(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7408518号
(P7408518)

(45)発行日　令和6年1月5日(2024.1.5)

(24)登録日　令和5年12月22日(2023.12.22)

(51)国際特許分類

Ｇ０６Ｎ	*20/00*	*(2019.01)*	Ｇ０６Ｎ	20/00	１３０
Ｇ０６Ｎ	*3/08*	*(2023.01)*	Ｇ０６Ｎ	3/08	
Ｇ１０Ｌ	*15/10*	*(2006.01)*	Ｇ１０Ｌ	15/10	５００Ｚ
Ｇ１０Ｌ	*15/16*	*(2006.01)*	Ｇ１０Ｌ	15/16	

ＦＩ

請求項の数　6　（全24頁）

(21)出願番号　特願2020-155830(P2020-155830)
(22)出願日　令和2年9月16日(2020.9.16)
(65)公開番号　特開2022-49570(P2022-49570A)
(43)公開日　令和4年3月29日(2022.3.29)
審査請求日　令和3年8月19日(2021.8.19)

(73)特許権者　500257300
ＬＩＮＥヤフー株式会社
東京都千代田区紀尾井町１番３号
(74)代理人　110002147
弁理士法人酒井国際特許事務所
(72)発明者　藤田 悠哉
東京都千代田区紀尾井町１番３号 ヤフー株式会社内
審査官　金田 孝之

最終頁に続く

(54)【発明の名称】　情報処理装置、情報処理方法、情報処理プログラム、端末装置、推論方法、及び推論プログラム

(57)【特許請求の範囲】

【請求項１】

機械学習のモデルの学習に用いる音声データである入力用データと、当該入力用データに含まれる認識対象を示す正解データと、前記入力用データに含まれるノイズの多寡に基づく分類結果を示す分類ラベルとを含む学習用データを取得する取得部と、

前記学習用データを用いて、データの入力に応じて、前記正解データに対応する第１出力と前記分類ラベルに対応し、当該データに含まれるノイズの多寡に基づく第２出力とを出力する前記モデルを学習する学習部と、

を備え、

前記取得部は、

前記入力用データが検知された場所の分類結果を示す前記分類ラベルを含む前記学習用データを取得し、

前記学習部は、

前記第１出力と、入力されたデータが検知された場所の分類結果を示す前記第２出力とを出力する前記モデルを学習する

ことを特徴とする情報処理装置。

【請求項２】

コンピュータが実行する情報処理方法であって、

機械学習のモデルの学習に用いる音声データである入力用データと、当該入力用データに含まれる認識対象を示す正解データと、前記入力用データに含まれるノイズの多寡に基

づく分類結果を示す分類ラベルとを含む学習用データを取得する取得工程と、
前記学習用データを用いて、データの入力に応じて、前記正解データに対応する第１出力と前記分類ラベルに対応し、当該データに含まれるノイズの多寡に基づく第２出力とを出力する前記モデルを学習する学習工程と、
を含み、
前記取得工程は、
前記入力用データが検知された場所の分類結果を示す前記分類ラベルを含む前記学習用データを取得し、
前記学習工程は、
前記第１出力と、入力されたデータが検知された場所の分類結果を示す前記第２出力とを出力する前記モデルを学習する
ことを特徴とする情報処理方法。
【請求項３】
機械学習のモデルの学習に用いる音声データである入力用データと、当該入力用データに含まれる認識対象を示す正解データと、前記入力用データに含まれるノイズの多寡に基づく分類結果を示す分類ラベルとを含む学習用データを取得する取得手順と、
前記学習用データを用いて、データの入力に応じて、前記正解データに対応する第１出力と前記分類ラベルに対応し、当該データに含まれるノイズの多寡に基づく第２出力とを出力する前記モデルを学習する学習手順と、
をコンピュータに実行させ、
前記取得手順は、
前記入力用データが検知された場所の分類結果を示す前記分類ラベルを含む前記学習用データを取得し、
前記学習手順は、
前記第１出力と、入力されたデータが検知された場所の分類結果を示す前記第２出力とを出力する前記モデルを学習する
ことを特徴とする情報処理プログラム。
【請求項４】
機械学習のモデルの学習に用いる音声データである入力用データと、当該入力用データに含まれる認識対象を示す正解データと、前記入力用データに含まれるノイズの多寡に基づく分類結果を示す分類ラベルとを含む学習用データであって、前記入力用データが検知された場所の分類結果を示す前記分類ラベルを含む前記学習用データを用いて生成されたモデルであって、データの入力に応じて、前記正解データに対応する第１出力と前記分類ラベルに対応し、当該データに含まれるノイズの多寡に基づく第２出力とを出力するモデルを受信する受信部と、
前記受信部により受信された前記モデルにデータを入力することにより、当該データに対応する前記第１出力と、入力されたデータが検知された場所の分類結果を示す前記第２出力とを生成する推論処理を行う推論部と、
を備えたことを特徴とする端末装置。
【請求項５】
機械学習のモデルの学習に用いる音声データである入力用データと、当該入力用データに含まれる認識対象を示す正解データと、前記入力用データに含まれるノイズの多寡に基づく分類結果を示す分類ラベルとを含む学習用データであって、前記入力用データが検知された場所の分類結果を示す前記分類ラベルを含む前記学習用データを用いて生成されたモデルであって、データの入力に応じて、前記正解データに対応する第１出力と前記分類ラベルに対応し、当該データに含まれるノイズの多寡に基づく第２出力とを出力するモデルを受信する受信工程と、
前記受信工程により受信された前記モデルにデータを入力することにより、当該データに対応する前記第１出力と、入力されたデータが検知された場所の分類結果を示す前記第２出力とを生成する推論処理を行う推論工程と、

を含んだことを特徴とする推論方法。
【請求項６】
機械学習のモデルの学習に用いる音声データである入力用データと、当該入力用データに含まれる認識対象を示す正解データと、前記入力用データに含まれるノイズの多寡に基づく分類結果を示す分類ラベルとを含む学習用データであって、前記入力用データが検知された場所の分類結果を示す前記分類ラベルを含む前記学習用データを用いて生成されたモデルであって、データの入力に応じて、前記正解データに対応する第１出力と前記分類ラベルに対応し、当該データに含まれるノイズの多寡に基づく第２出力とを出力するモデルを受信する受信手順と、
前記受信手順により受信された前記モデルにデータを入力することにより、当該データに対応する前記第１出力と、入力されたデータが検知された場所の分類結果を示す前記第２出力とを生成する推論処理を行う推論手順と、
を端末装置に実行させることを特徴とする推論プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、情報処理装置、情報処理方法、情報処理プログラム、端末装置、推論方法、及び推論プログラムに関する。
【背景技術】
【０００２】
機械学習の技術により、様々な用途に用いられる学習モデル（以下「モデル」ともいう）を学習する技術が提供されている。例えば、モデルを音声認識に用いる音声認識システムが提供されている。
【先行技術文献】
【特許文献】
【０００３】
【文献】特開２０１９－１５９０５８号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
しかしながら、上記の従来技術では、多様な出力を行うモデルを利用可能にすることが難しい。例えば、従来技術では音声認識の結果としてその文字データ（テキスト）の１つの種別の出力のみを行っているに過ぎない。そのため、複数種別の出力を行うモデルを利用可能にすることができるとは限らない。
【０００５】
本願は、上記に鑑みてなされたものであって、複数種別の出力を行うモデルを利用可能にする情報処理装置、情報処理方法、情報処理プログラム、端末装置、推論方法、及び推論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】
【０００６】
本願に係る情報処理装置は、機械学習のモデルの学習に用いる入力用データと、当該入力用データに含まれる認識対象を示す正解データと、前記認識対象に関連する分類結果を示す分類ラベルとを含む学習用データを取得する取得部と、前記学習用データを用いて、データの入力に応じて、前記正解データに対応する第１出力と前記分類ラベルに対応する第２出力とを出力する前記モデルを学習する学習部と、を備えたことを特徴とする。
【発明の効果】
【０００７】
実施形態の一態様によれば、複数種別の出力を行うモデルを利用可能に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】

【０００８】
【図１】図１は、実施形態に係る情報処理システムによる処理の一例を示す図である。
【図２】図２は、実施形態に係る情報処理装置の構成例を示す図である。
【図３】図３は、実施形態に係る学習用データ記憶部の一例を示す図である。
【図４】図４は、実施形態に係るモデル情報記憶部の一例を示す図である。
【図５】図５は、実施形態に係る端末装置の構成例を示す図である。
【図６】図６は、実施形態に係る情報処理装置による処理の一例を示すフローチャートである。
【図７】図７は、実施形態に係る端末装置による処理の一例を示すフローチャートである。
【図８】図８は、ハードウェア構成の一例を示す図である。
【発明を実施するための形態】
【０００９】
以下に、本願に係る情報処理装置、情報処理方法、情報処理プログラム、端末装置、推論方法、及び推論プログラムを実施するための形態（以下、「実施形態」と呼ぶ）について図面を参照しつつ詳細に説明する。なお、この実施形態により本願に係る情報処理装置、情報処理方法、情報処理プログラム、端末装置、推論方法、及び推論プログラムが限定されるものではない。また、以下の各実施形態において同一の部位には同一の符号を付し、重複する説明は省略される。
【００１０】
〔１．はじめに〕
近年、音声認識等に利用されるモデル（「音声認識モデル」ともいう）に、Ｅｎｄ－ｔｏ－Ｅｎｄモデル（「Ｅ２Ｅモデル」ともいう）が用いられている。Ｅ２Ｅモデルは、例えば１つのニューラルネットワークで構成されるモデルである。Ｅ２Ｅモデルは、ユーザが利用するデバイス（端末装置１０等）で完結する音声認識の処理に適している。音声データを入力してその音声データに対応する文字データを出力させる音声文字変換等の音声認識モデルにおいて、入力された音声データに対応する認識結果が出力される。このような音声認識モデルでは、従来は入力データに対応する文字データ（テキスト）等１つの種別の出力を行う。
【００１１】
一方で、以下に示す情報処理装置１００が学習するモデルは、入力用データに含まれる認識対象の認識結果を示す出力（「第１出力」ともいう）と、認識対象に関連する分類結果を示す出力（「第２出力」ともいう）との複数の種別の出力を行う。これにより、情報処理装置１００が学習するモデルは、認識対象に関連する分類結果に関連する特徴を加味して学習される。そのため、情報処理装置１００が学習するモデルは、認識対象に関連する分類を加味しつつ、入力用データの特徴を抽出し、第１出力を出力するため、第１出力に関する認識精度を向上させることができる。
【００１２】
また、入力に音声データとその音声データが検知された場所を示すラベル等の複数種別の情報の入力を用いる場合、推論時にも複数種別の情報を入力する必要となる。そのため、モデルの利用する場面において利便性が低く、その情報を入力として用意できない場合、モデルを利用できなかったり、推論の精度が低下したりする。一方で、情報処理装置１００が学習するモデルは、入力用データとして、以下に示すように例えば音声データ等の１つのデータのみでよいため、推論時にも複数種別の情報を入力する必要がない。そのため、情報処理装置１００は、認識精度を向上させつつ、利便性の高いモデルを学習することができる。
【００１３】
（実施形態）
〔２．情報処理〕
ここから、図１を用いて、実施形態に係る情報処理の一例について説明する。図１は、実施形態に係る情報処理システムによる処理の一例を示す図である。まず、情報処理シス

テム１の構成について説明する。

【００１４】

図１に示すように、情報処理システム１は、端末装置１０と、情報処理装置１００とが含まれる。端末装置１０と、情報処理装置１００とは図示しない所定の通信網を介して、有線または無線により通信可能に接続される。なお、図１に示した情報処理システム１には、複数台の端末装置１０や、複数台の情報処理装置１００が含まれてもよい。

【００１５】

情報処理装置１００は、機械学習のモデルの学習に用いる入力用データと、入力用データに含まれる認識対象を示す正解データと、認識対象に関連する分類結果を示す分類ラベルとを含む学習用データを用いて、正解データに対応する第１出力と分類ラベルに対応する第２出力とを出力するモデルを学習する情報処理装置である。情報処理装置１００は、音声文字変換結果である第１出力と、分類結果を示す第２出力との２つの種別の出力を行うモデルＭ１を学習し、端末装置１０に提供する。なお、モデルＭ１のネットワーク構成は、第１出力及び第２出力の出力が可能であればどのようなネットワーク構成であってもよく、Ｅ２Ｅモデルであってもよい。

【００１６】

端末装置１０は、ユーザによって利用されるデバイス（コンピュータ）である。端末装置１０は、ユーザによる音声入力を受け付ける。端末装置１０は、ユーザによる操作を受け付ける。端末装置１０は、情報処理装置１００から提供されたモデルを用いて推論を行う。

【００１７】

また、以下では、端末装置１０をユーザと表記する場合がある。すなわち、以下では、ユーザを端末装置１０と読み替えることもできる。なお、端末装置１０は、例えば、スマートフォンや、タブレット型端末や、ノート型ＰＣ（Personal Computer）や、デスクトップＰＣや、携帯電話機や、ＰＤＡ（Personal Digital Assistant）等により実現される。図１の例では、端末装置１０がタッチパネル機能を有するスマートフォンである場合を示す。

【００１８】

以下、図１を用いて、情報処理の一例を説明する。図１では、ユーザがユーザＩＤ「Ｕ１」により識別されるユーザ（以下、「ユーザＵ１」とする場合がある）である場合を示す。また、図１では、情報処理装置１００が音声データの入力に対して、その音声データが変換された文字データである第１出力、及びその音声データが検知（収集）された場所を分類する分類ラベルである第２出力を出力するモデルＭ１を学習する場合を一例として説明する。なお、分類ラベルは場所の分類に限らず、様々な対象の分類を示すものであってもよいが、この点については後述する。

【００１９】

まず、情報処理装置１００は、機械学習に用いる学習用データ群ＤＳ１を取得する（ステップＳ１１）。学習用データ群ＤＳ１には、モデルの入力として用いられる入力用データと、その入力用データに対応する正解データ及び分類ラベルとのセット（組合せ）が複数含まれる。例えば、入力用データである音声データＤＴ１は、正解データＲＴ１と分類ラベルＣＬ１とが対応付けられている。この場合、正解データＲＴ１は、音声データＤＴ１が文字変換された文字データ（文字列）であり、分類ラベルＣＬ１は、音声データＤＴ１が検知（収集）された場所の分類が「カフェ」であることを示す値（例えば１等）であるものとする。

【００２０】

そして、情報処理装置１００は、学習用データ群ＤＳ１を用いて、文字データである第１出力と、音声データの検知場所を示す分類ラベルである第２出力との２つの種別の出力を行うモデルＭ１を学習する（ステップＳ１２）。図１では、音声データである入力用データＩＮの入力層への入力に応じて、出力層から文字データである第１出力ＯＴ１と、音声データの検知場所を示す分類ラベルである第２出力ＯＴ２を出力するモデルＭ１を概念

的に示す。

【００２１】

情報処理装置１００は、音声データＤＴ１が入力された場合に、第１出力ＯＴ１として正解データＲＴ１が出力され、第２出力ＯＴ２として分類ラベルＣＬ１が出力されるようにモデルＭ１の重み等のパラメータを学習する。また、音声データＤＴ２が入力された場合に、第１出力ＯＴ１として正解データＲＴ２が出力され、第２出力ＯＴ２として分類ラベルＣＬ２が出力されるようにモデルＭ１の重み等のパラメータを学習する。これにより、情報処理装置１００は、第１出力と第２出力との２つの種別の出力を行うモデルＭ１を生成する。モデルＭ１の学習処理には、任意の手法が採用可能である。

【００２２】

例えば、情報処理装置１００は、バックプロパゲーション（誤差逆伝播法）等の手法により学習処理を行う。例えば、情報処理装置１００は、学習処理により、ノード間で値が伝達する際に考慮される重み（すなわち、接続係数）の値を調整する。このように、情報処理装置１００は、モデルＭ１における出力（第１出力及び第２出力）と、入力に対応する正解（正解データ及び分類ラベル）との誤差が少なくなるようにパラメータ（接続係数）を補正するバックプロパゲーション等の処理によりモデルＭ１を学習する。例えば、情報処理装置１００は、所定の損失（ロス）関数を最小化するようにバックプロパゲーション等の処理を行うことによりモデルＭ１を生成する。これにより、情報処理装置１００は、モデルＭ１のパラメータを学習する学習処理を行うことができる。

【００２３】

そして、情報処理装置１００は、学習したモデルＭ１をユーザＵ１が利用する端末装置１０に提供する（ステップＳ１３）。モデルＭ１を受信した端末装置１０は、モデルＭ１を利用してユーザＵ１の発話も文字に書き起こす音声文字変換処理（推論）を実行する。この点について以下説明する。

【００２４】

まず、ユーザＵ１が「ＸＸＸＸ」と発話する。なお、「ＸＸＸＸ」は具体的な内容を含む発話であるものとする。端末装置１０は、ユーザＵ１の発話ＰＡを検知し、ユーザＵ１の発話ＰＡである「ＸＸＸＸ」の音声データを入力として受け付ける（ステップＳ１４）。

【００２５】

そして、端末装置１０は、入力として受け付けた「ＸＸＸＸ」の音声データと、モデルＭ１とを利用して推論処理を行う（ステップＳ１５）。端末装置１０は、「ＸＸＸＸ」の音声データをモデルＭ１に入力し、モデルＭ１に文字データ及び分類ラベルを出力させることにより、音声を文字に変換するとともに、ユーザＵ１が発話ＰＡを行った場所の分類を推定する処理（推論処理）を行う。図１では、「ＸＸＸＸ」の音声データが入力されたモデルＭ１は、「ＸＸＸＸ」の文字データ（第１出力）と、発話ＰＡを行った場所の分類結果がカフェであることを示す分類ラベル（第２出力）を出力する。

【００２６】

そして、端末装置１０は、推論結果である「ＸＸＸＸ」の文字データを表示してもよい（ステップＳ１６）。例えば、端末装置１０は、文字列「ＸＸＸＸ」を画面に表示する。なお、端末装置１０は、分類ラベルを表示してもよい。

【００２７】

また、端末装置１０は、音声データ、その音声データに対応する文字データである正解データ、及びその音声データに対応する分類ラベルを学習用データとして情報処理装置１００に送信してもよい（ステップＳ１７）。この場合、情報処理装置１００は、端末装置１０から受信した学習用データを用いて、モデルＭ１のパラメータを更新してもよい。

【００２８】

上述したように、情報処理装置１００は、音声データを入力として、その音声データが変換された文字データと、音声データの検知場所を示す分類ラベルとの２つの種別の出力を行うモデルＭ１を適切に学習することができる。図１の例では、情報処理装置１００は、分類ラベルを出力するＥ２Ｅの音声認識モデルであるモデルＭ１を適切に学習すること

ができる。したがって、情報処理装置１００は、複数種別の出力を行うモデルを利用可能にすることができる。また、情報処理装置１００は、分類ラベルを出力することで音声文字変換の精度が改善することができる。例えば、情報処理装置１００は、認識対象に関連する分類を加味しつつ、入力用データの特徴を抽出し、音声文字変換の結果を出力するモデルＭ１を学習するため、分類ラベルの出力により音声文字変換の精度を向上させたモデルＭ１を学習することができる。また、端末装置１０は、音声データが変換された文字データと、音声データの検知場所を示す分類ラベルとの２つの種別の出力するモデルを用いて推論処理を行うことができるため、適切なモデルを利用した処理を行うことができる。したがって、端末装置１０は、複数種別の出力を行うモデルを利用した処理を行うことができる。

【００２９】

〔２－１．分類ラベル〕

なお、上記の例では、分類ラベルが音声データが検知（収集）された場所の分類を示す場合を示したが、分類ラベルは、認識対象に関連する分類結果を示すものであればどのような対象の分類であってもよい。分類ラベルは、入力用データに含まれる情報のうち、認識対象以外の情報の種別を示すものである。つまり、分類ラベルは、入力用データのうち認識対象となるデータ以外のデータの分類結果を示すものであってもよい。また、分類ラベルは、入力用データのうち、認識対象となるデータから認識される認識結果以外の各種情報であってもよい。また、分類ラベルは、入力用データが取得された際の各種コンテキストを示すものであってもよい。このように、分類ラベルは、認識対象となるデータから認識結果を認識する認識処理において影響を及ぼしうる任意の要素であって、入力用データから取得もしくは推定可能な要素、もしくは入力用データに付随する各種の要素の分類結果が採用可能である。この点について以下例示を列挙する。なお、図１と同様の点については適宜説明を省略する。また、以下に示す各モデルのネットワーク構成はモデルＭ１と同様であってもよい。

【００３０】

〔２－１－１．ユーザ関連〕

例えば、モデルが出力する分類ラベル（第２出力）は、入力用データが検知された場所に限らず、入力用データに関連する様々なコンテキストの分類結果であってもよい。例えば、モデルが出力する分類ラベル（第２出力）は、入力用データに含まれる発話を行ったユーザに関連するコンテキストの分類結果であってもよい。

【００３１】

〔２－１－２．周囲〕

モデルが出力する分類ラベル（第２出力）は、発話を行ったユーザの発話以外の分類結果を示してもよい。例えば、モデルが出力する分類ラベル（第２出力）は、入力用データに含まれる発話を行ったユーザの周囲の状況の分類結果であってもよい。この場合、情報処理装置１００は、ユーザの周囲の状況の分類結果を示す分類ラベルを含む学習用データを用いて、ユーザの周囲の状況の分類結果を示す第２出力と第１出力とを出力するモデル（「モデルＭ２」とする）を学習する。

【００３２】

情報処理装置１００は、ユーザの周囲が騒がしいか否かを示す分類ラベル、すなわちユーザの発話以外のノイズが多いか否かを示す分類ラベルを第２出力として出力するモデルＭ２を学習する。この場合、分類ラベルは、ユーザの周囲が騒がしい（ノイズが多い）程、大きい値となってもよい。

【００３３】

情報処理装置１００は、モデルＭ２の入力として用いられる入力用データと、その入力用データに対応する正解データ及び分類ラベルとのセット（組合せ）を複数含む学習用データ（「学習用データＤＳ２」とする）を用いて、モデルＭ２を学習する。例えば、学習用データＤＳ２の正解データは、入力用データである音声データが変換された文字データであり、学習用データＤＳ２の分類ラベルは、入力用データである音声データにおいて発

話したユーザの周囲の騒がしさの度合いを示す値である。

【００３４】

情報処理装置１００は、学習用データＤＳ２を用いて、各入力用データが入力された場合に、その入力用データに対応する正解データ及び分類ラベルが出力されるようにモデルＭ２のパラメータを学習する。例えば、情報処理装置１００は、音声データが入力された場合に、音声データが変換された文字データが第１出力として出力され、音声データに含まれるノイズの多寡を示す分類ラベルを第２出力として出力されるようにモデルＭ２を学習する。

【００３５】

〔２－１－３．ユーザ属性〕

モデルが出力する分類ラベル（第２出力）は、入力用データに含まれる発話を行ったユーザの属性の分類結果であってもよい。この場合、情報処理装置１００は、ユーザの属性の分類結果を示す分類ラベルを含む学習用データを用いて、ユーザの属性の分類結果を示す第２出力と第１出力とを出力するモデル（「モデルＭ３」とする）を学習する。以下では、ユーザの年齢をユーザの属性の一例として説明するが、ユーザの属性は、年齢に限らず、性別、身長、出身地等の様々な属性（要素）であってもよい。

【００３６】

情報処理装置１００は、ユーザの年齢を示す分類ラベル（例えば１０代、２０代等の段階的な分類）を第２出力として出力するモデルＭ３を学習する。情報処理装置１００は、モデルＭ３の入力として用いられる入力用データと、その入力用データに対応する正解データ及び分類ラベルとのセット（組合せ）を複数含む学習用データ（「学習用データＤＳ３」とする）を用いて、モデルＭ３を学習する。例えば、学習用データＤＳ３の正解データは、入力用データである音声データが変換された文字データであり、学習用データＤＳ３の分類ラベルは、入力用データである音声データにおいて発話したユーザの年齢を示す値である。

【００３７】

情報処理装置１００は、学習用データＤＳ３を用いて、各入力用データが入力された場合に、その入力用データに対応する正解データ及び分類ラベルが出力されるようにモデルＭ３のパラメータを学習する。例えば、情報処理装置１００は、音声データが入力された場合に、音声データが変換された文字データが第１出力として出力され、音声データに含まれる発話を行ったユーザの年齢を示す分類ラベルを第２出力として出力されるようにモデルＭ３を学習する。

【００３８】

〔２－１－４．端末装置〕

モデルが出力する分類ラベル（第２出力）は、ユーザが利用する端末装置１０の分類結果であってもよい。例えば、モデルが出力する分類ラベル（第２出力）は、ユーザの発話（音声データ）を検知（収集）した端末装置１０の機種の分類結果であってもよい。この場合、情報処理装置１００は、端末装置１０の機種の分類結果を示す分類ラベルを含む学習用データを用いて、端末装置１０の機種の分類結果を示す第２出力と第１出力とを出力するモデル（「モデルＭ４」とする）を学習する。

【００３９】

情報処理装置１００は、端末装置１０の機種を示す分類ラベル（例えば製品Ａ、製品Ｂ等の製品の分類）を第２出力として出力するモデルＭ４を学習する。情報処理装置１００は、モデルＭ４の入力として用いられる入力用データと、その入力用データに対応する正解データ及び分類ラベルとのセット（組合せ）を複数含む学習用データ（「学習用データＤＳ４」とする）を用いて、モデルＭ４を学習する。例えば、学習用データＤＳ４の正解データは、入力用データである音声データが変換された文字データであり、学習用データＤＳ４の分類ラベルは、入力用データである音声データを検知した端末装置１０の機種を示す値である。

【００４０】

情報処理装置１００は、学習用データＤＳ４を用いて、各入力用データが入力された場合に、その入力用データに対応する正解データ及び分類ラベルが出力されるようにモデルＭ４のパラメータを学習する。例えば、情報処理装置１００は、音声データが入力された場合に、音声データが変換された文字データが第１出力として出力され、音声データを検知した端末装置１０の機種を示す分類ラベルを第２出力として出力されるようにモデルＭ４を学習する。

【００４１】

なお、上述は一例に過ぎず、分類できるものであれば、どのような対象の分類ラベルを用いてもよい。

【００４２】

また、情報処理装置１００は、複数の第２出力を出力するようにモデルを学習してもよい。すなわち、情報処理装置１００は、１つの第１出力と、２つ以上の第２出力を出力するモデルを生成してもよい。この場合、情報処理装置１００は、上述した各種の分類ラベルから選択された２つ以上の分類ラベルを出力とするモデルを学習する。例えば、情報処理装置１００は、音声データが入力された場合に、音声データが変換された文字データと、音声データが検知（収集）された場所を示す第１分類ラベルと、音声データに含まれるユーザの属性を示す第２分類ラベルと出力するモデルを学習する。

【００４３】

〔２－２．推論対象〕

なお、学習するモデルの用途は、音声文字変換に限らず、他の音声認識に関する様々な用途であってもよい。また、モデルの入力は、音声データに限らず、画像データ等様々な種別のデータが対象であってもよい。例えば、モデルの入力が画像データである場合、学習されるモデルの用途は、一般物体認識等の各種の画像認識に関する用途であってもよい。この場合、分類ラベルは、画像に含まれる物体以外の分類を示すものであってもよい。例えば、分類ラベルは、画像に含まれる人（ユーザ）の年齢等のユーザの属性であってもよく、画像が示すシーンの状況（昼、夜、室内、屋外等）などのコンテキストであってもよい。

【００４４】

〔３．情報処理装置の構成〕

次に、図２を用いて、実施形態に係る情報処理装置１００の構成について説明する。図２は、実施形態に係る情報処理装置１００の構成例を示す図である。図２に示すように、情報処理装置１００は、通信部１１０と、記憶部１２０と、制御部１３０とを有する。なお、情報処理装置１００は、情報処理装置１００の管理者等から各種操作を受け付ける入力部（例えば、キーボードやマウス等）や、各種情報を表示するための表示部（例えば、液晶ディスプレイ等）を有してもよい。

【００４５】

（通信部１１０）

通信部１１０は、例えば、ＮＩＣ（Network Interface Card）等によって実現される。そして、通信部１１０は、所定の通信網（ネットワーク）と有線または無線で接続され、端末装置１０との間で情報の送受信を行う。

【００４６】

（記憶部１２０）

記憶部１２０は、例えば、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。実施形態に係る記憶部１２０は、図２に示すように、学習用データ記憶部１２１と、モデル情報記憶部１２２とを有する。

【００４７】

（学習用データ記憶部１２１）

実施形態に係る学習用データ記憶部１２１は、学習に用いるデータに関する各種情報を記憶する。学習用データ記憶部１２１は、学習に用いる学習データ（データセット）を記

憶する。図３は、本開示の実施形態に係る学習用データ記憶部の一例を示す図である。例えば、学習用データ記憶部１２１は、学習に用いる学習データや精度評価（測定）に用いる評価用データ等の種々のデータに関する各種情報を記憶する。図３に、実施形態に係る学習用データ記憶部１２１の一例を示す。図３の例では、学習用データ記憶部１２１は、「データセットＩＤ」、「データＩＤ」、「データ」、「正解データ」、「分類ラベル」といった項目が含まれる。

【００４８】

「データセットＩＤ」は、データセットを識別するための識別情報を示す。「データＩＤ」は、各学習用データを識別するための識別情報を示す。また、「データ」は、データＩＤにより識別されるデータを示す。「データ」は、モデルの入力として用いられるデータ（入力用データ）を示す。図３の例では、入力用データは、種別が「音声」である音声データの場合を示す。

【００４９】

「正解データ」は、対応するデータ（入力用データ）に対応する正解を示す。図３の例では、「正解データ」は、入力用データである音声データが変換された文字データ（文字列）を示す。「正解データ」は、対応するデータ（入力用データ）がモデルに入力された場合に、モデルが出力することが期待される第１出力（文字データ）を示す。

【００５０】

「分類ラベル」は、対応するデータ（入力用データ）に対応する分類結果を示す。図３の例では、「分類ラベル」は、入力用データである音声データが検知された際のコンテキストを推定するための分類を示す。「分類ラベル」は、対応するデータ（入力用データ）がモデルに入力された場合に、モデルが出力することが期待される第２出力（分類ラベル）を示す。

【００５１】

例えば、「分類ラベル」は、入力用データである音声データが検知された場所を推定するための分類を示す。例えば、分類ラベルが「１」の場合は「カフェ」であることを示し、分類ラベルが「２」の場合は「自宅」であることを示してもよい。記憶部１２０は、分類ラベルと各コンテキストの対応付けを示す情報を記憶してもよい。なお、「分類ラベル」は、１つに限らず、モデルが出力する分類ラベルの数に応じた数であってもよい。例えば、場所と発話ユーザの属性を推定する場合、場所を示すラベルを登録する「分類ラベル＃１」と、ユーザの属性を示すラベルを登録する「分類ラベル＃２」の複数の項目が含まれてもよい。すなわち、「分類ラベル」は、２つ以上のコンテキストの各々に対応する分類ラベルが記憶されてもよい。なおユーザの属性は、年齢や性別等のデモグラフィック属性やサイコグラフィック属性の様々なユーザの属性を示す情報であってもよい。

【００５２】

図３の例では、データセットＩＤ「ＤＳ１」により識別されるデータセット（データセットＤＳ１）には、データＩＤ「ＤＩＤ１」、「ＤＩＤ２」、「ＤＩＤ３」等により識別される複数のデータが含まれることを示す。

【００５３】

データＩＤ「ＤＩＤ１」により識別されるデータＤＴ１は、正解データが「ＲＴ１」であることを示す。図３の例では「ＲＴ１」のように抽象的に図示するが、「正解データ」には、音声データ（入力用データ）に含まれる認識対象（ユーザの発話）が変換された文字データ（文字列）であるものとする。

【００５４】

データＤＴ１は、分類ラベルが「ＣＬ１」であることを示す。図３の例では「ＣＬ１」のように抽象的に図示するが、「分類ラベル」には、認識対象に関連する分類結果を示す分類ラベル（値）であるものとする。

【００５５】

なお、学習用データ記憶部１２１は、上記に限らず、目的に応じて種々の情報を記憶してもよい。例えば、学習用データ記憶部１２１は、音声や画像等の教師データの種別を示

す情報を各データに対応付けて記憶する。例えば、学習用データ記憶部１２１は、データの種別を示す情報を各データに対応付けて記憶する。

【００５６】

例えば、学習用データ記憶部１２１は、各データが学習データであるか、評価用データであるか等を特定可能に記憶してもよい。例えば、学習用データ記憶部１２１は、学習データと評価用データとを区別可能に記憶する。学習用データ記憶部１２１は、各データが学習データや評価用データであるかを識別する情報を記憶してもよい。情報処理装置１００は、学習データとして用いられる各データと正解データと分類ラベルとに基づいて、モデルを学習する。情報処理装置１００は、評価用データとして用いられる各データと正解データと分類ラベルとに基づいて、モデルの精度を測定する。情報処理装置１００は、評価用データを入力した場合にモデルが出力する出力結果（第１出力、第２出力）と、正解データ及び分類ラベルとを比較した結果を収集することにより、モデルの精度を測定する。

【００５７】

（モデル情報記憶部１２２）

実施形態に係るモデル情報記憶部１２２は、モデルに関する情報を記憶する。例えば、モデル情報記憶部１２２は、学習処理により学習（生成）された学習済みモデル（モデル）の情報（モデルデータ）を記憶する。図４は、本開示の第１の実施形態に係るモデル情報記憶部の一例を示す図である。図４に、第１の実施形態に係るモデル情報記憶部１２２の一例を示す。図４に示した例では、モデル情報記憶部１２２は、「モデルＩＤ」、「用途」、「モデルデータ」といった項目が含まれる。

【００５８】

「モデルＩＤ」は、モデルを識別するための識別情報を示す。「用途」は、対応するモデルの用途を示す。「モデルデータ」は、モデルのデータを示す。図４等では「モデルデータ」に「ＭＤＴ１」といった概念的な情報が格納される例を示したが、実際には、モデルの構成（ネットワーク構成）の情報やパラメータに関する情報等、そのモデルを構成する種々の情報が含まれる。例えば、「モデルデータ」には、ネットワークの各層におけるノードと、各ノードが採用する関数と、ノードの接続関係と、ノード間の接続に対して設定される接続係数とを含む情報が含まれる。

【００５９】

図４に示す例では、モデルＩＤ「Ｍ１」により識別されるモデル（モデルＭ１）は、用途が「音声文字変換」、「コンテキスト推定」であることを示す。すなわち、モデルＭ１は、入力用データを文字起こしした文字データと、その入力用データに関連するコンテキストを推定する情報とを出力するモデルであることを示す。また、モデルＭ１のモデルデータは、モデルデータＭＤＴ１であることを示す。

【００６０】

なお、モデル情報記憶部１２２は、上記に限らず、目的に応じて種々の情報を記憶してもよい。

【００６１】

（制御部１３０）

図２の説明に戻って、制御部１３０は、コントローラ（controller）であり、例えば、ＣＰＵ（Central Processing Unit）やＭＰＵ（Micro Processing Unit）等によって、情報処理装置１００内部の記憶装置に記憶されている各種プログラム（情報処理プログラムの一例に相当）がＲＡＭを作業領域として実行されることにより実現される。また、制御部１３０は、コントローラであり、例えば、ＡＳＩＣ（Application Specific Integrated Circuit）やＦＰＧＡ（Field Programmable Gate Array）等の集積回路により実現される。

【００６２】

図２に示すように、制御部１３０は、取得部１３１と、決定部１３２と、学習部１３３と、提供部１３４とを有し、以下に説明する情報処理の機能や作用を実現または実行する。なお、制御部１３０の内部構成は、図２に示した構成に限られず、後述する情報処理を

行う構成であれば他の構成であってもよい。また、制御部１３０が有する各処理部の接続関係は、図２に示した接続関係に限られず、他の接続関係であってもよい。

【００６３】

（取得部１３１）

取得部１３１は、記憶部１２０から各種の情報を取得する。取得部１３１は、学習用データ記憶部１２１から学習に用いるデータを取得する。取得部１３１は、モデル情報記憶部１２２からモデルの情報を取得する。

【００６４】

取得部１３１は、機械学習のモデルの学習に用いる入力用データと、当該入力用データに含まれる認識対象を示す正解データと、認識対象に関連する分類結果を示す分類ラベルとを含む学習用データを取得する。取得部１３１は、音声データである入力用データと、当該入力用データに含まれる音声の音声認識の結果を示す正解データとを含む学習用データを取得する。取得部１３１は、入力用データに対応する文字データである正解データを含む学習用データを取得する。取得部１３１は、入力用データに関連するコンテキストの分類結果を示す分類ラベルを含む学習用データを取得する。

【００６５】

取得部１３１は、入力用データが検知された場所の分類結果を示す分類ラベルを含む学習用データを取得する。取得部１３１は、入力用データに含まれる発話を行ったユーザに関連するコンテキストの分類結果を示す分類ラベルを含む学習用データを取得する。取得部１３１は、ユーザの周囲の状況の分類結果を示す分類ラベルを含む学習用データを取得する。取得部１３１は、ユーザの発話以外の分類結果を示す分類ラベルを含む学習用データを取得する。取得部１３１は、ユーザの属性の分類結果を示す分類ラベルを含む学習用データを取得する。取得部１３１は、ユーザが利用する端末装置の分類結果を示す分類ラベルを含む学習用データを取得する。取得部１３１は、入力用データに含まれる情報のうち、認識対象以外の情報の種別を示す分類ラベルを取得する。

【００６６】

取得部１３１は、通信部１１０を介して、端末装置１０から情報を受信する。取得部１３１は、端末装置１０から学習用データを取得する。取得部１３１は、端末装置１０において、音声文字変換処理の対象となった音声データと、その音声データに対応する第１出力及び第２出力、またユーザが修正した修正結果とのセット（組合せ）を学習用データとして端末装置１０から収集する。

【００６７】

（決定部１３２）

決定部１３２は、種々の情報を決定する。例えば、決定部１３２は、分類ラベルを決定する。決定部１３２は、認識対象に関連する分類結果を決定することにより、モデルに学習させる分類ラベルを決定する。決定部１３２は、入力用データに関連するコンテキストをモデルに分類させる対象に決定する。決定部１３２は、音声データが検知された場所をモデルに分類させる対象に決定する。

【００６８】

（学習部１３３）

学習部１３３は、モデルを学習する。学習部１３３は、外部の情報処理装置からの情報や記憶部１２０に記憶された情報に基づいて、各種情報を学習する。学習部１３３は、学習用データ記憶部１２１に記憶された情報に基づいて、各種情報を学習する。学習部１３３は、学習により生成したモデルをモデル情報記憶部１２２に格納する。

【００６９】

学習部１３３は、学習用データを用いて、データの入力に応じて、正解データに対応する第１出力と分類ラベルに対応する第２出力とを出力するモデルを学習する。学習部１３３は、入力された音声データに対する音声認識の結果を示す第１出力と第２出力とを出力するモデルを学習する。学習部１３３は、入力された音声データが変換された文字データである第１出力と第２出力とを出力するモデルを学習する。

【００７０】
学習部１３３は、第１出力と、入力されたデータに関連するコンテキストの分類結果を示す第２出力とを出力するモデルを学習する。学習部１３３は、第１出力と、入力されたデータが検知された場所の分類結果を示す第２出力とを出力するモデルを学習する。学習部１３３は、第１出力と、入力されたデータに含まれる発話を行ったユーザに関連するコンテキストの分類結果を示す第２出力とを出力するモデルを学習する。学習部１３３は、第１出力と、ユーザの周囲の状況の分類結果を示す第２出力とを出力するモデルを学習する。学習部１３３は、第１出力と、ユーザの発話以外の分類結果を示す第２出力とを出力するモデルを学習する。学習部１３３は、第１出力と、ユーザの属性の分類結果を示す第２出力とを出力するモデルを学習する。学習部１３３は、第１出力と、ユーザが利用する端末装置の分類結果を示す第２出力とを出力するモデルを学習する。
【００７１】
学習部１３３は、モデル（ネットワーク）のパラメータを学習する。学習部１３３は、接続されたノード間の接続係数（重み）等のパラメータを学習する。学習部１３３は、種々の機械学習に関する技術を用いて、モデルを学習する。学習部１３３は、モデルに入力するデータと、そのデータが入力された場合の出力を示す正解データ及び分類ラベルとを用いて行う学習処理、すなわち教師有り学習の手法によりモデルのパラメータを学習する。なお、上記は一例であり、学習部１３３は、モデルのパラメータを学習可能であれば、どのような学習処理により、モデルのパラメータを学習してもよい。
【００７２】
（提供部１３４）
提供部１３４は、通信部１１０を介して、端末装置１０へ情報を送信する。提供部１３４は、端末装置１０へモデルを提供する。例えば、提供部１３４は、端末装置１０へ音声文字変換に用いるモデルＭ１を送信する。
【００７３】
〔４．端末装置の構成〕
次に、図５を用いて、実施形態に係る端末装置１０の構成について説明する。図５は、実施形態に係る端末装置１０の構成例を示す図である。図５に示すように、端末装置１０は、通信部１１と、記憶部１２と、入力部１３と、表示部１４と、制御部１５とを有する。なお、端末装置１０は、各種情報を音声出力するための音声出力部（例えばスピーカ等）を有してもよい。
【００７４】
（通信部１１）
通信部１１は、例えば、通信回路等によって実現される。そして、通信部１１は、図示しない所定の通信網と有線または無線で接続され、情報処理装置１００との間で情報の送受信を行う。
【００７５】
（記憶部１２）
記憶部１２は、例えば、ＲＡＭ、フラッシュメモリ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。記憶部１２は、例えば、端末装置１０にインストールされているアプリケーション（例えば音声文字変換アプリ等）に関する情報、例えばプログラム等を記憶する。また、記憶部１２は、情報処理装置１００から提供されたモデルを記憶する。例えば、記憶部１２は、モデルＭ１を記憶する。
【００７６】
（入力部１３）
入力部１３は、ユーザからの各種操作を受け付ける。入力部１３は、音声を検知する機能を有し、ユーザの発話による音声入力を受け付ける。入力部１３は、音声を検知するマイクにより検知されたユーザによる発話を入力として受け付ける。
【００７７】
また、入力部１３は、タッチパネル機能により表示面を介してユーザからの各種操作を

受け付けてもよい。また、入力部１３は、端末装置１０に設けられたボタンや、端末装置１０に接続されたキーボードやマウスからの各種操作を受け付けてもよい。

【００７８】

例えば、入力部１３は、端末装置１０の表示部１４を介してユーザの指定操作等の操作を受け付ける。例えば、入力部１３は、タッチパネルの機能によりユーザの操作を受け付ける受付部として機能する。この場合、入力部１３と受付部１５２とは一体であってもよい。なお、入力部１３によるユーザの操作の検知方式には、タブレット端末では主に静電容量方式が採用されるが、他の検知方式である抵抗膜方式、表面弾性波方式、赤外線方式、電磁誘導方式など、ユーザの操作を検知できタッチパネルの機能が実現できればどのような方式を採用してもよい。

【００７９】

（表示部１４）

表示部１４は、例えば液晶ディスプレイや有機ＥＬ（Electro-Luminescence）ディスプレイ等によって実現されるタブレット端末等の表示画面であり、各種情報を表示するための表示装置である。

【００８０】

（制御部１５）

制御部１５は、コントローラであり、例えば、ＣＰＵやＭＰＵ等によって、端末装置１０内部の記憶部１２などの記憶装置に記憶されている各種プログラムがＲＡＭを作業領域として実行されることにより実現される。例えば、この各種プログラムは、インストールされているアプリケーション（例えばメッセージアプリ等）のプログラムが含まれる。また、制御部１５は、コントローラであり、例えば、ＡＳＩＣやＦＰＧＡ等の集積回路により実現される。

【００８１】

図５に示すように、制御部１５は、受信部１５１と、受付部１５２と、推論部１５３と、処理部１５４と、送信部１５５とを有し、以下に説明する情報処理の機能や作用を実現または実行する。なお、制御部１５の内部構成は、図５に示した構成に限られず、後述する情報処理を行う構成であれば他の構成であってもよい。

【００８２】

（受信部１５１）

受信部１５１は、通信部１１を介して、情報処理装置１００から情報を受信する。受信部１５１は、情報処理装置１００から提供されたモデルを受信する。

【００８３】

受信部１５１は、機械学習のモデルの学習に用いる入力用データと、当該入力用データに含まれる認識対象を示す正解データと、認識対象に関連する分類結果を示す分類ラベルとを含む学習用データを用いて生成されたモデルであって、データの入力に応じて、正解データに対応する第１出力と分類ラベルに対応する第２出力とを出力するモデルを受信する。受信部１５１は、音声認識に関するモデルを受信する。受信部１５１は、音声データの入力に応じて、当該音声データに対応する文字データを第１出力として出力するモデルを受信する。

【００８４】

（受付部１５２）

受付部１５２は、各種情報を受け付ける。例えば、受付部１５２は、入力部１３を介してユーザによる入力を受け付ける。受付部１５２は、ユーザによる操作を受け付ける。受付部１５２は、表示部１４により表示された情報に対するユーザの操作を受け付ける。受付部１５２は、ユーザによる発話を入力として受け付ける。例えば、受付部１５２は、ユーザＵ１による「ＸＸＸＸ」という発話を入力として受け付ける。

【００８５】

（推論部１５３）

推論部１５３は、推論処理を行う。推論部１５３は、記憶部１２に記憶されたモデルを

用いて、推論処理を行う。推論部１５３は、受信部１５１により受信されたモデルを用いて推論を行う。推論部１５３は、受信部により受信されたモデルにデータを入力することにより、当該データに対応する第１出力と第２出力とを生成する推論処理を行う。推論部１５３は、モデルに音声データを入力することにより、当該音声データに対応する推論処理を行う。推論部１５３は、モデルに音声データを入力することにより、当該音声データに対応する文字データである第１出力と第２出力とを生成する推論処理を行う。

【００８６】

（処理部１５４）

処理部１５４は、推論部１５３の推論結果を用いて各種の処理を実行する。処理部１５４は、推論部１５３の推論結果を表示部１４に表示する。また、処理部１５４は、推論において入力に用いた音声データと、その音声データを書き起こした文字データと分類ラベルとのセットを学習用データとして、情報処理装置１００に提供する。処理部１５４は、推論において入力に用いた音声データと、その音声データに対応する出力結果をユーザが修正したデータとのセットを学習用データとして、情報処理装置１００に提供する。処理部１５４は、学習用データを送信部１５５に送信することを要求する。

【００８７】

（送信部１５５）

送信部１５５は、通信部１１を介して、情報処理装置１００へ情報を送信する。送信部１５５は、処理部１５４からの要求に応じて、通信部１１を介して、学習用データを情報処理装置１００に送信する。送信部１５５は、推論において入力に用いた音声データと、その音声データを書き起こした文字データと分類ラベルとのセットを学習用データとして、情報処理装置１００に送信する。送信部１５５は、推論において入力に用いた音声データと、その音声データに対応する出力結果をユーザが修正したデータとのセットを学習用データとして、情報処理装置１００に送信する。

【００８８】

なお、上述した制御部１５による各処理は、例えば、ＪａｖａＳｃｒｉｐｔ（登録商標）などにより実現されてもよい。また、上述した表示処理が所定のアプリケーション（例えば音声文字変換アプリ等）により行われる場合や推論処理等の処理が専用アプリにより行われる場合、制御部１５は、例えば、所定のアプリや専用アプリを制御するアプリ制御部を有してもよい。

【００８９】

〔５．処理フロー〕

次に、図６を用いて、実施形態に係る情報処理システム１による情報処理の手順について説明する。図６は、実施形態に係る情報処理装置による処理の一例を示すフローチャートである。

【００９０】

図６に示すように、情報処理装置１００は、機械学習のモデルの学習に用いる入力用データと、正解データと、分類ラベルとを含む学習用データを取得する（ステップＳ１０１）。すなわち、情報処理装置１００は、機械学習のモデルの学習に用いる入力用データと、当該入力用データに含まれる認識対象を示す正解データと、認識対象に関連する分類結果を示す分類ラベルとを含む学習用データを取得する。

【００９１】

情報処理装置１００は、データの入力に応じて、正解データに対応する第１出力と分類ラベルに対応する第２出力とを出力するモデルを学習する（ステップＳ１０２）。情報処理装置１００は、学習したモデルを端末装置１０へ提供する（ステップＳ１０３）。

【００９２】

次に、図７を用いて端末装置１０におけるモデルを用いた推論等の処理の流れを示す。図７は、実施形態に係る端末装置による処理の一例を示すフローチャートである。

【００９３】

図７に示すように、端末装置１０は、モデルを受信していない場合（ステップＳ２０１

：Ｎｏ）、モデルを受信するまで待機する。端末装置１０は、モデルを受信した後（ステップＳ２０１：Ｙｅｓ）、音声入力を受け付けていない場合（ステップＳ２０２：Ｎｏ）、音声入力を受け付けるまで待機する。

【００９４】

端末装置１０は、音声入力を受け付けた場合（ステップＳ２０２：Ｙｅｓ）、モデルに音声入力に対応する音声データを入力することにより、音声データに対応する第１出力と第２出力とを生成する推論処理を実行する（ステップＳ２０３）。

【００９５】

そして、端末装置１０は、推論結果を表示する（ステップＳ２０４）。また、端末装置１０は、音声データと正解データと分類ラベルとのセットを、学習用データとして情報処理装置１００へ送信してもよい。

【００９６】

〔６．効果〕

上述してきたように、実施形態に係る情報処理装置１００は、取得部１３１と、学習部１３３とを有する。取得部１３１は、機械学習のモデルの学習に用いる入力用データと、当該入力用データに含まれる認識対象を示す正解データと、認識対象に関連する分類結果を示す分類ラベルとを含む学習用データを取得する。学習部１３３は、学習用データを用いて、データの入力に応じて、正解データに対応する第１出力と分類ラベルに対応する第２出力とを出力するモデルを学習する。

【００９７】

これにより、実施形態に係る情報処理装置１００は、入力用データに含まれる認識対象を示す第１出力と、認識対象に関連する分類結果を示す第２出力との複数の種別の出力を行うモデルを学習することができる。したがって、情報処理装置１００は、複数種別の出力を行うモデルを利用可能にすることができる。

【００９８】

また、実施形態に係る情報処理装置１００において、取得部１３１は、音声データである入力用データと、当該入力用データに含まれる音声の音声認識の結果を示す正解データとを含む学習用データを取得する。学習部１３３は、入力された音声データに対する音声認識の結果を示す第１出力と第２出力とを出力するモデルを学習する。

【００９９】

これにより、実施形態に係る情報処理装置１００は、音声認識に関して複数の種別の出力を行うモデルを柔軟に学習することができ、複数種別の出力を行うモデルを利用可能にすることができる。

【０１００】

また、実施形態に係る情報処理装置１００において、取得部１３１は、入力用データに対応する文字データである正解データを含む学習用データを取得する。学習部１３３は、入力された音声データが変換された文字データである第１出力と第２出力とを出力するモデルを学習する。

【０１０１】

これにより、実施形態に係る情報処理装置１００は、音声文字変換に関して複数の種別の出力を行うモデルを柔軟に学習することができ、複数種別の出力を行うモデルを利用可能にすることができる。

【０１０２】

また、実施形態に係る情報処理装置１００において、取得部１３１は、入力用データに関連するコンテキストの分類結果を示す分類ラベルを含む学習用データを取得する。学習部１３３は、第１出力と、入力されたデータに関連するコンテキストの分類結果を示す第２出力とを出力するモデルを学習する。

【０１０３】

これにより、実施形態に係る情報処理装置１００は、認識結果を示す第１出力と、入力用データに関連するコンテキストの分類結果を示す第２出力を行うモデルを柔軟に学習す

ることができ、複数種別の出力を行うモデルを利用可能にすることができる。
【０１０４】
　また、実施形態に係る情報処理装置１００において、取得部１３１は、入力用データが検知された場所の分類結果を示す分類ラベルを含む学習用データを取得する。学習部１３３は、第１出力と、入力されたデータが検知された場所の分類結果を示す第２出力とを出力するモデルを学習する。
【０１０５】
　これにより、実施形態に係る情報処理装置１００は、認識結果を示す第１出力と、入力用データが検知された場所の分類結果を示す第２出力を行うモデルを柔軟に学習することができ、複数種別の出力を行うモデルを利用可能にすることができる。
【０１０６】
　また、実施形態に係る情報処理装置１００において、取得部１３１は、入力用データに含まれる発話を行ったユーザに関連するコンテキストの分類結果を示す分類ラベルを含む学習用データを取得する。学習部１３３は、第１出力と、入力されたデータに含まれる発話を行ったユーザに関連するコンテキストの分類結果を示す第２出力とを出力するモデルを学習する。
【０１０７】
　これにより、実施形態に係る情報処理装置１００は、認識結果を示す第１出力と、発話を行ったユーザに関連するコンテキストの分類結果を示す第２出力を行うモデルを柔軟に学習することができ、複数種別の出力を行うモデルを利用可能にすることができる。
【０１０８】
　また、実施形態に係る情報処理装置１００において、取得部１３１は、ユーザの発話以外の分類結果を示す分類ラベルを含む学習用データを取得する。学習部１３３は、第１出力と、ユーザの発話以外の分類結果を示す第２出力とを出力するモデルを学習する。
【０１０９】
　これにより、実施形態に係る情報処理装置１００は、認識結果を示す第１出力と、ユーザの発話以外の分類結果を示す第２出力を行うモデルを柔軟に学習することができ、複数種別の出力を行うモデルを利用可能にすることができる。
【０１１０】
　また、実施形態に係る情報処理装置１００において、取得部１３１は、ユーザの周囲の状況の分類結果を示す分類ラベルを含む学習用データを取得する。学習部１３３は、第１出力と、ユーザの周囲の状況の分類結果を示す第２出力とを出力するモデルを学習する。
【０１１１】
　これにより、実施形態に係る情報処理装置１００は、認識結果を示す第１出力と、発話を行ったユーザの周囲の状況の分類結果を示す第２出力を行うモデルを柔軟に学習することができ、複数種別の出力を行うモデルを利用可能にすることができる。
【０１１２】
　また、実施形態に係る情報処理装置１００において、取得部１３１は、ユーザの属性の分類結果を示す分類ラベルを含む学習用データを取得する。学習部１３３は、第１出力と、ユーザの属性の分類結果を示す第２出力とを出力するモデルを学習する。
【０１１３】
　これにより、実施形態に係る情報処理装置１００は、認識結果を示す第１出力と、発話を行ったユーザの属性の分類結果を示す第２出力を行うモデルを柔軟に学習することができ、複数種別の出力を行うモデルを利用可能にすることができる。
【０１１４】
　また、実施形態に係る情報処理装置１００において、取得部１３１は、ユーザが利用する端末装置の分類結果を示す分類ラベルを含む学習用データを取得する。学習部１３３は、第１出力と、ユーザが利用する端末装置の分類結果を示す第２出力とを出力するモデルを学習する。
【０１１５】

これにより、実施形態に係る情報処理装置１００は、認識結果を示す第１出力と、発話を行ったユーザが利用する端末装置の分類結果を示す第２出力を行うモデルを柔軟に学習することができ、複数種別の出力を行うモデルを利用可能にすることができる。

【０１１６】

また、実施形態に係る情報処理装置１００において、取得部１３１は、入力用データに含まれる情報のうち、認識対象以外の情報の種別を示す分類ラベルを取得する。

【０１１７】

これにより、実施形態に係る情報処理装置１００は、入力用データに含まれる情報のうち、認識対象以外の情報の種別を示す第２出力を行うモデルを柔軟に学習することができ、複数種別の出力を行うモデルを利用可能にすることができる。

【０１１８】

また、実施形態に係る端末装置１０は、受信部１５１と、推論部１５３とを有する。受信部１５１は、機械学習のモデルの学習に用いる入力用データと、当該入力用データに含まれる認識対象を示す正解データと、認識対象に関連する分類結果を示す分類ラベルとを含む学習用データを用いて生成されたモデルであって、データの入力に応じて、正解データに対応する第１出力と分類ラベルに対応する第２出力とを出力するモデルを受信する。推論部１５３は、受信部により受信されたモデルにデータを入力することにより、当該データに対応する第１出力と第２出力とを生成する推論処理を行う。

【０１１９】

これにより、実施形態に係る端末装置１０は、入力したデータに含まれる認識対象を示す第１出力と、認識対象に関連する分類結果を示す第２出力との複数の種別の出力を行うモデルを用いて推論処理を行うことができる。したがって、端末装置１０は、複数種別の出力を行うモデルを利用した処理を行うことができる。

【０１２０】

また、実施形態に係る端末装置１０おいて、受信部１５１は、音声認識に関するモデルを受信する。推論部１５３は、モデルに音声データを入力することにより、当該音声データに対応する推論処理を行う。

【０１２１】

これにより、実施形態に係る情報処理装置１００は、音声認識に関する第１出力及び第２出力を出力するモデルを用いて推論処理を行うことができるため、複数種別の出力を行うモデルを利用した処理を行うことができる。

【０１２２】

また、実施形態に係る端末装置１０おいて、受信部１５１は、音声データの入力に応じて、当該音声データに対応する文字データを第１出力として出力するモデルを受信する。推論部１５３は、モデルに音声データを入力することにより、当該音声データに対応する文字データである第１出力と第２出力とを生成する推論処理を行う。

【０１２３】

これにより、実施形態に係る情報処理装置１００は、音声文字変換の結果である第１出力と第２出力との２つの異なる種別の情報を出力するモデルを用いて推論処理を行うことができるため、複数種別の出力を行うモデルを利用した処理を行うことができる。

【０１２４】

〔７．ハードウェア構成〕

また、上述した実施形態に係る端末装置１０や情報処理装置１００は、例えば図８に示すような構成のコンピュータ１０００によって実現される。以下、情報処理装置１００を例に挙げて説明する。図８は、ハードウェア構成の一例を示す図である。コンピュータ１０００は、出力装置１０１０、入力装置１０２０と接続され、演算装置１０３０、一次記憶装置１０４０、二次記憶装置１０５０、出力Ｉ／Ｆ（Interface）１０６０、入力Ｉ／Ｆ１０７０、ネットワークＩ／Ｆ１０８０がバス１０９０により接続された形態を有する。

【０１２５】

演算装置１０３０は、一次記憶装置１０４０や二次記憶装置１０５０に格納されたプロ

グラムや入力装置１０２０から読み出したプログラム等に基づいて動作し、各種の処理を実行する。演算装置１０３０は、例えばＣＰＵ（Central Processing Unit）、ＭＰＵ（Micro Processing Unit）、ＡＳＩＣ（Application Specific Integrated Circuit）やＦＰＧＡ（Field Programmable Gate Array）等により実現される。

【０１２６】

一次記憶装置１０４０は、ＲＡＭ（Random Access Memory）等、演算装置１０３０が各種の演算に用いるデータを一次的に記憶するメモリ装置である。また、二次記憶装置１０５０は、演算装置１０３０が各種の演算に用いるデータや、各種のデータベースが登録される記憶装置であり、ＲＯＭ（Read Only Memory）、ＨＤＤ（Hard Disk Drive）、ＳＳＤ（Solid State Drive）、フラッシュメモリ等により実現される。二次記憶装置１０５０は、内蔵ストレージであってもよいし、外付けストレージであってもよい。また、二次記憶装置１０５０は、ＵＳＢメモリやＳＤ（Secure Digital）メモリカード等の取り外し可能な記憶媒体であってもよい。また、二次記憶装置１０５０は、クラウドストレージ（オンラインストレージ）やＮＡＳ（Network Attached Storage）、ファイルサーバ等であってもよい。

【０１２７】

出力Ｉ／Ｆ１０６０は、ディスプレイ、プロジェクタ、及びプリンタ等といった各種の情報を出力する出力装置１０１０に対し、出力対象となる情報を送信するためのインターフェイスであり、例えば、ＵＳＢ（Universal Serial Bus）やＤＶＩ（Digital Visual Interface）、ＨＤＭＩ（登録商標）（High Definition Multimedia Interface）といった規格のコネクタにより実現される。また、入力Ｉ／Ｆ１０７０は、マウス、キーボード、キーパッド、ボタン、及びスキャナ等といった各種の入力装置１０２０から情報を受信するためのインターフェイスであり、例えば、ＵＳＢ等により実現される。

【０１２８】

また、出力Ｉ／Ｆ１０６０及び入力Ｉ／Ｆ１０７０はそれぞれ出力装置１０１０及び入力装置１０２０と無線で接続してもよい。すなわち、出力装置１０１０及び入力装置１０２０は、ワイヤレス機器であってもよい。

【０１２９】

また、出力装置１０１０及び入力装置１０２０は、タッチパネルのように一体化していてもよい。この場合、出力Ｉ／Ｆ１０６０及び入力Ｉ／Ｆ１０７０も、入出力Ｉ／Ｆとして一体化していてもよい。

【０１３０】

なお、入力装置１０２０は、例えば、ＣＤ（Compact Disc）、ＤＶＤ（Digital Versatile Disc）、ＰＤ（Phase change rewritable Disk）等の光学記録媒体、ＭＯ（Magneto-Optical disk）等の光磁気記録媒体、テープ媒体、磁気記録媒体、又は半導体メモリ等から情報を読み出す装置であってもよい。

【０１３１】

ネットワークＩ／Ｆ１０８０は、ネットワークＮを介して他の機器からデータを受信して演算装置１０３０へ送り、また、ネットワークＮを介して演算装置１０３０が生成したデータを他の機器へ送信する。

【０１３２】

演算装置１０３０は、出力Ｉ／Ｆ１０６０や入力Ｉ／Ｆ１０７０を介して、出力装置１０１０や入力装置１０２０の制御を行う。例えば、演算装置１０３０は、入力装置１０２０や二次記憶装置１０５０からプログラムを一次記憶装置１０４０上にロードし、ロードしたプログラムを実行する。

【０１３３】

例えば、コンピュータ１０００が情報処理装置１００として機能する場合、コンピュータ１０００の演算装置１０３０は、一次記憶装置１０４０上にロードされたプログラムを実行することにより、制御部１３０の機能を実現する。また、コンピュータ１０００の演算装置１０３０は、ネットワークＩ／Ｆ１０８０を介して他の機器から取得したプログラ

ムを一次記憶装置１０４０上にロードし、ロードしたプログラムを実行してもよい。また、コンピュータ１０００の演算装置１０３０は、ネットワークＩ／Ｆ１０８０を介して他の機器と連携し、プログラムの機能やデータ等を他の機器の他のプログラムから呼び出して利用してもよい。

【０１３４】

〔８．その他〕

以上、本願の実施形態を説明したが、これら実施形態の内容により本発明が限定されるものではない。また、前述した構成要素には、当業者が容易に想定できるもの、実質的に同一のもの、いわゆる均等の範囲のものが含まれる。さらに、前述した構成要素は適宜組み合わせることが可能である。さらに、前述した実施形態の要旨を逸脱しない範囲で構成要素の種々の省略、置換又は変更を行うことができる。

【０１３５】

また、上記実施形態において説明した各処理のうち、自動的に行われるものとして説明した処理の全部又は一部を手動的に行うこともでき、あるいは、手動的に行われるものとして説明した処理の全部又は一部を公知の方法で自動的に行うこともできる。この他、上記文書中や図面中で示した処理手順、具体的名称、各種のデータやパラメータを含む情報については、特記する場合を除いて任意に変更することができる。例えば、各図に示した各種情報は、図示した情報に限られない。

【０１３６】

また、図示した各装置の各構成要素は機能概念的なものであり、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各装置の分散・統合の具体的形態は図示のものに限られず、その全部又は一部を、各種の負荷や使用状況などに応じて、任意の単位で機能的又は物理的に分散・統合して構成することができる。

【０１３７】

例えば、上述した情報処理装置１００は、複数のサーバコンピュータで実現してもよく、また、機能によっては外部のプラットホーム等をＡＰＩ（Application Programming Interface）やネットワークコンピューティング等で呼び出して実現するなど、構成は柔軟に変更できる。また、情報処理装置１００と端末装置１０とは一体であってもよい。この場合、例えばユーザが利用する端末装置１０が情報処理装置１００としての機能を有してもよい。

【０１３８】

また、上述してきた実施形態及び変形例は、処理内容を矛盾させない範囲で適宜組み合わせることが可能である。

【０１３９】

また、上述してきた「部（section、module、unit）」は、「手段」や「回路」などに読み替えることができる。例えば、取得部は、取得手段や取得回路に読み替えることができる。

【符号の説明】

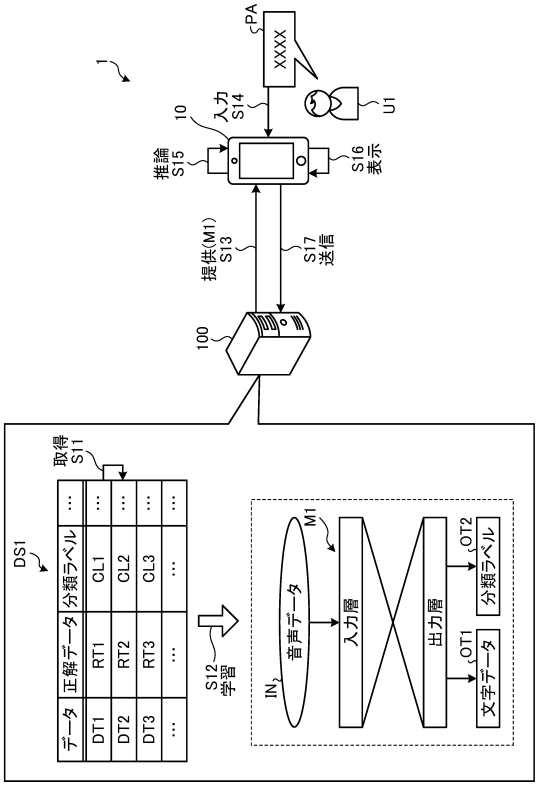
【０１４０】

１　情報処理システム
１００　情報処理装置
１２０　記憶部
１２１　学習用データ記憶部
１２２　モデル情報記憶部
１３０　制御部
１３１　取得部
１３２　決定部
１３３　学習部
１３４　提供部
１０　端末装置

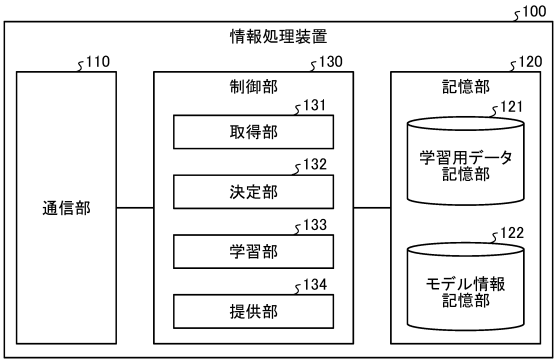
１１　通信部
１２　記憶部
１３　入力部
１４　表示部
１５　制御部
１５１　受信部
１５２　受付部
１５３　推論部
１５４　処理部
１５５　送信部

【図面】

【図１】



【図２】



【図３】

121

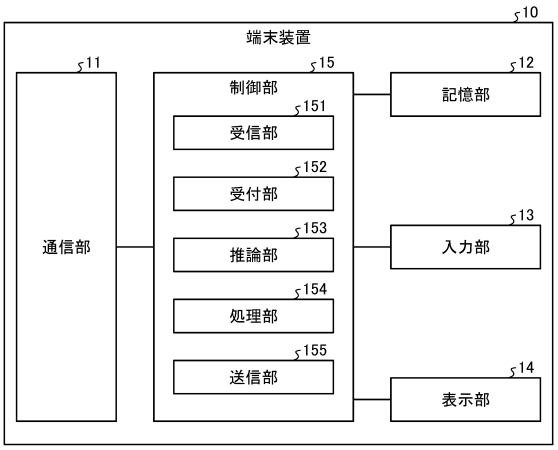
データセットID	データID	データ	正解データ	分類ラベル	…
DS1	DID1	DT1	RT1	CL1	…
	DID2	DT2	RT2	CL2	…
	DID3	DT3	RT3	CL3	…
	DID4	DT4	RT4	CL4	…
	DID5	DT5	RT5	CL5	…
	DID6	DT6	RT6	CL6	…
	DID7	DT7	RT7	CL7	…
	DID8	DT8	RT8	CL8	…
	…	…	…	…	…
…	…	…	…	…	…

【図４】

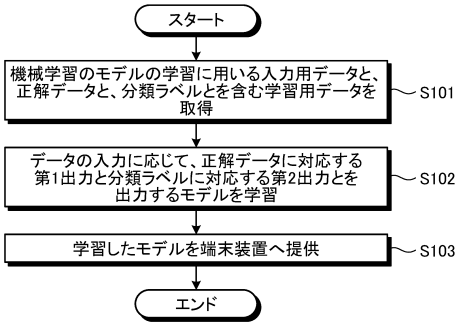
122

モデルID	用途	モデルデータ	…
M1	音声文字変換 コンテキスト推定	MDT1	…
…	…	…	…

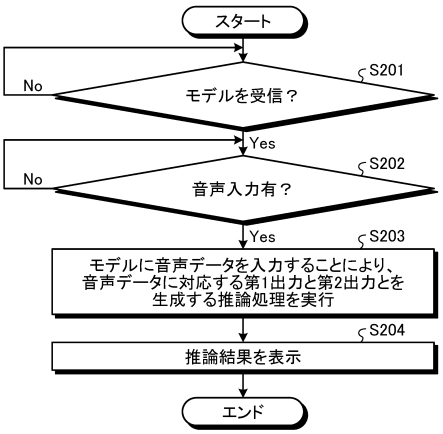
【図５】



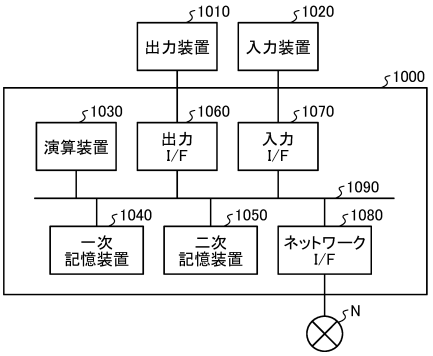
【図６】



【図７】


スタート
S201
No
モデルを受信？
Yes
S202
No
音声入力有？
Yes
S203
モデルに音声データを入力することにより、
音声データに対応する第1出力と第2出力とを
生成する推論処理を実行
S204
推論結果を表示
エンド

【図８】


1010
出力装置
1020
入力装置
1000
1030
演算装置
1060
出力
I/F
1070
入力
I/F
1090
1040
一次
記憶装置
1050
二次
記憶装置
1080
ネットワーク
I/F
N

フロントページの続き

(56)参考文献 特開２０１９－０８７２２９（ＪＰ，Ａ）
特開２０２０－０６４２５３（ＪＰ，Ａ）
特開２０２０－１４０６７３（ＪＰ，Ａ）
大町 基，単語の表記と素性を同時出力するｅｎｄ－ｔｏ－ｅｎｄ音声認識，日本音響学会２０２０年 秋季研究発表会講演論文集ＣＤ－ＲＯＭ［ＣＤ－ＲＯＭ］，一般社団法人日本音響学会，2020年08月26日，pp.815-818
早川 友瑛，Ｅｎｄ－ｔｏ－Ｅｎｄ複数言語音声認識モデルにおける様々なマルチタスク学習の検討，日本音響学会 ２０２０年 秋季研究発表会講演論文集ＣＤ－ＲＯＭ［ＣＤ－ＲＯＭ］，一般社団法人日本音響学会，2020年08月26日，pp.833-834
松原 拓未，ＣＮＮ Ａｕｔｏｅｎｃｏｄｅｒから抽出したボトルネック特徴量を用いた環境音分類，マルチメディア，分散，協調とモバイル（ＤＩＣＯＭＯ２０１９）シンポジウム論文集［ＣＤ－ＲＯＭ］，一般社団法人情報処理学会，2019年06月26日，Vol.2019, No.1，p.339-346，ISSN: 1882-0840

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｎ ２０／００
Ｇ０６Ｎ ３／０８
Ｇ１０Ｌ １５／１０
Ｇ１０Ｌ １５／１６