

(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
2022年6月16日(16.06.2022)



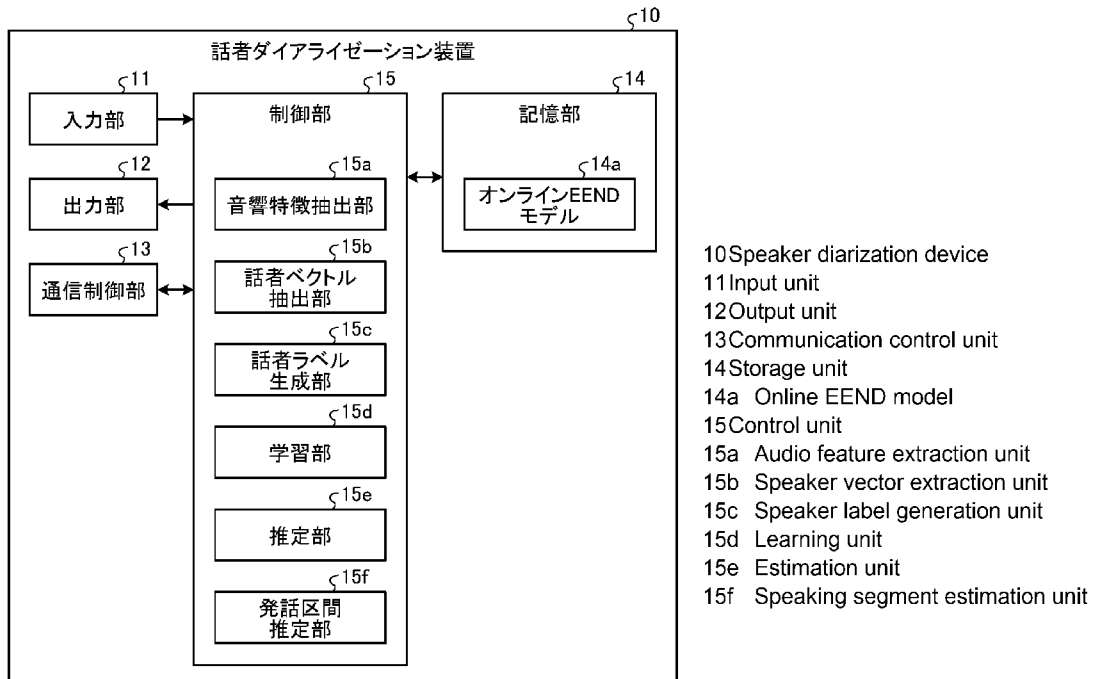
(10) 国際公開番号

WO 2022/123742 A1

- (51) 国際特許分類:
G10L 17/00 (2013.01) G10L 17/18 (2013.01)
G10L 17/02 (2013.01) G10L 15/04 (2013.01)
G10L 17/04 (2013.01)
- (21) 国際出願番号: PCT/JP2020/046117
- (22) 国際出願日: 2020年12月10日(10.12.2020)
- (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: 安藤 厚志 (ANDO, Atsushi); 〒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). 村田 有実子 (MURATA, Yumiko); 〒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). 森 岳至 (MORI, Takeshi); 〒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: 特許業務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 虎の門三井ビルディング Tokyo (JP).

(54) Title: SPEAKER DIARIZATION METHOD, SPEAKER DIARIZATION DEVICE, AND SPEAKER DIARIZATION PROGRAM

(54) 発明の名称: 話者ダイアライゼーション方法、話者ダイアライゼーション装置および話者ダイアライゼーションプログラム



(57) Abstract: A speaker vector extraction unit (15b) uses a time series of audio features for each frame of a most recent audio signal to extract speaker vectors representing a speaker feature for each frame. A learning unit (15d) uses a speaker vector and a speaker label representing an estimated speaker for the speaker vector to generate, through learning, an online EEND model (14a) that estimates a speaker label for the speaker vector for each frame.

(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, IT, 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, 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: 話者ベクトル抽出部(15b)が、直近の音響信号のフレームごとの音響特徴の系列を用いて、各フレームの話者特徴を表す話者ベクトルを抽出する。学習部(15d)が、話者ベクトルと推定された該話者ベクトルの話者を表す話者ラベルとを用いて、各フレームの話者ベクトルの話者ラベルを推定するオンラインEENDモデル(14a)を学習により生成する。

明 細 書

発明の名称：

話者ダイアライゼーション方法、話者ダイアライゼーション装置および話者ダイアライゼーションプログラム

技術分野

[0001] 本発明は、話者ダイアライゼーション方法、話者ダイアライゼーション装置および話者ダイアライゼーションプログラムに関する。

背景技術

[0002] 近年、音響信号を入力とし、音響信号に含まれる全ての話者の発話区間を同定する話者ダイアライゼーション技術が期待されている。話者ダイアライゼーション技術によれば、例えば、会議において誰がいつ発言したかを記録する自動書き起こしや、コンタクトセンタにおいて通話からオペレータと顧客との発話の自動切り出し等、様々な応用が可能となる。

[0003] 従来、話者ダイアライゼーション技術として、深層学習に基づく E E N D (End-to-End Neural Diarization) と呼ばれる技術が開示されている（非特許文献 1 参照）。E E N D では、音響信号をフレームごとに分割し、各フレームから抽出した音響特徴から、当該フレームにおいて特定の話者が存在するか否かを表す話者ラベルをフレームごとに推定する。音響信号内の最大話者数 S である場合に、フレームごとの話者ラベルは S 次元のベクトルであり、当該フレームにおいて、ある話者が発話している場合に 1、発話していない場合に 0 となる。すなわち、E E N D では、話者数のマルチラベル二値分類を行うことにより、話者ダイアライゼーションを実現している。

[0004] E E N D でフレームごとの話者ラベル系列の推定に用いられる E E N D モデルは、誤差逆伝搬可能な層で構成される深層学習に基づくモデルであって、音響特徴系列からフレームごとの話者ラベル系列を一気通貫で推定できる。E E N D モデルには、時系列モデル化を行う R N N (Recurrent Neural Network) 層が含まれる。これにより、E E N D では当該フレームだけでなく

周囲のフレームの音響特徴量を用いて、フレームごとの話者ラベルを推定することが可能となる。このRNN層には、双方向LSTM (Long Short-Term Memory) -RNNやTransformer Encoderが用いられる。

[0005] なお、非特許文献2には、RNN Transducerについて記載されている。また、非特許文献3には、音響特徴量について記載されている。

先行技術文献

非特許文献

[0006] 非特許文献1: Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, Shinji Watanabe, “END-TO-END NEURAL SPEAKER DIARIZATION WITH SELF-ATTENTION”, Proc. ASRU, 2019年, pp. 296-303

非特許文献2: Alex Graves, “Sequence Transduction with Recurrent Neural Networks”, Proc. ICML, 2012

非特許文献3: 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, 山本 幹雄, “音声認識システム”, オーム社, 2001年, pp.13-14

発明の概要

発明が解決しようとする課題

[0007] しかしながら、従来技術では、オンラインでの話者ダイアライゼーションが困難であった。つまり、従来のEENDモデルは、音響特徴系列の全体を参照する双方向LSTM-RNNやTransformerを用いるため、オンラインで話者ダイアライゼーションを実現することが困難であった。

[0008] 本発明は、上記に鑑みてなされたものであって、オンラインでの話者ダイアライゼーションを行うことを目的とする。

課題を解決するための手段

[0009] 上述した課題を解決し、目的を達成するために、本発明に係る話者ダイアライゼーション方法は、直近の音響信号のフレームごとの音響特徴の系列を

用いて、各フレームの話者特徴を表す話者ベクトルを抽出する抽出工程と、前記話者ベクトルと推定された該話者ベクトルの話者を表す話者ラベルとを用いて、各フレームの話者ベクトルの話者ラベルを推定するモデルを学習により生成する学習工程と、を含んだことを特徴とする。

発明の効果

[0010] 本発明によれば、オンラインでの話者ダイアライゼーションが可能となる。

図面の簡単な説明

[0011] [図1]図1は、話者ダイアライゼーション装置の概要を説明するための図である。

[図2]図2は、話者ダイアライゼーション装置の概略構成を例示する模式図である。

[図3]図3は、話者ダイアライゼーション装置の処理を説明するための図である。

[図4]図4は、話者ダイアライゼーション処理手順を示すフローチャートである。

[図5]図5は、話者ダイアライゼーション処理手順を示すフローチャートである。

[図6]図6は、話者ダイアライゼーションプログラムを実行するコンピュータを例示する図である。

発明を実施するための形態

[0012] 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。

[0013] [話者ダイアライゼーション装置の概要]

図1は、話者ダイアライゼーション装置の概要を説明するための図である。図1に示すように、本実施形態の話者ダイアライゼーション装置のE E N Dモデル（オンラインE E N Dモデル）は、直近の音響信号のフレームごと

の音響特徴の系列を入力として、最新のフレームの話者の特徴を表す話者ベクトルを出力するオンラインE E N Dモデル14 aを構築する。具体的には、オンラインE E N Dモデル14 aは、現在のtフレーム目から連続して遡った(t - N)フレーム目までの各フレームの音響特徴を用いて、tフレーム目の話者ラベルを推定する。

[0014] このオンラインE E N Dモデル14 aは、話者特徴抽出ブロックと、話者特徴更新ブロックと、話者ラベル推定ブロックとを有する。ここで、話者特徴抽出ブロックは、(t - N)フレーム目~tフレーム目の各フレームの音響特徴を用いて、tフレーム目の話者の特徴を表す話者ベクトルを抽出する。なお、図1に示す例では、話者特徴抽出ブロックは、L i n e a r (全結合)層とR N N層とを含むが、これに限定されず、例えばR N N層の代わりに入力されたベクトルを平均化する層が含まれてもよい。

[0015] 話者特徴更新ブロックは、tフレーム目の話者ベクトルと、この話者ベクトルに対して後述する話者ラベル推定ブロックが推定した話者ラベルの推定値とをベクトル結合して記憶する。また、話者特徴更新ブロックは、記憶した話者ベクトルと話者ラベルの推定値とをベクトル結合したベクトルの入力に対し、話者を識別する情報を含む話者ベクトルを記憶話者ベクトルとして出力するモデルのパラメータを更新する。図1に示す例では、モデルはL i n e a r (全結合)層とR N N層とを含む。

[0016] 話者ラベル推定ブロックは、話者ベクトルと記憶話者ベクトルとを用いて、tフレーム目の話者ラベルの推定値を出力する。図1に示す例では、話者ラベル推定ブロックは、L i n e a r (全結合)層とs i g m o i d層とを含む。話者ダイアライゼーション装置は、例えば、出力された話者ラベルの推定値を閾値判定することにより、話者ラベルを推定する。

[0017] このように、話者ダイアライゼーション装置は、自己回帰構造をもつオンラインE E N Dモデル14 aを用いて、1フレームずつ話者ラベルを推定する。これにより、話者ダイアライゼーション装置は、フレームが入力されるたびに記憶話者ベクトルを更新しながら話者ラベルを推定することが可能と

なる。したがって、オンラインでの話者ダイアライゼーションを実現することが可能となる。

[0018] [話者ダイアライゼーション装置の構成]

図2は、話者ダイアライゼーション装置の概略構成を例示する模式図である。また、図3は、話者ダイアライゼーション装置の処理を説明するための図である。まず、図2に例示するように、本実施形態の話者ダイアライゼーション装置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。

[0019] 入力部11は、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、実施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、情報通信装置等によって実現される。通信制御部13は、NIC (Network Interface Card) 等で実現され、サーバや、音響信号を取得する装置等の外部の装置と制御部15とのネットワークを介した通信を制御する。

[0020] 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。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。本実施形態において、記憶部14には、例えば、後述する話者ダイアライゼーション処理に用いられるオンラインEENDモデル14a等が記憶される。

[0021] 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2に例示するように、音響特徴抽出部15a、話者ベクトル抽出部15b、話者ラベル生成部15c、学習部15d、推定部15eおよび発話区間推定部15fとして機能する。なお、これらの機能部は、それぞれが異なるハードウェアに実装されてもよい。例えば、学習部15dは学習装

置として実装され、推定部 15 e は、推定装置として実装されてもよい。また、制御部 15 は、その他の機能部を備えてもよい。

[0022] 音響特徴抽出部 15 a は、話者の発話を含む音響信号のフレームごとの音響特徴を抽出する。例えば、音響特徴抽出部 15 a は、入力部 11 を介して、あるいは音響信号を取得する装置等から通信制御部 13 を介して、音響信号の入力を受け付ける。また、音響特徴抽出部 15 a は、音響信号をフレームごとに分割し、各フレームからの信号に対して離散フーリエ変換やフィルタバンク乗算を行うことにより音響特徴ベクトルを抽出し、フレーム方向に結合した音響特徴系列を出力する。本実施形態では、フレーム長は 25 ms、フレームシフト幅は 10 ms とする。

[0023] ここで、音響特徴ベクトルは、例えば、24次元のMFCC (Mel Frequency Cepstral Coefficient) であるが、これに限定されず、例えば、メルフィルタバンク出力等の他のフレームごとの音響特徴量でもよい。

[0024] 話者ベクトル抽出部 15 b は、直近の音響信号のフレームごとの音響特徴系列を用いて、各フレームの話者特徴を表す話者ベクトルを抽出する。具体的には、話者ベクトル抽出部 15 b は、音響特徴抽出部 15 a から取得した音響特徴系列を、図 1 に示した話者特徴抽出ブロックに入力することにより、話者ベクトルを生成する。

[0025] なお、話者ベクトル抽出部 15 b は、後述する学習部 15 d および推定部 15 e に内包されてもよい。例えば、後述する図 3 では、学習部 15 d および推定部 15 e が話者ベクトル抽出部 15 b の処理を行う例が示されている。

[0026] 話者ラベル生成部 15 c は、音響特徴系列を用いて、各フレームの話者ラベルを生成する。具体的には、話者ラベル生成部 15 c は、図 3 に示すように、音響特徴系列と話者の発話区間の正解ラベルとを用いて、フレームごとの話者ラベルを生成する。これにより、後述する学習部 15 d の処理に用いられる教師データとして、音響特徴系列とフレームごとの話者ラベルとの組が生成される。

[0027] ここで、話者数が S である（話者1、話者2、…、話者 S ）場合に、 t フレーム目（ $t = 0, 1, \dots, T$ ）の話者ラベルは S 次元のベクトルとなる。例えば、時刻 $t \times$ フレームシフト幅のフレームがいずれかの話者の発話区間に含まれる場合には、当該話者に対応する次元の値が1、それ以外の次元の値が0となる。したがって、フレームごとの話者ラベルは、 $T \times S$ 次元の二値 $[0, 1]$ のマルチラベルとなる。

[0028] 図2の説明に戻る。学習部15dは、話者ベクトルと推定された該話者ベクトルの話者を表す話者ラベルとを用いて、各フレームの話者ベクトルの話者ラベルを推定するオンラインEENDモデル14aを学習により生成する。具体的には、学習部15dは、図3に示すように、音響特徴系列とフレームごとの話者ラベルとの組を教師データとして用いて、オンラインEENDモデル14aの学習を行う。

[0029] ここで、オンラインEENDモデル14aは、図1に示したようにRNN層を含む複数の層で構成されている。本実施形態において、RNN層としては単方向LSTM-RNNが適用される。また、オンラインEENDモデル14aには、 $N = 10$ として、 t フレーム目から $(t - N)$ フレーム目までの各フレームの音響特徴ベクトルを統合したスーパーベクトルが入力されるものとする。ただし、 $t - N$ が負の値である場合には、音響特徴ベクトルはゼロベクトルとする。

[0030] また、オンラインEENDモデル14aは、 $T \times S$ 次元のフレームごとの話者ラベルの事後確率を出力する。学習部15dは、フレームごとの話者ラベルの事後確率と、フレームごとの話者ラベルとのマルチラベル二値交差エントロピーを損失関数として、誤差逆伝搬法により、オンラインEENDモデル14aの各層のパラメータの最適化を行う。学習部15dは、パラメータの最適化には、確率的勾配降下法を用いたオンライン最適化アルゴリズムを用いる。

[0031] すなわち、学習部15dは、話者特徴抽出ブロックである話者ベクトル抽出部15bが、教師データの $(t - N)$ フレーム目～ t フレーム目の各フレ

ームの音響特徴を用いて抽出した t フレーム目の話者ベクトルと、この話者ベクトルに対して話者ラベル推定ブロックが推定した話者ラベルの推定値とをベクトル結合して記憶する。また、学習部 15 d は、記憶した話者ベクトルと話者ラベルの推定値とをベクトル結合したベクトルを話者特徴更新ブロックに入力し、話者を識別する情報を含む記憶話者ベクトルを出力するモデルのパラメータを更新する。また、学習部 15 d は、 t フレーム目の話者ベクトルと記憶話者ベクトルとを話者ラベル推定ブロックに入力し、 t フレーム目の話者ラベルの推定値を出力するモデルのパラメータを更新する。

[0032] このように、学習部 15 d は、話者ベクトルと推定された該話者ベクトルの話者ラベルとの記憶された複数の組み合わせを用いて、オンライン E E N D モデル 14 a を生成する。これにより、フレームが入力されるたびに記憶話者ベクトルを更新しながら話者ラベルを推定することが可能となる。

[0033] 図 2 の説明に戻る。推定部 15 e は、生成されたオンライン E E N D モデル 14 a を用いて、音響信号のフレームごとの話者ラベルを推定する。具体的には、推定部 15 e は、図 3 に示すように、話者ベクトル抽出部 15 b が音響特徴系列の現在の t フレーム目から連続して遡った ($t - N$) フレーム目までの各フレームの音響特徴を用いて抽出した t フレーム目の話者ベクトルを、オンライン E E N D モデル 14 a に順伝搬させる。

[0034] オンライン E E N D モデル 14 a は、自己回帰構造を持つことから、音響特徴系列の先頭フレームから逐次順伝搬させることにより、音響特徴系列のフレームごとの話者ラベル事後確率（話者ラベルの推定値）を出力する。

[0035] 発話区間推定部 15 f は、出力された話者ラベル事後確率を用いて、音響信号中の話者の発話区間を推定する。具体的には、発話区間推定部 15 f は、複数のフレームの移動平均を用いて、話者ラベルを推定する。すなわち、発話区間推定部 15 f は、まず、フレームごとの話者ラベル事後確率に対し、自フレームとその直前の 5 フレームとの長さ 6 での移動平均を算出する。これにより、1 フレームしかない発話等、現実的ではない短い発話区間の誤検出を防止することが可能となる。

[0036] 次に、発話区間推定部 15 f は、算出した移動平均の値が 0.5 より大きい場合に、当該フレームが、当該次元の話者の発話区間と推定する。また、発話区間推定部 15 f は、各話者について、連続する発話区間フレーム群を 1 つの発話とみなし、所定の時刻までの発話区間の開始時刻と終了時刻とをフレームから逆算する。これにより、話者ごとの発話ごとの所定の時刻までの発話開始時刻と発話終了時刻とを得ることができる。

[0037] [話者ダイアライゼーション処理]

次に、話者ダイアライゼーション装置 10 による話者ダイアライゼーション処理について説明する。図 4 よび図 5 は、話者ダイアライゼーション処理手順を示すフローチャートである。本実施形態の話者ダイアライゼーション処理は、学習処理と推定処理とを含む。まず、図 4 は、学習処理手順を示す。図 4 のフローチャートは、例えば、学習処理の開始を指示する入力があったタイミングで開始される。

[0038] まず、音響特徴抽出部 15 a が、話者の発話を含む音響信号のフレームごとの音響特徴を抽出し、音響特徴系列を出力する（ステップ S 1）。

[0039] 次に、話者ベクトル抽出部 15 b が、直近の音響信号のフレームごとの音響特徴系列を用いて、各フレームの話者特徴を表す話者ベクトルを抽出する。（ステップ S 2）。

[0040] そして、学習部 15 d が、話者ベクトルと推定された該話者ベクトルの話者を表す話者ラベルとを用いて、自己回帰構造を持ち、各フレームの話者ベクトルの話者ラベルを推定するオンライン E E N D モデル 14 a を、学習により生成する（ステップ S 3）。これにより、一連の学習処理が終了する。

[0041] 次に、図 5 は、推定処理手順を示す。図 5 のフローチャートは、例えば、推定処理の開始を指示する入力があったタイミングで開始される。

[0042] まず、音響特徴抽出部 15 a が、話者の発話を含む音響信号のフレームごとの音響特徴を抽出し、音響特徴系列を出力する（ステップ S 1）。

[0043] また、話者ベクトル抽出部 15 b が、直近の音響信号のフレームごとの音響特徴系列を用いて、各フレームの話者特徴を表す話者ベクトルを抽出する

(ステップS2)。

[0044] 次に、推定部15eが、生成されたオンラインEENDモデル14aを用いて、音響信号のフレームごとの話者ラベルを推定する(ステップS4)。具体的には、推定部15eは、音響特徴系列のフレームごとの話者ラベル事後確率(話者ラベルの推定値)を出力する。

[0045] そして、発話区間推定部15fが、出力された話者ラベル事後確率を用いて、音響信号中の話者の発話区間を推定する(ステップS5)。これにより、一連の推定処理が終了する。

[0046] 以上、説明したように、本実施形態の話者ダイアライゼーション装置10において、話者ベクトル抽出部15bが、直近の音響信号のフレームごとの音響特徴系列を用いて、各フレームの話者特徴を表す話者ベクトルを抽出する。また、学習部15dが、話者ベクトルと推定された該話者ベクトルの話者を表す話者ラベルとを用いて、各フレームの話者ベクトルの話者ラベルを推定するオンラインEENDモデル14aを学習により生成する。

[0047] このように、話者ダイアライゼーション装置10は、自己回帰構造を持つオンラインEENDモデル14aにより、フレームが入力されるたびに話者ラベルを推定することが可能となる。したがって、オンラインでの話者ダイアライゼーションを実現することが可能となる。

[0048] また、学習部15dは、話者ベクトルと推定された該話者ベクトルの話者ラベルとの記憶された複数の組み合わせを用いて、オンラインEENDモデル14aを生成する。これにより、話者ダイアライゼーション装置10は、フレームが入力されるたびに記憶話者ベクトルを更新しながら話者ラベルを推定することが可能となる。したがって、オンラインでの話者ダイアライゼーションがより高精度に実現可能となる。

[0049] また、推定部15eが、生成されたオンラインEENDモデル14aを用いて、音響信号のフレームごとの話者ラベルを推定する。これにより、オンラインでの話者ダイアライゼーションが可能となる。

[0050] また、発話区間推定部15fが、複数のフレームの移動平均を用いて、話

者ラベルを推定する。これにより、現実的ではない短い発話区間の誤検出を防止することが可能となる。

[0051] [プログラム]

上記実施形態に係る話者ダイアライゼーション装置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、話者ダイアライゼーション装置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話者ダイアライゼーション処理を実行する話者ダイアライゼーション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、上記の話者ダイアライゼーションプログラムを情報処理装置に実行させることにより、情報処理装置を話者ダイアライゼーション装置10として機能させることができる。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) 等の移動体通信端末、さらには、PDA (Personal Digital Assistant) 等のスレート端末等がその範疇に含まれる。また、話者ダイアライゼーション装置10の機能を、クラウドサーバに実装してもよい。

[0052] 図6は、話者ダイアライゼーション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。

[0053] メモリ1010は、ROM (Read Only Memory) 1011およびRAM 1012を含む。ROM 1011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。

[0054] ここで、ハードディスクドライブ 1031 は、例えば、OS 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 1010 に記憶される。

[0055] また、話者ダイアライゼーションプログラムは、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話者ダイアライゼーション装置 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。

[0056] また、話者ダイアライゼーション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、CPU 1020 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 RAM 1012 に読み出して、上述した各手順を実行する。

[0057] なお、話者ダイアライゼーション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 CPU 1020 によって読み出されてもよい。あるいは、話者ダイアライゼーション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 (Local Area Network) 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。

[0058] 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

- [0059] 10 話者ダイアライゼーション装置
- 11 入力部
 - 12 出力部
 - 13 通信制御部
 - 14 記憶部
 - 14a オンラインEENDモデル
 - 15 制御部
 - 15a 音響特徴抽出部
 - 15b 話者ベクトル抽出部
 - 15c 話者ラベル生成部
 - 15d 学習部
 - 15e 推定部
 - 15f 発話区間推定部

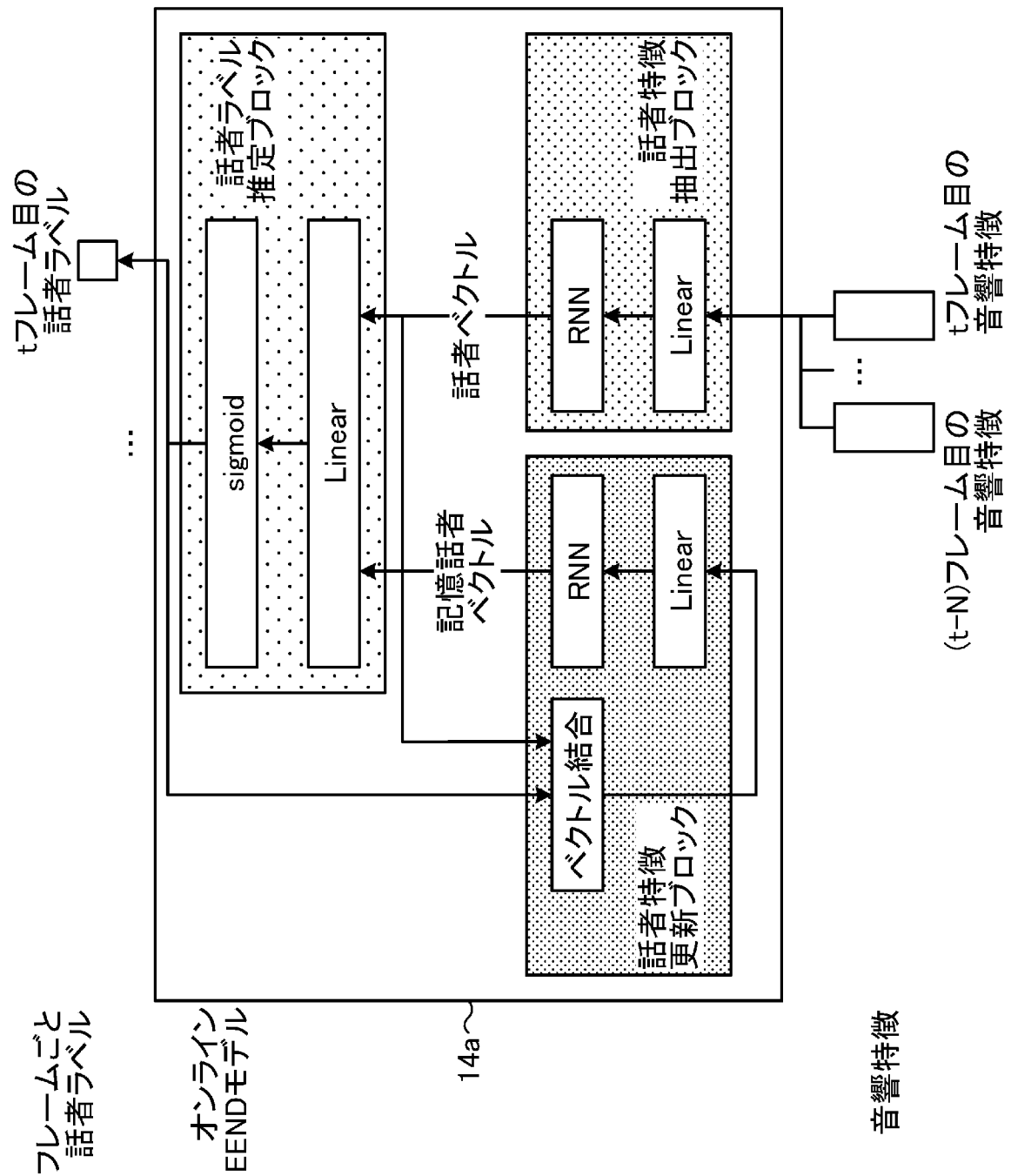
請求の範囲

- [請求項1] 話者ダイアライゼーション装置が実行する話者ダイアライゼーション方法であって、
- 直近の音響信号のフレームごとの音響特徴の系列を用いて、各フレームの話者特徴を表す話者ベクトルを抽出する抽出工程と、
- 前記話者ベクトルと推定された該話者ベクトルの話者を表す話者ラベルとを用いて、各フレームの話者ベクトルの話者ラベルを推定するモデルを学習により生成する学習工程と、
- を含んだことを特徴とする話者ダイアライゼーション方法。
- [請求項2] 前記学習工程は、前記話者ベクトルと推定された該話者ベクトルの話者ラベルとの記憶された複数の組み合わせを用いて、前記モデルを生成することを特徴とする請求項1に記載の話者ダイアライゼーション方法。
- [請求項3] 生成された前記モデルを用いて、音響信号のフレームごとの話者ラベルを推定する推定工程を、さらに含んだことを特徴とする請求項1に記載の話者ダイアライゼーション方法。
- [請求項4] 前記推定工程は、複数のフレームの移動平均を用いて、前記話者ラベルを推定することを特徴とする請求項3に記載の話者ダイアライゼーション方法。
- [請求項5] 直近の音響信号のフレームごとの音響特徴の系列を用いて、各フレームの話者特徴を表す話者ベクトルを抽出する抽出部と、
- 前記話者ベクトルと推定された該話者ベクトルの話者を表す話者ラベルとを用いて、各フレームの話者ベクトルの話者ラベルを推定するモデルを学習により生成する学習部と、
- を有することを特徴とする話者ダイアライゼーション装置。
- [請求項6] 直近の音響信号のフレームごとの音響特徴の系列を用いて、各フレームの話者特徴を表す話者ベクトルを抽出する抽出ステップと、
- 前記話者ベクトルと推定された該話者ベクトルの話者を表す話者ラ

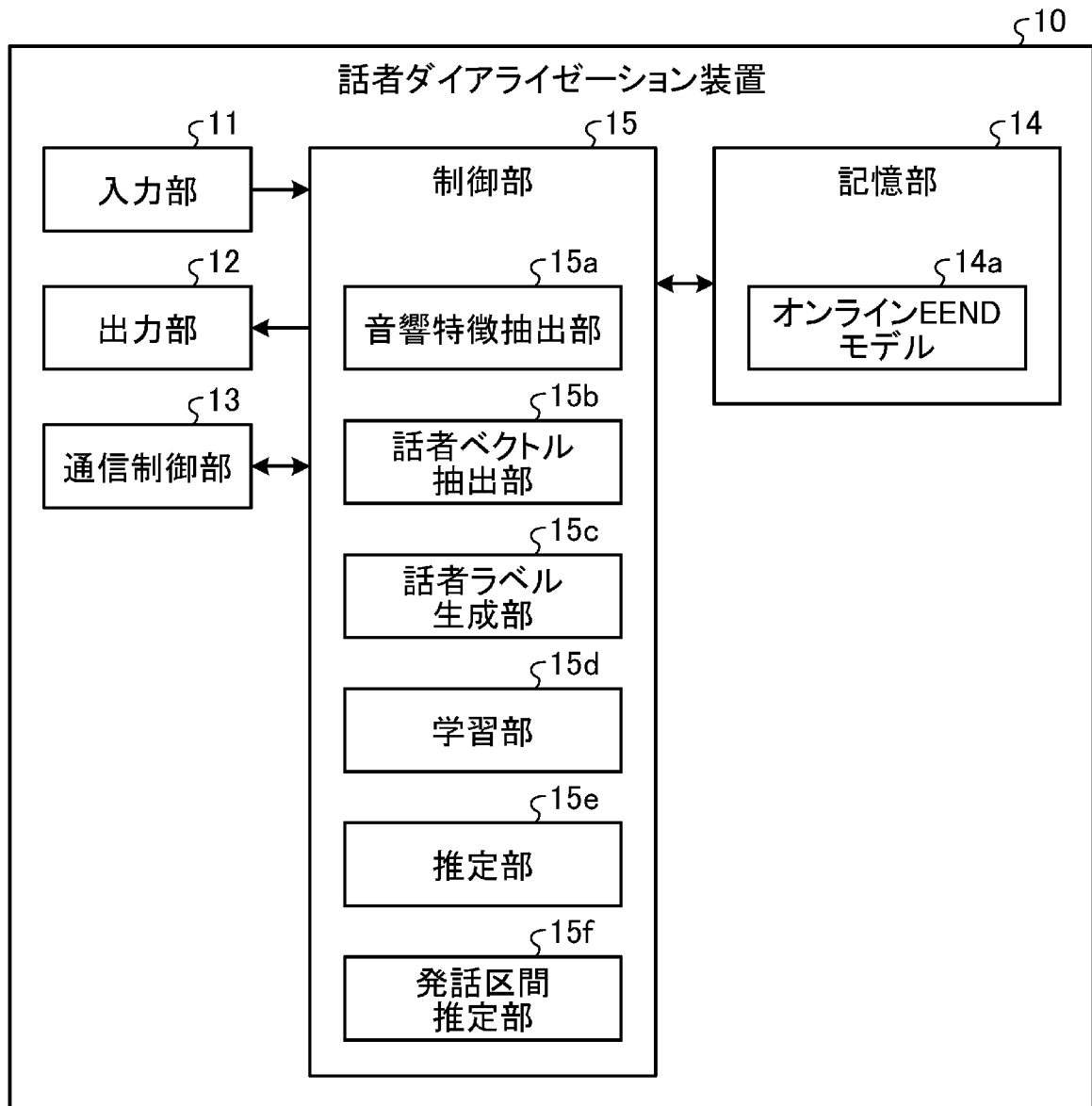
ベルとを用いて、各フレームの話者ベクトルの話者ラベルを推定するモデルを学習により生成する学習ステップと、

をコンピュータに実行させるための話者ダイアライゼーションプログラム。

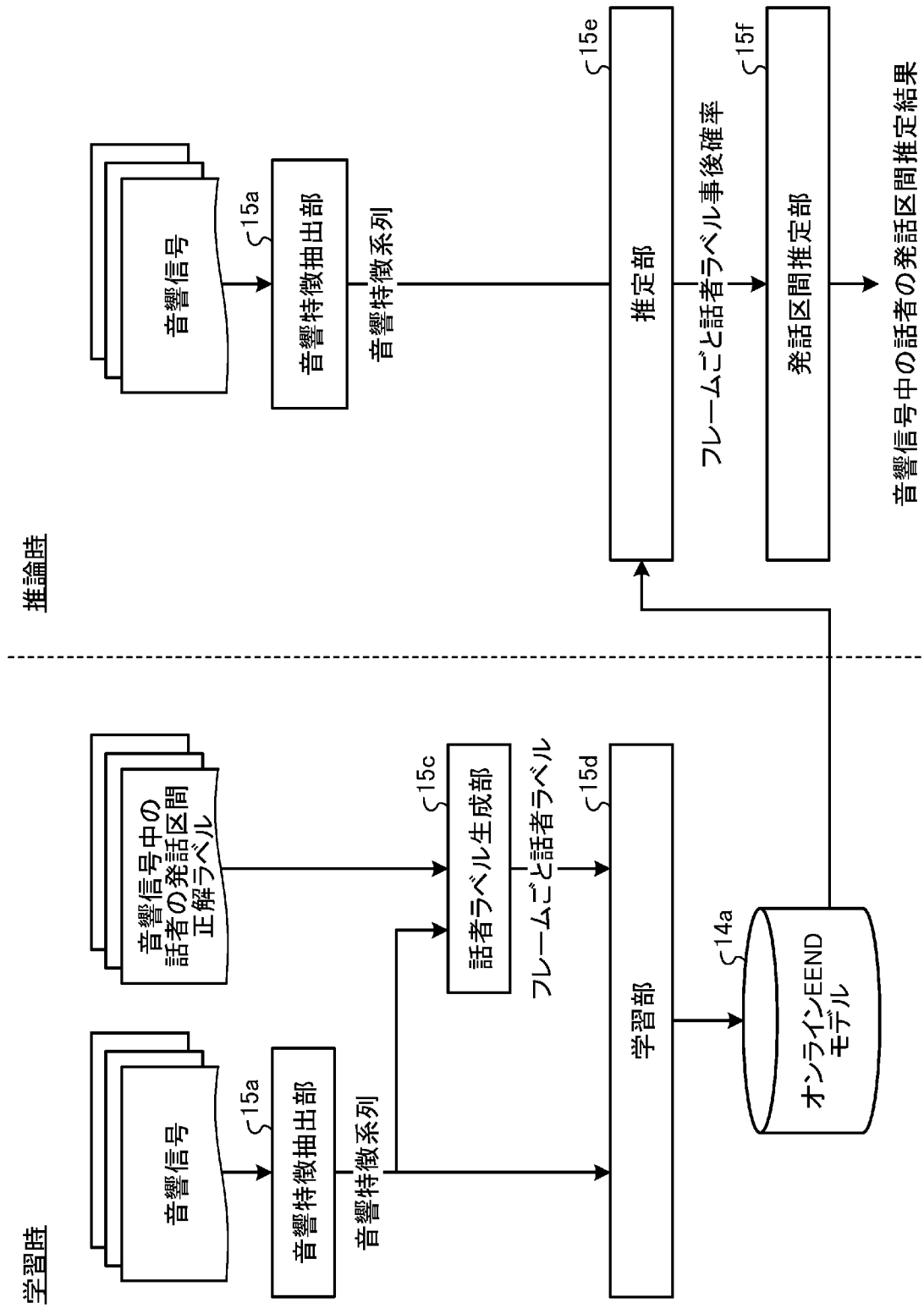
[図1]



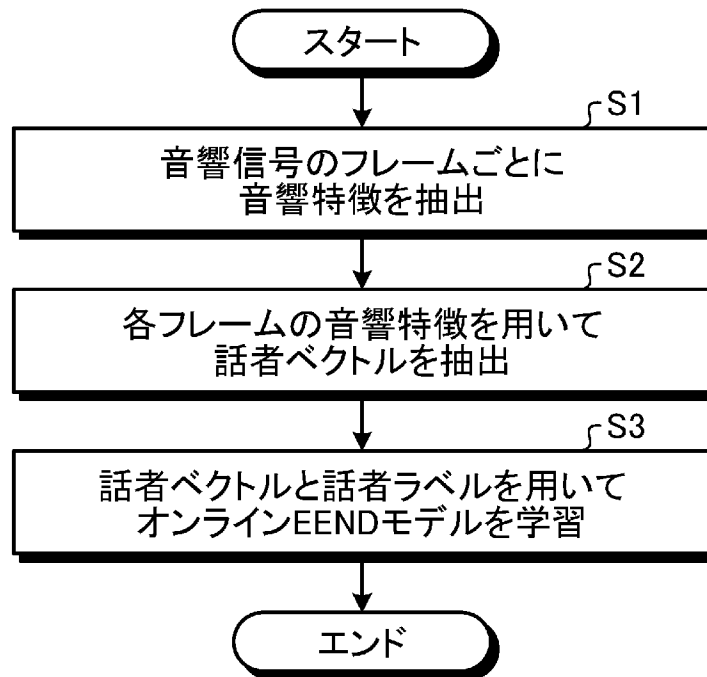
[図2]



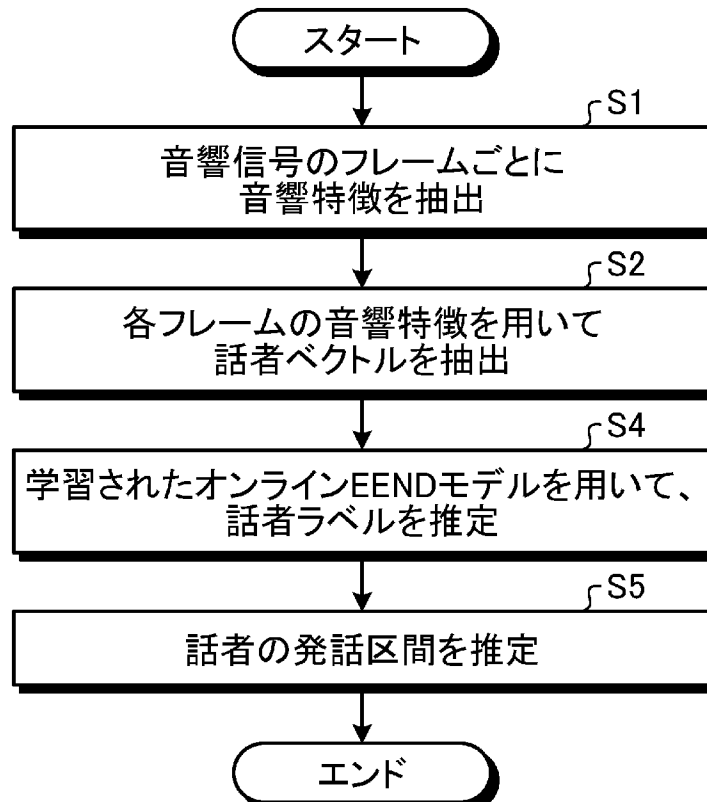
[図3]



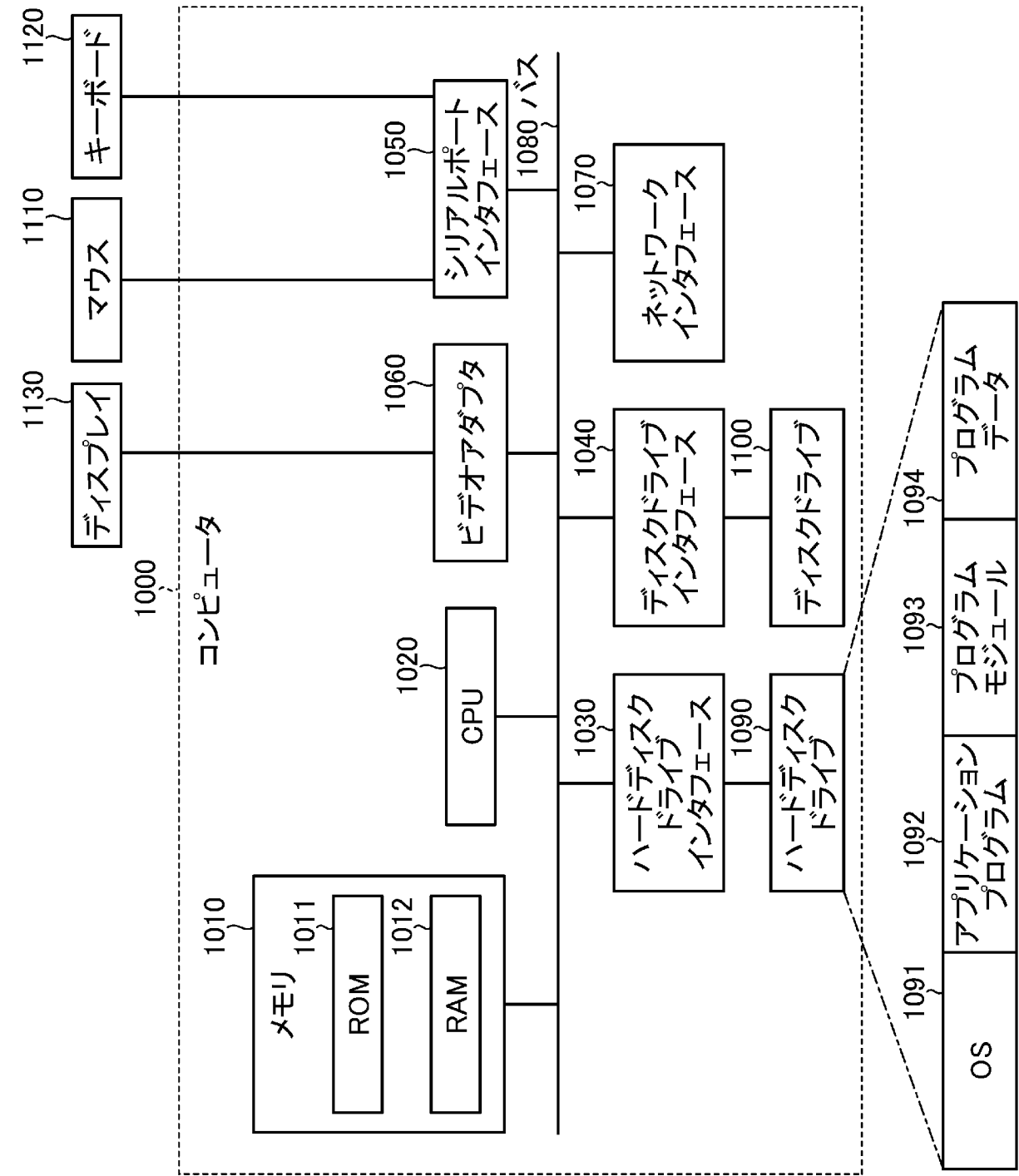
[図4]



[図5]



[図6]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/046117

A. CLASSIFICATION OF SUBJECT MATTER

G10L 17/00(2013.01)i; G10L 17/02(2013.01)i; G10L 17/04(2013.01)i; G10L 17/18(2013.01)i; G10L 15/04(2013.01)i

FI: G10L17/00 200C; G10L17/18; G10L17/04; G10L17/02; G10L15/04 300B

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)

G10L17/00; G10L17/02; G10L17/04; G10L17/18; G10L15/04

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 1971-2021

Registered utility model specifications of Japan 1996-2021

Published registered utility model applications of Japan 1994-2021

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 2019-527370 A (BEIJING DIDI INFINITY TECHNOLOGY AND DEVELOPMENT CO., LTD.) 26 September 2019 (2019-09-26) entire text, all drawings	1-6

☐	Further documents are listed in the continuation of Box C.	☒	See patent family annex.
--------------------------	--	-------------------------------------	--------------------------

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 16 March 2021 (16.03.2021)	Date of mailing of the international search report 30 March 2021 (30.03.2021)
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
Information on patent family members

International application No.

PCT/JP2020/046117

Patent Documents
referred in the
Report

Publication
Date

Patent Family

Publication
Date

JP 2019-527370 A

26 Sep. 2019

US 2018/0358020 A1
entire text, all
drawings
WO 2018/227381 A1
EP 3433854 A
CN 109429523 A

A. 発明の属する分野の分類（国際特許分類（IPC）） G10L 17/00(2013.01)i; G10L 17/02(2013.01)i; G10L 17/04(2013.01)i; G10L 17/18(2013.01)i; G10L 15/04(2013.01)i FI: G10L17/00 200C; G10L17/18; G10L17/04; G10L17/02; G10L15/04 300B		
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）） G10L17/00; G10L17/02; G10L17/04; G10L17/18; G10L15/04 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 1971-2021年 日本国実用新案登録公報 1996-2021年 日本国登録実用新案公報 1994-2021年 国際調査で利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 2019-527370 A（ペイジン ディディ インフィニティ テクノロジー アンド ディベロップメント カンパニー リミテッド）26.09.2019（2019-09-26） 全文、全図	1-6
☐ 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	
16.03.2021	30.03.2021	
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/JP2020/046117

引用文献			公表日	パテントファミリー文献			公表日
JP	2019-527370	A	26.09.2019	US	2018/0358020	A1	
				全文、全図			
				WO	2018/227381	A1	
				EP	3433854	A	
				CN	109429523	A	