

(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8月27日(27.08.2020)



(10) 国際公開番号

WO 2020/170907 A1

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

(21) 国際出願番号: PCT/JP2020/005332

(22) 国際出願日: 2020年2月12日(12.02.2020)

(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 2019-026853 2019年2月18日(18.02.2019) JP

(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).

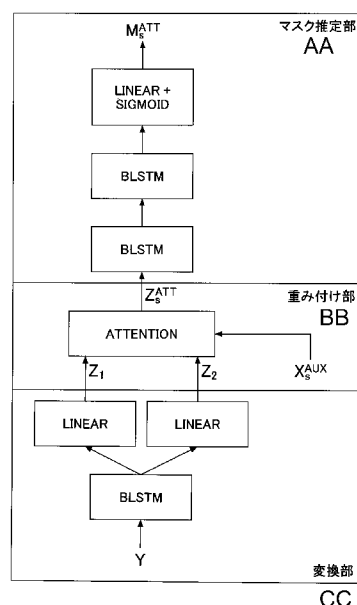
(72) 発明者: 落合 翼(OCHIAI, Tsubasa); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP). デルクロア マーク(DELacroix, Marc); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP). 木下 慶介(KINOSHITA, Keisuke); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP). 小川 厚徳(OGAWA, Atsunori); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (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: 伊東 忠重, 外(ITO, Tadashige et al.); 〒1000005 東京都千代田区丸の内二丁目1

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

(54) 発明の名称: 信号処理装置、学習装置、信号処理方法、学習方法及びプログラム

[図4]



AA Mask estimation unit
BB Weighting unit
CC Conversion unit

(57) Abstract: The signal processing device according to an embodiment of the present invention comprises: a conversion unit that converts an inputted mixed acoustic signal into a plurality of first internal states; a weighting unit that, when auxiliary information relating to an acoustic signal of a target sound source is inputted, generates a second internal state being the weighted sum of the plurality of first internal states on the basis of the auxiliary information, and when the auxiliary information is not inputted, generates the second internal state by selecting any of the plurality of first internal

番 1 号 丸の内 M Y P L A Z A (明治安
田生命ビル) 1 6 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, 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, 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:

- 国際調査報告 (条約第21条(3))

states; and a mask estimation unit that estimates a mask on the basis of the second internal state.

(57) 要約: 本発明の一形態に係る信号処理装置は、入力された混合音響信号を複数の第1内部状態に変換する変換部と、目的音源の音響信号に関する補助情報が入力された場合、前記補助情報に基づいて前記複数の第1内部状態の重み付け和である第2内部状態を生成し、前記補助情報が入力されない場合、前記複数の第1内部状態のいずれかを選択することによって第2内部状態を生成する重み付け部と、前記第2内部状態に基づいてマスクを推定するマスク推定部とを有する。

明 細 書

発明の名称：

信号処理装置、学習装置、信号処理方法、学習方法及びプログラム

技術分野

[0001] 本発明は、複数音源の音響信号が混合された混合音響信号から、音源毎の音響信号に分離したり、特定音源の音響信号を抽出したりする信号処理技術に関する。

背景技術

[0002] 近年、モノラル音声を対象とした話者分離技術の研究が盛んに行われている。話者分離技術には、事前情報を用いないブラインド音源分離（非特許文献1）と、話者の音声に関する補助情報を用いた目的話者抽出（非特許文献2）の大きく2つの手法が知られている。

先行技術文献

非特許文献

[0003] 非特許文献1：Morten Kolbaek, etc., "Multitalker speech separation with utterance-level permutation invariant training of deep re-current neural networks", Trans. on TASLP, 2017.

非特許文献2：Marc Delcroix, etc., "Single Channel Target Speaker Extraction and Recognition with Speaker Beam", Proc. on ICASSP, 2018.

発明の概要

発明が解決しようとする課題

[0004] ブラインド音源分離は、事前情報なしで話者分離が可能であるという利点があるが、発話間でパーミュテーション問題が生じるという問題点がある。ここでパーミュテーション問題とは、処理対象の長時間の音声を単位時間毎にブラインド音源分離で処理をしていく際に、各時間区間において分離信号の音源の順序が異なってしまう（入れ替わってしまう）問題である。

[0005] また、目的話者抽出では、補助情報を用いて話者をトラッキングすること

で、ブラインド音源分離において発生する発話間のパーミュテーション問題を解決できる可能性があるが、混合音声に含まれる話者が事前に分からない場合には、手法そのものを適用できないという問題点が存在する。

[0006] 上記のように、ブラインド音源分離と目的話者抽出はそれぞれ利点及び問題点を有するため、状況に応じて両者を使い分ける必要がある。しかし、これまでブラインド音源分離と目的話者抽出はそれぞれ目的に応じてモデル学習が行われ、独立したシステムとして構築されてきた。そのため、1つのモデルで両者を使い分けることができなかった。

[0007] 以上の問題に鑑み、本発明は、ブラインド音源分離と目的話者抽出とを統合的に扱うための手法を提供することを目的とする。

課題を解決するための手段

[0008] 本発明の一形態に係る信号処理装置は、
入力された混合音響信号を複数の第1内部状態に変換する変換部と、
目的音源の音響信号に関する補助情報が入力された場合、前記補助情報に基づいて前記複数の第1内部状態の重み付け和である第2内部状態を生成し、前記補助情報が入力されない場合、前記複数の第1内部状態のいずれかを選択することによって第2内部状態を生成する重み付け部と、
前記第2内部状態に基づいてマスクを推定するマスク推定部と、
を有することを特徴とする。

[0009] また、本発明の一形態に係る学習装置は、
ニューラルネットワークを用いて、入力された学習用の混合音響信号を複数の第1内部状態に変換する変換部と、
目的音源の音響信号に関する補助情報が入力された場合、ニューラルネットワークを用いて、前記複数の第1内部状態の重み付け和である第2内部状態を生成し、前記補助情報が入力されない場合、前記複数の第1内部状態のいずれかを選択することによって第2内部状態を生成する重み付け部と、
ニューラルネットワークを用いて、前記第2内部状態に基づいてマスクを推定するマスク推定部と、

前記学習用の混合音響信号に対して前記推定されたマスクを適用して得られる音響信号と、前記学習用の混合音響信号に含まれる正解の音源の音響信号との比較結果に基づいて、前記変換部、前記重み付け部及び前記マスク推定部のそれぞれにおいて用いられるニューラルネットワークのパラメータを更新するパラメータ更新部と、

を有することを特徴とする。

- [0010] また、本発明の一形態に係る信号処理方法は、
信号処理装置が実行する信号処理方法であって、
入力された混合音響信号を複数の第1内部状態に変換するステップと、
目的音源の音響信号に関する補助情報が入力された場合、前記複数の第1内部状態の重み付け和である第2内部状態を生成し、前記補助情報が入力されない場合、前記複数の第1内部状態のいずれかを選択することによって第2内部状態を生成するステップと、
前記第2内部状態に基づいてマスクを推定するステップと、
を有することを特徴とする。

- [0011] また、本発明の一形態に係る学習方法は、
学習装置が実行する学習方法であって、
ニューラルネットワークを用いて、入力された学習用の混合音響信号を複数の第1内部状態に変換するステップと、
目的音源の音響信号に関する補助情報が入力された場合、ニューラルネットワークを用いて、前記複数の第1内部状態の重み付け和である第2内部状態を生成し、前記補助情報が入力されない場合、前記複数の第1内部状態のいずれかを選択することによって第2内部状態を生成するステップと、
ニューラルネットワークを用いて、前記第2内部状態に基づいてマスクを推定するステップと、

前記学習用の混合音響信号に対して前記推定されたマスクを適用して得られる音響信号と、前記学習用の混合音響信号に含まれる正解の音源の音響信号との比較結果に基づいて、前記変換するステップ、前記生成するステップ

及び前記推定するステップのそれぞれにおいて用いられるニューラルネットワークのパラメータを更新するステップと、
を有することを特徴とする。

[0012] また、本発明の一形態に係るプログラムは、上記の装置としてコンピュータを機能させることを特徴とする。

発明の効果

[0013] 本発明によれば、ブラインド音源分離と目的話者抽出とを統合的に扱うことが可能になる。

図面の簡単な説明

[0014] [図1]本発明の実施例におけるシステム構成例を示す図である。

[図2]従来のブラインド音源分離を行うニューラルネットワークの構成を示す図である。

[図3]本発明の実施例における信号処理装置の原理を説明するための図（その1）である。

[図4]本発明の実施例における信号処理装置の原理を説明するための図（その2）である。

[図5]本発明の実施例における信号処理装置の構成を示す図である。

[図6]信号処理装置の変換部の構成を示す図である。

[図7]本発明の実施例における学習装置の構成を示す図である。

[図8]本発明の実施例の評価結果を示す図である。

[図9]本発明の実施例における各装置のハードウェア構成例を示す図である。

発明を実施するための形態

[0015] 以下、図面を参照して本発明の実施例について説明する。

[0016] 図1は、本発明の実施例におけるシステム構成例を示す図である。図1において、マイクMICは、複数の音源（以下、少なくとも一部の音源を話者とも言う） $Y_1 \sim Y_L$ からの音響信号（音声等）を收音することができる。マイクMICは、收音された音声を混合音声信号 Y として信号処理装置100に出力する。なお、以下で「音声」と称している信号は、人の声に限るものではなく、特

定の音源が発する音響信号を意味するものとする。つまり、混合音声信号は、複数の音源からの音響信号が混合された混合音響信号であってもよい。また、本実施例の信号処理装置100は、マイクで收音した音声信号を直接、信号処理装置100に入力する場合に限定されるものではなく、例えば、マイク等により收音した音声信号がメディアやハードディスク等に記憶されており、その記憶された音声信号を信号処理装置100に読み込むことで実行される形式であってもよい。

[0017] 信号処理装置100は、混合音声信号Yを入力として受け取り、事前情報なしに特定の音源の信号を分離（ブラインド音源分離）することができ、また、目的とする話者（以下、目的話者と呼ぶ）の音声に関する補助情報を用いて特定の音源の信号を抽出（目的話者抽出）することもできる装置である。上述の通り、目的話者は『目的とする音源』であれば人に限るものではないので、補助情報は目的とする音源が発する音響信号に関する補助情報を意味する。信号処理装置100は、特定の音源の信号を分離したり抽出したりするためにマスクを用いる。信号処理装置100は、マスクを推定するために双方向長短期記憶（BLSTM: bi-directional long short-term memory）等のニューラルネットワークを利用する。

[0018] ここで、音源数が2の場合を例に挙げて非特許文献1のブラインド音源分離について説明する。

[0019] 図2は、非特許文献1の従来のブラインド音源分離を行うニューラルネットワークの構成を示す図である。従来のブラインド音源分離では、入力された混合音声信号Yを複数のBLSTM層により内部状態に変換し、最後に混合音声信号に含まれる音源数（ここでは2個）分だけ用意された線形変換層（LINEAR+SIGMOID）により、内部状態をそれぞれ線形変換することで各音源に対応するマスク M_1 、 M_2 を得る。なお、線形変換層では、内部状態を線形変換した後、シグモイド関数を適用することで出力情報を決定する。

[0020] 次に、本発明の実施例における信号処理装置100の原理について説明する。

[0021] 図3及び図4は、本発明の実施例における信号処理装置100の原理を説明するための図である。

[0022] ブラインド音源分離と目的話者抽出とを統合的に扱うためには、ブラインド音源分離の枠組みの中に目的話者抽出の機能を取り込む必要がある。そのために、図2においてニューラルネットワークの後段部分におかれていた音源毎に分離して線形変換を行う線形変換層を、図3のようにニューラルネットワークの前半部分の変換部に移動させることを考える。後述するように、変換部は混合音声信号 Y をニューラルネットワークにより変換し、分離された信号に対応する内部状態 Z_1 、 Z_2 に変換する。内部状態の数は混合音声信号 Y に含まれると想定される最大の音源数（ここでは2）以上とすることが好ましい。このとき、線形変換層以降のマスク推定部におけるBLSTM層と線形変換層は共有化することができる。

[0023] さらに、図4のように、変換部とマスク推定部の間に重み付け部（ATTENTION層）を追加し、目的話者の音声に関する補助情報 X_s^{AUX} に応じて内部状態を変換するような構成とする。補助情報 X_s^{AUX} が入力された場合には、重み付け部は、入力された補助情報に基づいて複数の内部状態 Z_1 、 Z_2 から目的話者に対応する内部状態を Z_s^{ATT} として求めて後段のマスク推定部を実行させることで、目的話者抽出のためのマスクを推定させることが可能となる。また、補助情報が入力されない場合には、重み付け部は Z_s^{ATT} を Z_1 として後段のマスク推定部を実行させ、同様に Z_s^{ATT} を Z_2 として後段のマスク推定部を実行させることで、ブラインド音源分離のマスクを推定させることが可能となる。つまり、補助情報の有無に応じて内部状態を変換することで、ブラインド音源分離と目的話者抽出を切り替えて使うことが可能となる。

[0024] 以下に説明するように、信号処理装置100の変換部、重み付け部及びマスク推定部のそれぞれはニューラルネットワークを用いて構成される。信号処理装置100は、学習時には、予め用意された学習用データ（個別の音源からの正解の音声信号が知られているものとする）を用いて、ニューラルネットワークのパラメータを学習させる。信号処理装置100は、運用時には

、学習時に学習されたパラメータ設定済みのニューラルネットワークを用いて、マスクを計算する。

[0025] なお、信号処理装置 100 におけるニューラルネットワークのパラメータの学習は別の装置で行われても、同じ装置で行われてもよい。以下の実施例では、ニューラルネットワークの学習が学習装置という別の装置で行われるものとして説明する。

[0026] <実施例 1：信号処理装置>

実施例 1 では、話者の音声に関する補助情報の有無に応じて、ブラインド音源分離と目的話者抽出とを統合的に扱うための信号処理装置 100 について説明する。

[0027] 図 5 は、本発明の実施例における信号処理装置 100 の構成を示す図である。信号処理装置 100 は、変換部 110 と、補助情報入力部 120 と、重み付け部 130 と、マスク推定部 140 とを有する。変換部 110、重み付け部 130、及びマスク推定部 140 がそれぞれニューラルネットワークの層（複数層）に対応する。ニューラルネットワークの各パラメータは、予め用意された学習用データを用いて後述の学習装置により予め学習されているものとし、具体的には、学習用データに対してマスク推定部 140 で推定されたマスクを適用して得られる音声信号と、学習用データに含まれる正解の音声信号との誤差が小さくなるように学習されているものとする。

[0028] [変換部]

変換部 110 は、混合音声信号を入力とし、 I 個の内部状態を表すベクトル $z_1 \sim z_I$ を出力するニューラルネットワークである。ここで I は入力混合音声に含まれる音源数以上とすることが好ましい。ニューラルネットワークの種類は特に限定されるものではないが、例えば、非特許文献 1 及び 2 に記載の BLSTM が用いられてもよい。以下の説明では、BLSTM を例に挙げて説明する。

[0029] 具体的には、変換部 110 は図 6 のような層から構成される。まず、BLSTM 層により、入力された混合音声信号を内部状態 z に変換する。次に、内部状態 z を I 個の線形変換層（第 1 LINEAR 層～第 I -LINEAR 層）によりそれぞれ異なる

線形変換を行い、 I 個の内部状態である埋め込みベクトル $Z_1 \sim Z_I$ を得る。ここで、 t ($t=1, \dots, T$) を処理対象の時間フレームのインデックスとすると、埋め込みベクトル $Z_1 \sim Z_I$ は $Z_i = \{z_{it}\}_{t=1}^T$ ($i=1, \dots, I$) のように表現できる。

[0030] 〔補助情報入力部〕

補助情報入力部 120 は、目的話者抽出を行う場合に、目的話者の音声に関する補助情報 X_s^{AUX} を入力とし、重み付け部 130 に対して補助情報 X_s^{AUX} を出力する入力部である。

[0031] 補助情報入力部 120 には、目的話者抽出を行う場合には、目的話者の音声の特徴を表す補助情報 X_s^{AUX} が入力される。ここで、 s は目的話者を表すインデックスである。例えば、補助情報 X_s^{AUX} は、例えば、非特許文献 2 に記載の目的話者の音声信号を短時間フーリエ変換 (STFT: short-time Fourier transform) 等で特徴抽出することによって得られたベクトル $A^{(s)}(t, f)$ を変換して得た話者ベクトル等を用いてもよい。目的話者抽出を行わない場合 (すなわち、ブラインド音源分離を行う場合) には、補助情報入力部 120 には何も入力されない。

[0032] 〔重み付け部〕

重み付け部 130 は、変換部 110 から出力された内部状態 $Z_1 \sim Z_I$ を入力するとともに、目的話者抽出を行う場合には補助情報入力部 120 から出力された補助情報 X_s^{AUX} を入力とし、マスク推定のための内部状態 $Z_s^{ATT} = \{z_t^{ATT}\}_{t=1}^T$ を出力する処理部である。上記の通り、 t ($t=1, \dots, T$) は処理対象の時間フレームのインデックスである。

[0033] 重み付け部 130 は、補助情報 X_s^{AUX} の有無に応じて、入力された I 個の内部状態 $Z_1 \sim Z_I$ を重み付けすることにより、内部状態 z_t^{ATT} を得て出力する。例えば、 $I=2$ の場合は、補助情報の有無に応じてアテンション重み a_t を以下のように設定する。

[0034]

[数1]

$$\mathbf{a}_t = \begin{cases} [1, 0] \text{ or } [0, 1] & (\text{補助なし}) \\ \text{MLPAttention}(\mathbf{Z}_i, \mathbf{X}_s^{\text{AUX}}) & (\text{補助あり}) \end{cases}$$

ここで、MLPAttentionは、内部状態 $\mathbf{Z}_i$ と補助情報 $\mathbf{X}_s^{\text{AUX}}$ とに基づいてI次元の重みベクトルを得るためのニューラルネットワークである。ニューラルネットワークの種類は特に限定されるものではないが、例えば、多層パーセプトロン (MLP: multilayer perceptron) が用いられてもよい。

[0035] 次に、重み付け部130は以下により内部状態 $\mathbf{z}_t^{\text{ATT}}$ を得る。

[0036] [数2]

$$\mathbf{z}_t^{\text{ATT}} = \underbrace{\sum_{i=1}^I a_{ti} \mathbf{z}_{it}}_{\text{Attention}}, \quad (t = 1, \dots, T)$$

つまり、アテンション重み $\mathbf{a}_t$ はI次元のベクトルであって、補助情報が入力されない場合には、アテンション重み $\mathbf{a}_t$ を i ($i=1, 2, 3, \dots, I$) 番目の要素のみが1で、その他の要素が0となるような単位ベクトルとする。そして、重み付け部130は、I個の内部状態 $\mathbf{Z}_1 \sim \mathbf{Z}_I$ にアテンション重み $\mathbf{a}_t$ を適用することで i 番目の内部状態 $\mathbf{Z}_i$ を選択して内部状態 $\mathbf{z}_t^{\text{ATT}}$ として出力する。アテンション重み $\mathbf{a}_t$ としてI個の単位ベクトルをそれぞれ設定することで、混合音声に含まれる全ての話者の音声をブラインドな形で分離するマスクを推定させることが可能となる。言い換えれば、補助情報が入力されない場合には、重み付け部130は、I個の内部状態 $\mathbf{Z}_1 \sim \mathbf{Z}_I$ のいずれか1つを選択する演算 (hard alignment

）を実行することになる。

[0037] 補助情報が入力された場合は、内部状態 Z_i 及び補助情報 X_s^{AUX} に基づいて推定されたアテンション重み a_t を用いる。重み付け部130は、 I 個の内部状態 $Z_1 \sim Z_I$ にアテンション重み a_t を適用することで、 I 個の内部状態 $Z_1 \sim Z_I$ から目的話者 s に対応する内部状態を計算し、 z_t^{ATT} として出力する。言い換えれば、補助情報が入力された場合は、重み付け部130は、補助情報 X_s^{AUX} に基づいて、 I 個の内部状態 $Z_1 \sim Z_I$ の重み付け和(soft alignment)により、内部状態 z_t^{ATT} を得て出力する。

[0038] なお、この重み付け部130において各内部状態に乗ずる重みは時刻毎に異なる。つまり、重み付け部130は、時刻毎に重み付け和の算出(hard alignmentもしくはsoft alignment)を行う。

[0039] アテンション重みの推定には、例えばDzmitry Bahdanau, etc., "Neural machine translation by jointly learning to align and translate", Proc on ICLR, 2015に開示されているMLP attentionを用いることができる。ここでは、MLP attentionの構成として、キーをFeature(Z_i)、クエリをFeature(X_s^{AUX})、バリューを Z_i と設定した。なお、Feature($\cdot$)は、入力された系列 $\cdot$ から特徴抽出を行うMLPを表す。

〔マスク推定部〕

マスク推定部140は、重み付け部130から出力された内部状態 Z^{ATT} （時刻毎の内部状態 z_t^{ATT} を並べた時系列情報）を入力とし、マスクを出力するニューラルネットワークである。ニューラルネットワークの種類は特に限定されるものではないが、例えば、非特許文献1及び2に記載のBLSTMが用いられてもよい。

[0040] マスク推定部140は、例えば、BLSTMと全結合層により構成され、内部状態 Z^{ATT} を時間周波数マスク M_{ATT} に変換して出力する。

[0041] <実施例2：学習装置>

実施例2では、実施例1の信号処理装置100を構成するニューラルネットワークのパラメータを学習させる学習装置200について説明する。

[0042] 図7は、本発明の実施例における学習装置200の構成を示す図である。

学習装置200は、変換部210と、補助情報入力部220と、重み付け部230と、マスク推定部240と、パラメータ更新部250とを有する。変換部210、補助情報入力部220、重み付け部230、及びマスク推定部240の機能は実施例1と同じである。

[0043] ニューラルネットワークのパラメータを学習させるための学習用データとして、混合音声信号と、当該混合音声信号に含まれる各音源のクリーン信号（すなわち、正解の音声信号）と、目的話者の音声に関する補助情報（存在する場合と存在しない場合がある）を対応付けたデータの集合が与えられるものとする。

[0044] 学習用データの中の混合音声信号及び補助情報を入力として、変換部210、重み付け部230、及びマスク推定部240が実施例1と同様の処理を行うことで、マスクの推定値が得られる。ただし、ニューラルネットワークの各パラメータには適当な初期値を設定しておくものとする。

[0045] [パラメータ更新部]

パラメータ更新部250は、マスク推定部240から出力されたマスクと学習用データを入力として、ニューラルネットワークの各パラメータを出力する処理部である。

[0046] パラメータ更新部250は、学習用データの中の入力混合音声信号にマスク推定部240で推定したマスクを適用して得られる音声信号と、学習用データの中のクリーン信号との比較結果に基づいて、変換部210、重み付け部230、及びマスク推定部240におけるニューラルネットワークの各パラメータを誤差逆伝播法等により更新する。

[0047] ニューラルネットワークの各パラメータの更新のため、パラメータ更新部250は、補助情報を用いないブラインド音源分離と、補助情報を用いる目的話者抽出との双方の損失を考慮したマルチタスク学習を行う。例えば、補助情報を用いないブラインド音源分離のための損失関数を L_{uninfo} とし、補助情報を用いた目的話者抽出のための損失関数を L_{info} とし、 ε を所定の補間係数（

予め値を設定しておくものとする)として、マルチタスク学習に基づく損失関数 L_{multi} を以下のように定義し、これに基づいてパラメータ更新部250は誤差逆伝播学習を行う。

[0048]
$$L_{\text{multi}} = \varepsilon L_{\text{uninfo}} + (1 - \varepsilon) L_{\text{info}}$$

パラメータ更新部250は、誤差が閾値より小さくなるという収束条件等の所定の条件を満たすまで、マスクの推定とパラメータの更新を繰り返し、最終的に得られたパラメータを学習済みのニューラルネットワークパラメータとして用いる。

[0049] <本発明の実施例の効果>

本発明の実施例における信号処理装置100は、先に入力混合音声信号を複数の内部状態に分離した後、補助情報の有無に応じて、複数の内部状態の何れかを選択するか、複数の内部状態の重み付け和である内部状態を生成するかの何れかを行った後、選択又は生成された内部状態を変換することでマスクの推定を行う。そのため、1つのニューラルネットワークのモデルを用いて、ブラインド音源分離と目的話者抽出を切り替えて行うことが可能となる。

[0050] また、本発明の実施例における学習装置200は、ブラインド音源分離と目的話者抽出の双方の損失を考慮したマルチタスク学習を行う。そのため、個別の学習よりも良い分離性能を持つ信号処理装置を学習できる。

[0051] 本発明の実施例における信号処理装置100の性能を評価するため、実験データセットを用いて、ブラインド音源分離手法であるPIT (Permutation Invariant Training) と、目的話者抽出手法であるSpeakerBeamと、本発明の実施例(本方式)との性能評価を行った。3つの手法のいずれも3層のBLSTMに基づくニューラルネットワーク構造を使用した。図8は、本発明の実施例の評価結果を示す図であり、未処理の混合音声信号と、3つの手法のSDR (Signal to Distortion Ratio) (dB単位)を示す。図8から、補助情報を用いない場合、本発明の実施例はマルチタスク学習の効果によってPITと比較してより良い分離性能を示していることが分かる。また、補助情報を用いる場合で

あっても、その用途に特化して設計されたSpeakerBeamと同等の分離性能を示していることが分かる。

[0052] <ハードウェア構成例>

図9に、本発明の実施例における各装置（信号処理装置100、学習装置200）のハードウェア構成例を示す。各装置は、CPU（Central Processing Unit）151等のプロセッサ、RAM（Random Access Memory）やROM（Read Only Memory）等のメモリ装置152、ハードディスク等の記憶装置153等から構成されたコンピュータでもよい。例えば、各装置の機能及び処理は、記憶装置153又はメモリ装置152に格納されているデータやプログラムをCPU151が実行することによって実現される。また、各装置に必要な情報は、入出力インタフェース装置154から入力され、各装置において求められた結果は、入出力インタフェース装置154から出力されてもよい。

[0053] <補足>

説明の便宜上、本発明の実施例に係る信号処理装置及び学習装置は機能的なブロック図を用いて説明しているが、本発明の実施例に係る信号処理装置及び学習装置は、ハードウェア、ソフトウェア又はそれらの組み合わせで実現されてもよい。例えば、本発明の実施例は、コンピュータに対して本発明の実施例に係る信号処理装置及び学習装置の機能を実現させるプログラム、コンピュータに対して本発明の実施例に係る方法の各手順を実行させるプログラム等により、実現されてもよい。また、各機能部が必要に応じて組み合わせて使用されてもよい。また、本発明の実施例に係る方法は、実施例に示す順序と異なる順序で実施されてもよい。

[0054] 以上、ブラインド音源分離と目的話者抽出とを統合的に扱うための手法について説明したが、本発明は、上記の実施例に限定されることなく、特許請求の範囲内において、種々の変更・応用が可能である。

符号の説明

[0055] 100 信号処理装置
110 変換部

- 1 2 0 補助情報入力部
- 1 3 0 重み付け部
- 1 4 0 マスク推定部
- 2 0 0 学習装置
- 2 1 0 変換部
- 2 2 0 補助情報入力部
- 2 3 0 重み付け部
- 2 4 0 マスク推定部
- 2 5 0 パラメータ更新部

請求の範囲

- [請求項1] 入力された混合音響信号を複数の第1内部状態に変換する変換部と、
- 、
- 目的音源の音響信号に関する補助情報が入力された場合、前記補助情報に基づいて前記複数の第1内部状態の重み付け和である第2内部状態を生成し、前記補助情報が入力されない場合、前記複数の第1内部状態のいずれかを選択することによって第2内部状態を生成する重み付け部と、
- 前記第2内部状態に基づいてマスクを推定するマスク推定部と、
- を有する信号処理装置。
- [請求項2] 前記変換部、前記重み付け部及び前記マスク推定部のそれぞれはニューラルネットワークを用いて構成され、
- 前記ニューラルネットワークのそれぞれは、予め用意された学習用の混合音響信号に対して前記マスク推定部によって推定されたマスクを適用して得られる音響信号と、前記学習用の混合音響信号に含まれる正解の音源の音響信号との誤差が小さくなるように学習されたものである、請求項1に記載の信号処理装置。
- [請求項3] 前記変換部は、前記入力された混合音響信号を、 I 個の第1内部状態に変換し、
- 前記重み付け部は、前記補助情報が入力された場合、前記 I 個の第1内部状態に、前記 I 個の第1内部状態及び前記補助情報に基づいて推定された I 次元の重みベクトルを適用することによって第2内部状態を生成し、前記補助情報が入力されない場合、前記 I 個の第1内部状態に、 i ($i=1, \dots, I$) 番目の要素が1であり他の要素が0である I 次元の単位ベクトルを適用することによって第2内部状態を生成する、請求項1又は2に記載の信号処理装置。
- [請求項4] ニューラルネットワークを用いて、入力された学習用の混合音響信号を複数の第1内部状態に変換する変換部と、

目的音源の音響信号に関する補助情報が入力された場合、ニューラルネットワークを用いて、前記複数の第1内部状態の重み付け和である第2内部状態を生成し、前記補助情報が入力されない場合、前記複数の第1内部状態のいずれかを選択することによって第2内部状態を生成する重み付け部と、

ニューラルネットワークを用いて、前記第2内部状態に基づいてマスクを推定するマスク推定部と、

前記学習用の混合音響信号に対して前記マスク推定部によって推定されたマスクを適用して得られる音響信号と、前記学習用の混合音響信号に含まれる正解の音源の音響信号との比較結果に基づいて、前記変換部、前記重み付け部及び前記マスク推定部のそれぞれにおいて用いられるニューラルネットワークのパラメータを更新するパラメータ更新部と、

を有する学習装置。

[請求項5] 前記パラメータ更新部は、前記補助情報が入力された場合の損失と、前記補助情報が入力されない場合の損失との双方を考慮して、パラメータを更新する、請求項4に記載の学習装置。

[請求項6] 信号処理装置が実行する信号処理方法であって、
入力された混合音響信号を複数の第1内部状態に変換するステップと、

目的音源の音響信号に関する補助情報が入力された場合、前記複数の第1内部状態の重み付け和である第2内部状態を生成し、前記補助情報が入力されない場合、前記複数の第1内部状態のいずれかを選択することによって第2内部状態を生成するステップと、

前記第2内部状態に基づいてマスクを推定するステップと、

を有する信号処理方法。

[請求項7] 学習装置が実行する学習方法であって、
ニューラルネットワークを用いて、入力された学習用の混合音響信

号を複数の第1内部状態に変換するステップと、

目的音源の音響信号に関する補助情報が入力された場合、ニューラルネットワークを用いて、前記複数の第1内部状態の重み付け和である第2内部状態を生成し、前記補助情報が入力されない場合、前記複数の第1内部状態のいずれかを選択することによって第2内部状態を生成するステップと、

ニューラルネットワークを用いて、前記第2内部状態に基づいてマスクを推定するステップと、

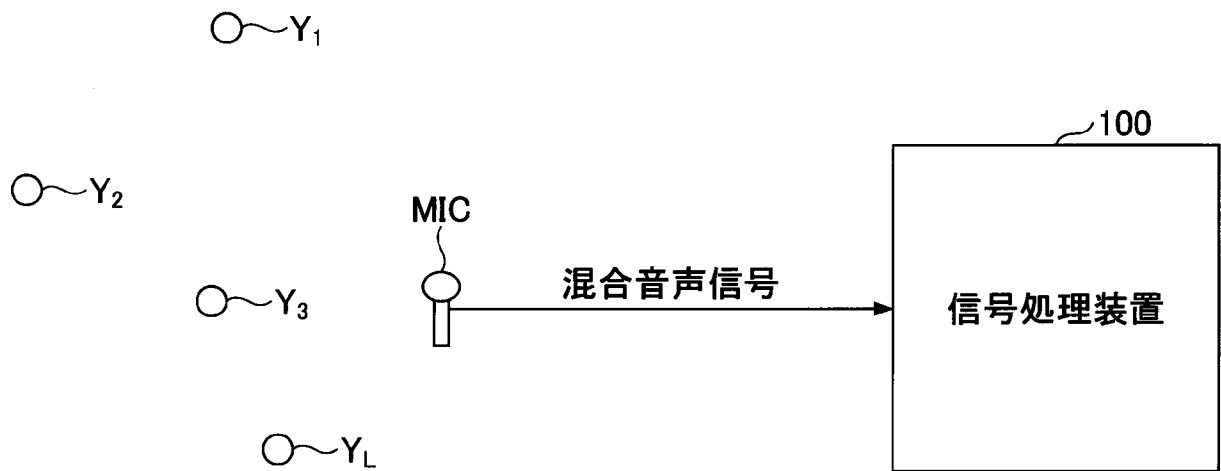
前記学習用の混合音響信号に対して前記推定されたマスクを適用して得られる音響信号と、前記学習用の混合音響信号に含まれる正解の音源の音響信号との比較結果に基づいて、前記変換するステップ、前記生成するステップ及び前記推定するステップのそれぞれにおいて用いられるニューラルネットワークのパラメータを更新するステップと、

を有する学習方法。

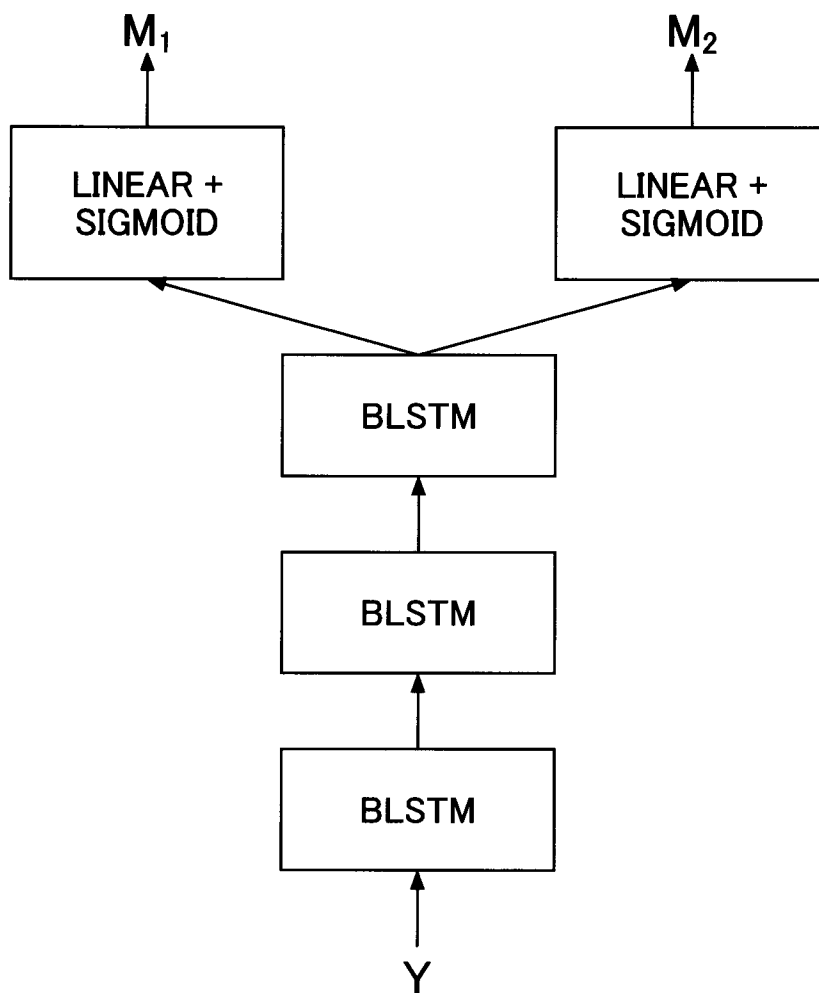
[請求項8]

請求項1乃至5のうちいずれか1項に記載の装置としてコンピュータを機能させるためのプログラム。

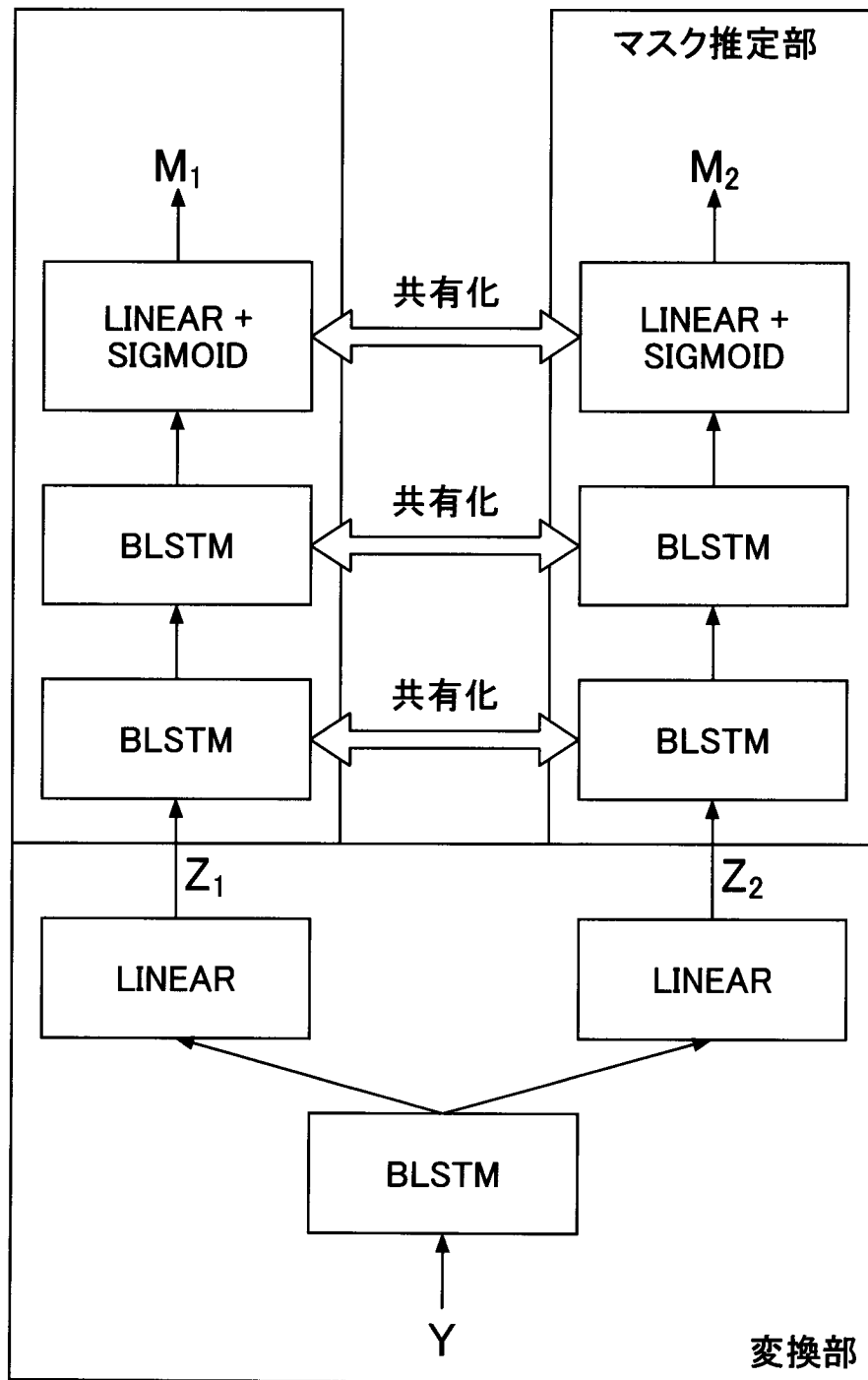
[図1]



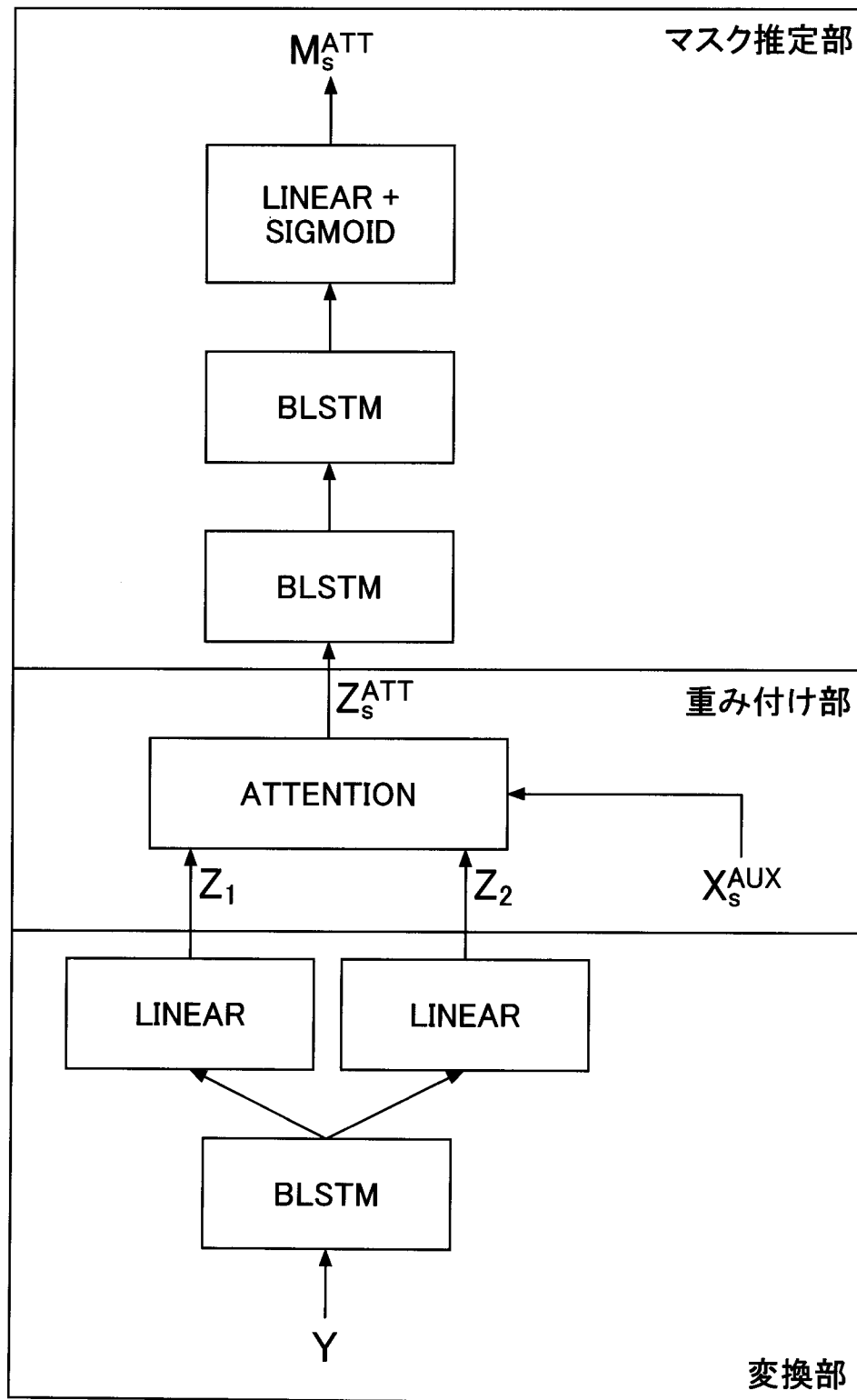
[図2]



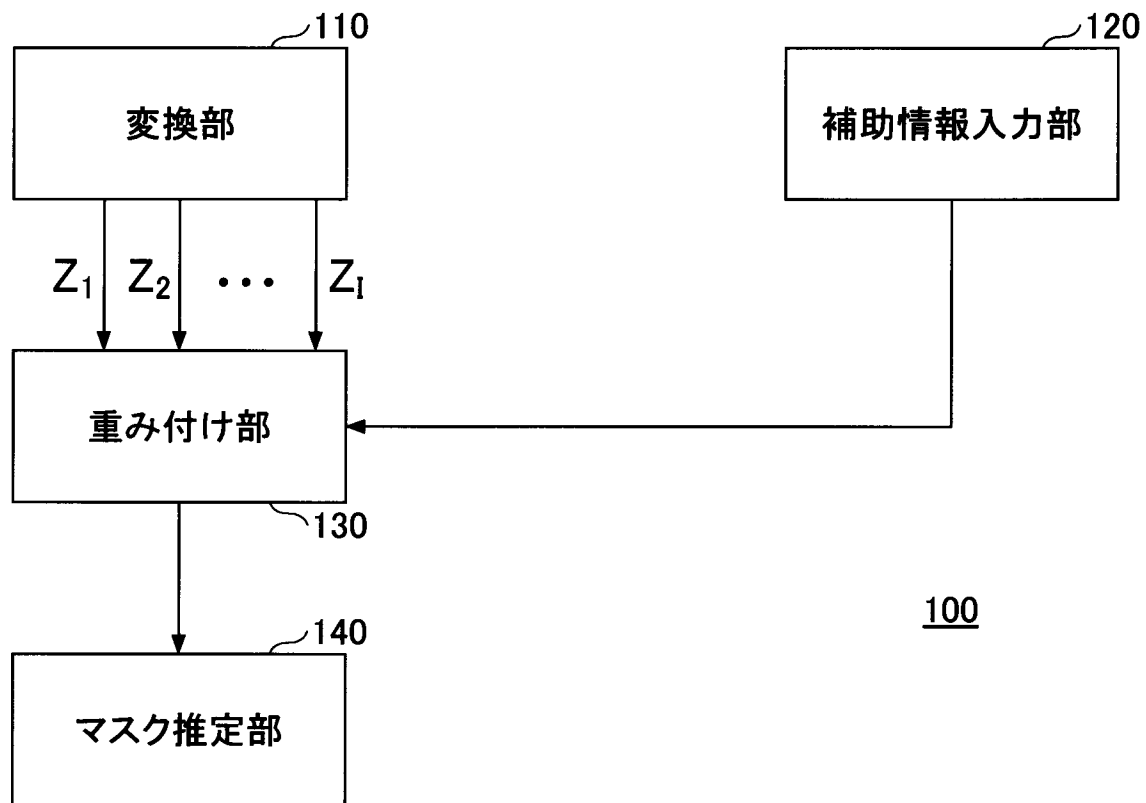
[図3]



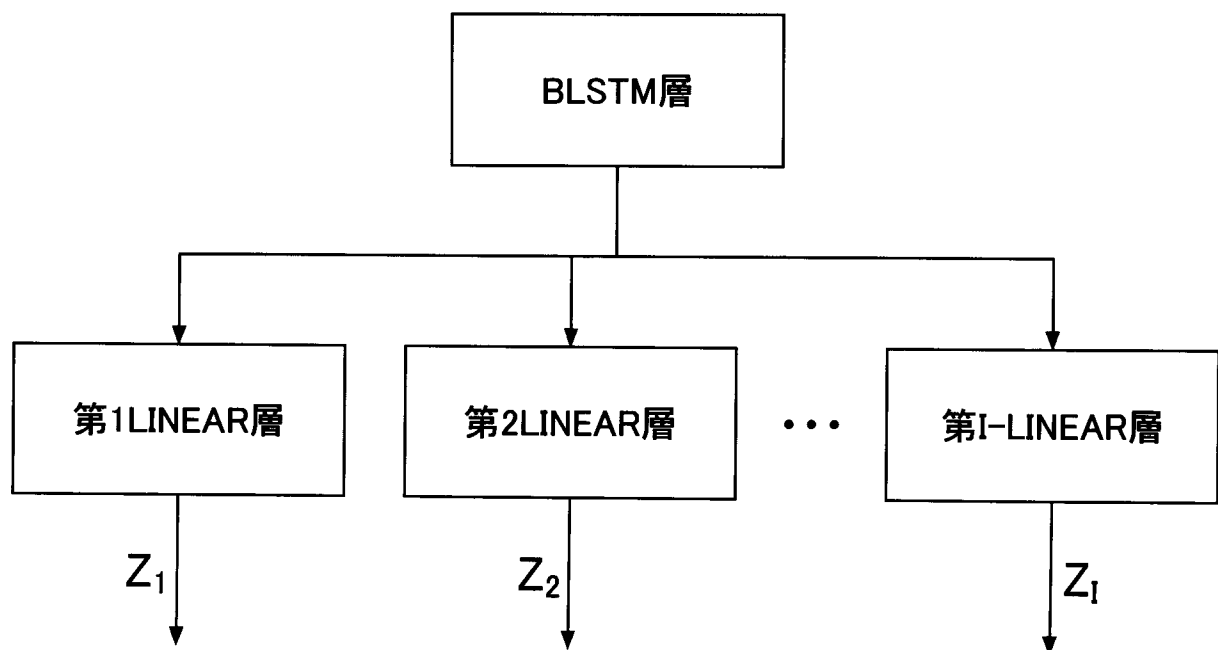
[図4]



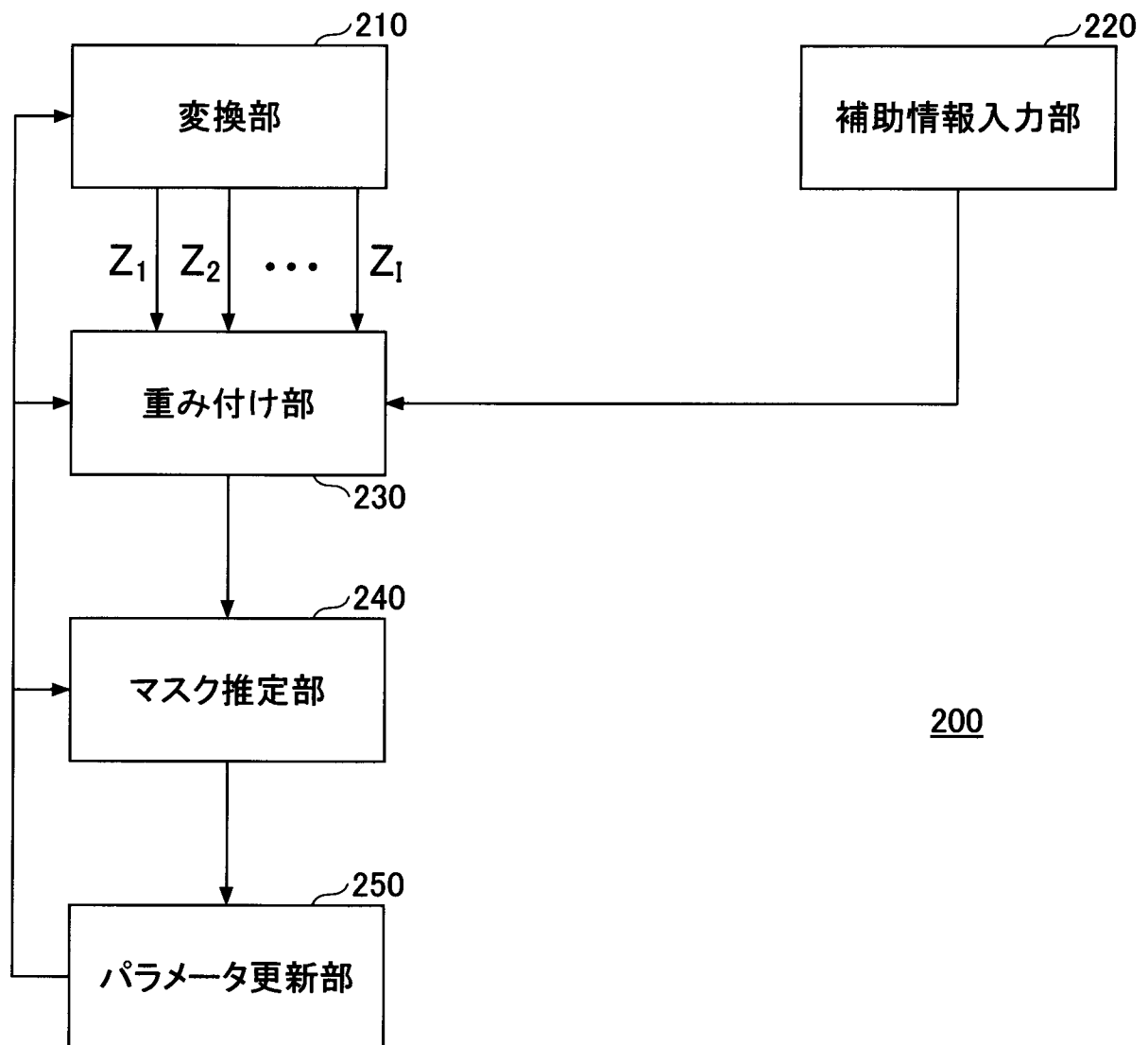
[図5]



[図6]



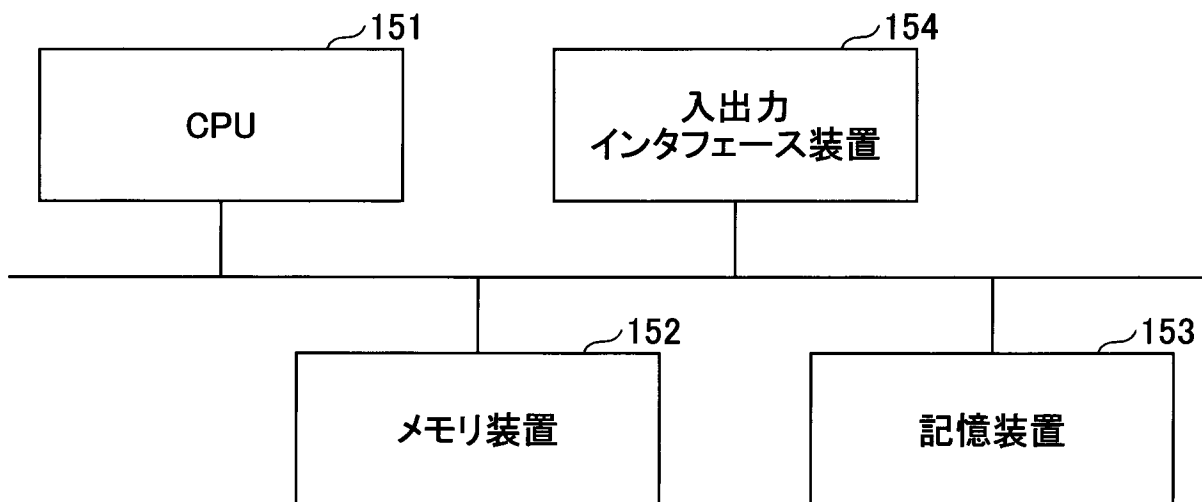
[図7]



[図8]

	補助情報なし	補助情報あり
混合音声信号	0.2	
PIT	8.6	—
SpeakerBeam	—	9.6
本方式	9.2	9.6

[図9]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/005332

A. CLASSIFICATION OF SUBJECT MATTER

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

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; 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	1971-2020
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)

IEEE Xplore

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2019/017403 A1 (NIPPON TELEGRAPH AND TELEPHONE CORP.) 24.01.2019 (2019-01-24) paragraphs [0056]-[0101], fig. 3-6	1-8
A	DELCROIX, Marc et al., SpeakerBeam: 聞きたい人の声に耳を傾けるコンピューター深層学習に基づく音声の選択的聴取, NTT 技術ジャーナル, September 2018, vol. 30, no. 9, pp. 12-15, https://www.ntt.co.jp/journal/1809/files/JN20180912.pdf , entire text, all drawings, non-official translation ("SpeakerBeam: Selective listening of voice-based on computer deep learning that listens to the voice of the person who wants to hear", NTT Technical Journal)	1-8

☒ 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

"I" 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
16 April 2020 (16.04.2020)

Date of mailing of the international search report
28 April 2020 (28.04.2020)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/005332

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CHEN, Xiuyi et al., "Distilled Binary Neural Network for Monaural Speech Separation", 2018 International Joint Conference on Neural Networks (IJCNN), July 2018, pp. 1-8, https://ieeexplore.ieee.org/document/8489456 , entire text, all drawings	1-8

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/JP2020/005332

Patent Documents
referred in the
Report

Publication
Date

Patent Family

Publication
Date

WO 2019/017403 A1

24 Jan. 2019

(Family: none)

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 25/30(2013.01)i FI: G10L21/0272 100B; G10L25/30										
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; 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 <table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922 - 1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971 - 2020年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2020年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2020年</td> </tr> </table>			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
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） IEEE Xplore										
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 2019/017403 A1（日本電信電話株式会社）24.01.2019（2019 - 01 - 24） [0056]-[0101], 図3-6	1-8								
A	DELCROIX, Marc et al., SpeakerBeam: 聞きたい人の声に耳を傾けるコンピュータ 深層学習に基づく音声の選択的聴取, NTT技術ジャーナル, 2018.09, Vol. 30, No. 9, 第12-15ページ, https://www.ntt.co.jp/journal/1809/files/JN20180912.pdf 全文, 全図	1-8								
A	CHEN, Xiuyi et al., Distilled Binary Neural Network for Monaural Speech Separation, 2018 International Joint Conference on Neural Networks (IJCNN), 2018.07, pp. 1-8, https://ieeexplore.ieee.org/document/8489456 全文, 全図	1-8								
☐ 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 16.04.2020	国際調査報告の発送日 28.04.2020									
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） 岩田 淳 5Z 4052 電話番号 03-3581-1101 内線 3591									

国際出願番号
PCT/JP2020/005332

引用文献	公表日	パテントファミリー文献	公表日
WO 2019/017403 A1	24.01.2019	(ファミリーなし)	