

(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11月25日(25.11.2021)



(10) 国際公開番号

WO 2021/234904 A1

- (51) 国際特許分類:
G10L 15/06 (2013.01) *G10L 15/16* (2006.01)
- (21) 国際出願番号: PCT/JP2020/020105
- (22) 国際出願日: 2020年5月21日(21.05.2020)
- (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: 芦原 孝典 (ASHIHARA, Takanori); 〒1808585 東京都武蔵野市緑町三丁目9番1号 NTT 知的財産センタ内 Tokyo (JP). 篠

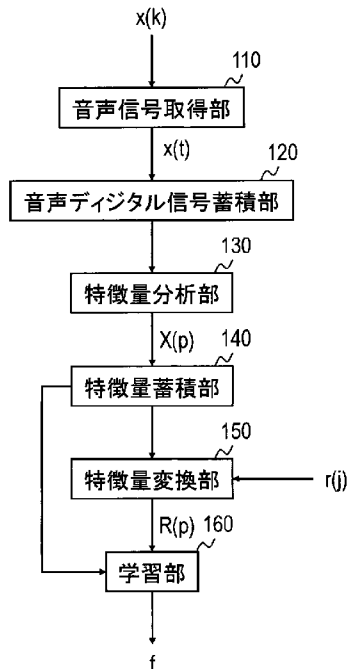
原 雄介 (SHINOHARA, Yusuke); 〒1808585 東京都武蔵野市緑町三丁目9番11号 NTT 知的財産センタ内 Tokyo (JP). 山口 義和 (YAMAGUCHI, Yoshikazu); 〒1808585 東京都武蔵野市緑町三丁目9番11号 NTT 知的財産センタ内 Tokyo (JP).

(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NS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,

(54) Title: TRAINING DATA GENERATION DEVICE, MODEL TRAINING DEVICE, TRAINING DATA GENERATION METHOD, AND PROGRAM

(54) 発明の名称: 学習データ生成装置、モデル学習装置、学習データ生成方法、およびプログラム



- 110... Speech signal acquisition unit
120... Speech digital signal accumulation unit
130... Feature amount analysis unit
140... Feature amount accumulation unit
150... Feature amount conversion unit
160... Training unit

図1

(57) Abstract: Provided is a training data generation device, etc., which generates training data for training an acoustic model that simulates robustness of speech perception of a human. The training data generation device generates training data to be used when training the acoustic model to be used in a speech recognition device. The training data generation device comprises a feature amount conversion unit which converts a first feature amount series, which is an acoustic feature amount series that can be obtained from a first training speech signal, into a second feature amount series, and a

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, 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)）

second training speech signal corresponding to the second feature amount series is a speech signal that can cause paracusis.

(57) 要約：ヒトの音声知覚に対する頑健性を模擬した音響モデルを学習するための学習データを生成する学習データ生成装置等を提供する。学習データ生成装置は、音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する。学習データ生成装置は、第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換部を含み、第二特徴量系列に対応する第二の学習用音声信号は、錯聴を生じ得る音声信号である。

明 細 書

発明の名称：

学習データ生成装置、モデル学習装置、学習データ生成方法、およびプログラム

技術分野

[0001] 本発明は、音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する学習データ生成装置、学習データを用いるモデル学習装置、学習データ生成方法、およびプログラムに関する。

背景技術

[0002] 音響モデルを用いた音声認識装置において、特許文献1は、実用レベルの音声認識性能を担保するために、認識対象とするタスクに対して音響モデルを適応させていく技術である。言い換えると、特許文献1は、話者や、雑音タイプ、喋り方などの音響的特徴が異なるタスクに対して元々の音響モデルを適応させていく技術である。一般的に、音声認識性能は、認識対象とするタスクの学習データ量や、音響的な網羅性に依存して上下する側面を持つ。そこで通常は、認識対象とするタスクの音声を十分に集め、その音声を書き起こしすることで、所望の学習データを収集する。

[0003] しかしながら、従来技術では、莫大な金銭的・時間的コストを要するという課題がある。

[0004] このような課題に対する解決技術の一つとしてData Augmentation(データ拡張)がある。データ拡張とは、オリジナルの学習データに対して何かしらの変動を加え、新しい学習データを生成し、学習データを水増しすることである。データ拡張により、同じ学習データで学習することを少なくし、一層の汎化性能を獲得することができる。

[0005] 例えば、非特許文献1では、話速をオリジナルデータに対して変換することで、様々な話者データを生成し、より広範な話者に対する汎化性能を向上させる。

[0006] また、非特許文献2では、雑音耐性や残響音声に対する認識性能を改善するために、オリジナルの学習データに対し雑音を重畳させ、更に残響の強い部屋のインパルス応答を畳み込むことで擬似的な残響音声を生成し、オリジナルの学習データに対し残響音を重畳させ、汎化性能を向上させる。

先行技術文献

特許文献

[0007] 特許文献1：特開2007-249051号公報

非特許文献

[0008] 非特許文献1：T. Ko, V. Peddinti, D. Povey and S. Khudanpur: "Audio augmentation for speech recognition", In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 3586-3589 (2015).

非特許文献2：R. Hsiao, J. Ma, W. Hartmann, M. Karafiat, F. Grezl, L. Burget, I. Szoke, J. H. Cernocky, S. Watanabe, Z. Chen, S. H. Mallidi, H. Hermansky, S. Tsakalidis and R. Schwartz: "Robust speech recognition in unknown reverberant and noisy conditions", In Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 533-538 (2015).

発明の概要

発明が解決しようとする課題

[0009] ここでは、音声認識装置に対し、長時間情報を捉えるためのData Augmentationを考える。まず、音声認識装置と長時間情報について説明する。長時間情報を音声認識装置に組み込むことで、様々な音響事象に頑健になり、音声認識精度が改善されるといった報告が多数存在している。

[0010] 例えば、recurrent neural network(RNN)モデルは、multi-layer perceptron(MLP)モデルなどとは異なり、長時間情報を陽に取り込むためにモデル自体に工夫をし、音声認識のような時系列情報を扱うタスクでは大きく精度改善

した。

[0011] また、参考文献 1 では、End-to-End音声認識モデルに言語的な長時間コンテキストを陽に組み込むことで、音声認識精度を改善している。

[0012] (参考文献 1) R. Masumura, T. Tanaka, T. Moriya, Y. Shinohara, T. Oba and Y. Aono, "Large Context End-to-end Automatic Speech Recognition via Extension of Hierarchical Recurrent Encoder-decoder Models", ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 5661-5665.

上述のように、モデルを工夫することで長時間情報を音声認識装置に組み込む手法は様々ある。しかしながら、学習データ自体を工夫することで長時間情報を組み込んだ音声認識装置を獲得するようなアプローチはこれまでに存在しない。

[0013] 本発明では、学習データ自体に工夫を施す。本発明に係る学習データ生成装置で生成した学習データを用いて音響モデルを学習することで、ヒトが先天的また後天的に獲得するような音声知覚に対する頑健性を獲得することができる。本実施形態では、錯聴を利用してData Augmentation処理を実行することで、学習データを生成する。

[0014] 更に、本発明では、錯聴に係る音声データをそのまま用いるのではなく、オリジナルの音声データを特徴量にしたあとで、錯聴を利用したData Augmentation処理を実行する。

[0015] 本発明は、ヒトの音声知覚に対する頑健性を模擬した音響モデルを学習するための学習データを生成する学習データ生成装置、その学習データを用いて音響モデルを学習するモデル学習装置、学習データ生成方法、およびプログラム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

[0016] 上記の課題を解決するために、本発明の一態様によれば、学習データ生成装置は、音声認識装置において用いる音響モデルを学習する際に用いる学習

データを生成する。学習データ生成装置は、第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換部を含み、第二特徴量系列に対応する第二の学習用音声信号は、錯聴を生じ得る音声信号である。

上記の課題を解決するために、本発明の他の態様によれば、学習データ生成装置は、音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する。学習データ生成装置は、第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換部を含み、特徴量変換部は、第一特徴量系列のあるセグメントを削除し、削除したセグメントの前後の特徴量の値以上の値を持つノイズを埋め込むことで、第一特徴量系列を第二特徴量系列に変換する。

上記の課題を解決するために、本発明の他の態様によれば、学習データ生成装置は、音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する。学習データ生成装置は、第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換部を含み、特徴量変換部は、第一特徴量系列を所定の時間長のセグメントに分割し、各セグメント内で分割した第一特徴量系列を時間的に反転させ、反転させた特徴量系列を連結することで、第一特徴量系列を第二特徴量系列に変換する。

発明の効果

[0017] 本発明により、ヒトの音声知覚に対する頑健性を模擬した音響モデルを学習することができるという効果を奏する。

図面の簡単な説明

[0018] [図1]第一実施形態に係るモデル学習装置の機能ブロック図。

[図2]第一実施形態に係るモデル学習装置の処理フローの例を示す図。

[図3]連続聴効果が得られるような音声信号に対応する特徴量系列に変換する例を示す図。

[図4]時間反転音声となるような音声信号に対応する特徴量系列に変換する例

を示す図。

[図5]本手法を適用するコンピュータの構成例を示す図。

発明を実施するための形態

[0019] 以下、本発明の実施形態について、説明する。なお、以下の説明に用いる図面では、同じ機能を持つ構成部や同じ処理を行うステップには同一の符号を記し、重複説明を省略する。以下の説明において、ベクトルや行列の各要素単位で行われる処理は、特に断りが無い限り、そのベクトルやその行列の全ての要素に対して適用されるものとする。

[0020] <第一実施形態のポイント>

本実施形態では、ヒトが獲得している音声知覚の頑健性を音声認識装置に獲得させるために、錯聴を利用したData Augmentation処理を実行する。

[0021] 錯聴とは、ヒトの聴覚特性により、物理的に提示された音刺激に対して必ずしもその通り知覚されないような錯覚現象であり、錯視の聴覚版といえる。

[0022] 例えば、連続聴効果(continuity illusion)では、周波数変化する純音や音声等の一部分を削除し、その削除された部分に、元の音を十分にマスキングするだけのノイズを重畳することで、物理的には削除されているはずの音区間が補完されて知覚される現象である(参考文献2参照)。

[0023] (参考文献2) R. M. Warren: "Perceptual restoration of missing speech sounds", Science, 167, pp. 392-393 (1970).

また、時間反転音声(Locally Time-reversed Speech)は、ある一定の短い時間セグメントごとに音声波形を切り分け、それぞれのセグメントにおいてその波形を時間軸上で反転させた後、その反転させた各セグメントを再度連結させた音声である(参考文献3参照)。

[0024] (参考文献3) K. Saberi and D. R. Perrott, "Cognitive restoration of reversed speech", Nature, 398, 6730, pp. 760-760 (1999).

ヒトがこのような時間反転音声を聴取した場合、その音声知覚の明瞭度はそのセグメント長が比較的短い場合、例えば25ms程度であれば、十分高いまま

維持される。しかしながら、セグメント長が長くなればなるほどその明瞭度はシグモイド関数的に低下していき、100ms程度でほぼ音声知覚が困難になることが実験的に示されている。つまりある程度までの局所的な時系列の破壊であれば、ヒトの音声知覚には影響を与えない（頑健である）ことが分かる。

- [0025] このような連続聴効果を生じ得るような音声信号や時間反転音声となるような音声信号を用いて、音響モデルを学習すれば、自ずと、削除またはマスキングした部分や、反転させる際のセグメントよりも長い時間間隔を考慮して音響モデルを学習することとなり、音響モデルは長時間情報を組み込んだものとなり、ヒトが獲得している音声知覚の頑健性を獲得したものとなる。
- [0026] 本実施形態では、上述のような錯聴を利用したData Augmentation処理を音声波形に対して実行するのではなく、特徴量空間上でData Augmentation処理を実行することで、学習データから長時間情報に頑健な音声認識装置を構築することが可能になる。もし音声波形に対してData Augmentation処理を実行する場合、学習データ量が単純に2倍になってしまい、そのデータを保管する場合、オリジナルデータ分も考慮して2倍の容量が必要になる。しかしながら、特徴量空間上でData Augmentation処理を実行することで、学習の最中に、その学習データとなる特徴量を変換処理することが可能になるため、データ容量もオリジナルデータ分だけで済む。
- [0027] 錯聴の中でも特徴量空間上で処理することが可能な例として、本実施形態では上述した連続長効果と時間反転音声を挙げる。
- [0028] 連続長効果では、一部分が欠如した状態でも音声知覚出来るヒトの頑健性を音声認識装置に獲得させることが可能になるため、結果として長時間情報に頑健な音声認識装置が構築される。また、特徴量空間上でも同様に似たような表現が可能である。例えば、特徴量において時間軸上のあるセグメントを削除し、そこにそのセグメントの前後の特徴量の大きさ以上の値を代わりに埋め込むことで連続長効果と同等の表現となる。
- [0029] 時間反転音声では、局所的に時系列が反転（破壊）された状態でも音声知

覚出来るヒトの頑健性を音声認識装置に獲得させることが可能になるため、結果として長時間情報に頑健な音声認識装置が構築される。こちらも同様に、特徴量空間上で似たような表現にするために、特徴量上の時間軸上における各セグメント内で特徴量系列を反転させ、それを全て再連結させたデータを拡張データとして用いる。

[0030] <第一実施形態>

図1は第一実施形態に係るモデル学習装置の機能ブロック図を、図2はその処理フローを示す。

[0031] モデル学習装置100は、音声信号取得部110と音声デジタル信号蓄積部120と特徴量分析部130と特徴量蓄積部140と特徴量変換部150と学習部160とを含む。

[0032] モデル学習装置は、例えば、中央演算処理装置（CPU: Central Processing Unit）、主記憶装置（RAM: Random Access Memory）などを有する公知又は専用のコンピュータに特別なプログラムが読み込まれて構成された特別な装置である。モデル学習装置は、例えば、中央演算処理装置の制御のもとで各処理を実行する。モデル学習装置に入力されたデータや各処理で得られたデータは、例えば、主記憶装置に格納され、主記憶装置に格納されたデータは必要に応じて中央演算処理装置へ読み出されて他の処理に利用される。モデル学習装置の各処理部は、少なくとも一部が集積回路等のハードウェアによって構成されていてもよい。モデル学習装置が備える各記憶部は、例えば、RAM（Random Access Memory）などの主記憶装置、ハードディスクや光ディスクもしくはフラッシュメモリ（Flash Memory）のような半導体メモリ素子により構成される補助記憶装置、またはリレーショナルデータベースやキーバリューストアなどのミドルウェアにより構成することができる。

[0033] モデル学習装置は、学習用のアナログの音声信号 $x(k)$ と、対応する正解ラベル $r(j)$ とを入力とし、これらの情報に基づき音響モデルを学習し、学習済みの音響モデル f を出力する。なお、 k は時刻を示すインデックスである。正解ラベルは例えば音素ラベルであり、 j は音素の順番を示すインデックスであ

る。なお、正解ラベル(音素ラベル)がアナログの音声信号のどこからどこまでに対応するかを示す情報は、予め学習データに含まれるものとし、以下で説明する音声デジタル信号、特徴量についても同様に紐付けられているものとする。

[0034] 以下、各部の処理内容について説明する。

[0035] <音声信号取得部 110>

入力：音声信号 $x(k)$

出力：音声デジタル信号 $x(t)$

処理：AD変換

音声信号取得部 110 は、アナログの音声信号 $x(k)$ を取得し、デジタルの音声デジタル信号 $x(t)$ に変換する (S 110)。なお、 t は音声デジタル信号のサンプル番号を示すインデックスである。

[0036] <音声デジタル信号蓄積部 120>

入力：音声デジタル信号 $x(t)$

処理：音声デジタル信号の蓄積

音声デジタル信号蓄積部 120 は、音声デジタル信号 $x(t)$ を蓄積する (S 120)。

[0037] <特徴量分析部 130>

入力：音声デジタル信号 $x(t)$

出力：特徴量系列 $X(p)$

処理：特徴量分析

特徴量分析部 130 は、音声デジタル信号 $x(t)$ に対して特徴量分析を行い、特徴量系列 $X(p)$ を得る。

[0038] 例えば、特徴量分析部 130 は、音声デジタル信号蓄積部 120 から発話 p 毎の音声デジタル信号 $x(t)$ を取り出し、音声デジタル信号 $x(t)$ をフレーム単位に分割し、フレーム毎に音響特徴量抽出を行い、発話 p 毎の(音響)特徴量系列 $X(p)$ を取得する。

[0039] 例えば、ある発話 p に含まれるフレーム総数を N_p とし、 $n_p=1_p, 2_p, \dots, N_p$ とし、

フレーム長を M とし、 $m=1, 2, \dots, M$ とし、シフト幅を D とすると、ある発話 p の n_p 番目のフレームの m 番目の音声デジタル信号 $x(t)$ は、 $x(D(n_p-1)+m)$ と表すことができる。ただし、下付き添え字 p は、発話 p に対応する値であることを示す。特徴量分析部130は、フレーム n_p 毎に音声デジタル信号 $x(D(n_p-1)+1), x(D(n_p-1)+2), \dots, x(D(n_p-1)+M)$ に対して、音響特徴量抽出を行い、特徴量 $X(n_p)$ を取得する。特徴量分析部130は、発話 p に含まれる全てのフレーム $1_p, 2_p, \dots, N_p$ に対して処理を行い、発話 p 毎の特徴量系列 $X(p)=\{X(1_p), X(2_p), \dots, X(N_p)\}$ を取得する。

[0040] 抽出する特徴量としては、例えば、音声信号の短時間フレーム分析に基づくMFCC(Mel-Frequency Cepstrum Coefficient)の1~12次元と、その動的特徴量である Δ MFCC、 $\Delta\Delta$ MFCCなどの動的参数や、パワー、 Δ パワー、 $\Delta\Delta$ パワー等を用いる。また、MFCCに対してはCMN(ケプストラム平均正規化)処理を行っても良い。特徴量は、MFCCやパワーに限定したものではなく、特殊発話の識別に用いられるパラメータ（例えば、自己相関ピーク値や群遅延など）を用いても良い。

[0041] <特徴量蓄積部140>

入力：特徴量系列 $X(p)$

処理：特徴量系列の蓄積

特徴量蓄積部140は、特徴量分析部130で分析した特徴量系列 $X(p)$ を蓄積する（S140）。

[0042] <特徴量変換部150>

入力：特徴量系列 $X(p)$

出力：特徴量系列 $R(p)$

処理：Data Augmentation処理

特徴量変換部150は、特徴量系列 $X(p)$ に対してData Augmentation処理を実行し、特徴量系列 $X(p)$ を特徴量系列 $R(p)$ に変換する（S150）。

[0043] なお、Data Augmentation処理は、後述する学習部160で学習する際に同時にオンラインで実行される。より詳しく説明すると、後述する学習部16

0で用いる全ての発話 p (ここでは $p=1, 2, \dots, P$)に対応する特徴量系列 $X(p)$ に対して予めData Augmentation処理を行うのではなく、ある発話 p' (p' は $1, 2, \dots, P$ の何れか)に対応する特徴量系列 $X(p')$ を用いて学習する際にある特徴量系列 $X(p')$ に対してData Augmentation処理を実行し、特徴量系列 $X(p')$ を特徴量系列 $R(p')$ に変換することを意味する。ただし、 P は、学習用のアナログの音声信号 $x(k)$ に含まれる発話の総数を表す。水増しした学習データは、学習時のみ利用し、保管する必要がないため、保管する学習データ量を減らすことができる。なお、入力の特徴量系列となるため、全て特徴量空間上でData Augmentation処理をすることになり、音声デジタル信号に対してData Augmentation処理を行う必要がなくなる。

[0044] 本実施形態におけるData Augmentation処理は、特徴量系列 $X(p)$ をある変換ルールに基づいて特徴量系列 $R(p)$ に変換する。この変換により、擬似的な水増し学習データを生成する。

[0045] 本実施形態では、変換後の特徴量系列 $R(p)$ に対応する音声信号が錯聴を生じ得る音声信号となるような変換ルールが採用される。なお、一般的に錯聴を生じ得る音声信号を生成する場合には音声波形に対して処理を行うが、本実施形態では特徴量系列上で変換処理を行う。

[0046] 変換後の特徴量系列 $R(p)$ に対応する音声信号が錯聴を生じ得る音声信号となるような変換ルールとして、本実施形態では、上述したような連続聴効果が得られるような音声信号や時間反転音声となるような音声信号に対応する特徴量系列に変換する変換ルールを採用する。

[0047] (i)連続聴効果が得られるような音声信号に対応する特徴量系列に変換する場合、特徴量変換部150は、特徴量系列 $X(p)$ のあるセグメントを削除し、削除したセグメントの前後の特徴量の値以上の値を持つ特徴量を削除した部分に埋め込むことで、特徴量系列 $X(p)$ を特徴量系列 $R(p)$ に変換する。セグメント長は、錯聴を生じ得る長さである。また、削除処理および埋め込み処理は、錯聴を生じ得る間隔で行う。また、埋め込む特徴量はノイズに対応する特徴量であり、ノイズは例えばホワイトノイズである。なお、処理S150に

先立ちノイズに対応する特徴量を予め用意しておく。例えば、特徴量系列 $\dots, X(s+1_p), X(s+2_p), X(s+3_p), X(s+4_p), X(s+5_p), X(s+6_p), X(s+7_p), X(s+8_p), X(s+9_p), X(s+10_p), \dots$ のうち、3つの特徴量 $X(s+3_p), X(s+4_p), X(s+5_p)$ を削除し、ノイズに対応する3つの特徴量 $X(1_n), X(2_n), X(3_n)$ を埋め込む処理を行う(図3参照)。 $X(1_n), X(2_n), X(3_n)$ の値は、前の特徴量 $X(s+2_p)$ と後ろの特徴量 $X(s+6_p)$ の値以上となるように設定する。例えば、この処理を、20フレーム毎に行う。

[0048] (ii) 時間反転音声となるような音声信号に対応する特徴量系列に変換する場合、特徴量変換部150は、特徴量系列 $X(p)$ を所定の時間長のセグメントに分割し、各セグメント内で分割した特徴量系列を時間的に反転させ、反転させた特徴量系列を連結することで、特徴量系列 $X(p)$ を特徴量系列 $R(p)$ に変換する。セグメント長は、錯聴を生じ得る長さである。例えば、特徴量変換部150は、特徴量系列 $\dots, X(s+1_p), X(s+2_p), X(s+3_p), X(s+4_p), X(s+5_p), X(s+6_p), X(s+7_p), X(s+8_p), X(s+9_p), X(s+10_p), \dots$ を5フレーム分の時間長のセグメント $\dots, s(1)=\{X(s+1_p), X(s+2_p), X(s+3_p), X(s+4_p), X(s+5_p)\}, s(2)=\{X(s+6_p), X(s+7_p), X(s+8_p), X(s+9_p), X(s+10_p)\}, \dots$ に分割する。さらに、特徴量変換部150は、各セグメント内の特徴量系列を時間的に反転させ、 $\dots, s'(1)=\{X(s+5_p), X(s+4_p), X(s+3_p), X(s+2_p), X(s+1_p)\}, s'(2)=\{X(s+10_p), X(s+9_p), X(s+8_p), X(s+7_p), X(s+6_p)\}, \dots$ とし、 $\dots, s'(1), s'(2), \dots$ の順番で連結する(図4参照)。

[0049] <学習部160>

入力：特徴量系列 $X(p)$ 、特徴量系列 $R(p)$ 、正解ラベル $r(j)$

出力：学習済み音響モデル f

処理：モデル学習

学習部160は、特徴量系列 $X(p)$ 、特徴量系列 $R(p)$ 、正解ラベル $r(j)$ を用いて、音響モデル f を学習する(S160)。音響モデル f は、特徴量系列を入力とし、音素ラベルを出力するモデルである。音声認識における音響モデルとしては、例えばGMM-HMMやDNN-HMMなどがしばしば用いられており、近年ではEnd-to-End音声認識モデルも用いられているが、本実施形態では、特に学習対象の音声認識モデルに制約はないため、GMM/DNN-HMMであってもEnd-to-E

nd音声認識モデルであってもよい。なお、正解ラベル $r(j)$ は、学習用のアナログの音声信号 $x(k)$ に対応するものであり、音声信号 $x(k)$ から得られる特徴量系列 $X(p)$ 、および、特徴量系列 $X(p)$ を変換して得られる特徴量系列 $R(p)$ にも対応する。

[0050] <効果>

以上の構成により、ヒトの音声知覚に対する頑健性を模擬した音響モデルを学習することができる。また、金銭的・時間的コストを低減することができる。さらに、Data Augmentation処理を音声波形に対して行うのではなく、特徴量空間上で行うことで、水増しした学習データに対する処理S 1 1 0～S 1 4 0を削減することができる。また、Data Augmentation処理を学習時に同時に行うことで、学習データの記憶容量を削減することができる。

[0051] <変形例>

モデル学習装置100の学習部160を含まない構成を、学習データ生成装置ともいう。つまり、学習データ生成装置は、音声信号取得部110と音声デジタル信号蓄積部120と特徴量分析部130と特徴量蓄積部140と特徴量変換部150とを含む。学習データ生成装置は、学習用のアナログの音声信号 $x(k)$ と正解ラベル $r(j)$ とを入力とし、音声信号 $x(k)$ から特徴量系列 $X(p)$ と特徴量系列 $R(p)$ とを生成し、特徴量系列 $X(p)$ 、特徴量系列 $R(p)$ および正解ラベル $r(j)$ の組合せを学習データとして出力する。

[0052] 第一実施形態では、特徴量系列 $R(p)$ に対応する音声信号は、錯聴を生じ得る音声信号であることを前提としているが、実験の結果、錯聴を生じ得ない音声信号であっても同様の効果を得ることができることが分かった。

[0053] 例えば、第一実施形態では、連続聴効果を得るために、特徴量変換部150は、特徴量系列 $X(p)$ のあるセグメントを削除し、削除したセグメントの前後の特徴量の値以上の値を持つ特徴量を削除した部分に埋め込むことで、特徴量系列 $X(p)$ を特徴量系列 $R(p)$ に変換する。このとき、削除する、または、埋め込むセグメントのセグメント長は、連続聴効果を生じ得ないほど長くとよい。また、削除処理および埋め込み処理の間隔は、連続聴効果を生じ得

ないほど短くともよい。このようなデータ拡張を実行した場合であっても、第一実施形態と同様の精度を持つ音響モデルを学習することができる。

[0054] また、例えば、第一実施形態では、時間反転音声となるような音声信号に変換するために、特徴量変換部150は、特徴量系列 $X(p)$ を所定の時間長のセグメントに分割し、各セグメント内で分割した特徴量系列を時間的に反転させ、反転させた特徴量系列を連結することで、特徴量系列 $X(p)$ を特徴量系列 $R(p)$ に変換する。このとき、セグメント長は、錯聴を生じ得ないほど長くともよい。このようなデータ拡張を実行した場合であっても、第一実施形態と同様の精度を持つ音響モデルを学習することができる。

[0055] <その他の変形例>

本発明は上記の実施形態及び変形例に限定されるものではない。例えば、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。

[0056] <プログラム及び記録媒体>

上述の各種の処理は、図5に示すコンピュータの記憶部2020に、上記方法の各ステップを実行させるプログラムを読み込ませ、制御部2010、入力部2030、出力部2040などに動作させることで実施できる。

[0057] この処理内容を記述したプログラムは、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としては、例えば、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等のようなものでもよい。

[0058] また、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。

[0059] 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、自己の記録媒体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。

[0060] また、この形態では、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。

請求の範囲

- [請求項1] 音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する学習データ生成装置であって、
- 第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換部を含み、
- 前記第二特徴量系列に対応する第二の学習用音声信号は、錯聴を生じ得る音声信号である、
- 学習データ生成装置。
- [請求項2] 音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する学習データ生成装置であって、
- 第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換部を含み、
- 前記特徴量変換部は、前記第一特徴量系列のあるセグメントを削除し、削除したセグメントの前後の特徴量の値以上の値を持つノイズを埋め込むことで、前記第一特徴量系列を前記第二特徴量系列に変換する、
- 学習データ生成装置。
- [請求項3] 音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する学習データ生成装置であって、
- 第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換部を含み、
- 前記特徴量変換部は、前記第一特徴量系列を所定の時間長のセグメントに分割し、各セグメント内で分割した第一特徴量系列を時間的に反転させ、反転させた特徴量系列を連結することで、前記第一特徴量系列を前記第二特徴量系列に変換する、
- 学習データ生成装置。
- [請求項4] 請求項1から請求項3の何れかの学習データ生成装置を含むモデル学習装置であって、

前記第一特徴量系列と、前記第二特徴量系列と、前記第一の学習用音声信号に対応する正解ラベルとを用いて、音響モデルを学習する学習部を含み、

前記特徴量変換部は、前記音響モデルの学習中に前記第一特徴量系列を前記第二特徴量系列に変換する、

モデル学習装置。

[請求項5] 音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する学習データ生成方法であって、

第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換ステップを含み、

前記第二特徴量系列に対応する第二の学習用音声信号は、錯聴を生じ得る音声信号である、

学習データ生成方法。

[請求項6] 音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する学習データ生成方法であって、

第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換ステップを含み、

前記特徴量変換ステップにおいて、前記第一特徴量系列のあるセグメントを削除し、削除したセグメントの前後の特徴量の値以上の値を持つノイズを埋め込むことで、前記第一特徴量系列を前記第二特徴量系列に変換する、

学習データ生成方法。

[請求項7] 音声認識装置において用いる音響モデルを学習する際に用いる学習データを生成する学習データ生成方法であって、

第一の学習用音声信号から得られる音響特徴量系列である第一特徴量系列を第二特徴量系列に変換する特徴量変換ステップを含み、

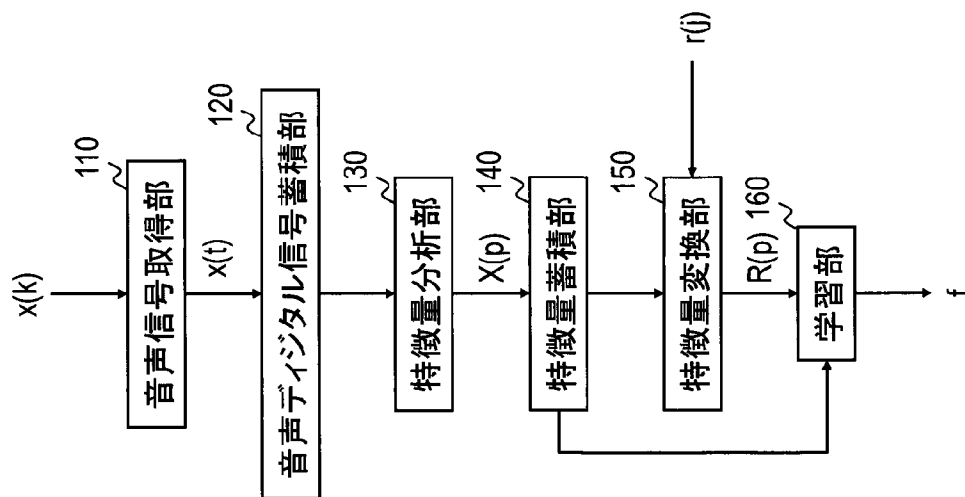
前記特徴量変換ステップにおいて、前記第一特徴量系列を所定の時間長のセグメントに分割し、各セグメント内で分割した第一特徴量系

列を時間的に反転させ、反転させた特徴量系列を連結することで、前記第一特徴量系列を前記第二特徴量系列に変換する、

学習データ生成方法。

[請求項8] 請求項 1 から請求項 3 の何れかの学習データ生成装置、または、請求項 4 のモデル学習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]

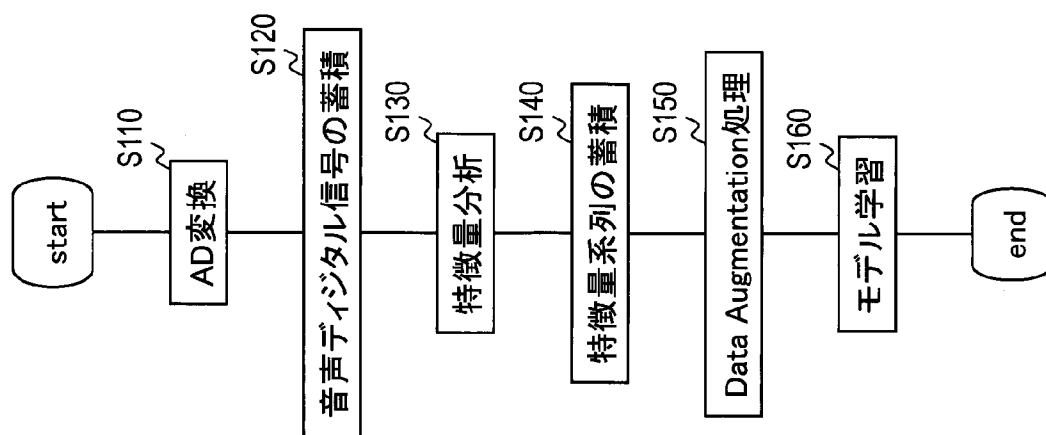
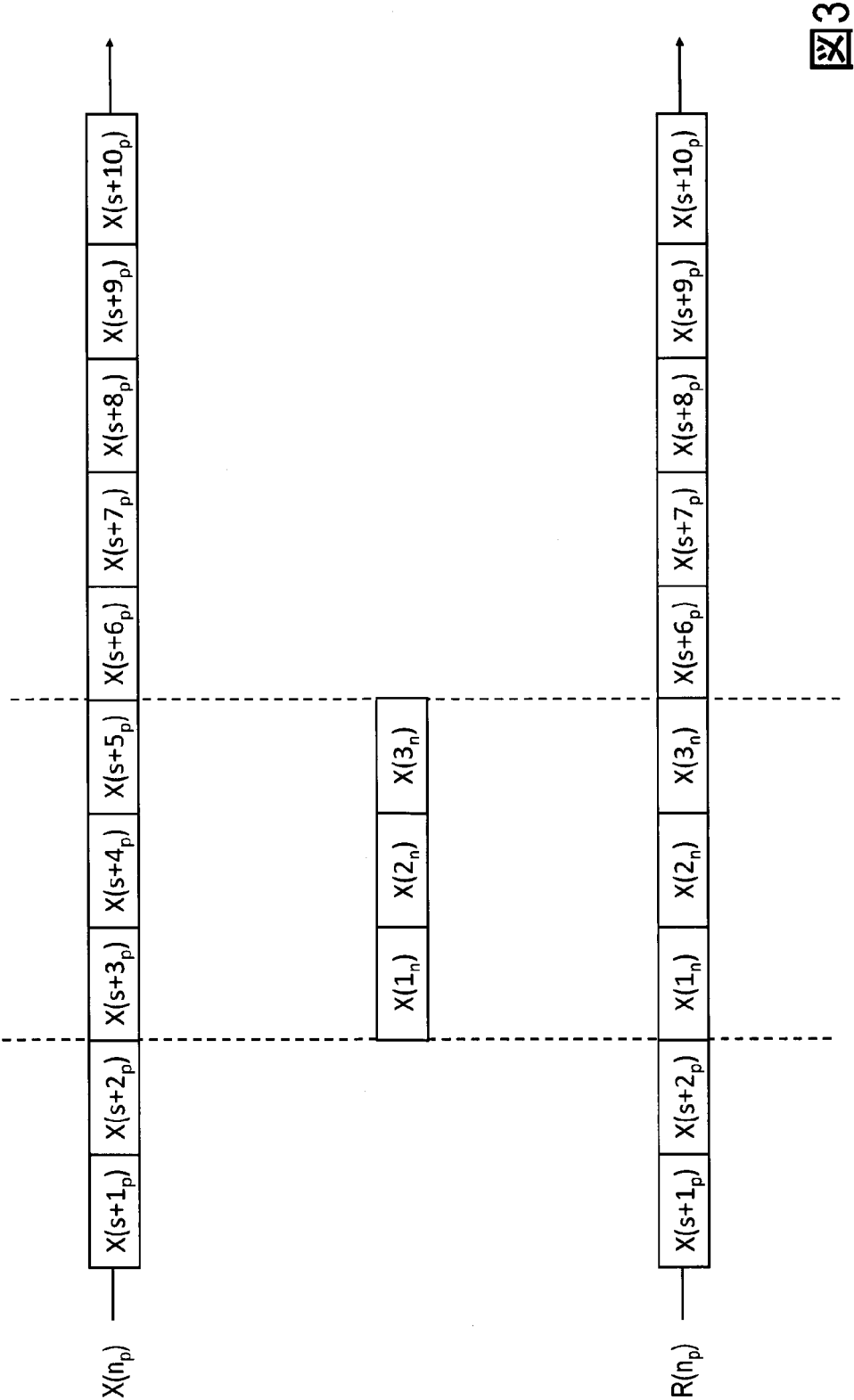


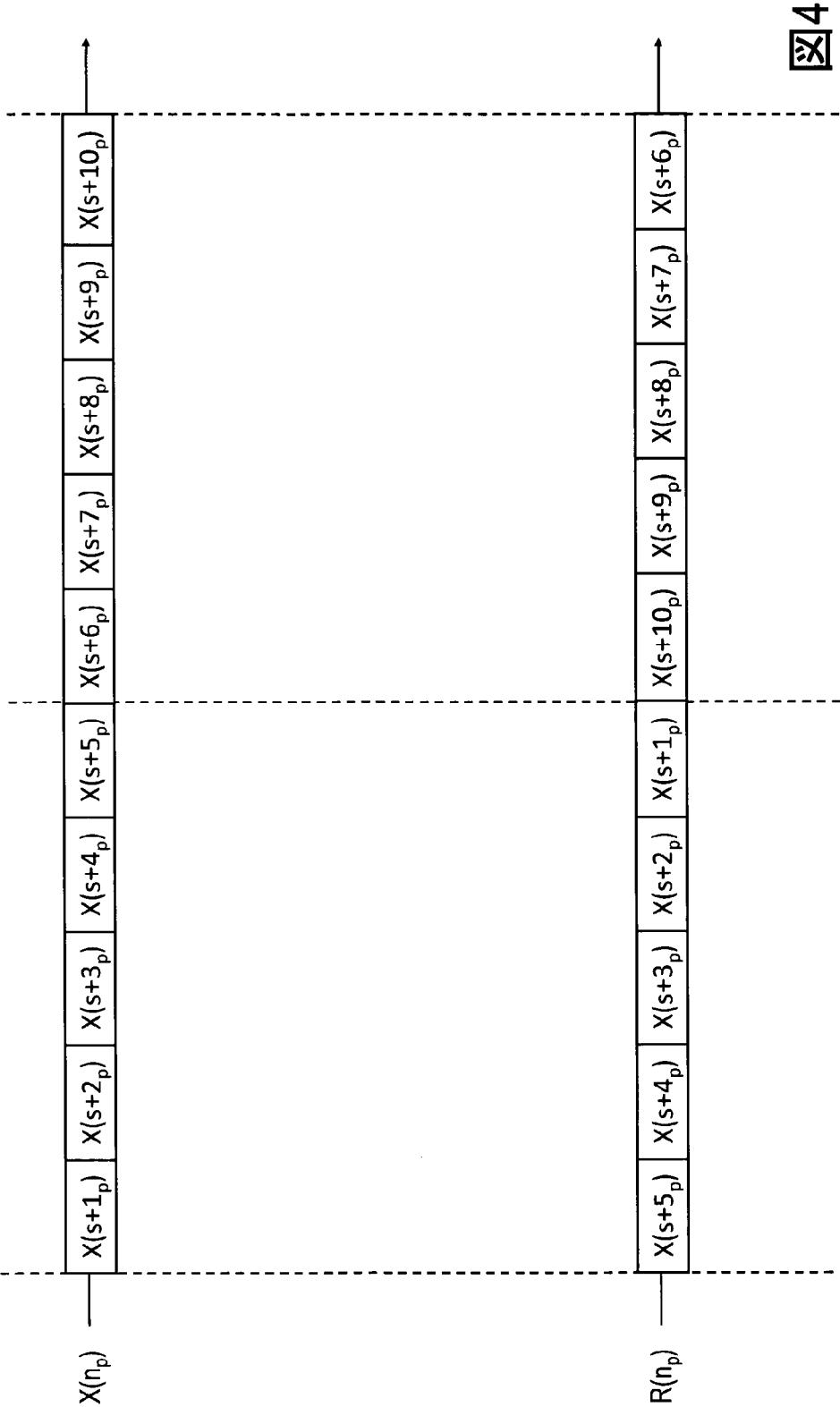
図2

[図3]



[図3]

[図4]



[図4]

[図5]

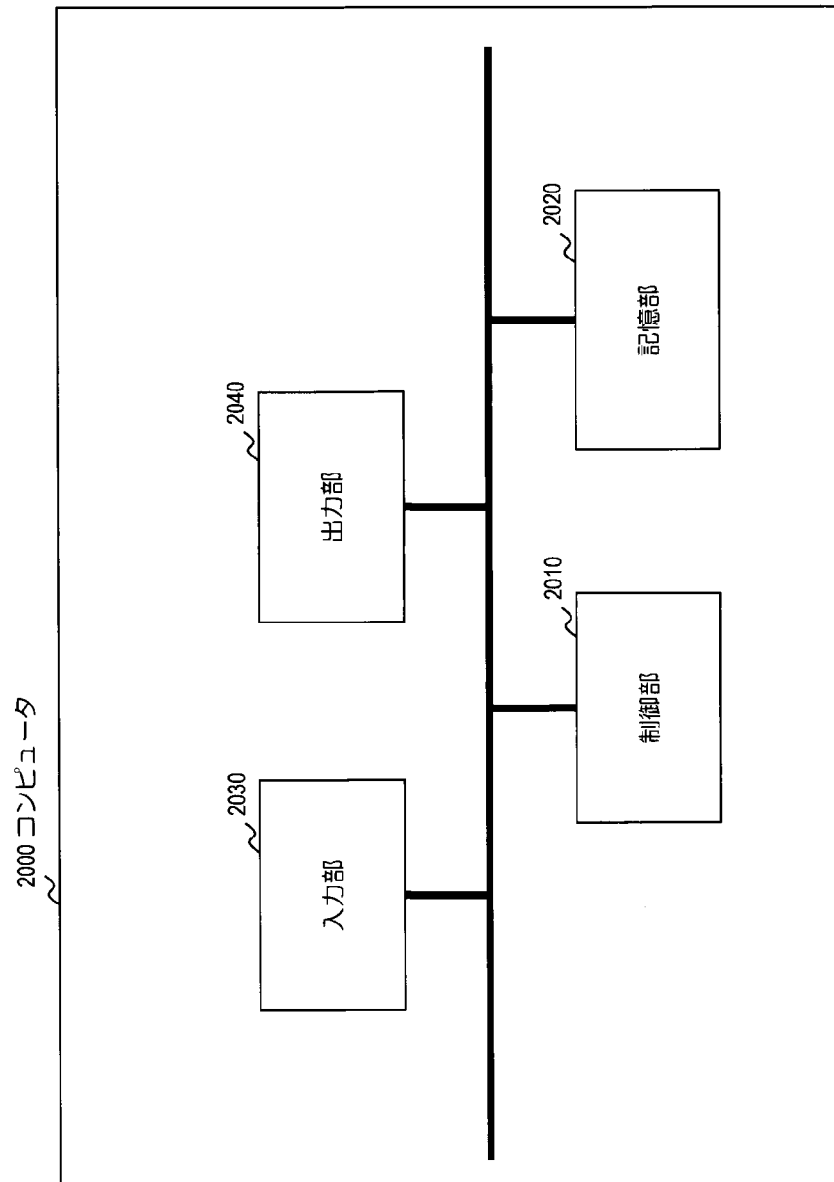


図5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/020105

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. G10L15/06(2013.01) i, G10L15/16(2006.01) i
FI: G10L15/06 300Z, G10L15/16

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. G10L15/06, G10L15/16

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 2016-161823 A (HITACHI, LTD.) 05 September 2016, entire text, all drawings	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

“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
31.07.2020

Date of mailing of the international search report
11.08.2020

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
Information on patent family members

International application No.
PCT/JP2020/020105

Patent Documents referred to in the Report	Publication Date	Patent Family	Publication Date
JP 2016-161823 A	05.09.2016	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G10L 15/06(2013.01)i; G10L 15/16(2006.01)i FI: G10L15/06 300Z; G10L15/16														
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）） G10L15/06; G10L15/16														
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）														
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 2016-161823 A（株式会社日立製作所）05.09.2016（2016 - 09 - 05） 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。														
<table border="0"> <tr> <td>* 引用文献のカテゴリー</td> <td>“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</td> </tr> <tr> <td>“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</td> <td>“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</td> </tr> <tr> <td>“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</td> <td>“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</td> </tr> <tr> <td>“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）</td> <td>“&” 同一パテントファミリー文献</td> </tr> <tr> <td>“O” 口頭による開示、使用、展示等に言及する文献</td> <td></td> </tr> <tr> <td>“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</td> <td></td> </tr> </table>			* 引用文献のカテゴリー	“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	“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	“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	“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	“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	“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）	“&” 同一パテントファミリー文献	“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	
* 引用文献のカテゴリー	“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													
“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	“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													
“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	“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													
“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）	“&” 同一パテントファミリー文献													
“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														
国際調査を完了した日 31.07.2020	国際調査報告の発送日 11.08.2020													
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													

国際出願番号
PCT/JP2020/020105

引用文献	公表日	パテントファミリー文献	公表日
JP 2016-161823 A	05.09.2016	(ファミリーなし)	