

(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
2021年7月22日(22.07.2021)



(10) 国際公開番号

WO 2021/144934 A1

(51) 国際特許分類:
G10L 21/0232 (2013.01) G10L 21/0264 (2013.01)

(21) 国際出願番号 : PCT/JP2020/001356

(22) 国際出願日 : 2020年1月16日(16.01.2020)

(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: 小泉 悠馬 (KOIZUMI, Yuma); 〒1808585 東京都武蔵野市緑町三丁目9番11号 NTT 知的財産センタ内 Tokyo (JP).

(74) 代理人: 中尾 直樹, 外 (NAKAO, Naoki et al.); 〒1600022 東京都新宿区新宿三丁目1番22号 新宿NSOビル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, 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.

(54) Title: VOICE ENHANCEMENT DEVICE, LEARNING DEVICE, METHODS THEREFOR, AND PROGRAM

(54) 発明の名称 : 音声強調装置、学習装置、それらの方法、およびプログラム

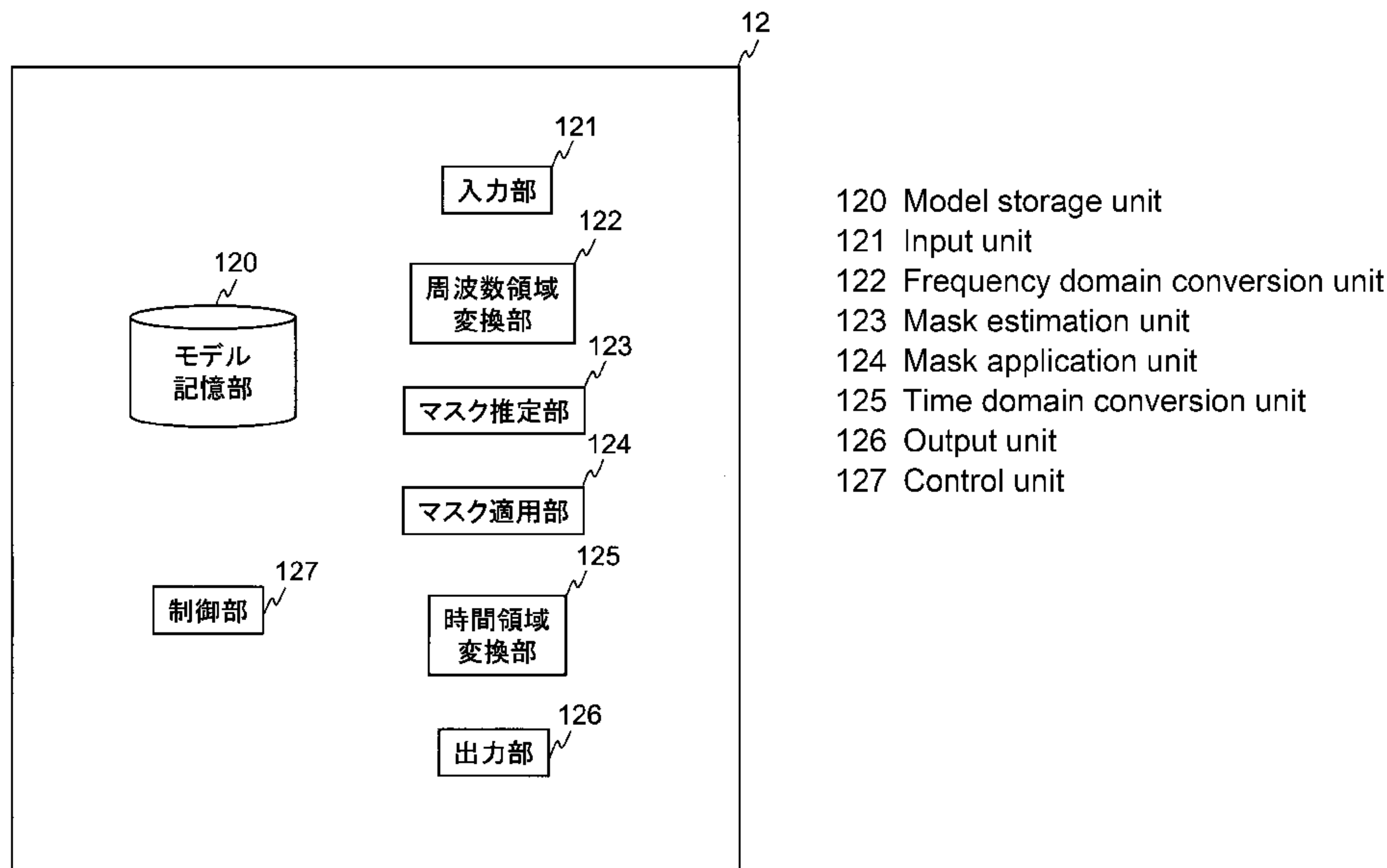


図2

(57) Abstract: A mask that enhances a voice emanating from a speaker is estimated from an observation signal, the mask is applied to the observation signal, and a post-mask voice signal is acquired. This mask is estimated from a characteristic quantity that results from combining a characteristic quantity for speaker recognition that is extracted from the observation signal and a characteristic quantity for generalization mask estimation that is extracted from the observation signal.

(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:

一 国際調査報告 (条約第21条(3))

(57) 要約: 話者から発せられた音声を強調するマスクを観測信号から推定し、観測信号にマスクを適用し、マスク後音声信号を取得する。このマスクは、観測信号から抽出された話者認識用の特徴量と、観測信号から抽出された汎化マスク推定用の特徴量と、を組み合わせた特徴量から推定される。

明 細 書

発明の名称：

音声強調装置、学習装置、それらの方法、およびプログラム

技術分野

[0001] 本発明は、音声強調技術に関する。

背景技術

[0002] 深層学習を利用した音声強調の代表的な手法に、深層ニューラルネットワーク（DNN: deep neural network）を利用して時間周波数（T-F: time-frequency）マスクを推定する方法がある（DNN音声強調）。これは、短時間フーリエ変換（STFT: short-time Fourier transform）などを利用して観測信号を時間周波数領域表現した観測信号を得、それに対してDNNを利用して推定した時間周波数マスクを乗じ、その結果を逆STFTして強調音を得る方法である（例えば、非特許文献1から5等参照）。

[0003] DNN音声強調を実現するうえで重要な機能要件として“汎化性能”がある。これは、あらゆる話者（例えば、既知話者であっても未知話者であっても、男性であっても女性であっても、幼児であっても老人であっても）の発話であっても音声強調が可能という性能である。これを実現するために、従来のDNN音声強調では、大量の話者が発話した大量の音声データを利用して一つのDNNを学習し、話者非依存モデルを学習することを是としてきた。

[0004] 一方で他の音声アプリケーションでは、モデルを“特化”する試みが成功を収めている。つまり、ある特定の話者にのみ高性能なDNNを学習する方法である。これを実現する代表的な方法が“モデル適応”である。

先行技術文献

非特許文献

[0005] 非特許文献1：C. Valentini-Botinho, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based Speech Enhancement methods for Noise-Rob

ust Text-to-Speech,” Proc. of 9th ISCA Speech Synth. Workshop (SSW), 2016.

非特許文献2 : S. Pascual, A. Bonafonte, and J. Serra, ”SEGAN: Speech Enhancement Generative Adversarial Network,” Proc. of Interspeech, 2017.

非特許文献3 : M. H. Soni, N. Shah, H. A. Patil, ”Time-Frequency Masking-Based Speech Enhancement Using Generative Adversarial Network,” Proc. of Int. Conf. on Acoust., Speech, and Signal Process. (ICASSP), 2018.

非特許文献4 : F. G. Germain, Q. Chen, and V. Koltun, ”Speech Denoising with Deep Feature Losses,” arXiv preprint, arXiv:1806.10522, 2018.

非特許文献5 : S. W. Fu, C. F. Liao, Y. Tsao, and S. D. Lin, ”MetricGAN : Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement,” Proc. of Int. Conf. on Machine Learning (ICML), 2019.

発明の概要

発明が解決しようとする課題

[0006] しかし、従来のモデルを“特化”する手法では、音声強調しようとする所望の話者（目的話者）の補助発話が必要であるという問題点がある。

[0007] 本発明はこのような点に鑑みてなされたものであり、音声強調しようとする目的話者の補助発話を必要とすることなく、目的話者に特化した音声強調を行うことを目的とする。

課題を解決するための手段

[0008] 話者から発せられた音声を強調するマスクを観測信号から推定し、観測信号にマスクを適用し、マスク後音声信号を取得する。このマスクは、観測信号から抽出された話者認識用の特徴量と、観測信号から抽出された汎化マスク推定用の特徴量と、を組み合わせた特徴量から推定される。

発明の効果

[0009] 以上のように、本発明では、音声強調しようとする目的話者の補助発話を必要とすることなく、目的話者に特化した音声強調を行うことができる。

図面の簡単な説明

[0010] [図1]図1は実施形態の学習装置の機能構成を例示したブロック図である。

[図2]図2は実施形態の音声強調装置の機能構成を例示したブロック図である。

[図3]図3は実施形態の学習方法を例示したフロー図である。

[図4]図4は実施形態の音声強調方法を例示したフロー図である。

[図5]図5は、ハードウェア構成を説明するためのブロック図である。

発明を実施するための形態

[0011] 以下、図面を参照して本発明の実施形態を説明する。

[原理]

まず原理を説明する。

< D N N 音声強調 >

問題設定： T サンプルの時間領域での観測信号 $x \in \mathbb{R}^T$ は、目的音声信号 s と雑音信号 n の混合信号 $x = s + n$ であるとする。音声強調の目的は、 x から s を高精度に推定することである。式 (1) に例示するように、 D N N 音声強調では、短時間フーリエ変換などの周波数領域変換処理 $Q : \mathbb{R}^T \rightarrow \mathbb{R}^{F \times K}$ によって観測信号 x を時間周波数領域表現した観測信号 $X = Q(x) \in \mathbb{C}^{F \times K}$ を得、 X に D N N を利用して推定した時間周波数 (T - F) マスク M を乗じてマスク後音声信号 $M(x; \theta) \odot Q(x)$ を得、さらにマスク後音声信号 $M(x; \theta) \odot Q(x)$ に対し、逆 S T F T などの時間領域変換処理 Q^+ を適用して強調音声 y を得る。

$$y = Q^+(M(x; \theta) \odot Q(x)) \quad (1)$$

ここで、 $\mathbb{R}$ は実数全体の集合を表し、 $\mathbb{C}$ は複素数全体の集合を表す。 T , F , K は正整数であり、 T は所定の時間区間に属する観測信号 x の個数 (時間長) を表し、 F は時間周波数領域の所定の帯域に属する離散周波数の個数 (

帯域幅)を表し、 K は時間周波数領域の所定の時間区間に属する離散時間の個数(時間長)を表す。 $M(x; \theta) \odot Q(x)$ は、 $Q(x)$ にT-Fマスク $M(x; \theta)$ を乗じることを表す。 θ はDNNのパラメータであり、通常は例えば以下の式(2)で表される信号対歪比(SDR: signal-to-distortion ratio) L^{SDR} を最小化するように学習される。

$$L^{SDR} = -(\text{clip}_{\beta}[\text{SDR}(s, y)] + \text{clip}_{\beta}[\text{SDR}(n, m)]) / 2 \quad (2)$$

ただし、

[数1]

$$\text{SDR}(s, y) = 10 \log_{10}(\|s\|_2^2 / \|s - y\|_2^2)$$

であり、

[数2]

$$\|\cdot\|_2$$

は L_2 ノルムであり、 $m = x - y$ であり、 $\text{clip}_{\beta}[x] = \beta \cdot \tanh(x / \beta)$ であり、 $\beta > 0$ はクリッピング定数である。例えば、 $\beta = 20$ である。

[0012] <DNN音声強調における“汎化”と“特化”>

着眼点：DNN音声強調を実現するうえで重要な機能要件として“汎化性能”がある。これは、あらゆる話者の発話であっても音声強調が可能という性能である。これを実現するために、従来のDNN音声強調では、大量の話者が発話した大量の音声データを利用して一つのDNNを学習し、話者非依存モデルを学習することを是としてきた。

[0013] 一方で他の音声アプリケーションでは、モデルを“特化”する試みが成功を収めている。つまり、ある特定の話者にのみ高性能なDNNを学習する方法である。これを実現する代表的な方法が“モデル適応”である。

[0014] 本実施形態では、このような話者適応の考え方をDNN音声強調に組み込むことで高精度化を実現する。その際、話者認識に関するマルチタスク学習を導入することで、補助発話が不要かつ、真の話者(目的話者)に特化した

DNN音声強調を実現する。例えば、DNNを利用したT-Fマスク推定器の内部に話者認識器を組み込み、そのボトルネック特徴をマスク推定に利用する。これを数式で記載すると以下ようになる。

$$M(x; \theta) = M_2(\Phi, \Psi; \theta_2) \quad (3)$$

$$\Phi = M_1(x; \theta_1) \in \mathbb{R}^{D_m \times K} \quad (4)$$

$$\Psi = Z_D(x; \theta_z) \in \mathbb{R}^{D_z \times K} \quad (5)$$

$$Z = (z_1, \dots, z_K) = W\Psi \in \mathbb{R}^{H \times K} \quad (6)$$

[数3]

$$\hat{z} = \text{softmax}\left(\frac{1}{K} \sum_{k=1}^K z_k\right) \in \mathbb{R}^H \quad (7)$$

ただし、 M_1 はパラメータ θ_1 を持つマスク推定特徴抽出DNNであり、観測信号 x から汎化マスク推定用（汎用マスク推定用）の特徴量 Φ を得て出力する。なお、汎化マスク（汎用マスク）とは特定の話者に特化されていないマスクを意味する。言い換えると、汎化マスクはすべての話者に共通するマスクである。 Z_D はパラメータ θ_z を持つ話者認識用特徴抽出DNNであり、観測信号 x から話者認識用の特徴量 Ψ を得て出力する。 M_2 はパラメータ θ_2 を持つマスク推定特徴抽出DNNであり、特徴量 Φ および Ψ からT-Fマスク $M(x; \theta)$ を推定して出力する。 $W \in \mathbb{R}^{H \times D_z}$ は行列である。softmaxはsoftmax関数である。 D_m 、 D_z 、 H 、 K は正整数である。 H は学習データセットが収録された環境における話者の数である。 θ はパラメータ θ_1 、 θ_2 、 θ_z の集合 $\{\theta_1, \theta_2, \theta_z\}$ を表す。

[0015] パラメータ θ_1 、 θ_2 、 θ_z は、観測信号 x および目的音声信号 s の学習データセットを用いた機械学習によって得られる。目的音声信号 s には当該目的音声信号 s を発話した話者を識別する情報 z が付与されている。 z の一例は、 s を発話した真の話者（目的話者）に対応する要素のみが1であって他の要素が0のベクトル（one-hot-vector）である。

[0016] 観測信号 x はマスク推定特徴抽出DNN M_1 および話者認識用特徴抽出D

$NN\ Z_D$ に入力され、マスク推定特徴抽出 $NN\ M_1$ および話者認識用特徴抽出 $NN\ Z_D$ は、それぞれ特徴量 $\Phi \in R^{D_m \times K}$ および $\Psi \in R^{D_z \times K}$ を得て出力する（式（４）（５））。 Φ と Ψ はマスク推定特徴抽出 $NN\ M_2$ に入力され（例えば、 Φ と Ψ は特徴量次元方向に結合されて M_2 に入力され）、マスク推定特徴抽出 $NN\ M_2$ はＴ－Ｆマスク $M(x; \theta)$ を得て出力する（式（３））。同時に、 Ψ に対して行列 $W \in R^{H \times D_z}$ が乗じられて $Z = (z_1, \dots, z_K)$ が得られ（式（６））、さらに式（７）を利用して推定話者を識別する情報 $\hat{z}$ が得られる。推定話者を識別する情報の種別は、推定話者を識別する情報の種別と同一である。推定話者を識別する情報の例は、推定話者に対応する要素のみが１であって他の要素が０のベクトル（one-hot-vector）である。また $\hat{z}$ の添え字「 $\hat{\cdot}$ 」は、式（７）のように「 z 」の真上に記載すべきであるが、記載表記の制約上「 z 」の右上に記載してある。パラメータ θ_1 、 θ_2 、 θ_z は、以下の、音声強調と話者認識のコスト関数が合成された、マルチタスク型のコスト関数 L を最小化するように学習される。

$$L = L^{SDR} + \alpha \text{CrossEntropy}(z, \hat{z}) \quad (8)$$

ここで $\alpha > 0$ は混合パラメータであり、例えば $\alpha = 1$ に設定できる。 $\text{CrossEntropy}(z, \hat{z})$ は z と $\hat{z}$ のクロスエントロピーである。特徴量 Ψ は話者認識のボトルネック特徴を表し、音声強調性能を向上させ、かつ、話者を判定するように抽出される。ゆえに、特徴量 Ψ には音声強調性能を向上させるための目的話者に関する情報を含んでおり、これをＴ－Ｆマスク M の推定に用いることで、目的話者の発話を強調する音声強調への特化が可能と期待できる。

[0017] [第１実施形態]

次に、図面を用いて本発明の第１実施形態を説明する。

<構成>

図１に例示するように、本実施形態の学習装置１１は、初期化部１１１、コスト関数計算部１１２、パラメータ更新部１１３、収束判定部１１４、出力部１１５、制御部１１６、記憶部１１７、１１８、およびメモリ１１９を

有する。初期化部 1 1 1、コスト関数計算部 1 1 2、パラメータ更新部 1 1 3、および収束判定部 1 1 4 が「学習部」に相当する。音声強調装置 1 1 は、制御部 1 1 6 の制御の下で各処理を実行する。図 2 に例示するように、本実施形態の音声強調装置 1 2 は、記憶部 1 2 0、入力部 1 2 1、周波数領域変換部 1 2 2、マスク推定部 1 2 3、マスク適用部 1 2 4、時間領域変換部 1 2 5、出力部 1 2 6、および制御部 1 2 7 を有する。音声強調装置 1 2 は制御部 1 2 7 の制御の下で各処理を実行する。

[0018] <学習処理>

学習処理の前提として、学習装置 1 1（図 1）の記憶部 1 1 7 に観測信号 x の学習データが格納され、記憶部 1 1 8 に目的音声信号 s の学習データが格納される。観測信号 x は時系列音響信号であり、目的音声信号 s と雑音信号 n の混合信号 $x = s + n$ である。目的音声信号 s も時系列音響信号であり、目的話者が発話したクリーンな音声信号である。目的音声信号 s には、目的話者を識別する情報（例えば、目的話者に対応する要素のみが 1 であって他の要素が 0 のベクトル）が付与されている。雑音信号 n は、目的話者が発話した音声信号以外の時系列音響信号である。

[0019] 図 3 に例示するように、学習処理では、まず学習装置 1 1（図 1）の初期化部 1 1 1 が擬似乱数などを利用して各パラメータ θ_1 , θ_2 , θ_z を初期化してメモリ 1 1 9 に格納する（ステップ S 1 1 1）。

[0020] 次に、コスト関数計算部 1 1 2 に、記憶部 1 1 7 から抽出した観測信号 x の学習データ、記憶部 1 1 8 から抽出した目的音声信号 s の学習データ、およびメモリ 1 1 9 から抽出したパラメータ θ_1 , θ_2 , θ_z が入力される。コスト関数計算部 1 1 2 は、例えば、式（1）～（8）に従って式（8）に示すコスト関数 L を計算して出力する（ステップ S 1 1 2）。式（2）（8）より、式（8）のコスト関数は以下のように変形できる。

$$L = -(\text{clip}_\beta[\text{SDR}(s, y)] + \text{clip}_\beta[\text{SDR}(n, m)]) / 2 + \alpha \text{CrossEntropy}(z, \hat{z}) \quad (9)$$

すなわち、コスト関数 L は、T-F マスクを観測信号 x に適用して得られる

マスク後音声信号に対応する音声強調信号 y と観測信号 x に含まれた目的音声信号 s との距離に対応する第 1 関数 ($-\text{clip}_\beta[\text{SDR}(s, y)]/2$) と、観測信号 x に含まれた雑音信号 n と観測信号 x から音声強調信号 y を除いた残存信号 m との距離に対応する第 2 関数 ($-\text{clip}_\beta[\text{SDR}(n, m)]/2$) と、推定話者を識別する情報 $\hat{z}$ と目的音声信号を発した話者を識別する情報 z との距離に対応する第 3 関数 ($\alpha \text{CrossEntropy}(z, \hat{z})$) とを加算したものである。ここで、第 1 関数の関数値が小さいほどコスト関数 L の関数値は小さく、第 2 関数の関数値が小さいほどコスト関数 L の関数値は小さく、第 3 関数の関数値が小さいほどコスト関数 L の関数値は小さい。

[0021] パラメータ更新部 113 には、コスト関数 L およびパラメータ θ_1 , θ_2 , θ_z が入力される。パラメータ更新部 113 は、コスト関数 L を最小化するようにパラメータ θ_1 , θ_2 , θ_z を更新する。例えば、パラメータ更新部 113 は、コスト関数 L に関する勾配を計算して勾配法によってコスト関数 L を最小化するようにパラメータ θ_1 , θ_2 , θ_z を更新する。パラメータ更新部 113 は、更新後のパラメータ θ_1 , θ_2 , θ_z でメモリ 119 に格納されたパラメータ θ_1 , θ_2 , θ_z を更新する（ステップ S113）。なお、パラメータ θ_1 , θ_2 , θ_z を更新することは、それぞれ、マスク推定特徴抽出 DNN M_1 , マスク推定特徴抽出 DNN M_2 , 話者認識用特徴抽出 DNN Z_D を更新することである。

[0022] 収束判定部 114 は、パラメータ θ_1 , θ_2 , θ_z の収束条件を満たしたか否かを判定する。収束条件の例は、ステップ S112～S114 の処理を所定回数繰り返したこと、ステップ S112～S114 の処理を実行する前後でのパラメータ θ_1 , θ_2 , θ_z やコスト関数 L の変化量が所定値以下であることなどである（ステップ S114）。

[0023] ここで収束条件を満たしていないと判定された場合には、処理がステップ S112 に戻される。一方、収束条件を満たしていると判定された場合、出力部 115 はパラメータ θ_1 , θ_2 , θ_z を出力する（ステップ S115）。このパラメータ θ_1 , θ_2 , θ_z は、例えば、収束条件を満たしていると判定され

た収束判定（ステップS 1 1 4）の直前のステップS 1 1 3で得られたものである。しかし、これに代え、それよりも前の時点で更新されたパラメータ θ_1 , θ_2 , θ_z が出力されてもよい。

[0024] 以上のステップS 1 1 1～S 1 1 5により、観測信号 x から話者認識用の特徴量 Ψ および汎化マスク推定用の特徴量 Φ を抽出し、話者認識用の特徴量 Ψ と汎化マスク推定用の特徴量 Φ とを組み合わせた特徴量からT-Fマスクを推定し、話者認識用の特徴量 Ψ から推定話者を識別する情報を得るモデル $M_1(x; \theta_1)$, $M_2(\Phi, \Psi; \theta_2)$, $Z_D(x; \theta_z)$ が学習される。

[0025] <音声強調処理>

上述のように学習されたモデル $M_1(x; \theta_1)$, $M_2(\Phi, \Psi; \theta_2)$, $Z_D(x; \theta_z)$ を特定する情報は、音声強調装置12（図2）のモデル記憶部120に格納される。例えば、ステップS 1 1 5で出力部115から出力されたパラメータ θ_1 , θ_2 , θ_z が、モデル記憶部120に格納される。この前提の下、以下のような音声強調処理が実行される。

[0026] 図4に例示するように、音声強調装置12（図2）の入力部121には、時間領域の時系列音響信号である観測信号 x が入力される（ステップS 1 2 1）。

[0027] 観測信号 x は周波数領域変換部122に入力される。周波数領域変換部122は、短時間フーリエ変換などの周波数領域変換処理 Q によって、観測信号 x を時間周波数領域表現した観測信号 $X=Q(x)$ を得て出力する（ステップS 1 2 2）。

[0028] 観測信号 x はマスク推定部123に入力される。マスク推定部123は、話者から発せられた音声を強調するT-Fマスク $M(x; \theta)$ を観測信号 x から推定して出力する。ここでマスク推定部123は、観測信号 x から抽出された話者認識用の特徴量 Ψ と、観測信号 x から抽出された汎化マスク推定用の特徴量 Φ と、を組み合わせた特徴量からT-Fマスク $M(x; \theta)$ の推定を行う。以下にこの処理を例示する。まずマスク推定部123は、モデル記憶部120からマスク推定特徴抽出DNN M_1 および話者認識用特徴抽出

DNN Z_D を特定するための情報（例えば、パラメータ θ_1 , θ_z ）を抽出し、観測信号 x を M_1 および Z_D に入力し、それぞれ特徴量 Φ および Ψ を得る（式（４）（５））。次にマスク推定部 123 は、モデル記憶部 120 からマスク推定特徴抽出 DNN M_2 を特定するための情報（例えば、パラメータ θ_2 ）を抽出し、 Φ と Ψ をマスク推定特徴抽出 DNN M_2 に入力して T-F マスク $M(x; \theta)$ を得て出力する（式（３））（ステップ S123）。

[0029] 観測信号 X および T-F マスク $M(x; \theta)$ はマスク適用部 124 に入力される。マスク適用部 124 は、時間周波数領域で観測信号 X に T-F マスク $M(x; \theta)$ を適用し（乗算し）、マスク後音声信号 $M(x; \theta) \odot X$ を得て出力する（ステップ S124）。

[0030] マスク後音声信号 $M(x; \theta) \odot X$ は、時間領域変換部 125 に入力される。時間領域変換部 125 は、マスク後音声信号 $M(x; \theta) \odot X$ に逆 STFT などの時間領域変換処理 Q^+ を適用し、時間領域の強調音声 y を得て出力する（式（１））（ステップ S126）。

[0031] <本実施形態の特徴>

以上のように、本実施形態の学習処理では、モデル学習装置 11 が、観測信号 x から話者認識用の特徴量 Ψ および汎化マスク推定用の特徴量 Φ を抽出し、話者認識用の特徴量 Ψ と汎化マスク推定用の特徴量 Φ とを組み合わせた特徴量から T-F マスクを推定し、話者認識用の特徴量 Ψ から推定話者を識別する情報を得るモデル $M_1(x; \theta_1)$, $M_2(\Phi, \Psi; \theta_2)$, $Z_D(x; \theta_z)$ を学習する。この学習は、T-F マスクを観測信号 x に適用して得られるマスク後音声信号に対応する音声強調信号 y と観測信号 x に含まれた目的音声信号 s との距離に対応する第 1 関数 $(-\text{clip}_\beta[\text{SDR}(s, y)]/2)$ と、観測信号 x に含まれた雑音信号 n と観測信号 x から音声強調信号 y を除いた残存信号 m との距離に対応する第 2 関数 $(-\text{clip}_\beta[\text{SDR}(n, m)]/2)$ と、推定話者を識別する情報 $z^{\wedge}$ と目的音声信号を発した話者を識別する情報 z との距離に対応する第 3 関数 $(\alpha \text{CrossEntropy}(z, z^{\wedge}))$ とを加算したコスト関数 L を最小化するように行われる。また、本実施形態の音声強調処理では、音声強調装置 12

が、観測信号 x から抽出された話者認識用の特徴量 Ψ と、観測信号 x から抽出された汎化マスク推定用の特徴量 Φ と、を組み合わせた特徴量から T-F マスク $M(x; \theta)$ の推定を行い、この T-F マスク $M(x; \theta)$ を当該観測信号 x に適用してマスク後音声信号 $M(x; \theta) \odot X$ を取得する。以上のように、T-F マスク $M(x; \theta)$ は、観測信号 x から抽出された話者認識用の特徴量 Ψ と、観測信号 x から抽出された汎化マスク推定用の特徴量 Φ とに基づくため、観測信号 x の話者に最適化されたものとなる。また、音声強調処理での T-F マスク $M(x; \theta)$ の推定のために、目的話者の補助発話を必要としない。そのため、本実施形態では、音声強調しようとする目的話者の補助発話を必要とすることなく、目的話者に特化した音声強調を行うことができる。

[0032] <学習と強調の実施結果例>

本実施形態の有効性を検証するために、音声強調の公開データセット（非特許文献 1）を用いて実験を行った。評価指標には、このデータセットの標準指標である、perceptual evaluation of speech quality (PESQ) と CSIG、CBAK、COVL を利用した。比較手法には、SEGAN（非特許文献 2）、MMSE-GAN（非特許文献 3）、DFL（非特許文献 4）、MetricGAN（非特許文献 5）を利用した。これらの手法は、話者情報を利用しておらず、大量の話者が発話した大量の音声データを利用して一つの DNN を学習し、話者非依存モデルを学習する方法である。また、音声強調処理が行われない場合の精度評価を Noisy として示した。表 1 に実験結果を示す。全ての指標で、本実施形態のスコアが上回っており、話者認識のマルチタスク学習を利用した音声強調の有効性が示された。

[表1]

Method	PESQ	CSIG	CBAK	COVL
Noisy	1.97	3.35	2.44	2.63
SEGAN	2.16	3.48	2.94	2.80
MMSE-GAN	2.53	3.80	3.12	3.14
DFL	n/a	3.86	3.33	3.22
MetricGAN	2.86	3.99	3.18	3.42
本実施形態	2.96	4.13	3.44	3.54

[0033] [ハードウェア構成]

各実施形態における学習装置 1 1 および音声強調装置 1 2 は、例えば、C P U (central processing unit) 等のプロセッサ (ハードウェア・プロセッサ) や R A M (random-access memory) ・ R O M (read-only memory) 等のメモリ等を備える汎用または専用のコンピュータが所定のプログラムを実行することで構成される装置である。このコンピュータは 1 個のプロセッサやメモリを備えていてもよいし、複数個のプロセッサやメモリを備えていてもよい。このプログラムはコンピュータにインストールされてもよいし、予め R O M 等に記録されていてもよい。また、C P U のようにプログラムが読み込まれることで機能構成を実現する電子回路 (circuitry) ではなく、単独で処理機能を実現する電子回路を用いて一部またはすべての処理部が構成されてもよい。また、1 個の装置を構成する電子回路が複数の C P U を含んでいてもよい。

[0034] 図 5 は、各実施形態における学習装置 1 1 および音声強調装置 1 2 のハードウェア構成を例示したブロック図である。図 5 に例示するように、この例の秘密計算装置 1, 2, 3 は、C P U (Central Processing Unit) 1 0 a、出力部 1 0 b、出力部 1 0 c、R A M (Random Access Memory) 1 0 d、R O M (Read Only Memory) 1 0 e、補助記憶装置 1 0 f 及びバス 1 0 g を有している。この例の C P U 1 0 a は、制御部 1 0 a a、演算部 1 0 a b 及びレジスタ 1 0 a c を有し、レジスタ 1 0 a c に読み込まれた各種プログラムに従って様々な演算処理を実行する。また、出力部 1 0 b は、データが出力される出力端子、ディスプレイ等である。また、出力部 1 0 c は、所定のプログラムを読み込んだ C P U 1 0 a によって制御される L A N カード等である。また、R A M 1 0 d は、S R A M (Static Random Access Memory)、D R A M (Dynamic Random Access Memory) 等であり、所定のプログラムが格納されるプログラム領域 1 0 d a 及び各種データが格納されるデータ領域 1 0 d b を有している。また、補助記憶装置 1 0 f は、例えば、ハードディスク、M O (Magneto-Optical disc)、半導体メモリ等であり、所定のプログラ

ムが格納されるプログラム領域 10 f a 及び各種データが格納されるデータ領域 10 f b を有している。また、バス 10 g は、CPU 10 a、出力部 10 b、出力部 10 c、RAM 10 d、ROM 10 e 及び補助記憶装置 10 f を、情報のやり取りが可能なように接続する。CPU 10 a は、読み込まれた OS (Operating System) プログラムに従い、補助記憶装置 10 f のプログラム領域 10 f a に格納されているプログラムを RAM 10 d のプログラム領域 10 d a に書き込む。同様に CPU 10 a は、補助記憶装置 10 f のデータ領域 10 f b に格納されている各種データを、RAM 10 d のデータ領域 10 d b に書き込む。そして、このプログラムやデータが書き込まれた RAM 10 d 上のアドレスが CPU 10 a のレジスタ 10 a c に格納される。CPU 10 a の制御部 10 a b は、レジスタ 10 a c に格納されたこれらのアドレスを順次読み出し、読み出したアドレスが示す RAM 10 d 上の領域からプログラムやデータを読み出し、そのプログラムが示す演算を演算部 10 a b に順次実行させ、その演算結果をレジスタ 10 a c に格納していく。このような構成により、学習装置 11 および音声強調装置 12 の機能構成が実現される。

[0035] 上述のプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体の例は非一時的な (non-transitory) 記録媒体である。このような記録媒体の例は、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等である。

[0036] このプログラムの流通は、例えば、そのプログラムを記録した DVD、CD-ROM 等の可搬型記録媒体を販売、譲渡、貸与等することによって行う。さらに、このプログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することにより、このプログラムを流通させる構成としてもよい。上述のように、このようなプログラムを実行するコンピュータは、例えば、まず、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、一旦、自己の記憶装置に格納する。

そして、処理の実行時、このコンピュータは、自己の記憶装置に格納されたプログラムを読み取り、読み取ったプログラムに従った処理を実行する。また、このプログラムの別の実行形態として、コンピュータが可搬型記録媒体から直接プログラムを読み取り、そのプログラムに従った処理を実行することとしてもよく、さらに、このコンピュータにサーバコンピュータからプログラムが転送されるたびに、逐次、受け取ったプログラムに従った処理を実行することとしてもよい。また、サーバコンピュータから、このコンピュータへのプログラムの転送は行わず、その実行指示と結果取得のみによって処理機能を実現する、いわゆるASP（Application Service Provider）型のサービスによって、上述の処理を実行する構成としてもよい。なお、本形態におけるプログラムには、電子計算機による処理の用に供する情報であってプログラムに準ずるもの（コンピュータに対する直接の指令ではないがコンピュータの処理を規定する性質を有するデータ等）を含むものとする。

[0037] 各実施形態では、コンピュータ上で所定のプログラムを実行させることにより、本装置を構成することとしたが、これらの処理内容の少なくとも一部をハードウェア的に実現することとしてもよい。

[0038] [その他の変形例]

なお、本発明は上述の実施形態に限定されるものではない。例えば、上述の実施形態では、音声強調装置12に時間領域の観測信号 x が入力され、周波数領域変換部122が観測信号 x を時間周波数領域表現した観測信号 $X = Q(x)$ に変換した。しかしながら、音声強調装置12に観測信号 x および観測信号 X が入力されてもよい。この場合、音声強調装置12から周波数領域変換部122が省略されてもよい。

[0039] 上述の実施形態では、音声強調装置12が、時間周波数領域のマスク後音声信号 $M(x; \theta) \odot X$ に時間領域変換処理 Q^+ を適用し、時間領域の強調音声 y を得て出力した。しかしながら、音声強調装置12がマスク後音声信号 $M(x; \theta) \odot X$ をそのまま出力してもよい。この場合、マスク後音声信号 $M(x; \theta) \odot X$ が他の処理の入力として使用されてもよい。この場合、音

声強調装置 1 2 から時間領域変換部 1 2 5 が省略されてもよい。

[0040] 上述の実施形態では、モデル M_1 , M_2 , Z_D としてDNNが用いられたが、モデル M_1 , M_2 , Z_D として確率モデルなどその他のモデルが用いられてもよい。モデル M_1 , M_2 , Z_D が1個または2個のモデルとして構成されてもよい。

[0041] 上述の実施形態では、所望の話者から発せられた音声を強調した。しかしながら、所望の音源から発せられた音声を強調する音声強調処理であってもよい。この場合、上述した「話者」を「音源」に置き換えた処理を実行すればよい。

[0042] また、上述の各種の処理は、記載に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されてもよい。その他、本発明の趣旨を逸脱しない範囲で適宜変更が可能であることはいうまでもない。

符号の説明

[0043] 1 1 学習装置

1 2 音声強調装置

請求の範囲

- [請求項1] 所望の話者から発せられた音声を強調する音声強調方法であって、
 前記話者から発せられた音声を強調するマスクを観測信号から推定するマスク推定ステップと、
 前記観測信号に前記マスクを適用し、マスク後音声信号を取得するマスク適用ステップと、を有し、
 前記マスク推定ステップは、
 前記観測信号から抽出された話者認識用の特徴量と、前記観測信号から抽出された汎化マスク推定用の特徴量と、を組み合わせた特徴量から前記マスクの推定を行う音声強調方法。
- [請求項2] 所望の音源から発せられた音声を強調する音声強調方法であって、
 前記音源から発せられた音声を強調するマスクを観測信号から推定するマスク推定ステップと、
 前記観測信号に前記マスクを適用し、マスク後音声信号を取得するマスク適用ステップと、を有し、
 前記マスク推定ステップは、
 前記観測信号から抽出された音源認識用の特徴量と、前記観測信号から抽出された汎化マスク推定用の特徴量と、を組み合わせた特徴量から前記マスクの推定を行う音声強調方法。
- [請求項3] 観測信号から話者認識用の特徴量および汎化マスク推定用の特徴量を抽出し、前記話者認識用の特徴量と前記汎化マスク推定用の特徴量とを組み合わせた特徴量からマスクを推定し、前記話者認識用の特徴量から推定話者を識別する情報を得るモデルを学習する学習ステップを有し、
 前記学習ステップは、前記マスクを前記観測信号に適用して得られるマスク後音声信号に対応する音声強調信号と前記観測信号に含まれた目的音声信号との距離に対応する第1関数と、前記観測信号に含まれた雑音信号と前記観測信号から前記音声強調信号を除いた残存信号

との距離に対応する第2関数と、前記推定話者を識別する情報と前記目的音声信号を発した話者を識別する情報との距離に対応する第3関数とを加算したコスト関数を最小化するように前記モデルを学習し、前記第1関数の関数値が小さいほど前記コスト関数の関数値は小さく、前記第2関数の関数値が小さいほど前記コスト関数の関数値は小さく、前記第3関数の関数値が小さいほど前記コスト関数の関数値は小さい、学習方法。

[請求項4]

観測信号から音源認識用の特徴量および汎化マスク推定用の特徴量を抽出し、前記音源認識用の特徴量と前記汎化マスク推定用の特徴量とを組み合わせた特徴量からマスクを推定し、前記音源認識用の特徴量から推定音源を識別する情報を得るモデルを学習する学習ステップを有し、

前記学習ステップは、前記マスクを前記観測信号に適用して得られるマスク後音声信号に対応する音声強調信号と前記観測信号に含まれた目的音声信号との距離に対応する第1関数と、前記観測信号に含まれた雑音信号と前記観測信号から前記音声強調信号を除いた残存信号との距離に対応する第2関数と、前記推定音源を識別する情報と前記目的音声信号を発した音源を識別する情報との距離に対応する第3関数とを加算したコスト関数を最小化するように前記モデルを学習し、前記第1関数の関数値が小さいほど前記コスト関数の関数値は小さく、前記第2関数の関数値が小さいほど前記コスト関数の関数値は小さく、前記第3関数の関数値が小さいほど前記コスト関数の関数値は小さい、学習方法。

[請求項5]

所望の話者から発せられた音声を強調する音声強調装置であって、前記話者から発せられた音声を強調するマスクを観測信号から推定するマスク推定部と、

前記観測信号に前記マスクを適用し、マスク後音声信号を取得するマスク部とを有し、

前記マスク推定部は、

前記観測信号から抽出された話者認識用の特徴量と、前記観測信号から抽出された汎化マスク推定用の特徴量と、を組み合わせた特徴量から前記マスクの推定を行う音声強調装置。

[請求項6]

観測信号から話者認識用の特徴量および汎化マスク推定用の特徴量を抽出し、前記話者認識用の特徴量と前記汎化マスク推定用の特徴量とを組み合わせた特徴量からマスクを推定し、前記話者認識用の特徴量から推定話者を識別する情報を得るモデルを学習する学習部を有し、

前記学習部は、前記マスクを前記観測信号に適用して得られるマスク後音声信号に対応する音声強調信号と前記観測信号に含まれた目的音声信号との距離に対応する第1関数と、前記観測信号に含まれた雑音信号と前記観測信号から前記音声強調信号を除いた残存信号との距離に対応する第2関数と、前記推定話者を識別する情報と前記目的音声信号を発した話者を識別する情報との距離に対応する第3関数とを加算したコスト関数を最小化するように前記モデルを学習し、前記第1関数の関数値が小さいほど前記コスト関数の関数値は小さく、前記第2関数の関数値が小さいほど前記コスト関数の関数値は小さく、前記第3関数の関数値が小さいほど前記コスト関数の関数値は小さい、学習装置。

[請求項7]

請求項1または2の音声強調方法をコンピュータに実行させるためのプログラム。

[請求項8]

請求項3または4の学習方法をコンピュータに実行させるためのプログラム。

[図1]

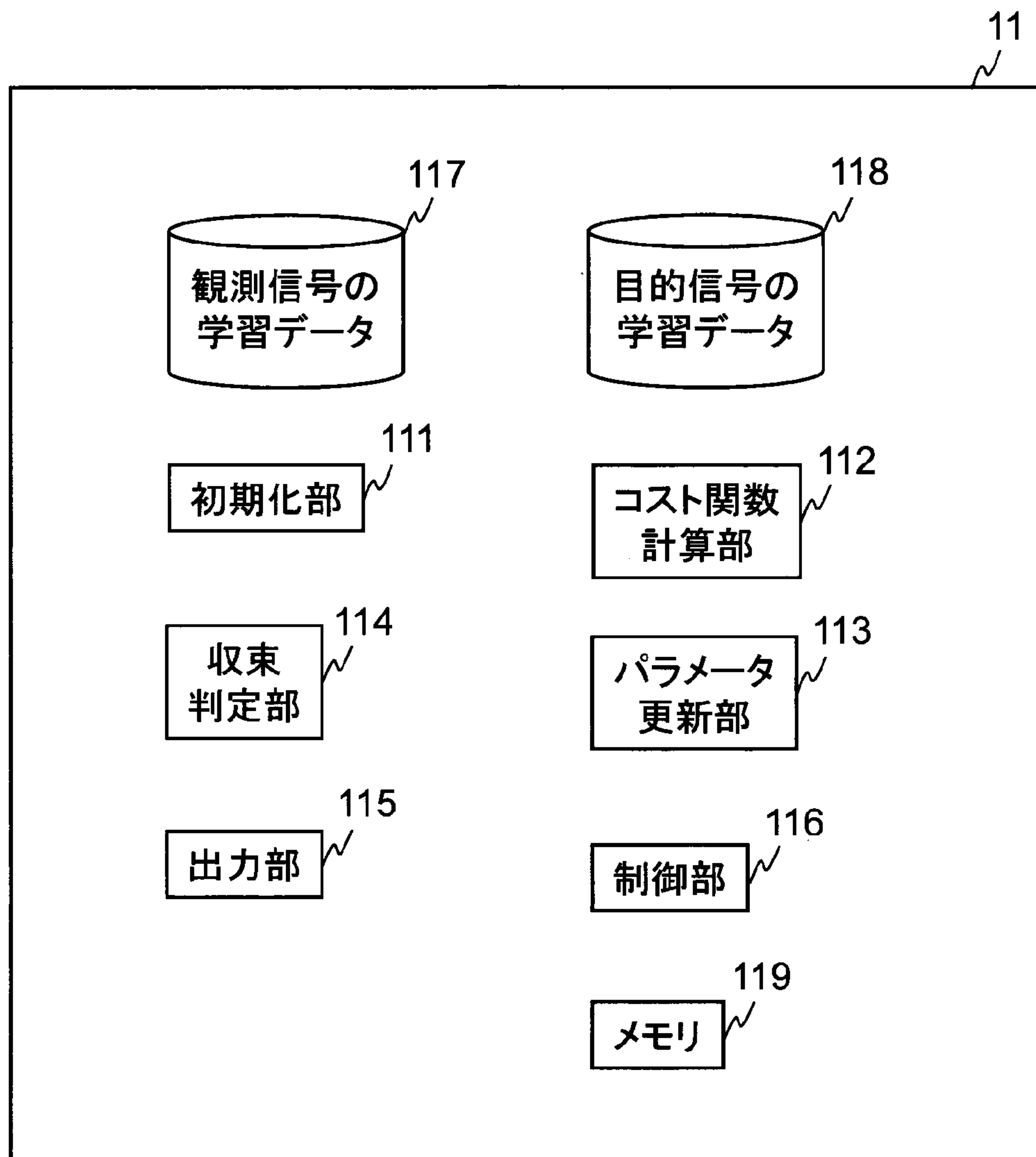


図 1

[図2]

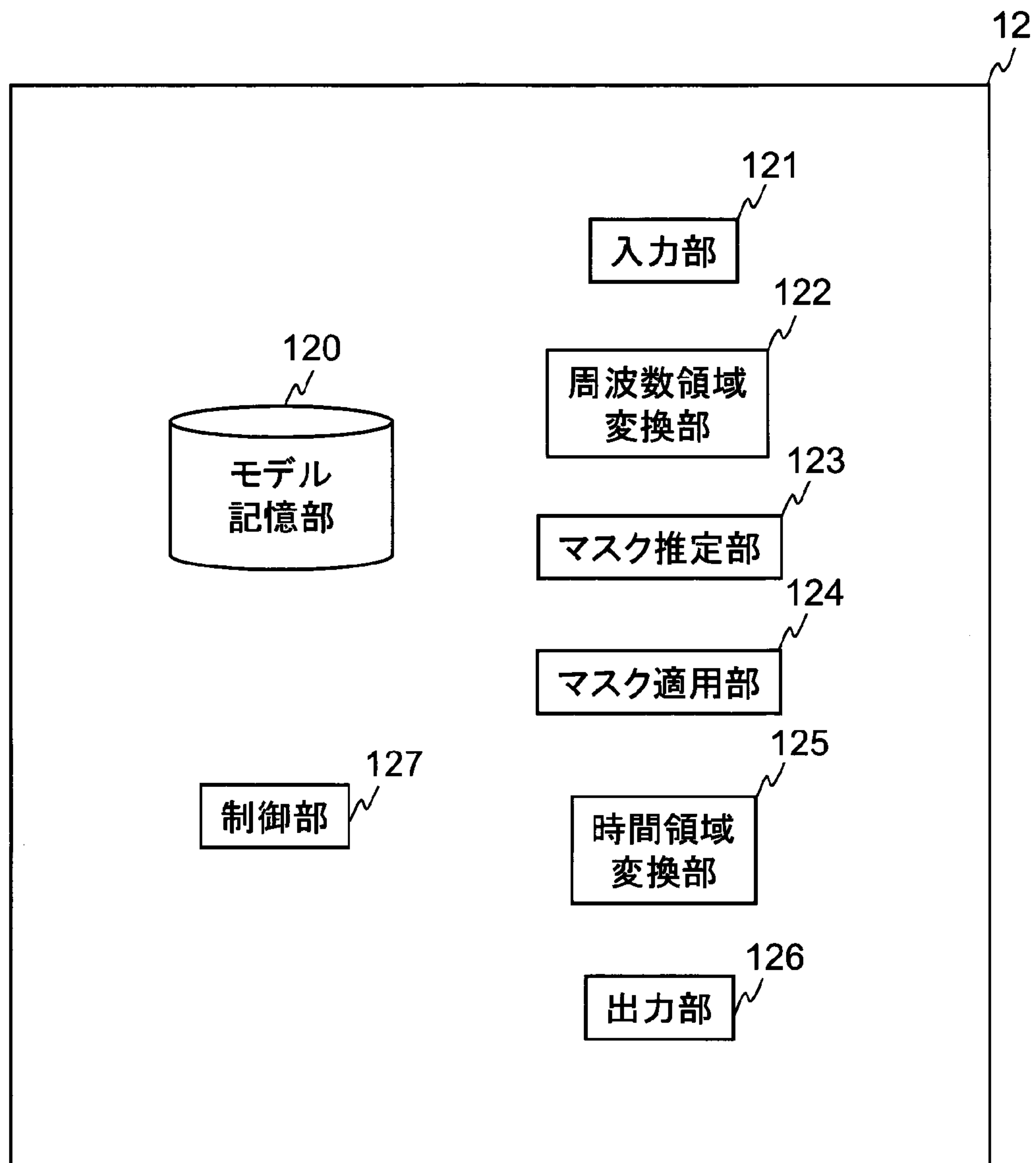


図2

[図3]

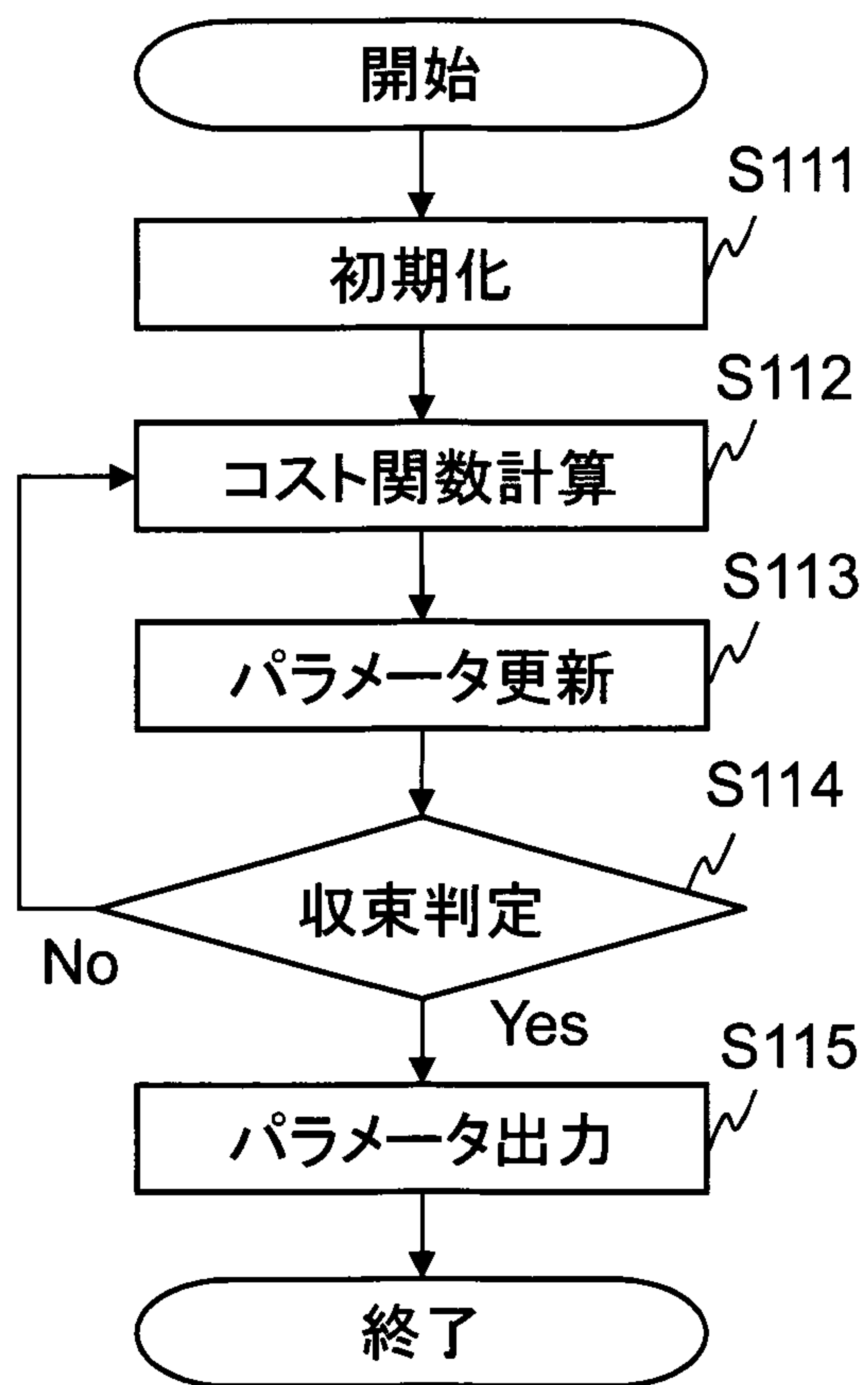


図3

[図4]

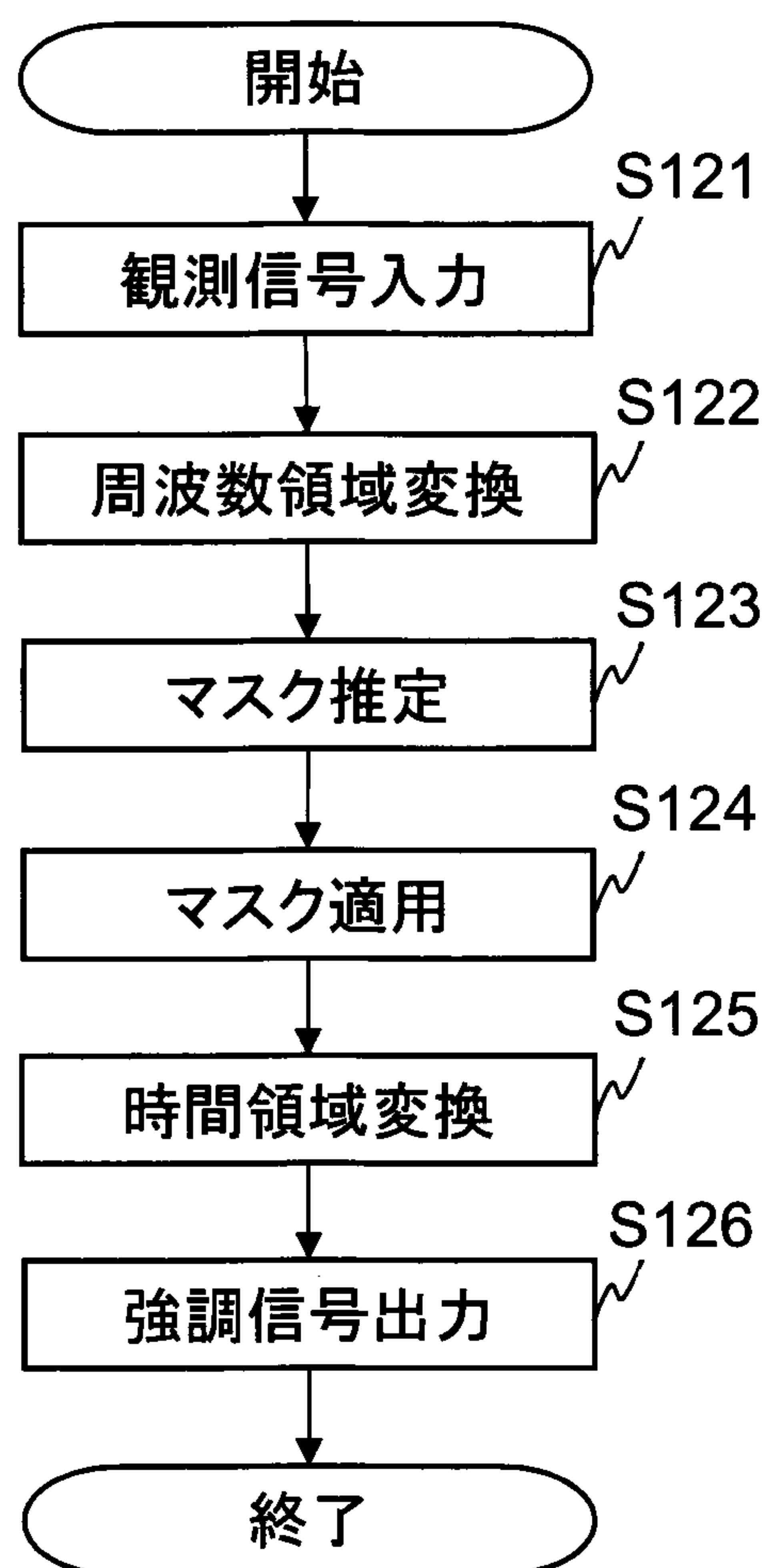


図4

[図5]

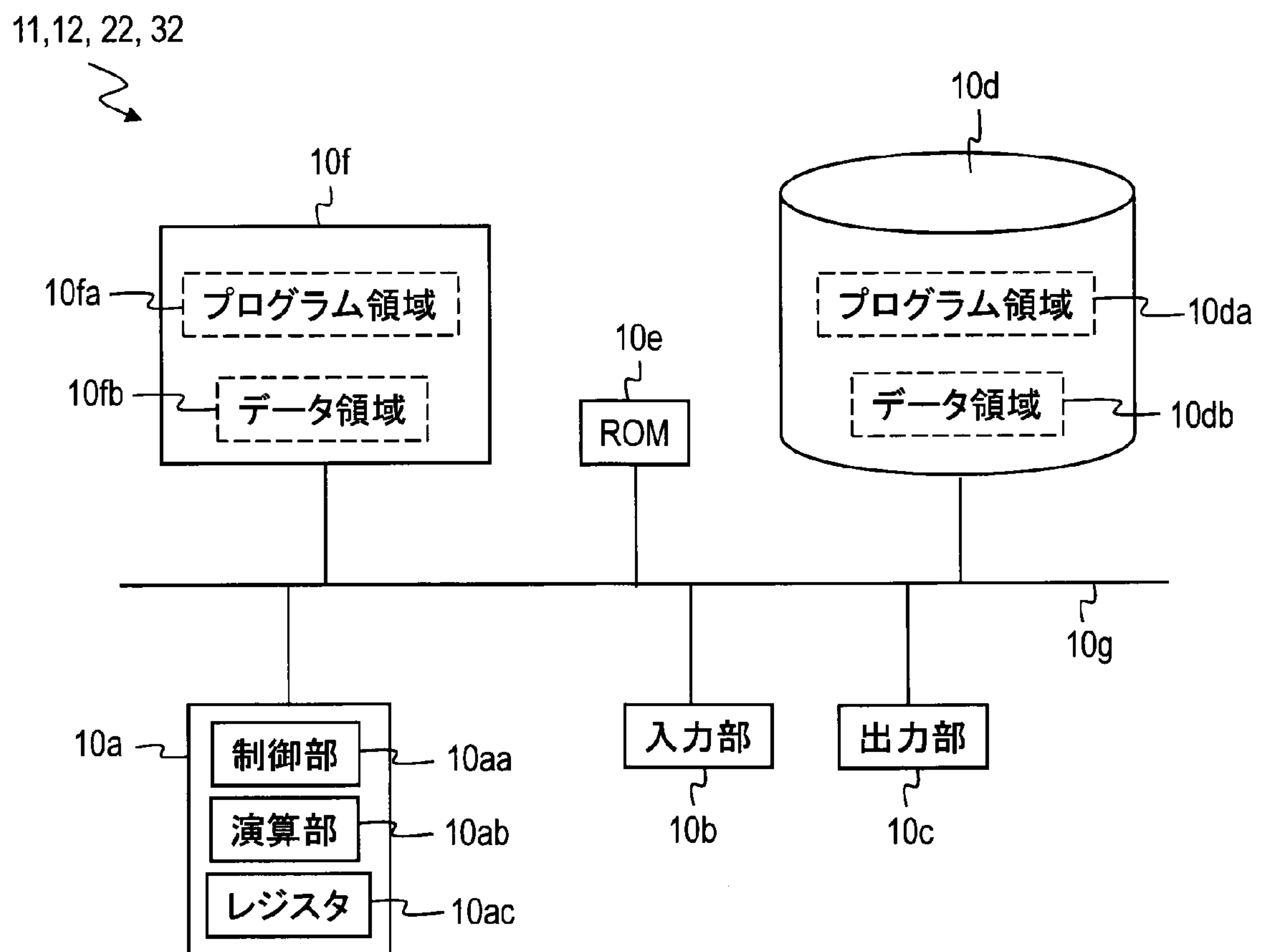


図5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/001356

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G10L21/0232(2013.01) i, G10L21/0264(2013.01) i
FI: G10L21/0232, G10L21/0264 Z
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. G10L21/0232, G10L21/0264
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	ZMOLIKOVA, K. et al., SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures, IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 4, IEEE, August 2019, pp. 800-814, pp. 802-804	1-8
A	WANG, Q. et al., VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking, Proc. INTERSPEECH 2019, ISCA, September 2019, pp. 2728-2732, pp. 2729-2730	1-5
A	XIAO, X. et al., Single-channel Speech Extraction Using Speaker Inventory and Attention Network, Proc. ICASSP 2019, IEEE, May 2019, pp. 86-90, pp. 87-88	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	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“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
“O” document referring to an oral disclosure, use, exhibition or other means	“&” document member of the same patent family
“P” document published prior to the international filing date but later than the priority date claimed	
Date of the actual completion of the international search	Date of mailing of the international search report
13.07.2020	21.07.2020
Name and mailing address of the ISA/	Authorized officer
Japan Patent Office	
3-4-3, Kasumigaseki, Chiyoda-ku,	
Tokyo 100-8915, Japan	Telephone No.

国際調査報告		国際出願番号 PCT/JP2020/001356
A. 発明の属する分野の分類（国際特許分類（I P C）） G10L 21/0232(2013.01)i; G10L 21/0264(2013.01)i FI: G10L21/0232; G10L21/0264 Z		
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（I P C）） G10L21/0232; G10L21/0264		
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 1 9 2 2 - 1 9 9 6 年 日本国公開実用新案公報 1 9 7 1 - 2 0 2 0 年 日本国実用新案登録公報 1 9 9 6 - 2 0 2 0 年 日本国登録実用新案公報 1 9 9 4 - 2 0 2 0 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	ZMOLIKOVA, K., et al., SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures, IEEE Journal of Selected Topics in Signal Processing, Vol. 13, No. 4, IEEE, 2019.08, pp. 800-814 pp. 802-804	1-8
A	WANG, Q., et al., VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking, Proc. INTERSPEECH 2019, ISCA, 2019.09, pp. 2728-2732 pp. 2729-2730	1-8
A	XIAO, X., et al., Single-channel Speech Extraction Using Speaker Inventory and Attention Network, Proc. ICASSP 2019, IEEE, 2019.05, pp. 86-90 pp. 87-88	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 13. 07. 2020	国際調査報告の発送日 21. 07. 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） 上田 雄 5Z 5095 電話番号 03-3581-1101 内線 3591	