(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1月2日(02.01.2020)

WIPO | PCT

(10) 国際公開番号
WO 2020/003413 A1

(51) 国際特許分類:
*G10L 17/00* (2013.01)

(21) 国際出願番号： PCT/JP2018/024391

(22) 国際出願日： 2018年6月27日(27.06.2018)

(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: 日本電気株式会社 (**NEC CORPORATION**) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 Tokyo (JP).

(72) 発明者: カク レイ (**GUO Ling**); 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP). 山本 仁 (**YAMAMOTO Hitoshi**); 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP). 越仲 孝文 (**KOSHINAKA Takafumi**); 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP).

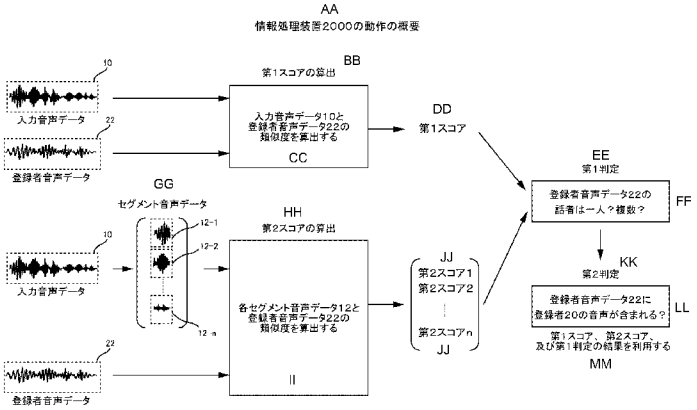
(74) 代理人: 速水 進治 (**HAYAMI Shinji**); 〒1410031 東京都品川区西五反田７丁目９番２号 ＫＤＸ五反田ビル９階 Tokyo (JP).

(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,

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

(54) 発明の名称： 情報処理装置、制御方法、及びプログラム

[図1]




10 Input voice data
22 Registrant voice data
AA Outline of operation of information processing device 2000
BB Calculation of first score
CC Calculate degree of similarity between input voice data 10 and registrant voice data 22
DD First score
EE First determination
FF Is speaker of registrant voice data 22 one or plural?
GG Segment voice data
HH Calculation of second score
II Calculate degree of similarity between each segment voice data 12 and registrant voice data 22
JJ Second score
KK Second determination
LL Does registrant voice data 22 include voice of registrant 20?
MM Use first score, second score, and result of first determination

(57) **Abstract:** An information processing device (2000) calculates a first score indicating degree of similarity between input voice data (10) and registrant voice data (22) of a registrant (20). The information processing device (2000) divides the input voice data (10) in a time direction to thereby obtain a plurality of segment voice data (12). The information processing device (2000) calculates a second score indicating degree of similarity between the segment voice data (12) and the registrant voice data (22) with regard to each of the segment voice data (12). The information processing device

WO 2020/003413 A1

［続葉有］

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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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))

2000 performs first determination that determines whether a speaker of voice included in the input voice data (10) is one speaker or plural speakers by using at least the second score. The information processing device (2000) performs second determination that determines whether the input voice data (10) includes voice of the registrant (20) on the basis of the first score, the second score, and the result of the first determination.

(57) 要約:情報処理装置(２０００)は、入力音声データ(１０)と、登録者(２０)の登録者音声データ(２２)との類似度合いを表す第１スコアを算出する。情報処理装置(２０００)は、入力音声データ(１０)を時間方向に分割することで、複数のセグメント音声データ(１２)を得る。情報処理装置(２０００)は、各セグメント音声データ(１２)について、セグメント音声データ(１２)と登録者音声データ(２２)との類似度合いを表す第２スコアを算出する。情報処理装置２０００は、少なくとも第２スコアを用いて、入力音声データ(１０)に含まれる音声の話者が、一人と複数のいずれであるかを判定する第１判定を行う。情報処理装置(２０００)は、第１スコア、第２スコア、及び第１判定の結果に基づいて、入力音声データ(１０)に登録者(２０)の音声が含まれるか否かを判定する第２判定を行う。

# 明　細　書

## 発明の名称 ： 情報処理装置、制御方法、及びプログラム

## 技術分野

[0001] 本発明は音声データに含まれる音声の話者を認識する技術に関する。

## 背景技術

[0002] 入力された音声からその話者を認識する技術が開発されている。例えば特許文献１には、入力された音声信号の特徴量を算出し、算出した特徴量を話者モデルに入力することで話者スコアを算出し、算出した話者スコアに基づいて話者を特定する技術が開示されている。

[0003] ここで、入力音声には、任意の環境下で収録した音声を用いられることが多いため、認識対象の話者の音声以外の音声も含まれうる。このような認識対象の話者の音声以外の音声が入力音声に含まれることにより、話者認識の精度が低下してしまう。

[0004] そこで、認識対象の話者の音声以外の音声が含まれる入力音声を対象として、話者認識の精度を向上させるための技術が開発されている。例えば非特許文献1には、背景雑音に頑健なスコア正規化手法が記載されている。この手法は、認識音声を音声区間と非音声区間の二つの部分に分ける。さらに、認識音声の SNR (signal noise ratio) 、登録音声のSNR、及び話者認識を行う際に求めたスコアを用いて、スコア正規化を行う。そして、正規化したスコアを用いて、認識音声の話者が登録者であるか否かが判定される。

[0005] ここで、認識音声の SNR は、主に非音声区間に含まれている雑音（すなわち、背景雑音）の量を表しており、SNR が低ければ低いほどスコアが小さくなる。このように背景雑音を考慮して話者認識を行うことで、雑音に対して頑健な話者認識を実現している。

## 先行技術文献

## 特許文献

[0006] 特許文献1：国際公開第２００８／１１７６２６号

## 非特許文献

[0007] 非特許文献1：Harmse Jorgen、Steven D. Beck、及び Hirotaka Nakasone、「Speaker recognition score-normalization to compensate for snr and duration」、Speaker and Language Recognition Workshop、2006年

非特許文献2：Ajmera Jitendra、Iain McCowan、及び Herve Bourlard、「Robust speaker change detection」、IEEE Signal Processing Letters、2004年.

## 発明の概要

## 発明が解決しようとする課題

[0008] 認識音声には、背景雑音だけでなく、認識対象の話者以外の他者の音声も混在しうる。非特許文献１の手法では、このような他者の音声が混在する入力音声について認識精度を向上させることが難しい。なぜなら、非音声区間に含められる背景雑音とは異なり、他者の音声は認識対象の話者の音声と共に音声区間に含まれてしまうため、他者の音声の混入具合を上述した SNR で表現できないためである。

[0009] 本発明は上述した課題に鑑みてなされたものであり、その目的の一つは、認識対象の話者以外の人の音声が入力音声に含まれるケースについて、話者認識の精度を向上させる技術を提供することである。

## 課題を解決するための手段

[0010] 本発明の情報処理装置は、１）入力音声データと、登録者の音声データである登録者音声データとの類似度合いを表す第１スコアを算出する第１算出部と、２）入力音声データを時間方向に分割することにより、入力音声データを複数のセグメント音声データに分け、各セグメント音声データについて、セグメント音声データと登録者音声データとの類似度合いを表す第２スコアを算出する第２算出部と、３）少なくとも第２スコアを用いて、入力音声データに含まれる音声の話者が一人と複数のいずれであるかを判定する第１判定部と、４）第１スコア、第２スコア、及び第１判定部による判定結果に

基づいて、入力音声データに登録者の音声が含まれるか否かを判定する第２判定部と、を有する。

[0011] 本発明の制御方法は、コンピュータによって実行される。当該制御方法は、１）入力音声データと、登録者の音声データである登録者音声データとの類似度合いを表す第１スコアを算出する第１算出ステップと、２）入力音声データを時間方向に分割することにより、入力音声データを複数のセグメント音声データに分け、各セグメント音声データについて、セグメント音声データと登録者音声データとの類似度合いを表す第２スコアを算出する第２算出ステップと、３）少なくとも第２スコアを用いて、入力音声データに含まれる音声の話者が一人と複数のいずれであるかを判定する第１判定ステップと、４）第１スコア、第２スコア、及び第１判定ステップによる判定結果に基づいて、入力音声データに登録者の音声が含まれるか否かを判定する第２判定ステップと、を有する。

[0012] 本発明のプログラムは、コンピュータに、本発明の制御方法が有する各ステップを実行させる。

## 発明の効果

[0013] 本発明によれば、認識対象の話者以外の人の音声が入力音声に含まれるケースについて、話者認識の精度を向上させる技術が提供される。

## 図面の簡単な説明

[0014] 上述した目的、およびその他の目的、特徴および利点は、以下に述べる好適な実施の形態、およびそれに付随する以下の図面によってさらに明らかになる。

[0015] [図1]本実施形態の情報処理装置が行う処理を概念的に示す図である。

[図2]実施形態１の情報処理装置の機能構成を例示する図である。

[図3]情報処理装置を実現するための計算機を例示する図である。

[図4]実施形態１の情報処理装置によって実行される処理の流れを例示するフローチャートである。

[図5]所定長に分割された入力音声データを例示する図である。

[図6]第１スコアと第２スコアをグラフで例示する図である。

[図7]第２スコアのヒストグラムを例示する図である。

## 発明を実施するための形態

[0016] 以下、本発明の実施の形態について、図面を用いて説明する。尚、すべての図面において、同様な構成要素には同様の符号を付し、適宜説明を省略する。また、特に説明する場合を除き、各ブロック図において、各ブロックは、ハードウエア単位の構成ではなく、機能単位の構成を表している。

[0017] ［実施形態１］

＜概要＞

図１は、本実施形態の情報処理装置２０００が行う処理の概要を概念的に示す図である。情報処理装置２０００は、入力音声データ１０に含まれる音声の話者の認識を行う。そのために、情報処理装置２０００は、入力音声データ１０と、登録者２０（図示せず）の音声を表す登録者音声データ２２との比較を行う。以下、情報処理装置２０００の動作をより具体的に説明する。

[0018] まず情報処理装置２０００は、入力音声データ１０と登録者音声データ２２との類似度合いを表す第１スコアを算出する。さらに情報処理装置２０００は、入力音声データ１０を時間方向に分割することで、複数のセグメント音声データ１２を得る。そして、情報処理装置２０００は、各セグメント音声データ１２について、セグメント音声データ１２と登録者音声データ２２との類似度合いを表す第２スコアを算出する。

[0019] 情報処理装置２０００は、少なくとも第２スコアを用いて、入力音声データ１０に含まれる音声の話者が、一人と複数のいずれであるかを判定する第１判定を行う。ただし、この判定には、第１スコアがさらに利用されてもよい。図１では、第１判定に第１スコア及び第２スコアが利用されるケースを例示している。そして、情報処理装置２０００は、第１スコア、第２スコア、及び第１判定の結果に基づいて、入力音声データ１０に登録者２０の音声が含まれるか否かを判定する第２判定を行う。

[0020] ここで、入力音声データ１０に含まれる音声の話者が複数であると判定された場合、情報処理装置２０００は、少なくとも第２スコアを用いて補正スコアを算出し、算出した補正スコアを閾値と比較することにより、第２判定を行う。一方、入力音声データ１０に含まれる音声の話者が一人であると判定された場合、情報処理装置２０００は、第１スコアを閾値と比較することにより、第２判定を行う。いずれの場合も、スコアが閾値以上であれば、入力音声データ１０に登録者２０の音声が含まれると判定され、スコアが閾値未満であれば、入力音声データ１０に登録者２０の音声が含まれないと判定される。

[0021] ＜作用効果＞

本実施形態の情報処理装置２０００によれば、入力音声データ１０を分割することで得られる複数のセグメント音声データ１２それぞれについて、登録者音声データ２２との類似度を表す第２スコアが算出され、少なくとも第２スコアを用いて、入力音声データ１０に含まれる音声の話者が一人と複数のいずれであるかが判定される。そして、この判定結果を利用して、入力音声データ１０に登録者２０の音声が含まれるか否かが判定される。このように、入力音声データ１０に含まれる音声の話者が一人と複数のいずれであるかを判定することにより、入力音声データ１０に登録者２０以外の人の音声も含まれているか否かを考慮して、入力音声データ１０に登録者２０の音声が含まれるか否かを判定することができるようになる。よって、情報処理装置２０００によれば、入力音声データ１０に登録者２０以外の人の音声も含まれているケースについて、話者認識の精度を向上させることができる。

[0022] より具体的には、入力音声データ１０に複数の人物の音声が含まれている場合において、補正スコアが算出され、その補正スコアを用いて、入力音声データ１０に登録者２０の音声が含まれるか否かが判定される。このように、複数の人物の音声が含まれる入力音声データ１０について、登録者２０の音声が含まれるか否かの判定を、入力音声データ１０全体について算出したスコア（すなわち、第１スコア）をそのまま用いて行うのではなく、補正し

たスコアを利用して行うようにすることで、より高い精度で判定が行えるようになる。

[0023] 情報処理装置２０００を利用した話者認識は、様々な場面で利用することができる。例えば、音声データを用いた生体認証に利用することが考えられる。具体的には、認証を行いたい人物が発した声を録音することで生成された音声データを利用して、話者認識を行う。

[0024] ここで、生体認証には高い精度が要求される。また、生体認証が行われる場所には、認識対象の人物以外の人物も存在する蓋然性が高い。

[0025] 本実施形態の情報処理装置２０００によれば、認識対象の人物が発した音声を録音した音声データに、その人物以外の人物の音声が混入してしまったとしても、高い精度で話者認識を行うことができる。よって、認識対象の人物以外の人物が存在する環境においても、音声データを用いた生体認証を高い精度で実現することができる。

[0026] なお、図１を参照した上述の説明は、情報処理装置２０００の理解を容易にするための例示であり、情報処理装置２０００の機能を限定するものではない。以下、本実施形態の情報処理装置２０００についてさらに詳細に説明する。

[0027] ＜情報処理装置２０００の機能構成の例＞

図２は、実施形態１の情報処理装置２０００の機能構成を例示する図である。情報処理装置２０００は、第１算出部２０２０、第２算出部２０４０、第１判定部２０６０、及び第２判定部２０８０を有する。第１算出部２０２０は、入力音声データ１０と登録者音声データ２２との類似度合いを表す第１スコアを算出する。第２算出部２０４０は、入力音声データ１０を複数のセグメント音声データ１２に分割し、各セグメント音声データ１２について、登録者音声データ２２との類似度合いを表す第２スコアを算出する。なお、入力音声データ１０は、時間方向に分割される。第１判定部２０６０は、少なくとも第２スコアを用いて、入力音声データ１０に含まれる音声の話者が一人と複数のいずれであるかを判定する。第２判定部２０８０は、第１ス

コア、第２スコア、及び第１判定部２０６０による判定の結果に基づいて、入力音声データ１０に登録者２０の音声が含まれるか否かを判定する。

[0028] ＜情報処理装置２０００のハードウエア構成＞

情報処理装置２０００の各機能構成部は、各機能構成部を実現するハードウエア（例：ハードワイヤードされた電子回路など）で実現されてもよいし、ハードウエアとソフトウエアとの組み合わせ（例：電子回路とそれを制御するプログラムの組み合わせなど）で実現されてもよい。以下、情報処理装置２０００の各機能構成部がハードウエアとソフトウエアとの組み合わせで実現される場合について、さらに説明する。

[0029] 図３は、情報処理装置２０００を実現するための計算機１０００を例示する図である。計算機１０００は任意の計算機である。例えば計算機１０００は、Personal Computer（PC）やサーバマシンなどである。計算機１０００は、情報処理装置２０００を実現するために設計された専用の計算機であってもよいし、汎用の計算機であってもよい。

[0030] 計算機１０００は、バス１０２０、プロセッサ１０４０、メモリ１０６０、ストレージデバイス１０８０、入出力インタフェース１１００、及びネットワークインタフェース１１２０を有する。バス１０２０は、プロセッサ１０４０、メモリ１０６０、ストレージデバイス１０８０、入出力インタフェース１１００、及びネットワークインタフェース１１２０が、相互にデータを送受信するためのデータ伝送路である。ただし、プロセッサ１０４０などを互いに接続する方法は、バス接続に限定されない。

[0031] プロセッサ１０４０は、CPU（Central Processing Unit）、GPU（Graphics Processing Unit）、FPGA（Field-Programmable Gate Array）などの種々のプロセッサである。メモリ１０６０は、RAM（Random Access Memory）などを用いて実現される主記憶装置である。ストレージデバイス１０８０は、ハードディスク、SSD（Solid State Drive）、メモリカード、又は ROM（Read Only Memory）などを用いて実現される補助記憶装置である。

[0032] 入出力インタフェース１１００は、計算機１０００と入出力デバイスとを

接続するためのインタフェースである。例えば入出力インタフェース１１００には、キーボードなどの入力装置や、ディスプレイ装置などの出力装置が接続される。ネットワークインタフェース１１２０は、計算機１０００を通信網に接続するためのインタフェースである。この通信網は、例えば LAN（Local Area Network）や WAN（Wide Area Network）である。ネットワークインタフェース１１２０が通信網に接続する方法は、無線接続であってもよいし、有線接続であってもよい。

[0033] ストレージデバイス１０８０は、情報処理装置２０００の各機能構成部を実現するプログラムモジュールを記憶している。プロセッサ１０４０は、これら各プログラムモジュールをメモリ１０６０に読み出して実行することで、各プログラムモジュールに対応する機能を実現する。

[0034] ストレージデバイス１０８０は、登録者音声データ２２をさらに記憶していてもよい。ただし、登録者音声データ２２は、計算機１０００から取得可能な情報であればよく、ストレージデバイス１０８０に記憶されていなければならないものではない。例えば登録者音声データ２２は、ネットワークインタフェース１１２０を介して計算機１０００と接続されているデータベースサーバに記憶させておくことができる。

[0035] また、登録者音声データ２２そのものではなく、登録者音声データ２２から抽出される特徴量を記憶装置に記憶させておいてもよい。この場合、登録者音声データ２２は、情報処理装置２０００から取得可能でなくてもよい。

[0036] ＜処理の流れ＞

図４は、実施形態１の情報処理装置２０００によって実行される処理の流れを例示するフローチャートである。第１算出部２０２０は、入力音声データ１０を取得する（Ｓ１０２）。第２算出部２０４０は第１スコアを算出する（Ｓ１０４）。第２算出部２０４０は、入力音声データ１０を複数のセグメント音声データ１２に分割する（Ｓ１０６）。第２算出部２０４０は、各セグメント音声データ１２について第２スコアを算出する（Ｓ１０８）。第１判定部２０６０は、第１判定（入力音声データ１０に含まれる音声の話者

が一人と複数のいずれであるかの判定）を行う（Ｓ１１０）。第２判定部２０８０は、第２判定（入力音声データ１０に登録者２０の音声が含まれるか否かの判定）を行う（Ｓ１１２）。

[0037] ＜入力音声データ１０の取得：Ｓ１０２＞

第１算出部２０２０は入力音声データ１０を取得する（Ｓ１０２）。入力音声データ１０は、話者認識の対象となる音声データである。第１算出部２０２０が入力音声データ１０を取得する方法は任意である。例えば第１算出部２０２０は、入力音声データ１０が記憶されている記憶装置から入力音声データ１０を取得する。入力音声データ１０が記憶されている記憶装置は、情報処理装置２０００の内部に設けられていてもよいし、外部に設けられていてもよい。その他にも例えば、第１算出部２０２０は、他の装置によって送信される入力音声データ１０を受信することで、入力音声データ１０を取得する。

[0038] なお、後述するように、第１スコアや第２スコアの算出には、入力音声データ１０から抽出される特徴量を利用する。そこで第１算出部２０２０は、登録者音声データ２２を取得する代わりに、登録者音声データ２２から予め抽出しておいた特徴量を取得してもよい。この場合、登録者音声データ２２から抽出した特徴量を予め任意の記憶装置に記憶させておく。

[0039] ＜第１スコアの算出：Ｓ１０４＞

第１算出部２０２０は、入力音声データ１０と登録者音声データ２２との比較により、第１スコアの算出を行う（Ｓ１０４）。より具体的には、第１算出部２０２０は、入力音声データ１０と登録者音声データ２２のそれぞれから抽出される特徴量の類似度を算出し、算出した類似度を第１スコアとする。

[0040] 第１スコアと第２スコアの算出に利用する特徴量には、音声データから抽出できる任意の特徴量を利用することができる。音声データから抽出できる特徴量は、例えば、声道情報を反映したスペクトルの包絡特性や、声帯情報を反映した基本周波数特性などの物理量を表す情報である。より具体的な例

としては、メル周波数ケプストラム係数（MFCC: Mel-Frequency Cepstrum Coefficients）を用いて算出した i-vector を利用できる。例えば、Probabilistic linear discriminant analysis（PLDA）により、i-vector 空間上で話者の識別に寄与しない情報を低減することにより、特徴量同士の類似度をより正確に表すスコアを算出することができる。なお、音声データから特徴量を抽出する具体的な技術、及び特徴量同士の類似度を算出する具体的な技術には、既存の技術を利用することができる。

[0041] ＜入力音声データ１０の分割：Ｓ１０６＞

第２算出部２０４０は、入力音声データ１０を時間方向に分割することで、入力音声データ１０を複数のセグメント音声データ１２に分ける（Ｓ１０６）。ここで、入力音声データ１０の分割の方法には、様々な方法を採用できる。以下、その方法の具体例を説明する。

[0042] ＜＜所定長の時間で分割する方法＞＞

例えば第２算出部２０４０は、入力音声データ１０を所定長（１０秒など）の音声データに分割することにより、入力音声データ１０を複数のセグメント音声データ１２に分ける。図５は、所定長に分割された入力音声データ１０を例示する図である。図５において、所定長、すなわちセグメント音声データ１２の長さは１０秒である。

[0043] ここで、図５（ｂ）に示されているように、隣接するセグメント音声データ１２同士は、それらの一部が互いにオーバーラップするように分割されてもよい。図５（ｂ）において、隣接する２つのセグメント音声データ１２は、互いに３秒間オーバーラップしている。

[0044] また、図５（ｃ）に示されているように、隣接するセグメント音声データ１２同士が、時間方向で離れていてもよい。図５（ｃ）において、隣接する２つのセグメント音声データ１２は、３秒間離れている。

[0045] ＜＜話者交換点で分割する方法＞＞

例えば第２算出部２０４０は、入力音声データ１０について話者交換点を検出し、話者交換点で入力音声データ１０を区切ることで、入力音声データ

１０を複数のセグメント音声データ１２に分割してもよい。話者交換点を検出する技術には、非特許文献２記載の技術などを利用することができる。

[0046] ＜第２スコアの算出：Ｓ１０８＞

第２算出部２０４０は、各セグメント音声データ１２について第２スコアを算出する（Ｓ１０８）。そのために第２算出部２０４０は、各セグメント音声データ１２から特徴量を抽出する。そして第２算出部２０４０は、セグメント音声データ１２から抽出された特徴量と、登録者音声データ２２から抽出された特徴量との類似度を算出し、算出された類似度を、そのセグメント音声データ１２の第２スコアとする。

[0047] ＜第１判定：Ｓ１１０＞

第１判定部２０６０は、少なくとも第２スコアを用いて、入力音声データ１０に含まれる音声の話者が一人と複数のいずれであるかを判定する（Ｓ１１０）。ただし前述したように、この判定には、第１スコアをさらに利用してもよい。例えば第１判定部２０６０は、第１スコアを第２スコアの最大値と比較する。具体的には、第１判定部２０６０は、第２算出部２０４０によって算出された複数の第２スコアの中から最大値を特定し、第１スコアがその最大値よりも小さければ、入力音声データ１０に含まれる音声の話者が複数であると判定する。一方、第１判定部２０６０は、第１スコアが第２スコアの最大値以上であれば、入力音声データ１０に含まれる音声の話者が一人であると判定する。

[0048] 以下、図６を用いて、上述した判定の根拠について説明する。図６は、第１スコアと第２スコアをグラフで例示する図である。図６の上段は、入力音声データ１０に登録者の音声のみが含まれているケースを示している。一方、図６の下段は、入力音声データ１０に登録者以外の音声が含まれているケースを例示している。

[0049] 一般的に、特徴量同士の類似度を表すスコアは、入力音声の長さの影響を受ける。具体的には、特徴量の抽出に用いられる情報の量が、入力音声が短くなるほど少なくなるため、入力音声が短いほど、抽出される特徴量の正確

性（特徴量が話者の特徴を表す度合い）が低下する。このことから、入力音声データ１０に登録者２０の音声しか含まれていなければ、第１スコアは、どの第２スコアよりも大きくなる。すなわち、第１スコアは第２スコアの最大値よりも大きくなる（図６上段参照）。

[0050] 一方、入力音声データ１０に登録者２０の音声以外の人物の音声も含まれていると、第２スコアが第１スコアよりも大きくなることがある（図６下段参照）。これは、入力音声データ１０全体には登録者２０以外の人物の音声が含まれていても、入力音声データ１０の一部であるセグメント音声データ１２の中には、登録者２０以外の人物の音声をほとんど含まないものが存在しうるためである。このようなセグメント音声データ１２から抽出される特徴量は、入力音声データ１０から抽出される特徴量と比較し、登録者音声データ２２から抽出される特徴量との類似度が高いと考えられる。そのため、第１スコアよりも大きい第２スコアが存在しうることとなる。すなわち、第２スコアの最大値が第１スコアよりも大きくなりうる。

[0051] 以上のことから、第１スコアが第２スコアの最大値よりも小さい場合には、入力音声データ１０に登録者２０の音声以外の人物の音声も含まれている蓋然性が高いと言える。そこで前述した様に、第１判定部２０６０は、第１スコアが第２スコアの最大値よりも小さければ、入力音声データ１０に含まれる音声の話者が複数であると判定する。

[0052] ただし、入力音声データ１０に含まれる音声の話者が一人と複数のいずれであるかを判定する方法は、第１スコアと第２スコアの最大値とを比較する方法に限定されない。例えば第１判定部２０６０は、複数の第２スコアの値のばらつきの大きさを表す指標値を算出し、その指標値を所定の閾値と比較することで、入力音声データ１０に含まれる音声の話者が一人と複数のいずれであるかを判定する。具体的には、第１判定部２０６０は、算出した指標値が所定の閾値以上であれば、入力音声データ１０に含まれる音声の話者が複数であると判定し、算出した指標値が所定の閾値未満であれば、入力音声データ１０に含まれる音声の話者が一人であると判定する。ここで、複数の

第２スコアの値のばらつきの大きさを表す指標値には、第２スコアの最大値と最小値の差分、第２スコアの分散、第２スコアの標準偏差などの値を利用することができる。

[0053] その他にも例えば、入力音声データ１０から算出された第１スコア及び第２スコアが入力されたことに応じて、その入力音声データ１０に含まれる音声の話者が一人と複数のいずれであるかを判定する予測モデルを予め構築しておいてもよい。このような予測モデルには、サポートベクタマシン、ニューラルネットワーク、及び線形分類器など、分類を実現する種々のモデルを利用することができる。第１判定部２０６０は、第１算出部２０２０によって算出された第１スコア、及び第２算出部２０４０によって算出された第２スコアを、学習済みの予測モデルに入力する。これにより、予測モデルの出力として、入力音声データ１０に含まれる音声の話者が一人と複数のいずれであるかを判定した結果が得られる。

[0054] 予測モデルの学習は、話者の人数が既知である音声データから生成される学習データを用いて、予め実行しておく。具体的には、話者の人数が既知である音声データ全体について第１スコアを算出し、なおかつその音声データを分割することで得られる複数のセグメント音声データそれぞれについて第２スコアを算出する。そして、「既知の話者人数、算出した第１スコア、算出した第２スコア」の組み合わせを教師データとして利用して、予測モデルの学習を行う。なお、予測モデルの学習に利用する音声には、登録者２０の音声が含まれている必要はない。

[0055] なお、予測モデルは、第１スコアを利用せず、第２モデルのみを利用するように構築してもよい。すなわち、入力音声データ１０から算出された第２スコアが入力されたことに応じて、その入力音声データ１０に含まれる音声の話者が一人と複数のいずれであるかを判定する予測モデルを構築しておく。採用可能なモデルの種類については、第１スコアを利用するケースと同様である。また、予測モデルの学習には、話者の人数が既知である音声データを分割することで得られる複数のセグメント音声データそれぞれについて算

出した第２スコアと、既知の話者人数とを対応づけた学習データを利用する。

[0056] ＜第２判定：Ｓ１１２＞

第２判定部２０８０は第２判定を行う（Ｓ１１２）。具体的には、第２判定部２０８０は、第１スコア、第２スコア、及び第１判定の結果に基づいて、入力音声データ１０に登録者２０の音声が含まれるか否かを判定する（Ｓ１１２）。ここで、第２判定の具体的な方法は、第１判定の結果によって異なる。以下、第１判定の結果ごとに、第２判定の具体的な方法を説明する。

[0057] ＜＜入力音声データ１０に含まれる音声の話者が一人である場合＞＞

入力音声データ１０に含まれる音声の話者が一人であると判定された場合、第２判定部２０８０は、第１スコアを閾値と比較する。第１スコアが閾値以上である場合、第２判定部２０８０は、入力音声データ１０に登録者２０の音声が含まれると判定する。一方、第１スコアが閾値未満である場合、第２判定部２０８０は、入力音声データ１０に登録者２０の音声が含まれないと判定する。この閾値は、情報処理装置２０００からアクセス可能な記憶装置に予め記憶させておく。

[0058] ＜＜入力音声データ１０に含まれる音声の話者が複数である場合＞＞

入力音声データ１０に含まれる音声の話者が複数であると判定された場合、第２判定部２０８０は、少なくとも第２スコアを用いて補正スコアを算出し、算出した補正スコアを上記閾値と比較する。補正スコアが閾値以上である場合、第２判定部２０８０は、入力音声データ１０に登録者２０の音声が含まれると判定する。一方、第１スコアが閾値未満である場合、第２判定部２０８０は、入力音声データ１０に登録者２０の音声が含まれないと判定する。

[0059] 補正スコアの算出方法には、様々な方法を採用できる。以下、補正スコアの算出方法を例示する。

[0060] ＜＜補正スコアの算出方法１＞＞

例えば第２判定部２０８０は、第２スコアの定義域を分割した複数の部分

範囲それぞれに含まれる第２スコアの数を表すヒストグラムを生成し、このヒストグラムを用いて補正スコアを算出する。例えば、第２スコア S2 の定義域が「0≦S2≦100」である場合において、この定義域が１０個の部分範囲（「0≦S2＜10」、・・・、「80≦S2＜90」、「90≦S2≦100」）に等分される。第２判定部２０８０は、これらの部分範囲それぞれについて、セグメント音声データ１２から算出された第２スコアの数を算出する。

[0061] 図７は、第２スコアのヒストグラムを例示する図である。図７の例では、前述した例の様に、第２スコア S2 の定義域が「0≦S2≦100」であり、この定義域が１０等分されている。そして、各部分範囲における第２スコアの数がグラフで表されている。

[0062] 第２判定部２０８０は、上述のヒストグラムの中から、ピークを示す部分範囲を１つ以上特定し、なおかつ特定した部分範囲の中で第２スコアが最大であるものを特定する。ここで特定された部分範囲を、注目範囲と呼ぶ。例えば図７の例において、ピークを示す部分範囲は「20≦S2＜30」と「60≦S2＜70」の２つである。このうち、第２スコアが最大である部分範囲は「60≦S2＜70」である。そこで、「60≦S2＜70」が注目範囲として特定される。

[0063] 第２スコアのヒストグラムにおいてピークを示す部分範囲では、セグメント音声データ１２に含まれている音声の発話者が主に一人であると考えられる。特に、会話では話者が交替で話すことが多いため、会話を収録した音声（電話の録音など）から抽出されたセグメント音声では、１つのセグメント音声に含まれる話者が一人であることが多くなる。そして、主な発話者が登録者２０であるセグメント音声データ１２では、主な発話者が登録者２０以外の人物であるセグメント音声データ１２と比較し、算出される第２スコアが高くなると考えられる。そのため、ピークを示す数値範囲のうちで第２スコアが最大である数値範囲、すなわち注目範囲に含まれるのは、主な発話者が登録者２０であるセグメント音声データ１２について算出された第２スコアとなる。

[0064] そこで第２判定部２０８０は、注目範囲を利用して補正スコアを算出する

。例えば第２判定部２０８０は、注目範囲に含まれる第２スコアの統計値（最大値や平均値など）を補正スコアとする。

[0065] その他にも例えば、第２判定部２０８０は、注目範囲に含まれるセグメント音声データ１２を結合して１つの音声データを生成し、生成した音声データから抽出される特徴量と、登録者音声データ２２から抽出された特徴量との類似度を、補正スコアとして算出する。一般に、音声データの長さが長い方が、精度の良いスコアを算出することができる。そのため、注目範囲に含まれるセグメント音声データ１２を結合して、セグメント音声データ１２よりも長い音声データを生成し、この音声データについてスコアを算出することにより、セグメント音声データ１２について算出されたスコアよりも精度の良いスコアを得ることができる。よって、このように算出したスコアを補正スコアとすることで、より精度の良いスコアを用いて、入力音声データ１０に登録者２０の音声が含まれるか否かを判定することができる。

[0066] ここで、第２判定部２０８０は、上述の様に結合するセグメント音声データ１２に、注目範囲に含まれるセグメント音声データ１２だけでなく、注目範囲よりも第２スコアが大きい各部分範囲に含まれるセグメント音声データ１２を含めるようにしてもよい。言い換えれば、第２判定部２０８０は、算出された第２スコアが注目範囲の下限値以上である全てのセグメント音声データ１２を結合して、１つの音声データを生成するようにする。例えば図７の例では、第２スコアが６０以上である８個のセグメント音声データ１２を結合して１つの音声データを生成し、この音声データについて算出する登録者音声データ２２との類似度を補正スコアとする。

[0067] ＜＜補正スコアの算出方法２＞＞

その他にも例えば、第１スコアと第２スコアを入力として受け付けて補正スコアを出力する予測モデルを用意しておいてもよい。第２判定部２０８０は、第１算出部２０２０によって算出された第１スコアと、第２算出部２０４０によって算出された第２スコアを予測モデルに入力することで、補正スコアを得る。

[0068] ここでは、予測モデルとして、１）全ての第２スコアの分布の中から、登録者２０である確率が最も高い話者のセグメント音声から得た第２スコアの分布（前述した注目範囲を含む分布）を抽出する予測モデルと、２）抽出した分布に基づいて補正スコアを算出する予測モデルという２つを用いる。

[0069] 前者の予測モデルには、例えば、ガウス混合モデル（GMM: Gaussian Mixture Model）を利用することができる。第２判定部２０８０は、第２算出部２０４０によって算出された複数の第２スコアを用いて GMM を構築する。GMM を利用することで、入力音声データ１０から得られた全ての第２スコアの分布を、複数のガウス分布に分割することができる。そして、これら複数のガウス分布のうち、第２スコアの平均値が最大である分布が、前述した注目範囲を含む分布であると考えられる。そこで第２判定部２０８０は、GMM を利用して得られる複数のガウス分布の中から、第２スコアの平均値が最大であるガウス分布を抽出する。なお、GMM の構築には、EM（Expectation Maximization） や MAP（Maximum A Posteriori） などの既知のアルゴリズムを利用できる。

[0070] 第２スコアの分布に基づいて補正スコアを算出する予測モデルには、サポートベクトル回帰（SVR: Support Vector Regression）やニューラルネットワークなど、回帰を実現する種々の予測モデルを利用することができる。この予測モデルには、話者が一人である音声において、第２スコアの分布と第１スコアとがどのように対応するかを学習させる。このような学習をすることで、予測モデルが、第２スコアの分布が入力されたことに応じて、その分布に対応すると予測される第１スコアを、補正スコアとして出力するようにする。

[0071] この予測モデルの学習に利用する学習データは、話者が一人である任意の音声を利用して生成できる。具体的には、話者が一人である音声データ全体について、第１スコアを算出する。また、その音声データを複数のセグメント音声データに分割し、各セグメント音声データについて第２スコアを算出する。こうすることで、話者が一人である音声における、第１スコアと複数

の第２スコア（第２スコアの分布）との対応関係を得ることができる。そこで、算出された第１スコアと複数の第２スコアとの対応を予測モデルに学習させる。このような学習により、予測モデルが、第２スコアの分布が入力されたことに応じて、対応する第１スコアを出力することができるようになる。

[0072] ＜登録者音声データ２２について＞

上述の説明では、入力音声データ１０との比較に用いる登録者音声データ２２が１つに特定されている。このように入力音声データ１０と比較すべき登録者音声データ２２が１つに特定できるケースとしては、例えば、登録者２０を特定する識別子（ユーザＩＤなど）の入力を別途受け付けるケースが考えられる。具体的には、情報処理装置２０００は、登録者２０を特定する識別子（例えば、文字列）の入力を受け付け、受け付けた識別子に対応づけて記憶装置に記憶されている登録者音声データ２２を取得する。そして、情報処理装置２０００は、この登録者音声データ２２を用いて、上述した一連の話者認識処理（図４のフローチャートに示した処理）を行う。このような話者認識は、例えば、ユーザＩＤとパスワードのペアを用いてユーザ認証を行う代わりに、ユーザＩＤとユーザの音声のペアを用いてユーザ認証を行うケースに利用できる。

[0073] 一方で、入力音声データ１０との比較に用いる登録者音声データ２２は、１つに特定されていなくてもよい。例えば情報処理装置２０００は、登録者音声データ２２が複数記憶されている記憶装置から１つずつ登録者音声データ２２を取得し、取得した登録者音声データ２２に対応する登録者２０について、上述した一連の話者認識処理を行う。

[0074] 取得した登録者音声データ２２について行った話者認識処理において、入力音声データ１０に登録者２０の音声が含まれていると判定されたとする。この場合、情報処理装置２０００は、話者認識処理を終了する。この場合、処理対象とした登録者音声データ２２に対応する登録者２０の音声が、入力音声データ１０に含まれていたと判定される。一方、取得した登録者音声デ

ータ２２について行った話者認識処理において、入力音声データ１０に登録者２０の音声が含まれていないと判定されたとする。この場合、情報処理装置２０００は、登録者音声データ２２が記憶されている記憶装置から、次の登録者音声データ２２を取得し、その登録者音声データ２２を対象として話者認識処理を行う。このような話者認識は、例えば、ユーザＩＤとパスワードのペアを用いてユーザ認証を行う代わりに、ユーザの音声のみを用いてユーザ認証を行うケースに利用できる。

[0075] ＜判定結果の出力＞

第２判定部２０８０は第２判定の結果、すなわち入力音声データ１０に登録者２０の音声が含まれているか否かを示す情報を出力してもよい。第２判定の結果の出力方法には、様々な方法を採用できる。例えば第２判定部２０８０は、第２判定の結果を表す情報を情報処理装置２０００に接続されているディスプレイ装置に出力する。その他にも例えば、第２判定部２０８０は、第２判定の結果を表す情報を情報処理装置２０００に接続されている記憶装置に記憶させてもよい。

[0076] 第２判定の結果を表す情報は、例えば、「入力音声データ１０に登録者２０の音声が含まれている」という情報、又は「入力音声データ１０に登録者２０の音声が含まれていない」という情報を表す文字列、画像、又は音声などである。なお、第２判定部２０８０は、第２判定の結果を表す情報に加え、入力音声データ１０に含まれている音声の話者が一人と複数のどちらであるかを示す情報（すなわち、第１判定の結果を表す情報）や、閾値と比較したスコア（第１スコア又は補正スコア）を示す情報を出力してもよい。こうすることで、情報処理装置２０００の利用者は、入力音声データ１０に登録者２０の音声が含まれているか否かという判定の結果だけでなく、その判定の根拠も把握することができる。

[0077] また、入力音声データ１０と比較する登録者音声データ２２が１つに特定されておらず、複数の登録者音声データ２２それぞれについて順次入力音声データ１０と比較する場合、情報処理装置２０００は、入力音声データ１０

に音声が含まれている登録者を特定する情報（例えば、登録者の識別子）を出力してもよい。

[0078] 以上、図面を参照して本発明の実施形態について述べたが、これらは本発明の例示であり、上記各実施形態の構成を組み合わせた構成や、上記以外の様々な構成を採用することもできる。

# 請求の範囲

[請求項1] 入力音声データと、登録者の音声データである登録者音声データとの類似度合いを表す第１スコアを算出する第１算出部と、

前記入力音声データを時間方向に分割することにより、前記入力音声データを複数のセグメント音声データに分け、各前記セグメント音声データについて、前記セグメント音声データと前記登録者音声データとの類似度合いを表す第２スコアを算出する第２算出部と、

少なくとも前記第２スコアを用いて、前記入力音声データに含まれる音声の話者が一人と複数のいずれであるかを判定する第１判定部と、

前記第１スコア、前記第２スコア、及び前記第１判定部による判定結果に基づいて、前記入力音声データに前記登録者の音声が含まれるか否かを判定する第２判定部と、を有する情報処理装置。

[請求項2] 前記入力音声データに含まれる音声の話者が複数であると判定された場合、前記第２判定部は、少なくとも前記第２スコアを用いて補正スコアを算出し、算出した補正スコアを閾値と比較することで、前記入力音声データに前記登録者の音声が含まれるか否かを判定し、

前記入力音声データに含まれる音声の話者が一人であると判定された場合、前記第２判定部は、前記第１スコアを閾値と比較することで、前記入力音声データに前記登録者の音声が含まれるか否かを判定する、請求項１に記載の情報処理装置。

[請求項3] 前記第２判定部は、

前記第２スコアの定義域に含まれる複数の部分範囲それぞれに含まれる第２スコアの数の分布を生成し、

前記分布におけるピークに対応する部分範囲のうちで前記第２スコアが最大の部分範囲である注目範囲を特定し、

前記注目範囲に含まれる第２スコアを用いて前記補正スコアを算出する、請求項２に記載の情報処理装置。

[請求項4] 前記第２判定部は、前記注目範囲に含まれる第２スコアの統計値を前記補正スコアとして算出する、請求項３に記載の情報処理装置。

[請求項5] 前記第２判定部は、

前記注目範囲に含まれる第２スコアが算出された複数の前記セグメント音声データを結合して１つの音声データを生成するか、又は、前記注目範囲の下限値以上の第２スコアが算出された複数の前記セグメント音声データを結合して１つの音声データを生成し、

前記生成された音声データと前記登録者音声データとの類似度を前記補正スコアとして算出する、請求項３に記載の情報処理装置。

[請求項6] 前記第２判定部は、前記第２スコアが入力されたことに応じて補正スコアを出力するように学習されている予測モデルに対して、前記第２算出部によって算出された各前記第２スコアを入力することにより、前記補正スコアを算出する、請求項２に記載の情報処理装置。

[請求項7] 前記第１判定部は、前記第１スコアが前記第２スコアの最大値よりも小さい場合に、前記入力音声データに含まれる音声の話者が複数であると判定する、請求項１乃至６いずれか一項に記載の情報処理装置。

[請求項8] 前記第１判定部は、複数の前記第２スコアのばらつきを表す指標値を算出し、前記算出した指標値が閾値以上である場合に、前記入力音声データに含まれる音声の話者が複数であると判定する、請求項１乃至６いずれか一項に記載の情報処理装置。

[請求項9] 前記第１判定部は、学習済みの予測モデルに対し、第２スコアのみ又は第１スコア及び第２スコアを入力することで、前記入力音声データに含まれる音声の話者が一人と複数のいずれであるかを判定し、

前記予測モデルは、前記第２スコアが入力されたこと、又は前記第１スコア及び前記第２スコアが入力されたことに応じて、前記入力音声データに含まれる音声の話者が一人と複数のいずれであるかを判定するように学習されている、請求項１乃至６いずれか一項に記載の情

報処理装置。

[請求項10] コンピュータによって実行される制御方法であって、

入力音声データと、登録者の音声データである登録者音声データとの類似度合いを表す第１スコアを算出する第１算出ステップと、

前記入力音声データを時間方向に分割することにより、前記入力音声データを複数のセグメント音声データに分け、各前記セグメント音声データについて、前記セグメント音声データと前記登録者音声データとの類似度合いを表す第２スコアを算出する第２算出ステップと、

少なくとも前記第２スコアを用いて、前記入力音声データに含まれる音声の話者が一人と複数のいずれであるかを判定する第１判定ステップと、

前記第１スコア、前記第２スコア、及び前記第１判定ステップによる判定結果に基づいて、前記入力音声データに前記登録者の音声が含まれるか否かを判定する第２判定ステップと、を有する制御方法。

[請求項11] 前記入力音声データに含まれる音声の話者が複数であると判定された場合、前記第２判定ステップにおいて、少なくとも前記第２スコアを用いて補正スコアを算出し、算出した補正スコアを閾値と比較することで、前記入力音声データに前記登録者の音声が含まれるか否かを判定し、

前記入力音声データに含まれる音声の話者が一人であると判定された場合、前記第２判定ステップにおいて、前記第１スコアを閾値と比較することで、前記入力音声データに前記登録者の音声が含まれるか否かを判定する、請求項１０に記載の制御方法。

[請求項12] 前記第２判定ステップにおいて、

前記第２スコアの定義域に含まれる複数の部分範囲それぞれに含まれる第２スコアの数の分布を生成し、

前記分布におけるピークに対応する部分範囲のうちで前記第２スコアが最大の部分範囲である注目範囲を特定し、

前記注目範囲に含まれる第２スコアを用いて前記補正スコアを算出する、請求項１１に記載の制御方法。

[請求項13] 前記第２判定ステップにおいて、前記注目範囲に含まれる第２スコアの統計値を前記補正スコアとして算出する、請求項１２に記載の制御方法。

[請求項14] 前記第２判定ステップにおいて、

前記注目範囲に含まれる第２スコアが算出された複数の前記セグメント音声データを結合して１つの音声データを生成するか、又は、前記注目範囲の下限値以上の第２スコアが算出された複数の前記セグメント音声データを結合して１つの音声データを生成し、

前記生成された音声データと前記登録者音声データとの類似度を前記補正スコアとして算出する、請求項１２に記載の制御方法。

[請求項15] 前記第２判定ステップにおいて、前記第２スコアが入力されたことに応じて補正スコアを出力するように学習されている予測モデルに対して、前記第２算出部によって算出された各前記第２スコアを入力することにより、前記補正スコアを算出する、請求項１１に記載の制御方法。

[請求項16] 前記第１判定ステップにおいて、前記第１スコアが前記第２スコアの最大値よりも小さい場合に、前記入力音声データに含まれる音声の話者が複数であると判定する、請求項１０乃至１５いずれか一項に記載の制御方法。

[請求項17] 前記第１判定ステップにおいて、複数の前記第２スコアのばらつきを表す指標値を算出し、前記算出した指標値が閾値以上である場合に、前記入力音声データに含まれる音声の話者が複数であると判定する、請求項１０乃至１５いずれか一項に記載の制御方法。

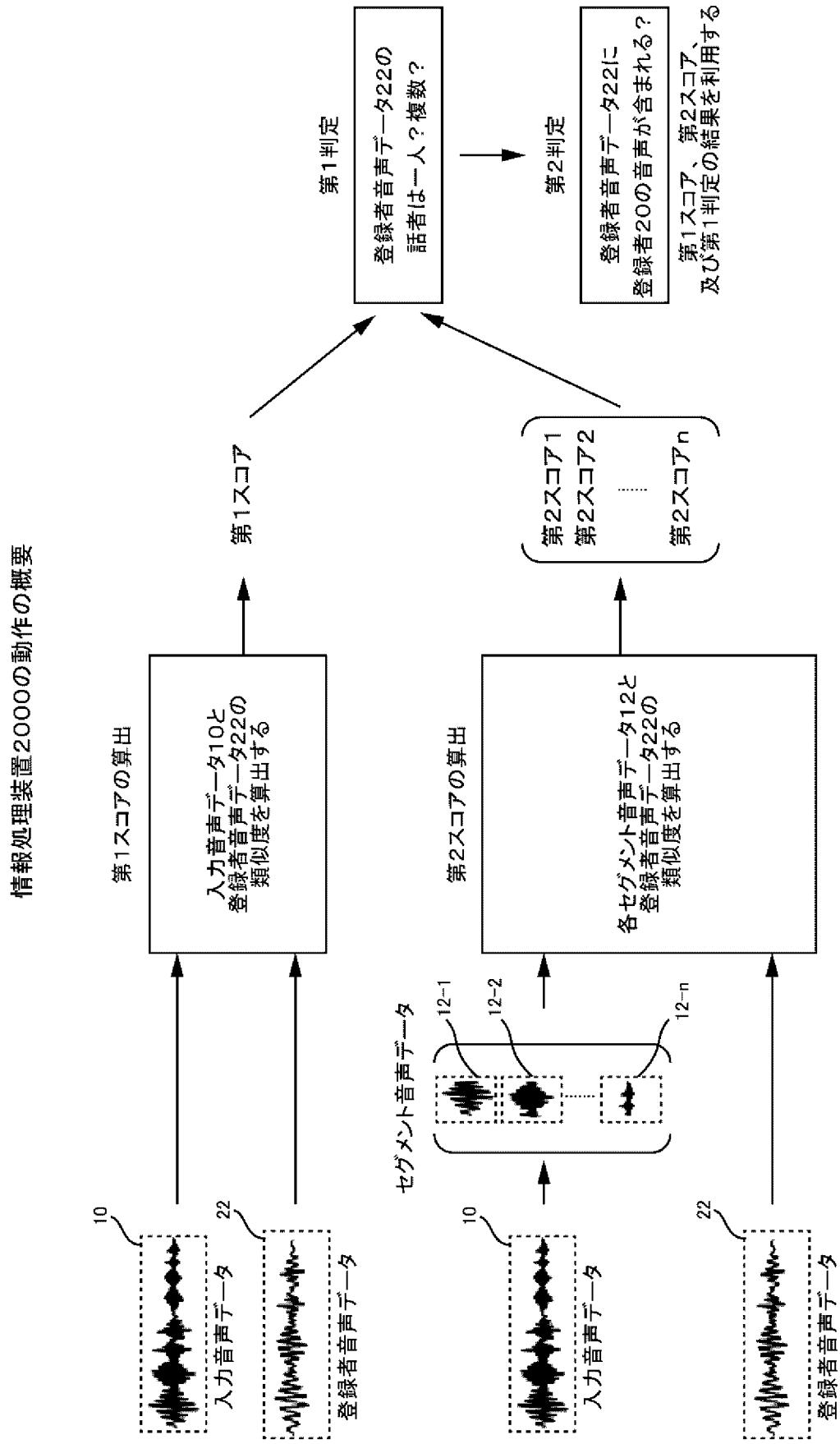
[請求項18] 前記第１判定ステップにおいて、学習済みの予測モデルに対し、第２スコアのみ又は第１スコア及び第２スコアを入力することで、前記入力音声データに含まれる音声の話者が一人と複数のいずれであるか

を判定し、

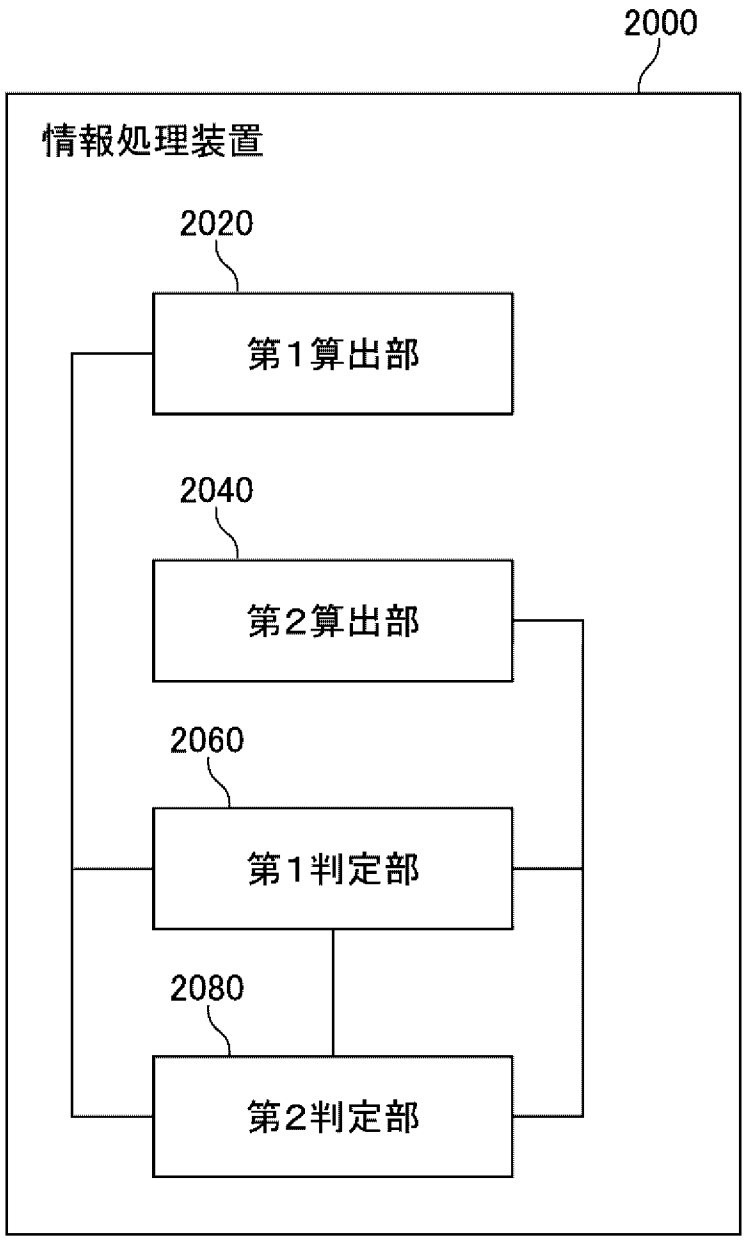
前記予測モデルは、前記第２スコアが入力されたこと、又は前記第１スコア及び前記第２スコアが入力されたことに応じて、前記入力音声データに含まれる音声の話者が一人と複数のいずれであるかを判定するように学習されている、請求項１０乃至１５いずれか一項に記載の制御方法。

[請求項19] 請求項１０乃至１８いずれか一項に記載の制御方法の各ステップをコンピュータに実行させるプログラム。

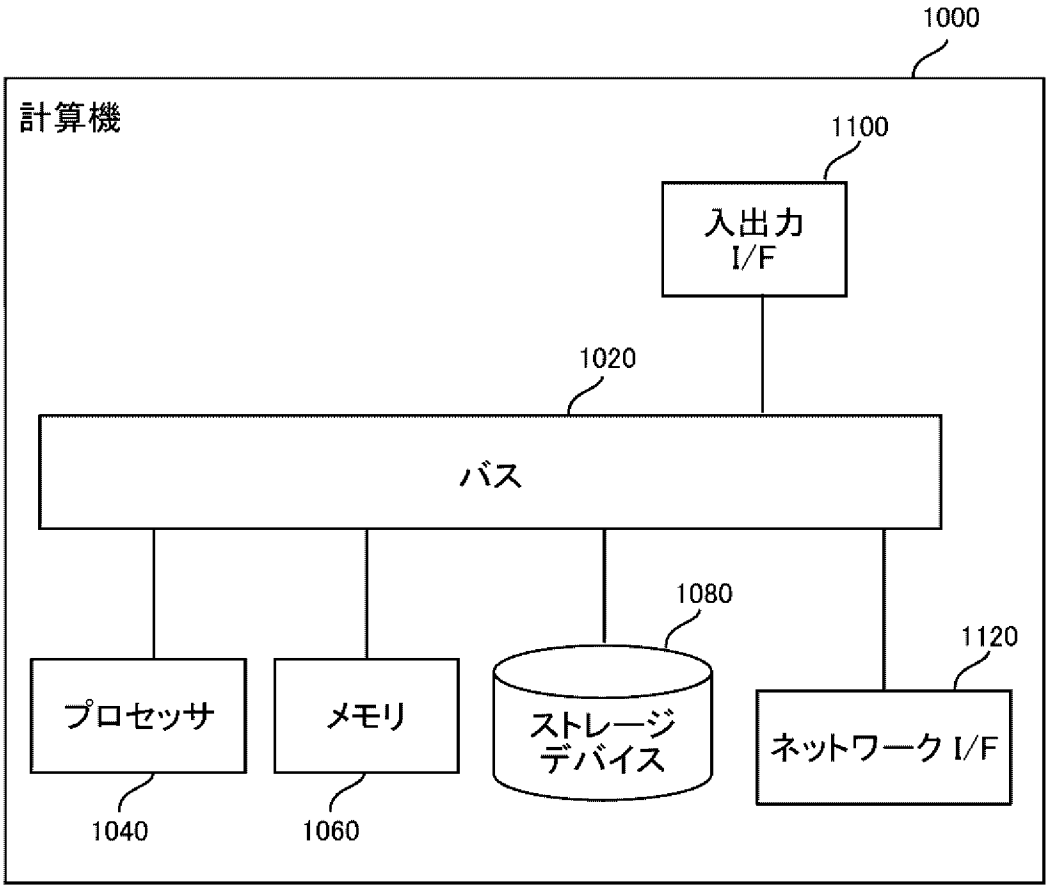
[図1]


情報処理装置2000の動作の概要
第1スコアの算出
入力音声データ10と
登録者音声データ22の
類似度を算出する
第1スコア
第1判定
登録者音声データ22の
話者は一人？複数？
第2判定
登録者音声データ22に
登録者20の音声が含まれる？
第1スコア、第2スコア、
及び第1判定の結果を利用する
第2スコアの算出
各セグメント音声データ12と
登録者音声データ22の
類似度を算出する
第2スコア1
第2スコア2
……
第2スコアn
セグメント音声データ
12-1
12-2
12-n
10
入力音声データ
22
登録者音声データ
10
入力音声データ
22
登録者音声データ

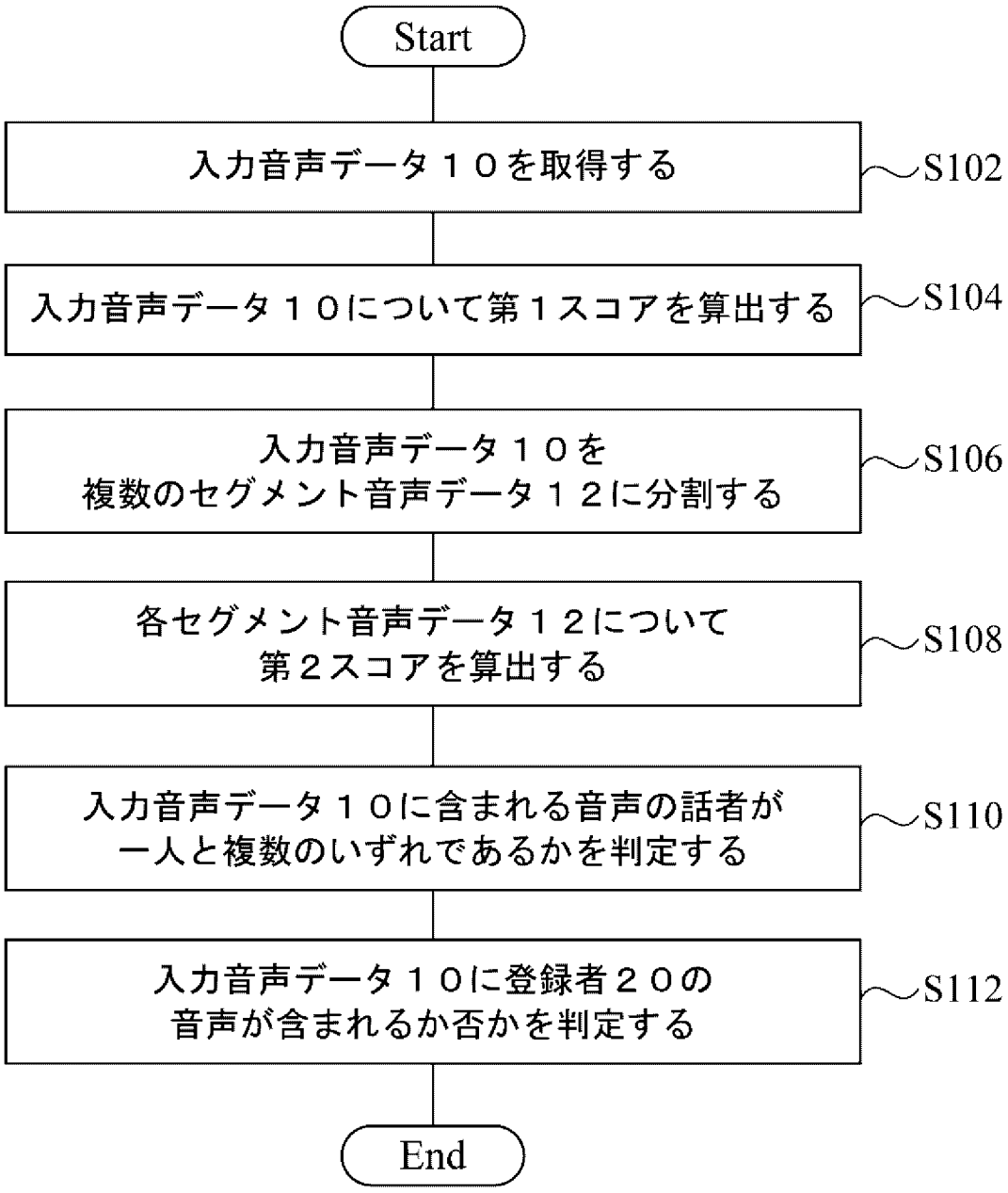
[図2]



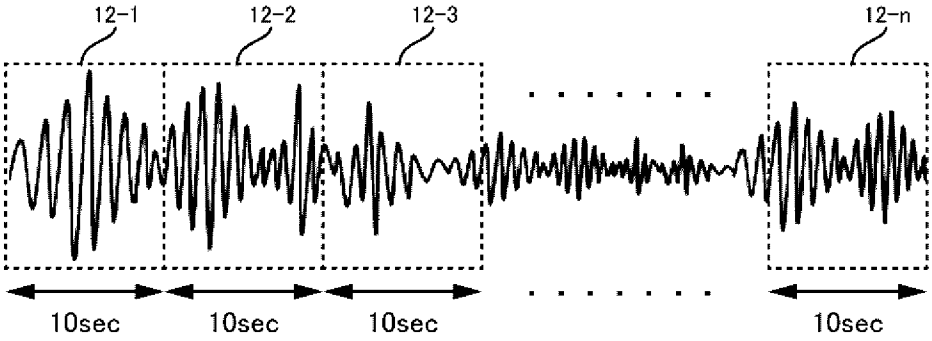
[図3]


1000
計算機
1100
入出力
I/F
1020
バス
1080
1120
プロセッサ
メモリ
ストレージ
デバイス
ネットワーク I/F
1040
1060

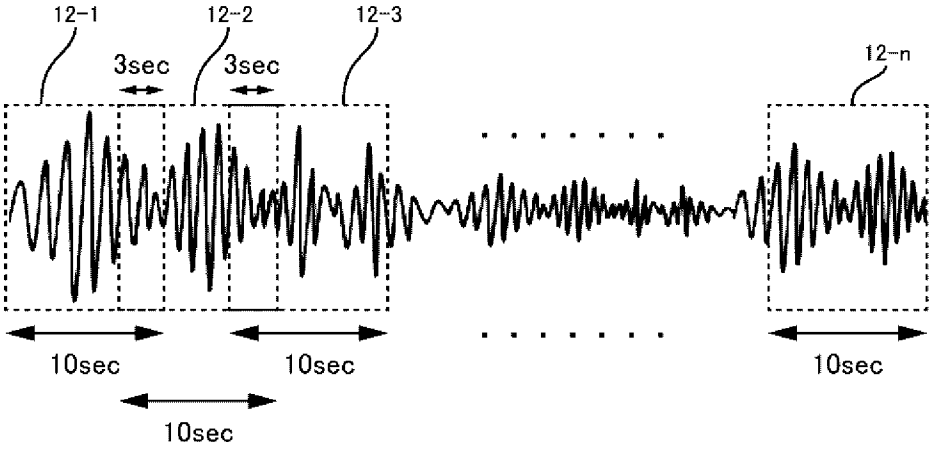
[図4]


Start
入力音声データ１０を取得する
S102
入力音声データ１０について第１スコアを算出する
S104
入力音声データ１０を
複数のセグメント音声データ１２に分割する
S106
各セグメント音声データ１２について
第２スコアを算出する
S108
入力音声データ１０に含まれる音声の話者が
一人と複数のいずれであるかを判定する
S110
入力音声データ１０に登録者２０の
音声が含まれるか否かを判定する
S112
End

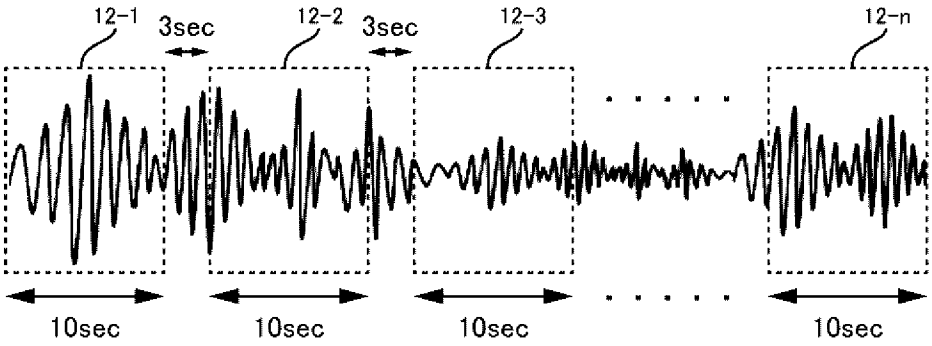
[図5]


12-1
12-2
12-3
12-n
10sec
10sec
10sec
10sec

(a)

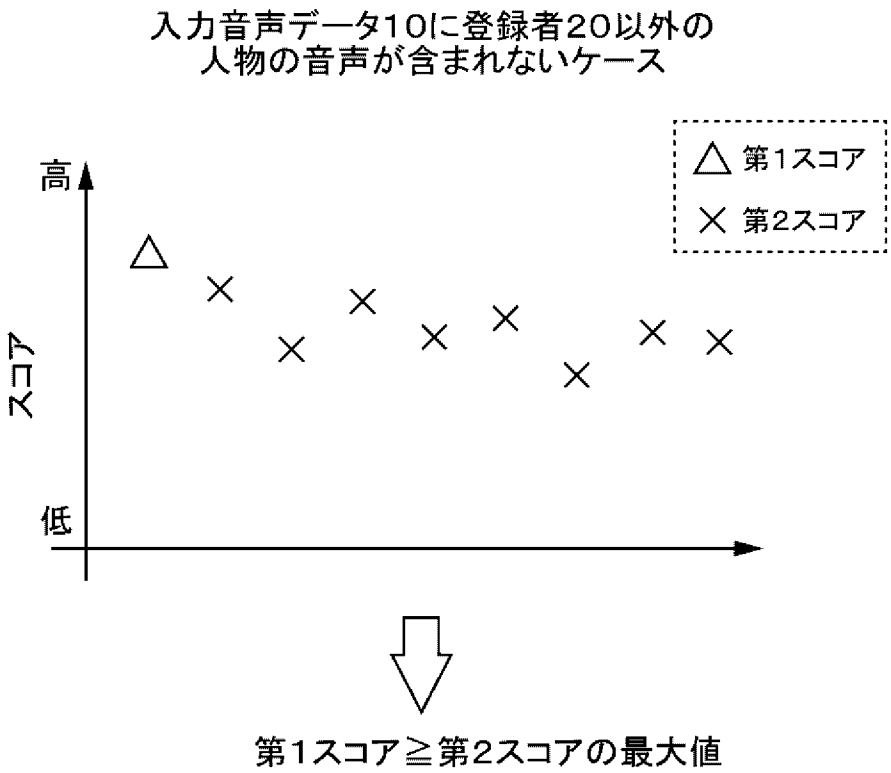

12-1
12-2
12-3
3sec
3sec
12-n
10sec
10sec
10sec
10sec

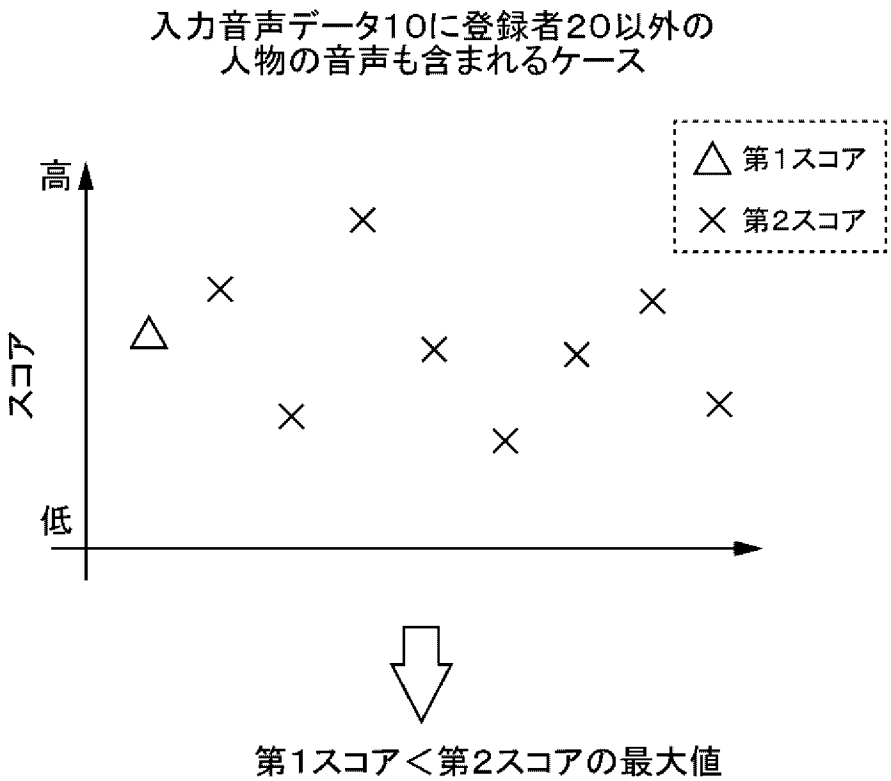
(b)


12-1
3sec
12-2
3sec
12-3
12-n
10sec
10sec
10sec
10sec

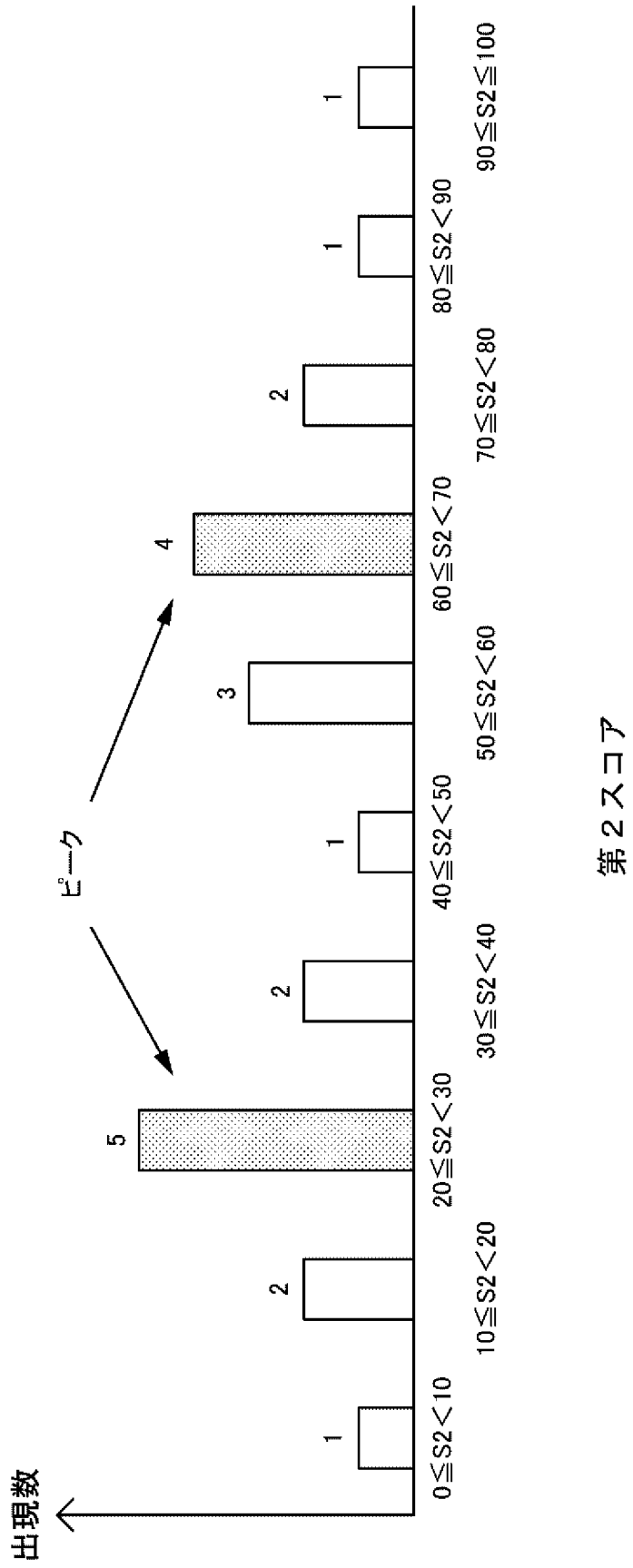
(c)

[図6]


入力音声データ10に登録者20以外の
人物の音声が含まれないケース
△ 第1スコア
× 第2スコア
高
スコア
低
第1スコア≧第2スコアの最大値


入力音声データ10に登録者20以外の
人物の音声も含まれるケース
△ 第1スコア
× 第2スコア
高
スコア
低
第1スコア＜第2スコアの最大値

[図7]


出現数
ピーク
1
2
5
2
1
3
4
1
2
1
0≦S2＜10
10≦S2＜20
20≦S2＜30
30≦S2＜40
40≦S2＜50
50≦S2＜60
60≦S2＜70
70≦S2＜80
80≦S2＜90
90≦S2≦100
第２スコア

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2018/024391

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G10L17/00(2013.01)i

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. G10L17/00

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 1971-2018
Registered utility model specifications of Japan 1996-2018
Published registered utility model applications of Japan 1994-2018

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 2017-517027 A (GOOGLE INC.) 22 June 2017, paragraphs [0066]-[0093] & US 2016/0019889 A1, paragraphs [0070]-[0097] & EP 0003129982 A1 & CN 106164921 A & KR 2016/0143680 A | 1-19 |
| A | WO 2008/117626 A1 (NEC CORP.) 02 October 2008, paragraphs [0024]-[0037] & US 2010/0114572 A1, paragraphs [0050]-[0064] | 1-19 |

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 04.09.2018 | 11.09.2018 |
| Name and mailing address of the ISA/<br>Japan Patent Office<br>3-4-3, Kasumigaseki, Chiyoda-ku,<br>Tokyo 100-8915, Japan | Authorized officer<br><br>Telephone No. |

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１８／０２４３９１

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G10L17/00(2013.01)i

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））

Int.Cl.　G10L17/00

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 | １９７１－２０１８年 |
| 日本国実用新案登録公報 | １９９６－２０１８年 |
| 日本国登録実用新案公報 | １９９４－２０１８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 2017-517027 A (グーグル　インコーポレイテッド)<br>2017.06.22, 段落[0066]-[0093]<br>& US 2016/0019889 A1, 段落[0070]-[0097]<br>& EP 003129982 A1<br>& CN 106164921 A & KR 2016/0143680 A | 1-19 |
| A | WO 2008/117626 A1 (日本電気株式会社)<br>2008.10.02, 段落[0024]-[0037]<br>& US 2010/0114572 A1, 段落[0050]-[0064] | 1-19 |

☐　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矛盾するものではなく、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 |
|---|---|
| ０４．０９．２０１８ | １１．０９．２０１８ |

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｚ　１１６２
上田　雄
電話番号　０３－３５８１－１１０１　内線　３５９１