ディスプレイ
1070
ネットワークインタフェース
1080
1091
OS
1092
アプリケーションプログラム
1093
プログラムモジュール
1094
プログラムデータ

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2023/026455

A. CLASSIFICATION OF SUBJECT MATTER

G10L 21/0272(2013.01)i; ***G10L 21/0208***(2013.01)i; ***G10L 25/30***(2013.01)i
FI: G10L21/0272 100Z; G10L25/30; G10L21/0208 100Z

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L21/0272; G10L21/0208; G10L25/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2023
Registered utility model specifications of Japan 1996-2023
Published registered utility model applications of Japan 1994-2023

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2018/047643 A1 (SONY CORPORATION) 15 March 2018 (2018-03-15) entire text, all drawings	1-4

☐ Further documents are listed in the continuation of Box C.

☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 August 2023	**05 September 2023**
Name and mailing address of the ISA/JP **Japan Patent Office (ISA/JP)** **3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915** **Japan**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2023/026455

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
WO 2018/047643 A1	15 March 2018	US 2019/0208320 A1 entire text, all drawings EP 3511937 A1 CN 109661705 A	

国際調査報告

国際出願番号
PCT/JP2023/026455

A.　発明の属する分野の分類（国際特許分類（ＩＰＣ））

G10L 21/0272(2013.01)i; G10L 21/0208(2013.01)i; G10L 25/30(2013.01)i
FI: G10L21/0272 100Z; G10L25/30; G10L21/0208 100Z

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

G10L21/0272; G10L21/0208; G10L25/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０２３年
日本国実用新案登録公報	１９９６－２０２３年
日本国登録実用新案公報	１９９４－２０２３年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C.　関連すると認められる文献

引用文献のカテゴリー*	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	WO 2018/047643 A1 (ソニー株式会社) 15.03.2018 (2018-03-15) 全文、全図	１－４

☐ Ｃ欄の続きにも文献が列挙されている。

☑ パテントファミリーに関する別紙を参照。

*　引用文献のカテゴリー
“A” 特に関連のある文献ではなく、一般的技術水準を示すもの
“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
“O” 口頭による開示、使用、展示等に言及する文献
“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの
“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
“Y” 特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
“&” 同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
08.08.2023	05.09.2023

名称及びあて先	権限のある職員（特許庁審査官）
日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目４番３号	大野　弘 5Z 9175 電話番号 03-3581-1101 内線 3591

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2023/026455

引用文献	公表日	パテントファミリー文献	公表日
WO 2018/047643 A1	15.03.2018	US 2019/0208320 A1 全文、全図 EP 3511937 A1 CN 109661705 A	