(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
2020年10月1日(01.10.2020)

WIPO | PCT

(10) 国際公開番号
WO 2020/195924 A1

(51) 国際特許分類:
G10L 21/0272 (2013.01) *G10L 21/028* (2013.01)

(21) 国際出願番号： PCT/JP2020/011008

(22) 国際出願日： 2020年3月13日(13.03.2020)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(30) 優先権データ：
特願 2019-059819 2019年3月27日(27.03.2019) JP

(71) 出願人: ソニー株式会社 (**SONY CORPORATION**) [JP/JP]; 〒1080075 東京都港区港南１丁目７番１号 Tokyo (JP).

(72) 発明者: 高橋 直也 (**TAKAHASHI Naoya**); 〒1080075 東京都港区港南１丁目７番１号 ソニー株式会社内 Tokyo (JP).

(74) 代理人: 西川 孝，外 (**NISHIKAWA Takashi** et al.); 〒1700013 東京都豊島区東池袋３丁目９番１０号 池袋ＦＮビル４階 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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, 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, GH, GM, KE, LR, LS,

(54) **Title:** SIGNAL PROCESSING DEVICE, METHOD, AND PROGRAM

(54) 発明の名称： 信号処理装置および方法、並びにプログラム

(57) **Abstract:** The present art relates to a signal processing device, method, and program that facilitate sound source separation. This signal processing device includes a sound source separating unit that recursively performs sound separation to an input sound signal according to a predetermined sound-source separation model that is learned in advance so as to separate a predetermined sound source from a training sound signal including the predetermined sound source. The present art is applicable to signal processing devices.

(57) 要約：本技術は、より簡単に音源分離をすることができるようにする信号処理装置および方法、並びにプログラムに関する。 信号処理装置は、所定の音源を含む学習用音響信号から所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離を、入力された音響信号に対して再帰的に行う音源分離部を備える。本技術は信号処理装置に適用することができる。

WO 2020/195924 A1

［続葉有］

MW, MZ, NA, RW, 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, 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)）

明　細　書

発明の名称 ： 信号処理装置および方法、並びにプログラム

技術分野

[0001] 本技術は、信号処理装置および方法、並びにプログラムに関し、特により簡単に音源分離をすることができるようにした信号処理装置および方法、並びにプログラムに関する。

背景技術

[0002] 例えば複数話者の音声認識（例えば、特許文献１参照）やキャプショニング、音声の明瞭化など、複数話者の同時発話を分離して扱いたいという状況は多く存在する。

[0003] 従来、複数の話者の発話が含まれた混合音声の音響信号を、各話者の音響信号に分離する音源分離手法として、方向情報を用いる手法（例えば、特許文献２参照）や、音源の独立性を仮定する手法が提案されている。

[0004] しかし、それらの手法では、単一のマイクロホンでの実現や、複数の音源からの音の到来方向が同じ方向である状況での対応が困難であった。

[0005] そこで、このような状況で同時に発話された音声を分離する手法として、Deep Clustering（例えば、非特許文献１参照）やPermutation Invariant Training（例えば、非特許文献２参照）が知られている。

先行技術文献

特許文献

[0006] 特許文献1：特表２０１７－５１５１４０号公報

特許文献2：特開２０１０－１１２９９５号公報

非特許文献

[0007] 非特許文献1：J. R. Hershey, Z. Chen, and J. Le Roux, “Deep Clustering : Discriminative Embeddings for Segmentation and Separation”

非特許文献2：M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant train

ing of deep recurrent neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1901-1913, 2017.

発明の概要

発明が解決しようとする課題

[0008] しかしながら、上述した技術では、話者数が未知である混合音声から各話者の発話を分離することは容易ではなかった。

[0009] 例えばDeep ClusteringやPermutation Invariant Trainingでは、同時に発話した話者の数が既知であることが前提となっている。

[0010] しかし、一般的に話者数が未知である状況は数多く存在する。そのような場合、これらの手法では話者数を推定するモデルが別途必要となり、話者数ごと話者の発話を分離する音源分離モデル（分離アルゴリズム）を用意するなど、アルゴリズムを切り替える必要がある。

[0011] そのため、これらの手法により話者数が未知の混合音声から話者ごとの発話を分離しようとすると、開発時間の増大や音源分離モデルを保持しておくためのメモリ量の増加が生じ、さらに話者数の推定が正しく行われなかったときには大幅に性能が悪化してしまう。

[0012] 本技術は、このような状況に鑑みてなされたものであり、より簡単に音源分離をすることができるようにするものである。

課題を解決するための手段

[0013] 本技術の一側面の信号処理装置は、所定の音源を含む学習用音響信号から前記所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離を、入力された音響信号に対して再帰的に行う音源分離部を備える。

[0014] 本技術の一側面の信号処理方法またはプログラムは、所定の音源を含む学習用音響信号から前記所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離を、入力された音響信号に対して再帰的に行うステップを含む。

[0015] 本技術の一側面においては、所定の音源を含む学習用音響信号から前記所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離が、入力された音響信号に対して再帰的に行われる。

図面の簡単な説明

[0016] [図1]再帰的な音源分離について説明する図である。

[図2]信号処理装置の構成例を示す図である。

[図3]音源分離処理について説明するフローチャートである。

[図4]再帰的な音源分離について説明する図である。

[図5]信号処理装置の構成例を示す図である。

[図6]音源分離処理について説明するフローチャートである。

[図7]コンピュータの構成例を示す図である。

発明を実施するための形態

[0017] 以下、図面を参照して、本技術を適用した実施の形態について説明する。

[0018] 〈第１の実施の形態〉

〈本技術について〉

まず、本技術の概要について説明する。ここでは、複数の話者が同時に、または異なるタイミングで発話したときの混合音声を１または複数のマイクロホンにより収音することで得られた入力音響信号から、単一の音源分離モデルを用いて各話者の発話（音声）を分離する例について説明する。

[0019] 特に、ここでは入力音響信号に基づく混合音声に含まれる話者数は未知であるものとする。本技術では、単一の音源分離モデルを用いて、入力音響信号に対して再帰的に音源分離を行うことで、より簡単に入力音響信号から不特定の未知数の各話者の発話（音声）を分離することができるようにした。

[0020] なお、ここでは音源の音が話者の発話である例について説明するが、これに限らず、動物の鳴き声や楽器の音など、どのようなものであってもよい。

[0021] 本技術で用いられる音源分離モデルは、入力音声を話者に応じて分離するように学習されたニューラルネットワーク等のモデルである。すなわち、音源分離モデルは、音源としての話者の発話を含む混合音声の学習用音響信号

から、話者の発話の音響信号を分離するように予め学習されたものである。

[0022] 音源分離モデルは、所定の音源分離アルゴリズムに従って演算係数を用いた演算を行うことで、入力音響信号を各音源（話者）の音響信号（以下、分離信号とも称する）に分離するものであり、音源分離アルゴリズムと演算係数により実現される。

[0023] 本技術では、話者数が未知または既知である混合音声の入力音響信号に対して音源分離モデルを用いた音源分離が行われる。

[0024] そして、得られた分離信号に基づいて、所定の終了条件が満たされるか否かが判定され、終了条件が満たされたと判定されるまで、分離信号に対して同一の音源分離モデルによる音源分離が再帰的に行われ、最終的に各音源（話者）の分離信号が得られる。

[0025] ここで、具体的な例として、音源としての２人の話者の発話が含まれる学習用音響信号を、一方の話者の発話を含む分離信号と、他方の話者の発話を含む分離信号とに分離するように学習された２話者分離モデルを音源分離モデルとして用いる場合について説明する。

[0026] このような音源分離モデルは、例えばDeep ClusteringやPermutation Invariant Trainingなどの学習手法による学習によって得ることができる。

[0027] ２話者分離モデルでは、話者数が２人である混合音声の入力音響信号が入力されたときには、各話者の発話（音声）の分離信号を音源分離結果として出力することが期待される。

[0028] また、２話者分離モデルでは、話者数が１人である音声の入力音響信号が入力されたときには、その１人の話者の発話の分離信号と、無音の分離信号とを音源分離結果として出力することが期待される。

[0029] 一方、２話者分離モデルの入力、つまり入力音響信号が３話者以上の複数話者の混合音声の信号である場合には、そのような混合音声は２話者分離モデルの学習時には現れなかった入力となる。

[0030] この場合、３話者の混合音声の入力に対して、例えば図１に示すように一方の分離信号に２話者の発話（音声）が含まれるように音源分離が行われる

。

[0031] 図１に示す例では、入力音響信号に基づく混合音声には話者PS1乃至話者PS3の３人の話者の発話が含まれている。

[0032] このような入力音響信号に対して、矢印Q11に示すように２話者分離モデルにより音源分離、すなわち話者分離を行った結果、一方の分離信号には話者PS1の発話のみが含まれ、他方の分離信号には話者PS2および話者PS3の発話のみが含まれるように混合音声が分離される。

[0033] また、例えば矢印Q12に示すように話者PS1の発話のみが含まれている分離信号に対して、２話者分離モデルによりさらに音源分離を行った結果、一方の分離信号には話者PS1の発話のみが含まれ、他方の分離信号は無音信号となるように音声が分離される。

[0034] 同様に、例えば矢印Q13に示すように話者PS2および話者PS3の発話のみが含まれている分離信号に対して、２話者分離モデルによりさらに音源分離を行った結果、一方の分離信号には話者PS2の発話のみが含まれ、他方の分離信号には話者PS3の発話のみが含まれるように混合音声が分離される。

[0035] このように入力音響信号に対して同一の２話者分離モデルにより再帰的に音源分離を行うと、話者PS1乃至話者PS3のそれぞれのみが含まれる分離信号が得られる。

[0036] この例では、矢印Q11に示す１回目の音源分離が行われた時点において、得られた分離信号には最大でも２話者の発話しか含まれておらず、殆どの場合、入力音響信号が３話者の発話の分離信号と無音の分離信号とに分離されることはない。

[0037] したがって、１回目の音源分離が行われた時点で、全ての分離信号は２話者分離モデルにより解くことが可能な音声、つまり話者ごとの分離信号を得ることができる信号となっており、そのような分離信号に対して矢印Q12や矢印Q13に示すように再帰的な音源分離を行うことで各話者の分離信号を得ることができる。

[0038] なお、入力音響信号が４以上の話者数の発話の混合音声である場合でも、

再帰的に行う音源分離の回数を増やすことによって、最終的に話者ごとの分離信号を得ることができる。

[0039] また、入力音響信号に対して再帰的に音源分離を行って各話者の分離信号を分離（抽出）する場合、入力音響信号の混合音声の話者数が未知（不明）であるときには、再帰的な音源分離を終了するための終了条件が必要となる。

[0040] この終了条件は、音源分離で得られた分離信号が１人の話者の発話のみが含まれる信号であるときに満たされる条件、換言すれば、分離信号に２以上の話者の発話が含まれていない状態となったときに満たされる条件とされる。

[0041] ここでは、一例として音源分離により得られた一方の分離信号が無音信号である場合、より詳細には一方の分離信号の平均レベル（エネルギ）が所定の閾値以下である場合に終了条件が満たされた、つまり各話者の分離信号が得られたとされるものとする。

[0042] 以上のような本技術によれば、入力音響信号の話者数が未知である場合であっても、話者数を推定するモデルや話者数ごとの音源分離モデル、音源の方向を示す方向情報などを必要とせずに簡単に音源分離を行い、各音源（話者）の分離信号を得ることができる。

[0043] したがって、本技術では音源分離モデル等の開発時間の増大や音源分離モデルを保持しておくためのメモリ量の増加を大幅に抑制することができる。

[0044] すなわち、本技術では、入力音響信号の話者数によらず、１つの音源分離モデルで各話者の分離信号を得ることができるので、システムの単純化や必要メモリ量の削減、音源分離モデルの開発の一本化などを実現することができる。

[0045] しかも、本技術では再帰的に音源分離を行うことで、各回の音源分離で解く問題（タスク）を簡単にすることができ、その結果、分離性能を向上させることができる。

[0046] なお、ここでは音源分離モデルとして２話者分離モデルを用いる例につい

て説明した。しかし、これに限らず、３話者分離モデルなど、入力音響信号を３以上の話者ごとの分離信号に分離する複数話者の話者分離モデルなどにより再帰的な音源分離を行うようにしてもよい。

[0047] 例えば３話者分離モデルは、音源として３人の話者の発話が含まれる学習用音響信号を、３人の話者のそれぞれの発話を含む３つの分離信号のそれぞれ、つまり３つの話者ごとの分離信号に分離するように学習された話者分離モデルである。

[0048] 〈信号処理装置の構成例〉

次に、本技術を適用した信号処理装置について説明する。

[0049] 本技術を適用した信号処理装置は、例えば図２に示すように構成される。

[0050] 図２に示す信号処理装置１１は、音源分離部２１および終了判定部２２を有している。

[0051] 音源分離部２１には、外部から入力音響信号が入力される。また、音源分離部２１は、予め学習により得られた音源分離モデルを保持している。

[0052] なお、この実施の形態では、入力音響信号は、話者数、特に同時に発話を行った話者数が未知である混合音声の音響信号であるものとして説明を行う。また、ここでは音源分離部２１が保持している音源分離モデルは２話者分離モデルであるとする。

[0053] 音源分離部２１は、終了判定部２２から供給される終了判定の結果に応じて、供給された入力音響信号に対して、保持している音源分離モデルに基づいて音源分離を再帰的に行い、その結果得られた分離信号を終了判定部２２に供給する。

[0054] 終了判定部２２は、音源分離部２１から供給された分離信号に基づいて、再帰的な音源分離を終了するか否か、すなわち終了条件が満たされるか否かを判定する終了判定を行い、その判定結果を音源分離部２１に供給する。

[0055] また、終了判定部２２は、終了条件が満たされたと判定された場合、音源分離により得られた分離信号を各話者の発話の音響信号として後段に出力する。

[0056] 〈音源分離処理の説明〉

続いて、図３のフローチャートを参照して、信号処理装置１１により行われる音源分離処理について説明する。

[0057] ステップＳ１１において音源分離部２１は、保持している音源分離モデルに基づいて、供給された入力音響信号に対する音源分離を行い、その結果得られた分離信号を終了判定部２２に供給する。

[0058] 具体的には、音源分離部２１は、音源分離モデルを構成する演算係数と、入力音響信号とに基づいて、音源分離モデルに対応する音源分離アルゴリズムに従って演算処理を行い、音源分離モデルの出力である２つの分離信号を得る。

[0059] ステップＳ１２において終了判定部２２は、音源分離部２１から供給された分離信号に基づいて、１度の音源分離により得られた２つの分離信号のペア（組）ごとに終了判定を行い、全てのペアが終了条件を満たすか否かを判定する。

[0060] 具体的には、例えば終了判定部２２は１つのペアについて、そのペアを構成する２つの分離信号のうちの１つの分離信号の平均レベルが所定の閾値以下である場合、そのペアは終了条件を満たしていると判定する。

[0061] ステップＳ１２において全てのペアが終了条件を満たしていないと判定された場合、終了判定部２２は、終了条件を満たしていないペアを示す情報を終了判定の結果として音源分離部２１に供給し、その後、処理はステップＳ１３へと進む。

[0062] ステップＳ１３において音源分離部２１は、終了判定部２２から供給された終了判定の結果に基づいて、終了条件を満たしていないペアを構成する各分離信号に対して音源分離モデルによる音源分離を行い、その結果得られた分離信号を終了判定部２２に供給する。

[0063] 例えばステップＳ１３では、ステップＳ１１において用いられた音源分離モデルと同一の音源分離モデルが用いられて音源分離が行われる。

[0064] なお、例えばステップＳ１１においては３話者分離モデルが用いられて音

源分離が行われ、ステップＳ１３においては２話者分離モデルが用いられて音源分離が行われるなど、互いに異なる複数の音源分離モデルが用いられて音源分離が再帰的に行われてもよい。

[0065] ステップＳ１３の処理で再帰的な音源分離が行われると、その後、処理はステップＳ１２に戻り、全てのペアが終了条件を満たすと判定されるまで、上述した処理が繰り返し行われる。

[0066] 例えば図１に示した例においては、矢印Q12に示した音源分離では一方の分離信号が無音信号となるので、矢印Q12に示した音源分離の結果として得られた分離信号のペアは、終了条件を満たすことになる。

[0067] これに対して、図１の矢印Q13に示した音源分離では無音の分離信号が得られないため、終了条件を満たすとは判定されず、矢印Q13に示した音源分離で得られた２つの分離信号のそれぞれに対して、ステップＳ１３で再帰的な音源分離が行われることになる。

[0068] また、図３のステップＳ１２において全てのペアが終了条件を満たすと判定された場合、入力音響信号が各話者の分離信号に分離されたので、処理はステップＳ１４へと進む。

[0069] ステップＳ１４において終了判定部２２は、これまでの音源分離により得られた話者ごとの分離信号を後段に出力し、音源分離処理は終了する。

[0070] 以上のようにして信号処理装置１１は、終了条件が満たされるまで、入力音響信号に対して再帰的に音源分離を行い、各話者の分離信号を得る。このようにすることで、より簡単に、かつ十分な分離性能で音源分離をすることができる。

[0071] 〈第２の実施の形態〉

〈分離結果の合成について〉

ところで、音源分離モデルとして話者分離モデルを用いて、入力音響信号に対して再帰的に音源分離を行った場合、ある話者の発話が異なる分離結果、つまり異なる分離信号に分散してしまうこともある。

[0072] 具体的には、例えば図１に示したように話者PS1乃至話者PS3の発話が含ま

れる混合音声の入力音響信号に対して２話者分離モデルを用いて音源分離を行ったとする。

[0073] この場合、例えば図１の矢印Q11に示した音源分離の結果のように、ある話者の発話が１つの分離信号にのみ現れるのではなく、図４に示すように、ある話者の発話が２つの分離信号に分散して現れることがある。なお、図４において図１における場合と対応する部分には同一の符号を付してあり、その説明は適宜省略する。

[0074] 図４に示す例では、話者PS1乃至話者PS3の発話が含まれる混合音声の入力音響信号に対して、２話者分離モデルにより再帰的に音源分離（話者分離）が行われている。

[0075] ここでは、まず矢印Q21に示すように入力音響信号に対して音源分離が行われる。

[0076] その結果、話者PS1の発話と話者PS2の発話の一部とが含まれる分離信号、および話者PS3の発話と話者PS2の発話の一部とが含まれる分離信号が得られている。

[0077] すなわち、話者PS1や話者PS3の発話は１つの分離信号にのみ現れているが、話者PS2の発話は２つの分離信号に分散している。

[0078] ここで、矢印Q21に示した音源分離の結果として得られた、話者PS1の発話と話者PS2の発話の一部とが含まれる分離信号に対して、矢印Q22に示すように２話者分離モデルによる再帰的な音源分離を行うと、話者ごとの分離信号が得られる。

[0079] すなわち、この例では矢印Q22に示す音源分離の結果として、話者PS1の発話のみが含まれる分離信号と、話者PS2の発話の一部のみが含まれる分離信号とが得られている。

[0080] 同様に、矢印Q21に示した音源分離の結果として得られた、話者PS3の発話と話者PS2の発話の一部とが含まれる分離信号に対して、矢印Q23に示すように２話者分離モデルによる再帰的な音源分離を行うと、話者ごとの分離信号が得られる。

[0081] すなわち、この例では矢印Q23に示す音源分離の結果として、話者PS3の発話のみが含まれる分離信号と、話者PS2の発話の一部のみが含まれる分離信号とが得られている。

[0082] このような例においても、結果として各分離信号には１人の話者の発話のみが含まれている。但し、ここでは話者PS2の発話が２つの分離信号に分散してしまっている。

[0083] そこで、２以上の複数の分離音声、つまり複数の分離信号に分散してしまった同一話者の分離音声（発話）を合成することにより、分散した話者の発話を１つにまとめるようにしてもよい。

[0084] そのような場合、分離信号を入力とし、話者の識別結果を出力とする話者識別モデルを利用することができる。

[0085] 具体的には、例えば予め任意の多数の話者を識別するニューラルネットワーク等が話者識別モデルとして学習される。ここで、話者識別モデルの学習時の話者は、学習時の話者数が多ければ、実際に音源分離を行いたい話者が含まれている必要はない。

[0086] このようにして話者識別モデルが用意されると、この話者識別モデルが用いられて、音源分離により得られた分離信号、すなわち分離信号に対応する話者のクラスタリングが行われる。

[0087] クラスタリング時には、各分離信号が話者識別モデルに入力されて話者識別が行われる。

[0088] このとき、話者識別モデルの出力、つまり話者識別の結果、または話者識別モデルの中間層のアクティベーション（出力）、つまり話者識別結果を得るための演算処理における途中までの演算結果が、入力とされた分離信号についての話者を表す特徴量（speaker embedding）とされる。

[0089] なお、話者を表す特徴量の算出時には、分離信号の無音区間を無視して計算を行うことが可能である。

[0090] 各分離信号（分離音声）について特徴量が得られると、それらの特徴量同士の距離、つまり特徴量間の距離が求められ、特徴量間の距離が閾値以下で

ある分離信号は同一話者の分離信号とされる。

[0091] さらに、クラスタリングの結果、同一話者のものであるとされた複数の分離信号が合成され、合成により得られた１つの分離信号が、その話者の最終的な分離信号とされる。

[0092] したがって、例えば図４の例では、矢印Q22に示した音源分離によって得られた話者PS2の発話の一部のみが含まれる分離信号と、矢印Q23に示した音源分離によって得られた話者PS2の発話の一部のみが含まれる分離信号とが同一話者のものであるとされる。

[0093] そして、それらの分離信号を加算することで分離信号が合成され、その結果得られた１つの信号が、話者PS2の発話を含む最終的な分離信号として出力される。

[0094] 〈信号処理装置の構成例〉

以上のように音源分離により得られた分離信号のクラスタリングが行われる場合、信号処理装置は、例えば図５に示すように構成される。なお、図５において図２における場合と対応する部分には同一の符号を付してあり、その説明は適宜省略する。

[0095] 図５に示す信号処理装置５１は、音源分離部２１、終了判定部２２、および同一話者判定部６１を有している。

[0096] この信号処理装置５１の構成は、新たに同一話者判定部６１を設けた点で信号処理装置１１の構成と異なっており、その他の点では信号処理装置１１と同じ構成となっている。

[0097] 同一話者判定部６１は、再帰的な音源分離により得られた複数の分離信号が同一話者の信号であるか否かを判定する同一話者判定を行い、その判定結果に応じて同一話者の複数の分離信号を合成し、最終的な話者の分離信号を生成する。

[0098] より具体的には、同一話者判定部６１は予め学習により求められた話者識別モデルを保持しており、保持している話者識別モデルと、終了判定部２２から供給された話者ごとの分離信号とに基づいてクラスタリングを行う。す

なわち、同一話者判定部６１はクラスタリングを行うことで同一話者判定を行う。

[0099] また、同一話者判定部６１は、クラスタリングにより同一話者のものであるとされた分離信号を合成して、その話者の最終的な分離信号とするとともに、最終的に得られた各話者の分離信号を後段に出力する。

[0100] 〈音源分離処理の説明〉

続いて、図６のフローチャートを参照して、信号処理装置５１により行われる音源分離処理について説明する。

[0101] なお、ステップＳ４１乃至ステップＳ４３の処理は、図３のステップＳ１１乃至ステップＳ１３の処理と同様であるので、その説明は省略する。

[0102] ステップＳ４１乃至ステップＳ４３で再帰的な音源分離が行われ、各話者の分離信号が得られると、それらの分離信号が終了判定部２２から同一話者判定部６１に供給され、その後、処理はステップＳ４４へと進む。すなわち、ステップＳ４２において全てのペアが終了条件を満たすと判定された場合、処理はステップＳ４４へと進む。

[0103] ステップＳ４４において同一話者判定部６１は、保持している話者識別モデルと、終了判定部２２から供給された分離信号とに基づいて、それらの分離信号ごとに、話者を表す特徴量を算出する。

[0104] すなわち、同一話者判定部６１は、分離信号を入力として話者識別モデルによる演算を行うことで、分離信号ごとに話者を表す特徴量を算出する。

[0105] ステップＳ４５において同一話者判定部６１は、ステップＳ４４で求めた特徴量に基づいて、同一話者の分離信号があるか否かを判定する。すなわち、同一話者判定が行われる。

[0106] 例えば同一話者判定部６１は、全ての分離信号のうちの任意の２つの分離信号について、それらの２つの分離信号の特徴量間の距離を求め、その距離が所定の閾値以下である場合、それらの２つの分離信号は同一話者のもの（信号）であると判定する。

[0107] 同一話者判定部６１は、全ての分離信号を対象とし、２つの分離信号の組

み合わせとして取り得る全ての組み合わせについて同一話者のものであるか否かの判定を行う。

[0108] そして同一話者判定部６１は、全ての組み合わせで同一話者のものではないとの判定結果が得られた場合、ステップＳ４５において同一話者の分離信号がないと判定する。

[0109] 同一話者判定部６１では、以上のステップＳ４４およびステップＳ４５の処理がクラスタリングの処理として行われる。

[0110] ステップＳ４５において同一話者の分離信号があると判定された場合、ステップＳ４６において同一話者判定部６１は、同一話者のものであるとされた複数の分離信号を合成し、その話者の最終的な分離信号とする。

[0111] 同一話者の分離信号が合成されて、最終的な各話者の分離信号が得られると、その後処理はステップＳ４７へと進む。

[0112] 一方、ステップＳ４５において同一話者の分離信号がないと判定された場合、既に各話者の分離信号が得られているので、ステップＳ４６の処理はスキップされ、処理はステップＳ４７へと進む。

[0113] ステップＳ４５において同一話者の分離信号がないと判定されたか、またはステップＳ４６の処理が行われると、ステップＳ４７において同一話者判定部６１は、最終的に得られた話者ごとの分離信号を後段に出力し、音源分離処理は終了する。

[0114] 以上のようにして信号処理装置５１は、終了条件が満たされるまで入力音響信号に対して再帰的に音源分離を行うとともに、分離信号のクラスタリングを行って同一話者の分離信号を合成し、最終的な話者ごとの分離信号を得る。

[0115] このようにすることで、より簡単に、かつ十分な分離性能で音源分離をすることができる。特に信号処理装置５１では、同一話者の分離信号を合成することで、信号処理装置１１における場合よりもさらに分離性能を向上させることができる。

[0116] 〈第３の実施の形態〉

〈１対多話者分離モデルについて〉

ところで、以上においてはｍ人（但し、ｍ≧２）の話者の発話が含まれている混合音声の音響信号を、話者ごとのｍ個の分離信号に分離させるように学習したｍ話者分離モデルを用いて音源分離を行う例について説明した。

[0117] 特に、音源分離時には、所定の話者の発話が複数の分離信号に分散して現れる可能性があるため、第２の実施の形態ではクラスタリングを行い、適宜、分離信号を合成する例について説明した。

[0118] しかし、このような話者分離モデルに限らず、その他、例えば不確定の話者数に対して学習を行うことで得られる話者分離モデル（以下、１対多話者分離モデルとも称する）を用いて音源分離を行うようにしてもよい。

[0119] １対多話者分離モデルは、任意の未知（不確定）の話者数の混合音声の学習用音響信号を、所定の１人の話者の発話（音声）のみを含む分離信号と、混合音声に含まれる複数の話者のうちの上記の所定の１人の話者を除く残りの話者の発話を含む分離信号とに分離するように学習されたニューラルネットワーク等の話者分離モデルである。

[0120] ここで、１対多話者分離モデルによる音源分離の分離結果、すなわち１対多話者分離モデルの出力をヘッドとも称することとする。

[0121] 特に、ここでは１人の話者の発話が含まれる分離信号が出力される側をヘッド１とも称し、その他の残りの話者の発話が含まれる分離信号が出力される側をヘッド２とも称することとする。また、ヘッド１とヘッド２とを特に区別する必要のない場合には、単にヘッドと称することとする。

[0122] １対多話者分離モデルの学習時には、学習用音響信号の話者数ｍをランダムに変化させながら、その話者数ｍの学習用音響信号が用いられて損失関数Ｌが最小となるように学習が行われる。

[0123] このとき、話者数ｍは最大話者数Ｍ以下となるように設定される。また、１対多話者分離モデルは、常に、学習用音響信号の混合音声に含まれるｍ人の話者のうち、損失が最も小さくなる１人の話者の発話のみを含む分離信号がヘッド１の出力とされ、残りの（ｍ-1）人の話者の発話を含む分離信号がヘ

ッド２の出力とされるように学習される。

[0124] また、１対多話者分離モデルの学習時の損失関数Ｌは、例えば次式（１）で表される。

[0125] [数1]

$$L=\sum_j \min_i L_i^{1j}+L_i^{2j} \quad \cdots (1)$$

[0126] なお、式（１）においてｊは学習用音響信号、つまり学習用の混合音声を示すインデックスであり、ｉはｊ番目の混合音声に含まれる発話の話者を示すインデックスである。

[0127] また、式（１）においてL_i^{1j}は、ｊ番目の混合音声の学習用音響信号x^jを音源分離したときのヘッド１の出力$s'^1(x^j)$と、ｉ番目の話者の発話の音響信号s_i^jとを比較したときの損失関数を示している。この損失関数L_i^{1j}は、例えば次式（２）に示す二乗誤差で定義できる。

[0128] [数2]

$$L_i^{1j}=\left\| s'^1(x^j)-s_i^j \right\|^2 \quad \cdots (2)$$

[0129] さらに、式（１）におけるL_i^{2j}は、ｊ番目の混合音声の学習用音響信号x^jを音源分離したときのヘッド２の出力$s'^2(x^j)$と、ｉ番目の話者以外の残りの話者ｋの音響信号s_k^jの和とを比較したときの損失関数を示している。この損失関数L_i^{2j}は、例えば次式（３）に示す二乗誤差で定義できる。

[0130] [数3]

$$L_i^{2j}=\frac{1}{m-1}\left\| s'^2(x^j)-\sum_{k\neq i} s_k^j \right\|^2 \quad \cdots (3)$$

[0131] 以上のような学習により得られた１対多話者分離モデルは、常にヘッド１の出力として１話者の発話のみの分離信号が得られ、ヘッド２の出力として残りの話者の発話の分離信号が得られることが期待される。

[0132] したがって、例えば図１に示した例と同様に、１対多話者分離モデルにより入力音響信号に対して再帰的に音源分離を行うだけで、各話者の発話のみを含む分離信号が順番に分離されていくことが期待できる。

[0133] このように１対多話者分離モデルを利用する場合、例えば信号処理装置１１の音源分離部２１は、予め学習により得られた１対多話者分離モデルを音源分離モデルとして保持している。そして、信号処理装置１１は、図３を参照して説明した音源分離処理を行って、各話者の分離信号を得る。

[0134] 但し、この場合、ステップＳ１１やステップＳ１３では音源分離部２１は、１対多話者分離モデルに基づいて音源分離を行う。このとき、ヘッド１の出力は１話者の発話の分離信号となっているので、ヘッド２の出力（分離信号）に対して１対多話者分離モデルによる音源分離が再帰的に行われることになる。

[0135] また、ステップＳ１２では、最後に行った音源分離のヘッド２の出力（分離信号）の平均レベルが所定の閾値以下である場合、終了条件が満たされたと判定され、処理はステップＳ１４へと進む。

[0136] なお、ここでは１つの入力音響信号を入力として、２ヘッド、つまりヘッド１とヘッド２の２つの出力が得られる１対多話者分離モデルを用いる例について説明した。

[0137] しかし、これに限らず、例えば３ヘッドの出力が得られる１対多話者分離モデルを用いて音源分離を行うようにしてもよい。

[0138] そのような場合、例えばヘッド１乃至ヘッド３のうち、ヘッド１とヘッド２の出力が１人の話者の発話のみを含む分離信号となり、ヘッド３の出力がその他の残りの話者の発話を含む分離信号となるように学習される。

[0139] 〈第４の実施の形態〉

〈１対多話者分離モデルとクラスタリングの組み合わせについて〉

さらに、音源分離モデルとして１対多話者分離モデルを用いる場合においても、必ずしも音源、つまり話者ごとの発話が完全に分離できないことがある。すなわち、例えばヘッド１に出力されるべき話者の発話が、僅かにヘッド２の出力に漏れ出てしまうことがある。

[0140] したがって、このような場合には、図４を参照して説明したように再帰的な音源分離により得られた複数の分離信号に同一話者の発話が分散してしま

うことになる。但し、この場合、一方の分離信号に含まれる話者の発話は僅かに漏れ出た成分であるため、他方の分離信号に含まれる話者の発話と比べてはるかに音量が小さくなっている。

[0141] そこで、音源分離モデルとして１対多話者分離モデルを用いる場合においても、第２の実施の形態と同様にクラスタリングを行うようにしてもよい。

[0142] そのような場合、例えば信号処理装置５１の音源分離部２１は、予め学習により得られた１対多話者分離モデルを音源分離モデルとして保持している。

[0143] そして、信号処理装置５１は、図６を参照して説明した音源分離処理を行って、各話者の分離信号を得る。

[0144] 但し、この場合、ステップＳ４１やステップＳ４３では、第３の実施の形態における場合と同様に、音源分離部２１は１対多話者分離モデルに基づいて音源分離を行う。

[0145] また、ステップＳ４４では、上述した話者識別モデルの出力等が話者を表す特徴量として算出され、２つの分離信号の特徴量間の距離が閾値以下であるときに、それらの２つの分離信号は同一話者のものであると判定される。

[0146] その他、例えば分離信号の時間的なエネルギ変動が話者を表す特徴量とされ、２つの分離信号の特徴量の相関、つまり分離信号のエネルギ変動の相関が閾値以上である場合に、それらの２つの分離信号は同一話者のものであると判定されてもよい。

[0147] 〈その他の変形例１〉

〈単一話者判定モデルの利用について〉

ところで、以上において説明した各実施の形態では、音源分離により得られた分離信号の平均レベル（エネルギ）が十分小さくなった場合、つまり平均レベルが閾値以下となった場合に、再帰的な音源分離の終了条件が満たされたと判定される例について説明した。

[0148] この場合、単一の話者の発話のみが含まれる分離信号に対して音源分離が行われたときに、無音の分離信号が得られて終了条件が満たされたと判定さ

れる。

[0149] そのため、本来であれば単一の話者の発話のみが含まれる分離信号が得られた時点で各話者の分離信号が得られているのにも関わらず、さらにもう１度音源分離を行わなければならないので、その分だけ音源分離の処理回数が多くなってしまう。このようなことは、例えば処理時間が限られているアプリケーション等では好ましいとはいえない。

[0150] そこで、分離信号を入力とし、その分離信号が単一話者の発話のみが含まれている音響信号であるか、または複数話者の発話が含まれている混合音声の音響信号であるかを判定する音響モデルである単一話者判定モデルを用いて終了判定を行うようにしてもよい。

[0151] 換言すれば、単一話者判定モデルは、入力された分離信号に含まれる発話の話者数が１人であるか否かを判定するための音響モデルである。

[0152] このような例では、例えば信号処理装置１１や信号処理装置５１の終了判定部２２には、予め学習により求められた単一話者判定モデルが保持されている。

[0153] そして、例えば図３のステップＳ１２や図６のステップＳ４２では、終了判定部２２は保持している単一話者判定モデルと、音源分離により得られた分離信号とに基づく演算を行い、分離信号に含まれる発話の話者数が１人であるか否かを判定する。換言すれば、分離信号が単一話者の発話のみを含む信号であるか否かが判定される。

[0154] そして終了判定部２２は、全ての分離信号に含まれる発話の話者数が１人である、つまり分離信号が単一話者の発話のみを含む信号であるとの判定結果が得られた場合に、終了条件が満たされたと判定する。

[0155] このような単一話者判定モデルによる判定では、分離信号に含まれている発話の話者数を推定する話者数推定モデルによる推定と比較してタスクが簡単になる。そのため、より小さいモデル規模で、より高性能な音響モデル（単一話者判定モデル）を得ることができるというメリットがある。すなわち、話者数推定モデルを用いる場合と比較して、より簡単に音源分離を行うこ

とができる。

[0156] 以上のように単一話者判定モデルを用いて終了条件が満たされたかを判定することで、図３や図６を参照して説明した音源分離処理の全体の処理量（処理回数）や処理時間を低減させることができる。

[0157] また、例えば単一話者判定モデル等を用いて終了判定を行う場合、図３や図６を参照して説明した音源分離処理において、まず終了判定、すなわち終了条件を満たすか否かを判定した後、その判定結果に応じて再帰的な音源分離を行うようにしてもよい。

[0158] この場合、例えば単一話者判定モデルが終了判定に用いられるときには、単一話者判定モデルにより、単一話者の発話のみを含む分離信号ではないとされた分離信号に対して、再帰的な音源分離が行われることになる。

[0159] その他、音源分離部２１が大まかな話者数を判定する話者数判定モデルを用いて、再帰的な音源分離に用いる音源分離モデルを選択するようにしてもよい。

[0160] 具体的には、例えば音源分離部２１が、入力された音響信号が２以下の話者の発話を含む信号であるか、または３以上の話者の発話を含む信号であるかを判定する話者数判定モデルと、２話者分離モデルと、３話者分離モデルとを保持しているとする。

[0161] この場合、音源分離部２１は入力音響信号や、音源分離により得られた分離信号に対して話者数判定モデルを用いた話者数の判定を行い、音源分離に用いる音源分離モデルとして、２話者分離モデルと３話者分離モデルの何れかを選択する。

[0162] すなわち、例えば音源分離部２１は、３以上の話者の発話を含む信号であると判定された入力音響信号や分離信号に対しては、３話者分離モデルによる音源分離を行う。

[0163] これに対して、音源分離部２１は２以下の話者の発話を含む信号であると判定された入力音響信号や分離信号に対しては、２話者分離モデルによる音源分離を行う。

[0164] このようにすることで、適切な音源分離モデルを選択的に用いて音源分離を行うことができる。

[0165] 〈その他の変形例２〉

〈言語情報の利用について〉

また、第２の実施の形態や第４の実施の形態において、複数の分離信号の言語情報に基づいて同一話者判定が行われるようにしてもよい。特に、ここでは言語情報として、分離信号に基づく音声（発話）の内容を示すテキスト情報が用いられる例について説明する。

[0166] そのような場合、例えば信号処理装置５１の同一話者判定部６１は、終了判定部２２から供給された話者ごとの分離信号に対して音声認識処理を行い、それらの話者ごとの分離信号の音声をテキスト化する。すなわち、音声認識処理により、分離信号に基づく発話の内容を示すテキスト情報が生成される。

[0167] そして、同一話者判定部６１は、任意の２以上の分離信号のテキスト情報により示されるテキスト、つまり発話内容をマージ（統合）したときに、マージ後のテキストが文として成立する場合には、それらの分離信号が同一話者のものであるとする。

[0168] 具体的には、例えば２つの分離信号の各テキスト情報により示される発話のタイミングと発話内容が同じである場合、それらの２つの分離信号は同一話者のものであるとされる。

[0169] また、例えば２つの分離信号のテキスト情報により示される発話のタイミングは異なるが、それらの発話を統合して１つの発話としたときに意味のある１つの文として成立する場合、それらの２つの分離信号は同一話者のものであるとされる。

[0170] このようにテキスト情報等の言語情報を用いれば、同一話者判定の判定精度を向上させることができ、これにより分離性能を向上させることができる。

[0171] 〈その他の変形例３〉

〈同一話者判定モデルの利用について〉

また、第２の実施の形態や第４の実施の形態において、任意の２つの分離信号のそれぞれに含まれている発話の話者が同一であるか否か、つまり２つの分離信号が同一話者の信号であるか否かを判別（判定）する同一話者判定モデルに基づいて同一話者判定が行われるようにしてもよい。

[0172] ここで、同一話者判定モデルは、２つの分離信号を入力とし、それらの分離信号のそれぞれに含まれている発話の話者が同一であるか、または互いに異なる話者であるかの判定結果を出力とする音響モデルである。

[0173] そのような場合、例えば信号処理装置５１の同一話者判定部６１は、予め学習により求められた同一話者判定モデルを保持している。

[0174] 同一話者判定部６１は、保持している同一話者判定モデルと、終了判定部２２から供給された話者ごとの分離信号とに基づいて、全ての取り得る組み合わせについて２つの分離信号のそれぞれに含まれる発話の話者が同一であるか否かを判定する。

[0175] このような同一話者判定モデルによる同一話者判定では、上述の話者識別モデルにおける場合と比較してタスクが簡単になる。そのため、より小さいモデル規模で、より高性能な音響モデル（同一話者判定モデル）を得ることができるというメリットがある。

[0176] なお、同一話者判定時においては、以上において説明した特徴量間の距離を用いる方法や、言語情報を用いる方法、同一話者判定モデルを用いる方法等の複数の任意の方法を組み合わせて同一話者の分離信号を特定するようにしてもよい。

[0177] 〈コンピュータの構成例〉

ところで、上述した一連の処理は、ハードウェアにより実行することもできるし、ソフトウェアにより実行することもできる。一連の処理をソフトウェアにより実行する場合には、そのソフトウェアを構成するプログラムが、コンピュータにインストールされる。ここで、コンピュータには、専用のハードウェアに組み込まれているコンピュータや、各種のプログラムをインス

トールすることで、各種の機能を実行することが可能な、例えば汎用のパーソナルコンピュータなどが含まれる。

[0178] 図７は、上述した一連の処理をプログラムにより実行するコンピュータのハードウェアの構成例を示すブロック図である。

[0179] コンピュータにおいて、CPU (Central Processing Unit) ５０１, ROM (Read Only Memory) ５０２, RAM (Random Access Memory) ５０３は、バス５０４により相互に接続されている。

[0180] バス５０４には、さらに、入出力インターフェース５０５が接続されている。入出力インターフェース５０５には、入力部５０６、出力部５０７、記録部５０８、通信部５０９、及びドライブ５１０が接続されている。

[0181] 入力部５０６は、キーボード、マウス、マイクロホン、撮像素子などよりなる。出力部５０７は、ディスプレイ、スピーカなどよりなる。記録部５０８は、ハードディスクや不揮発性のメモリなどよりなる。通信部５０９は、ネットワークインターフェースなどよりなる。ドライブ５１０は、磁気ディスク、光ディスク、光磁気ディスク、又は半導体メモリなどのリムーバブル記録媒体５１１を駆動する。

[0182] 以上のように構成されるコンピュータでは、CPU５０１が、例えば、記録部５０８に記録されているプログラムを、入出力インターフェース５０５及びバス５０４を介して、RAM５０３にロードして実行することにより、上述した一連の処理が行われる。

[0183] コンピュータ（CPU５０１）が実行するプログラムは、例えば、パッケージメディア等としてのリムーバブル記録媒体５１１に記録して提供することができる。また、プログラムは、ローカルエリアネットワーク、インターネット、デジタル衛星放送といった、有線または無線の伝送媒体を介して提供することができる。

[0184] コンピュータでは、プログラムは、リムーバブル記録媒体５１１をドライブ５１０に装着することにより、入出力インターフェース５０５を介して、記録部５０８にインストールすることができる。また、プログラムは、有線

または無線の伝送媒体を介して、通信部５０９で受信し、記録部５０８にインストールすることができる。その他、プログラムは、ROM５０２や記録部５０８に、あらかじめインストールしておくことができる。

[0185] なお、コンピュータが実行するプログラムは、本明細書で説明する順序に沿って時系列に処理が行われるプログラムであっても良いし、並列に、あるいは呼び出しが行われたとき等の必要なタイミングで処理が行われるプログラムであっても良い。

[0186] また、本技術の実施の形態は、上述した実施の形態に限定されるものではなく、本技術の要旨を逸脱しない範囲において種々の変更が可能である。

[0187] 例えば、本技術は、１つの機能をネットワークを介して複数の装置で分担、共同して処理するクラウドコンピューティングの構成をとることができる。

[0188] また、上述のフローチャートで説明した各ステップは、１つの装置で実行する他、複数の装置で分担して実行することができる。

[0189] さらに、１つのステップに複数の処理が含まれる場合には、その１つのステップに含まれる複数の処理は、１つの装置で実行する他、複数の装置で分担して実行することができる。

[0190] さらに、本技術は、以下の構成とすることも可能である。

[0191] （１）

所定の音源を含む学習用音響信号から前記所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離を、入力された音響信号に対して再帰的に行う音源分離部を備える

信号処理装置。

（２）

前記音源分離部は、前記音源分離により前記音響信号から話者の発話の分離信号を分離する

（１）に記載の信号処理装置。

（３）

前記音源分離部は、話者数が未知である前記音響信号に対して前記音源分離を行う

（２）に記載の信号処理装置。

（４）

前記音源分離モデルは、２人の話者の発話が含まれる前記学習用音響信号を、一方の話者の発話を含む分離信号と、他方の話者の発話を含む分離信号とに分離するように学習された話者分離モデルである

（２）または（３）に記載の信号処理装置。

（５）

前記音源分離モデルは、３人の話者の発話が含まれる前記学習用音響信号を、前記３人の話者のそれぞれの発話を含む３つの分離信号のそれぞれに分離するように学習された話者分離モデルである

（２）または（３）に記載の信号処理装置。

（６）

前記音源分離モデルは、任意の複数の話者の発話が含まれる前記学習用音響信号を、１人の話者の発話を含む分離信号と、前記複数の話者のうちの前記１人の話者を除く残りの話者の発話を含む分離信号とに分離するように学習された話者分離モデルである

（２）または（３）に記載の信号処理装置。

（７）

前記音源分離部は、前記所定の前記音源分離モデルとして互いに異なる複数の音源分離モデルを用いて前記音源分離を再帰的に行う

（２）乃至（６）の何れか一項に記載の信号処理装置。

（８）

前記音源分離により得られた前記分離信号に基づいて、再帰的な前記音源分離を終了するか否かを判定する終了判定部をさらに備える

（２）乃至（７）の何れか一項に記載の信号処理装置。

（９）

前記終了判定部は、前記音源分離により得られた１つの前記分離信号が無音信号である場合、再帰的な前記音源分離を終了すると判定する

（８）に記載の信号処理装置。

（１０）

前記終了判定部は、前記分離信号に含まれる発話の話者数が１人であるか否かを判定するための単一話者判定モデルと前記分離信号とに基づいて、前記音源分離により得られた前記分離信号に含まれる発話の話者数が１人であると判定された場合、再帰的な前記音源分離を終了すると判定する

（８）に記載の信号処理装置。

（１１）

再帰的な前記音源分離により得られた複数の前記分離信号が同一話者の信号であるか否かの同一話者判定を行い、同一話者の複数の前記分離信号を合成する同一話者判定部をさらに備える

（２）乃至（１０）の何れか一項に記載の信号処理装置。

（１２）

前記同一話者判定部は、前記分離信号のクラスタリングを行うことで前記同一話者判定を行う

（１１）に記載の信号処理装置。

（１３）

前記同一話者判定部は、前記分離信号の特徴量を算出し、２つの前記分離信号の前記特徴量間の距離が閾値以下である場合、前記２つの前記分離信号は同一話者の信号であると判定する

（１２）に記載の信号処理装置。

（１４）

前記同一話者判定部は、２つの前記分離信号の時間的なエネルギ変動の相関に基づいて前記同一話者判定を行う

（１２）に記載の信号処理装置。

（１５）

前記同一話者判定部は、複数の前記分離信号の言語情報に基づいて前記同一話者判定を行う

（１１）に記載の信号処理装置。

（１６）

前記同一話者判定部は、２つの前記分離信号が同一話者の信号であるかを判定する同一話者判定モデルに基づいて前記同一話者判定を行う

（１１）に記載の信号処理装置。

（１７）

信号処理装置が、

所定の音源を含む学習用音響信号から前記所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離を、入力された音響信号に対して再帰的に行う

信号処理方法。

（１８）

所定の音源を含む学習用音響信号から前記所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離を、入力された音響信号に対して再帰的に行う

ステップを含む処理をコンピュータに実行させるプログラム。

符号の説明

[0192] １１　信号処理装置，　２１　音源分離部，　２２　終了判定部，　５１　信号処理装置，　６１　同一話者判定部

請求の範囲

[請求項1] 所定の音源を含む学習用音響信号から前記所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離を、入力された音響信号に対して再帰的に行う音源分離部を備える

信号処理装置。

[請求項2] 前記音源分離部は、前記音源分離により前記音響信号から話者の発話の分離信号を分離する

請求項１に記載の信号処理装置。

[請求項3] 前記音源分離部は、話者数が未知である前記音響信号に対して前記音源分離を行う

請求項２に記載の信号処理装置。

[請求項4] 前記音源分離モデルは、２人の話者の発話が含まれる前記学習用音響信号を、一方の話者の発話を含む分離信号と、他方の話者の発話を含む分離信号とに分離するように学習された話者分離モデルである

請求項２に記載の信号処理装置。

[請求項5] 前記音源分離モデルは、３人の話者の発話が含まれる前記学習用音響信号を、前記３人の話者のそれぞれの発話を含む３つの分離信号のそれぞれに分離するように学習された話者分離モデルである

請求項２に記載の信号処理装置。

[請求項6] 前記音源分離モデルは、任意の複数の話者の発話が含まれる前記学習用音響信号を、１人の話者の発話を含む分離信号と、前記複数の話者のうちの前記１人の話者を除く残りの話者の発話を含む分離信号とに分離するように学習された話者分離モデルである

請求項２に記載の信号処理装置。

[請求項7] 前記音源分離部は、前記所定の前記音源分離モデルとして互いに異なる複数の音源分離モデルを用いて前記音源分離を再帰的に行う

請求項２に記載の信号処理装置。

[請求項8] 前記音源分離により得られた前記分離信号に基づいて、再帰的な前

記音源分離を終了するか否かを判定する終了判定部をさらに備える

請求項２に記載の信号処理装置。

[請求項9] 前記終了判定部は、前記音源分離により得られた１つの前記分離信号が無音信号である場合、再帰的な前記音源分離を終了すると判定する

請求項８に記載の信号処理装置。

[請求項10] 前記終了判定部は、前記分離信号に含まれる発話の話者数が１人であるか否かを判定するための単一話者判定モデルと前記分離信号とに基づいて、前記音源分離により得られた前記分離信号に含まれる発話の話者数が１人であると判定された場合、再帰的な前記音源分離を終了すると判定する

請求項８に記載の信号処理装置。

[請求項11] 再帰的な前記音源分離により得られた複数の前記分離信号が同一話者の信号であるか否かの同一話者判定を行い、同一話者の複数の前記分離信号を合成する同一話者判定部をさらに備える

請求項２に記載の信号処理装置。

[請求項12] 前記同一話者判定部は、前記分離信号のクラスタリングを行うことで前記同一話者判定を行う

請求項１１に記載の信号処理装置。

[請求項13] 前記同一話者判定部は、前記分離信号の特徴量を算出し、２つの前記分離信号の前記特徴量間の距離が閾値以下である場合、前記２つの前記分離信号は同一話者の信号であると判定する

請求項１２に記載の信号処理装置。

[請求項14] 前記同一話者判定部は、２つの前記分離信号の時間的なエネルギ変動の相関に基づいて前記同一話者判定を行う

請求項１２に記載の信号処理装置。

[請求項15] 前記同一話者判定部は、複数の前記分離信号の言語情報に基づいて前記同一話者判定を行う

請求項１１に記載の信号処理装置。

[請求項16] 前記同一話者判定部は、２つの前記分離信号が同一話者の信号であるかを判定する同一話者判定モデルに基づいて前記同一話者判定を行う

請求項１１に記載の信号処理装置。

[請求項17] 信号処理装置が、

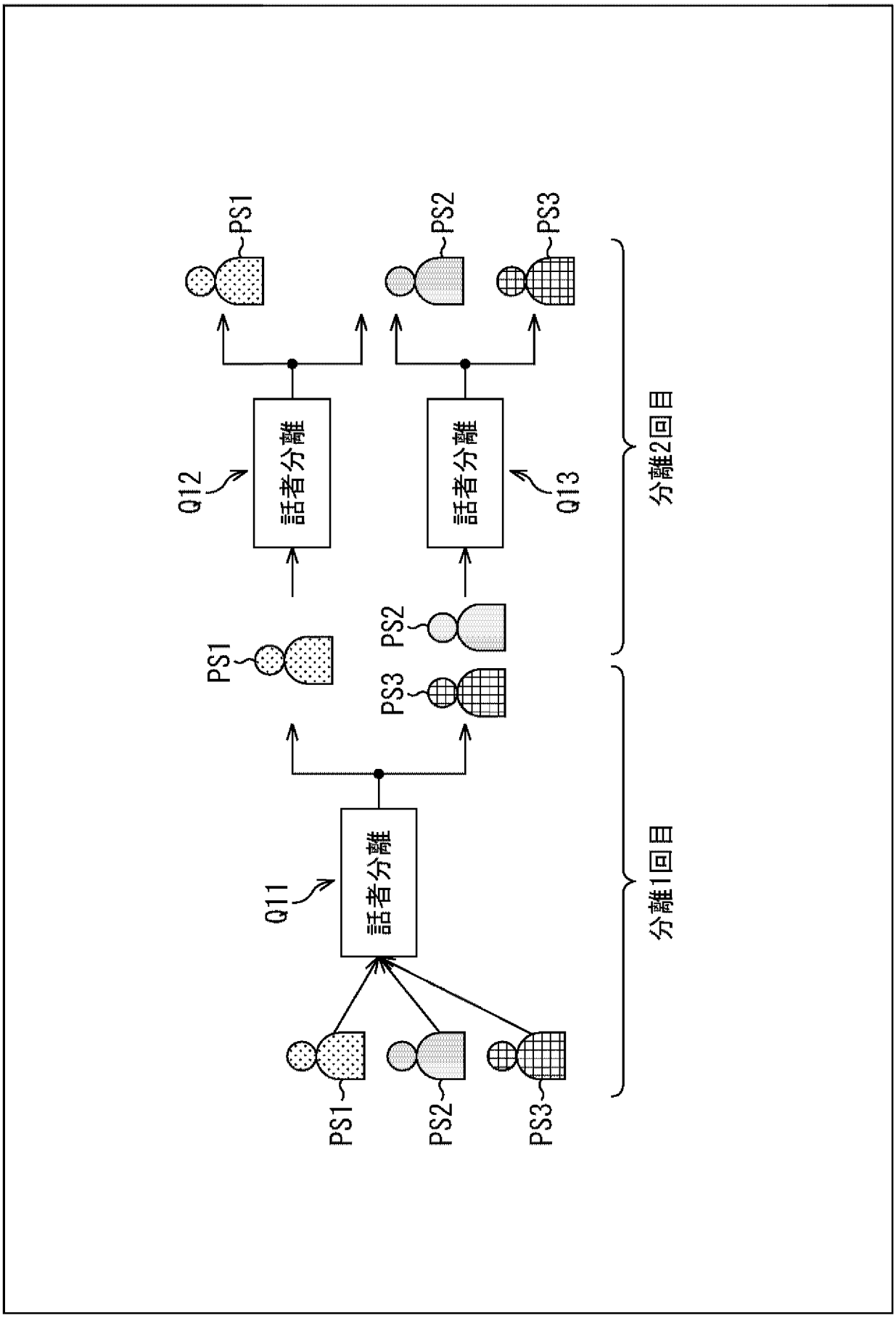
所定の音源を含む学習用音響信号から前記所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離を、入力された音響信号に対して再帰的に行う

信号処理方法。

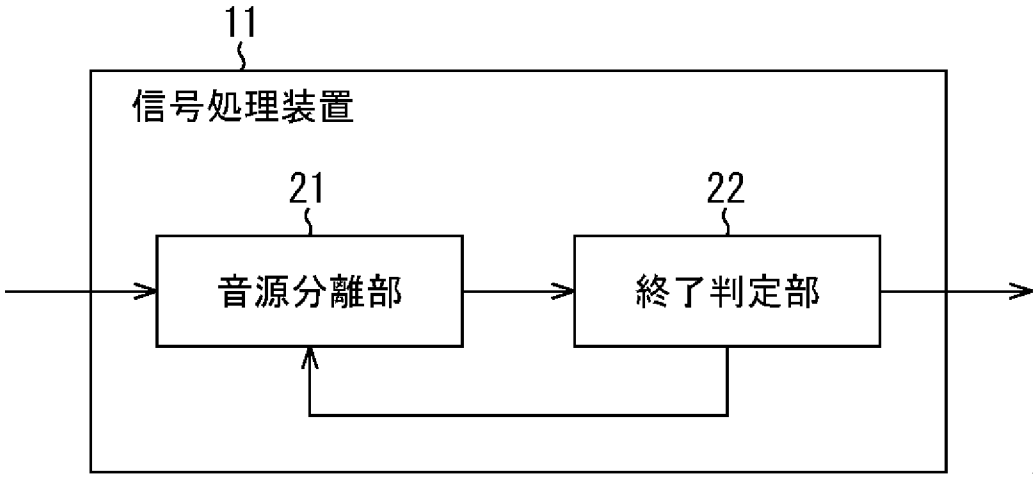
[請求項18] 所定の音源を含む学習用音響信号から前記所定の音源を分離するように予め学習された所定の音源分離モデルによる音源分離を、入力された音響信号に対して再帰的に行う

ステップを含む処理をコンピュータに実行させるプログラム。

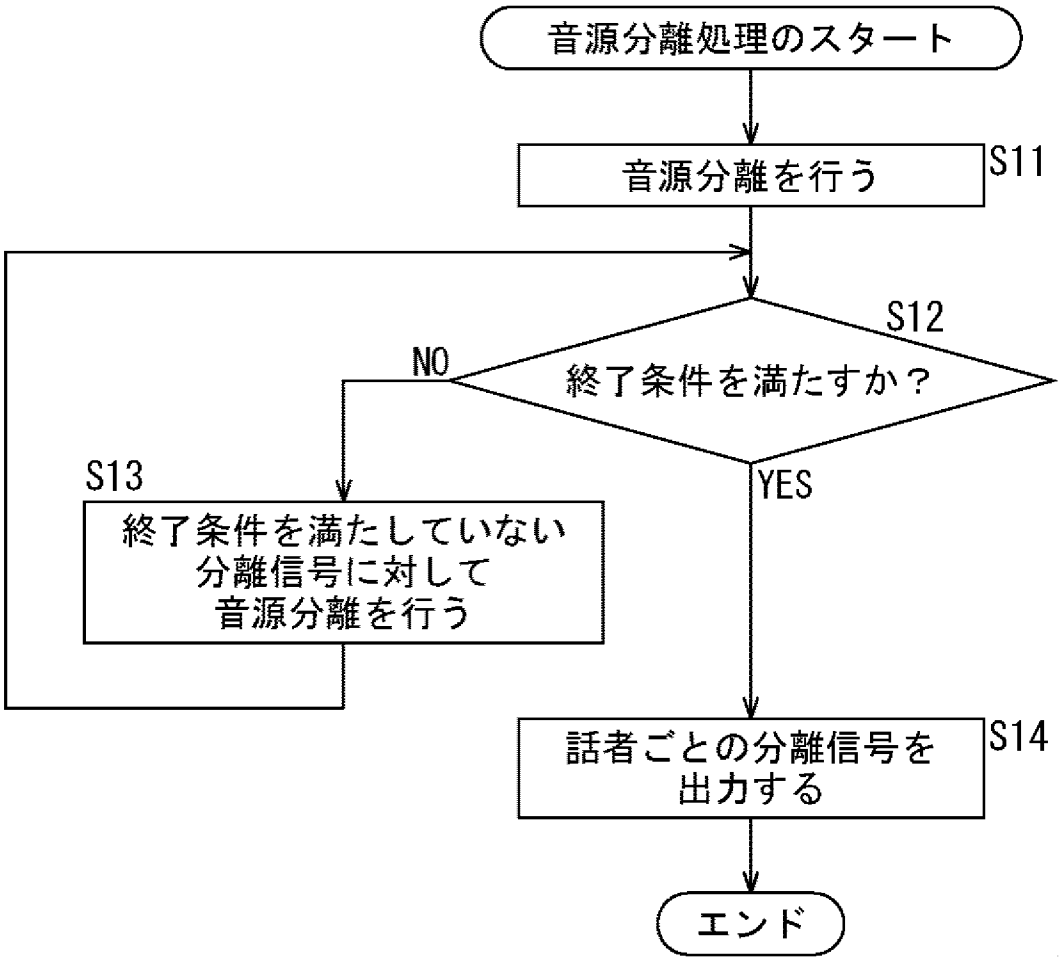
[図1]
FIG. 1


PS1
PS2
PS3
話者分離
Q12
話者分離
Q13
PS1
PS2
PS3
話者分離
Q11
PS1
PS2
PS3
分離2回目
分離1回目

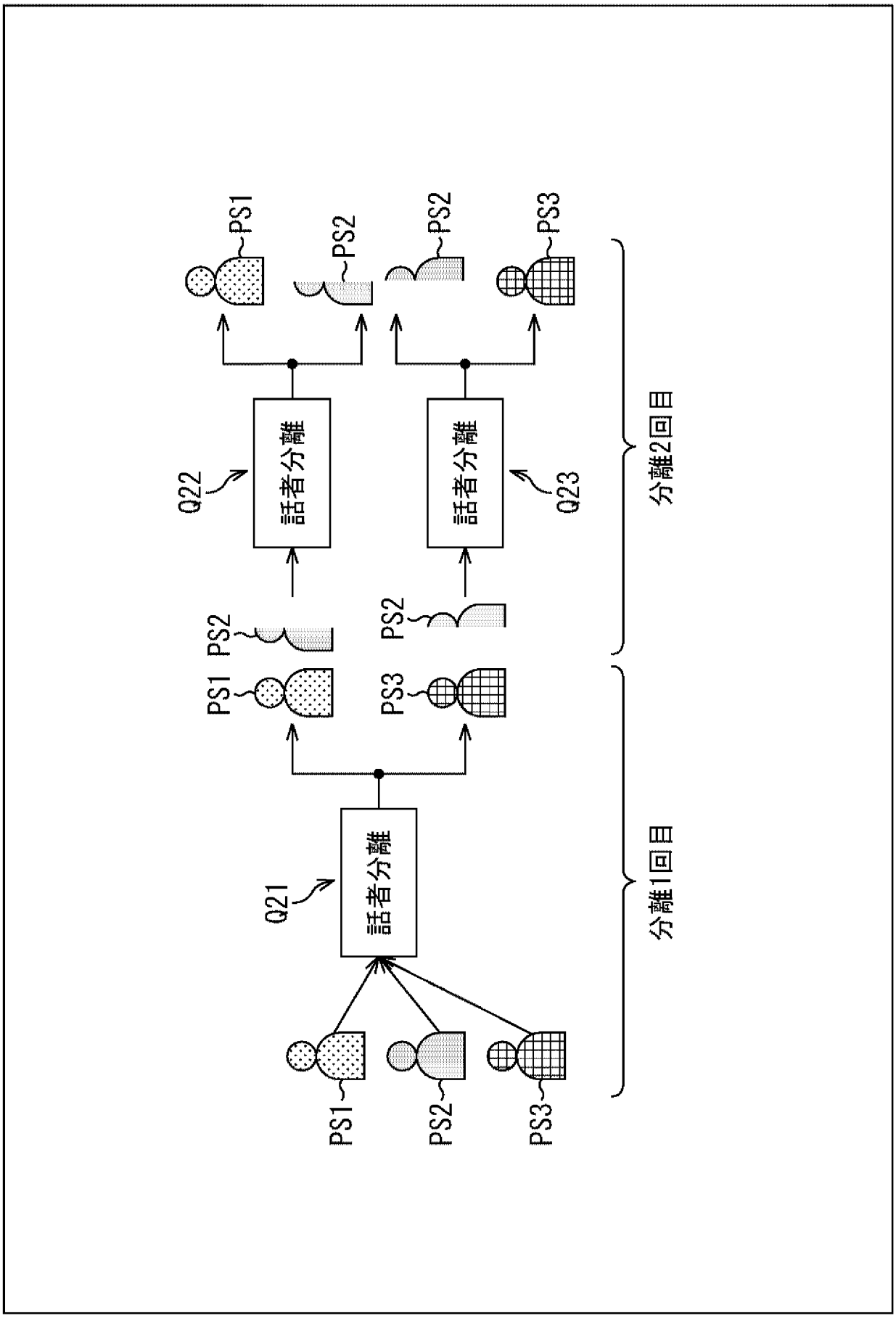
[図2]
FIG. 2



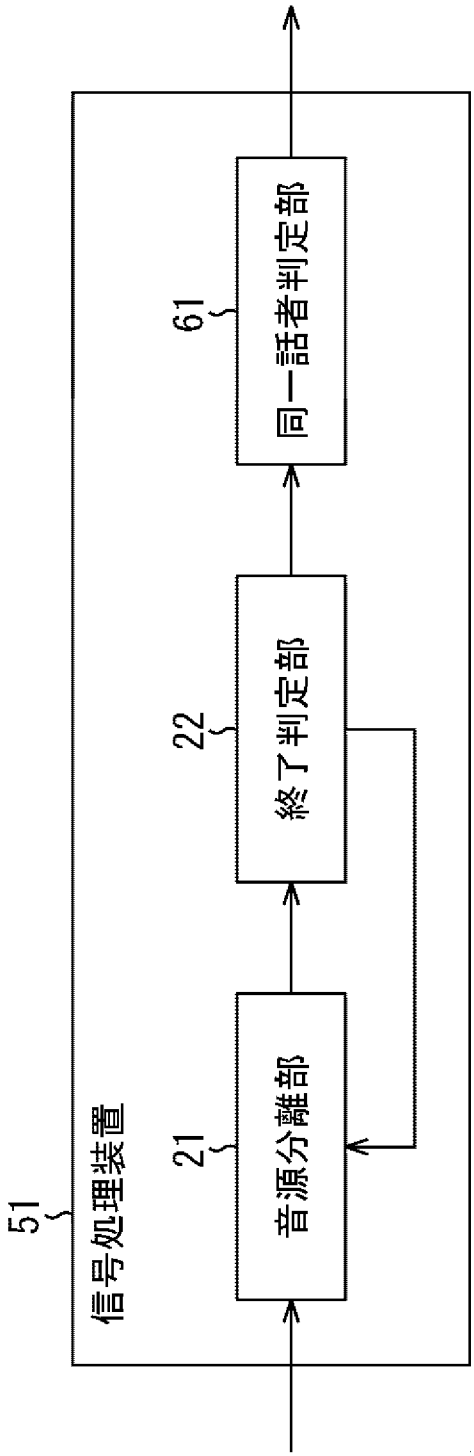
[図3]
FIG. 3



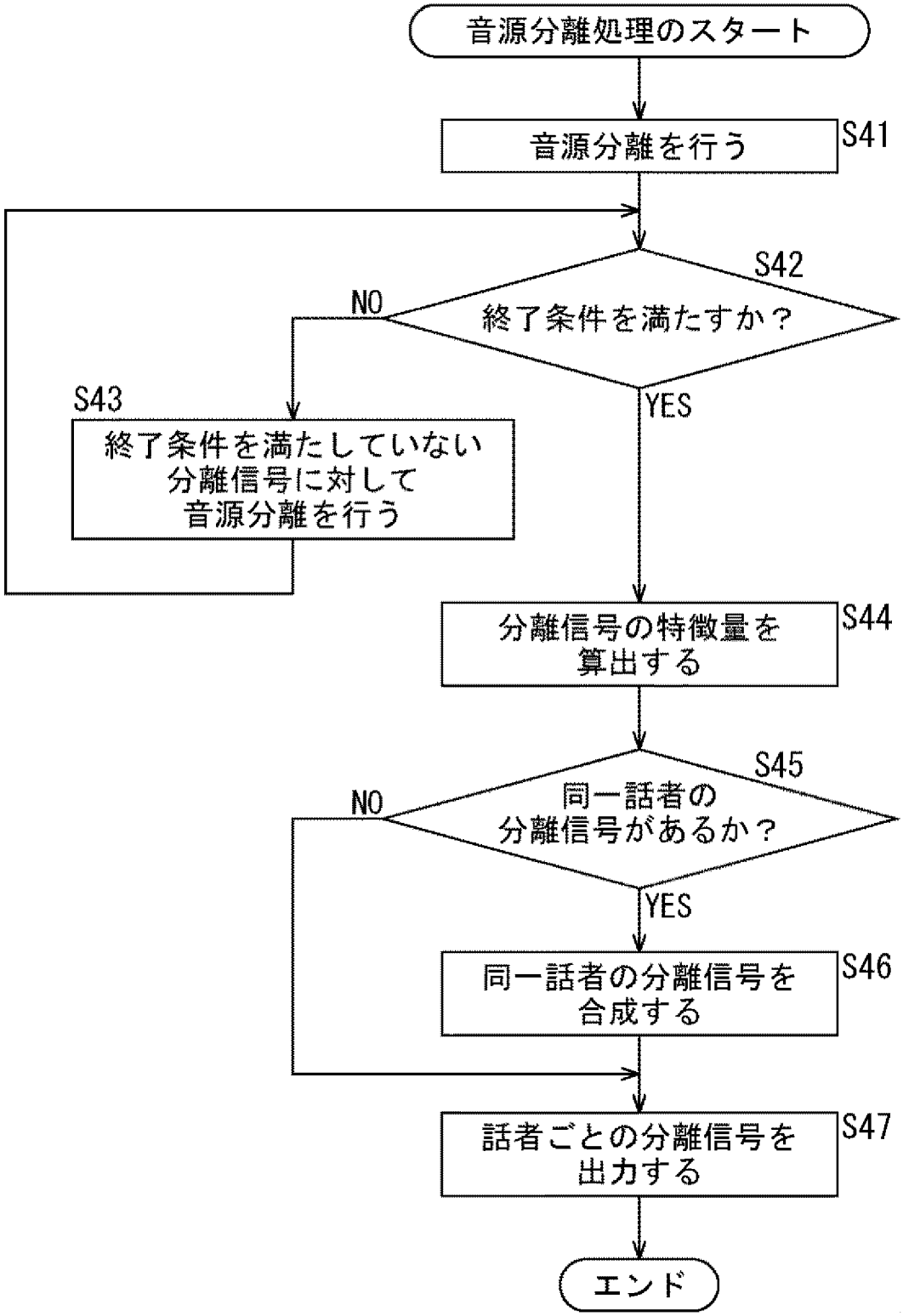
[図4]
FIG. 4


PS1
PS2
PS2
PS3
話者分離
Q22
話者分離
Q23
分離2回目
PS2
PS1
PS3
話者分離
Q21
分離1回目
PS1
PS2
PS3

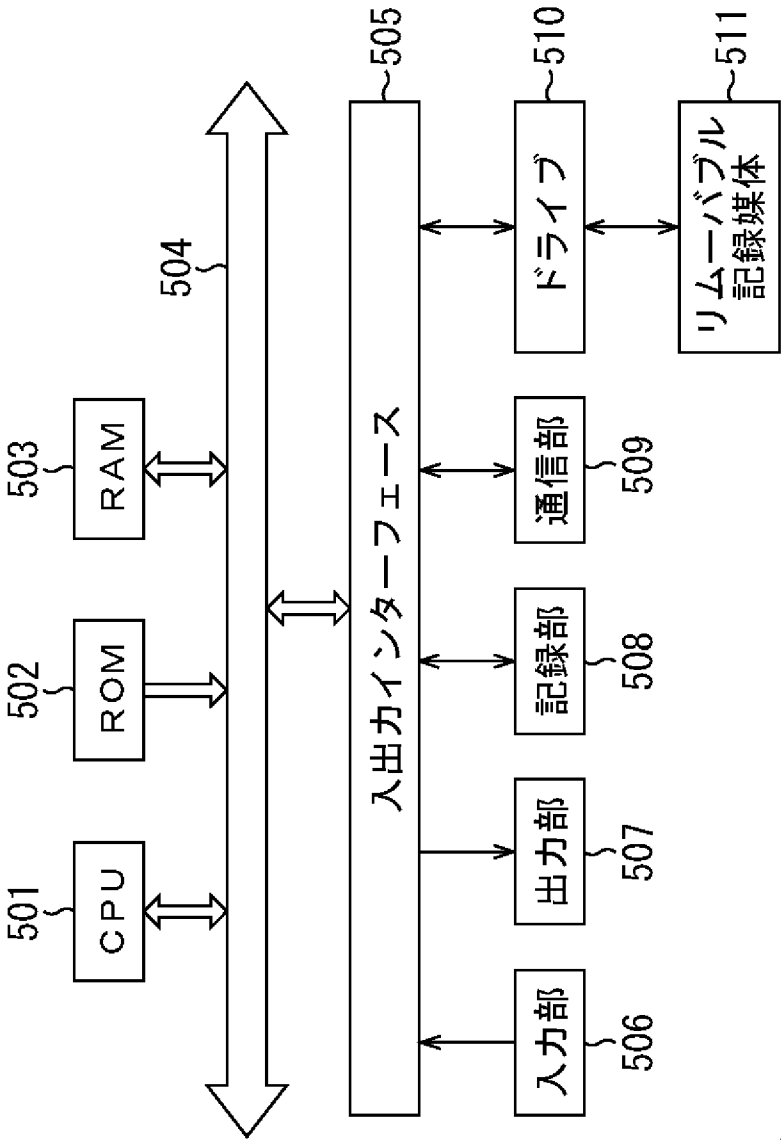
[図5]
FIG. 5



[図6]
FIG. 6


音源分離処理のスタート
音源分離を行う
S41
S42
終了条件を満たすか？
NO
YES
S43
終了条件を満たしていない
分離信号に対して
音源分離を行う
分離信号の特徴量を
算出する
S44
S45
同一話者の
分離信号があるか？
NO
YES
同一話者の分離信号を
合成する
S46
話者ごとの分離信号を
出力する
S47
エンド

[図7]
FIG. 7


501
CPU
502
ROM
503
RAM
504
505
入出力インターフェース
506
入力部
507
出力部
508
記録部
509
通信部
510
ドライブ
511
リムーバブル記録媒体

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/011008

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl. G10L21/0272(2013.01)i, G10L21/028(2013.01)i
FI: G10L21/0272100Z, G10L21/028B

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)
Int.Cl. G10L21/0272, G10L21/028

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	1971-2020
Registered utility model specifications of Japan	1996-2020
Published registered utility model applications of Japan	1994-2020

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	JP 2018-28620 A (HITACHI, LTD.) 22.02.2018 (2018-02-22), entire text, all drawings	1-18

☐ 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
29.05.2020	09.06.2020

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2020/011008

JP 2018-28620 A 22.02.2018 (Family: none)

国際調査報告

国際出願番号
PCT/JP2020/011008

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 21/028(2013.01)i
FI: G10L21/0272 100Z; G10L21/028 B

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; G10L21/028

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	１９７１－２０２０年
日本国実用新案登録公報	１９９６－２０２０年
日本国登録実用新案公報	１９９４－２０２０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	JP 2018-28620 A (株式会社日立製作所) 22.02.2018 (2018-02-22) 全文、全図	1-18

☐ 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
29.05.2020	09.06.2020

名称及びあて先
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

権限のある職員（特許庁審査官）
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/JP2020/011008

引用文献	公表日	パテントファミリー文献	公表日
JP 2018-28620 A	22.02.2018	（ファミリーなし）	