(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2018-173522
(P2018-173522A)
(43)公開日　平成30年11月8日(2018.11.8)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 15/22	(2006.01)	G10L 15/22	200V	
G10L 15/10	(2006.01)	G10L 15/10	500Z	

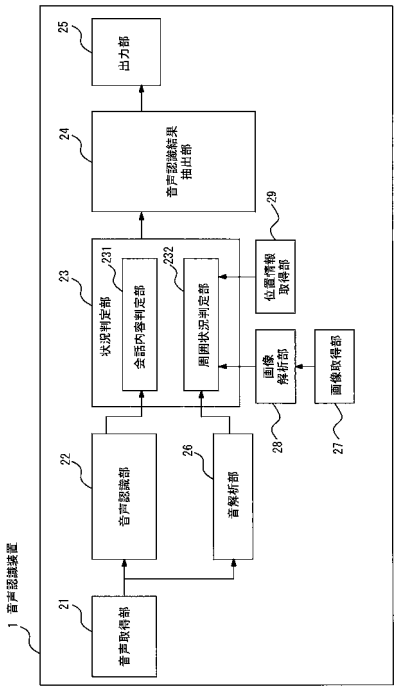
審査請求 未請求　請求項の数 9　OL　（全 12 頁）

(21)出願番号　特願2017-71168(P2017-71168)
(22)出願日　平成29年3月31日(2017.3.31)

(71)出願人　000002897
大日本印刷株式会社
東京都新宿区市谷加賀町一丁目１番１号
(74)代理人　100096091
弁理士　井上　誠一
(72)発明者　松本　征二
東京都新宿区市谷加賀町一丁目１番１号
大日本印刷株式会社内

(54)【発明の名称】音声認識装置、音声認識方法、及びプログラム

(57)【要約】
【課題】　状況に適した音声認識結果を得ることが可能な音声認識装置、音声認識方法、及びプログラムを提供する。
【解決手段】　音声認識装置１は、会話の音声データを取得し、音声認識部２２により音声データをテキストに変換し、音声認識結果として複数の変換候補を求める。また、状況判定部２３は、会話の内容のジャンルや、周囲の音の特徴等を解析したりすることにより、コンテキストや周囲の状況（場所、シーン、会話の目的）等を求め、会話の状況を判定する。音声認識結果抽出部２４は、複数の変換候補の中から会話の状況に適した変換候補を抽出し、出力する。
【選択図】図１



【特許請求の範囲】
【請求項１】
音声データを取得する音声取得手段と、
取得した音声データを認識し、音声認識結果として複数の変換候補を求める音声認識手段と、
音声データ取得時の状況を判定する状況判定手段と、
前記音声認識手段により求めた複数の変換候補の中から前記状況判定手段により判定した状況に適した変換候補を抽出する抽出手段と、
を備えることを特徴とする音声認識装置。
【請求項２】
状況と関連付けられた単語を記憶する記憶手段を備え、
前記状況判定手段は、前記単語を音声データから抽出し、前記単語に基づいて前記状況を判定することを特徴とする請求項１に記載の音声認識装置。
【請求項３】
前記状況判定手段は、更に、取得される周囲の音に基づいて前記状況を判定することを特徴とする請求項２に記載の音声認識装置。
【請求項４】
前記状況判定手段は、更に、取得される画像に基づいて前記状況を判定することを特徴とする請求項２または請求項３に記載の音声認識装置。
【請求項５】
前記状況判定手段は、更に、取得される位置情報に基づいて前記状況を判定することを特徴とする請求項２から請求項４のいずれかに記載の音声認識装置。
【請求項６】
前記状況判定手段により判定した会話の状況の履歴を記憶する記憶手段を更に備え、
前記抽出手段は、前記記憶手段に記憶された会話の状況の履歴に基づき、前記音声認識手段により求めた複数の変換候補の中から会話の状況に適した変換候補を抽出することを特徴とする請求項１から請求項５のいずれかに記載の音声認識装置。
【請求項７】
前記抽出手段は、設定された尤度に基づいて抽出候補に優先付けを行うことを特徴とする請求項１から請求項６のいずれかに記載の音声認識装置。
【請求項８】
コンピュータが、
音声データを取得するステップと、
取得した音声データを認識し、音声認識結果として複数の変換候補を求めるステップと、
音声データ取得時の状況を判定するステップと、
前記複数の変換候補の中から前記状況に適した変換候補を抽出するステップと、
を含むことを特徴とする音声認識方法。
【請求項９】
コンピュータを、
音声データを取得する音声取得手段、
取得した音声データを認識し、音声認識結果として複数の変換候補を求める音声認識手段、
音声データ取得時の状況を判定する状況判定手段、
前記音声認識手段により求めた複数の変換候補の中から前記状況判定手段により判定した状況に適した変換候補を抽出する抽出手段、
として機能させるためのプログラム。

【発明の詳細な説明】
【技術分野】

【０００１】
本発明は、音声認識装置、音声認識方法、及びプログラムに関し、詳細には、音声認識精度を向上するための技術に関する。
【背景技術】
【０００２】
従来より、スマートフォンやカーナビゲーションシステム等のユーザインターフェースとして音声入力を用いるものが利用されている。また、ＡＩ（Artificial Intelligence；人工知能）や対話型ロボットが普及し始め、機器と人とが音声によってコミュニケーションをとる機会が増加している。そのため、精度のよい音声認識システムが要望されている。例えば、特許文献１には、ユーザが発話する言葉に含まれるユーザの意図や感情を推定することで、ユーザとの間の対話をより円滑にする機能を有する音声対話装置について記載されている。
【０００３】
また、従来の音声認識処理では、音声認識の結果、複数の変換候補がある場合にはシステムが第1候補を自動的に選択する方法等が一般的である。例えば、「おすすめのたいけんはありますか」という音声が入力された際の音声認識結果としては、「Ｎｏ．１：おすすめの大剣は何ですか」、「Ｎｏ．２：おすすめの体験は何ですか」、「Ｎｏ．３：お勧めの体験はなんですか」、「Ｎｏ．４：お奨めの体験は何ですか」、「Ｎｏ．５：おすすめの大剣はなんですか」のような複数の変換候補が得られる。この場合、観光のシーンであればＮｏ．２、３、４が正しい変換候補となるが、ゲームについての会話中であれば、Ｎｏ．１、あるいは５が正しい変換となる。したがって適切な音声の文書変換を行うには、シーンや会話の流れ等の状況を把握し、それに応じた候補を出力する必要がある。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２００６－３１３２８７号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
例えば、上述の特許文献１には、ユーザの感情や生理状態等を音声、画像、生理信号等の非言語情報を用いて入力音声の意図（質問、詰問、疑い）を推定し判断する技術について開示されている。しかしながら、特許文献１は、ユーザの感情や生理状態によるユーザ情報のみから対話内容を推定するため、簡単な応答文など２者択一形式の認識には対応できるものの、前述のように多義的であって複雑な会話内容に対しては適切な認識を行うことが困難である。
【０００６】
本発明は、このような課題に鑑みてなされたもので、会話のシーンやコンテキストといった状況に適した音声認識結果を得ることが可能な音声認識装置、音声認識方法、及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】
【０００７】
前述した課題を解決するため第１の発明は、音声データを取得する音声取得手段と、取得した音声データを認識し、音声認識結果として複数の変換候補を求める音声認識手段と、音声データ取得時の状況を判定する状況判定手段と、前記音声認識手段により求めた複数の変換候補の中から前記状況判定手段により判定した状況に適した変換候補を抽出する抽出手段と、を備えることを特徴とする音声認識装置である。
【０００８】
第１の発明によれば、会話の音声データを認識し、音声認識結果として複数の変換候補を求め、会話の状況（シーンやコンテキスト）を判定し、複数の変換候補の中から会話の状況に適した変換候補を抽出する。これにより、シーンやコンテキストといった会話の状

況に適した音声認識結果を得ることが可能となり、音声認識精度の高い音声認識装置を提供できる。

【０００９】

第１の発明において、前記状況判定手段は、会話の内容に基づいて前記状況を求めることが望ましい。これにより、会話の内容そのものから状況を推定し、適切な音声認識結果を抽出可能となる。また、前記状況判定手段は、更に、取得される周囲の音に基づいて前記状況を求めてもよい。これにより、会話が行われている状況を周囲の音の特徴から取得できるため、より正確に音声認識結果を得ることができる。

【００１０】

また、前記状況判定手段は、更に、取得される画像に基づいて前記状況を求めてもよい。更に、前記状況判定手段は、更に、取得される位置情報に基づいて前記状況を求めてもよい。音の特徴のみならず、画像や位置情報等からも会話の状況を求めることで、状況を正確に求めることが可能となり、これにより正確な音声認識結果を得ることが可能となる。

【００１１】

また、前記状況判定手段により判定した会話の状況の履歴を記憶する記憶手段を更に備え、前記抽出手段は、前記記憶手段に記憶された会話の状況の履歴に基づき、前記音声認識手段により求めた複数の変換候補の中から会話の状況に適した変換候補を抽出することが望ましい。これにより、会話の状況の履歴から、話の流れを認識することが可能となり、コンテキストに適した音声認識結果をより正確に得ることが可能となる。

【００１２】

第２の発明は、コンピュータが、会話の音声データを取得するステップと、取得した音声データを認識し、音声認識結果として複数の変換候補を求めるステップと、会話の状況を判定するステップと、前記複数の変換候補の中から前記会話の状況に適した変換候補を抽出するステップと、を含むことを特徴とする音声認識方法である。

【００１３】

第２の発明によれば、コンピュータは、会話の音声データを認識し、音声認識結果として複数の変換候補を求め、会話の状況を判定し、複数の変換候補の中から会話の状況に適した変換候補を抽出する。これにより、シーンやコンテキストといった会話の状況に適した音声認識結果を得ることが可能となり、音声認識精度を向上させることが可能となる。

【００１４】

第３の発明は、コンピュータを、会話の音声データを取得する音声取得手段、
取得した音声データを認識し、音声認識結果として複数の変換候補を求める音声認識手段、会話の状況を判定する状況判定手段、前記音声認識手段により求めた複数の変換候補の中から前記状況判定手段により判定した会話の状況に適した変換候補を抽出する抽出手段、として機能させるためのプログラムである。

【００１５】

第３の発明により、コンピュータを第１の発明の音声認識装置として機能させることが可能となる。

【発明の効果】

【００１６】

本発明により、シーンやコンテキストといった状況に適した音声認識結果を得ることが可能な音声認識装置、音声認識方法、及びプログラムを提供できる。

【図面の簡単な説明】

【００１７】

【図１】音声認識装置１の構成図

【図２】音声認識装置１として機能するコンピュータ１０の構成図

【図３】音声認識装置１が実行する音声認識処理の流れを示すフローチャート

【図４】音声認識の変換候補と会話のジャンルとを関連づけたデータであるジャンルデータ５の例

【図５】会話内容による状況判定結果、及び音による状況判定結果の具体例
【図６】本発明に係る音声認識装置１を利用した音声認識システム１００の例
【発明を実施するための形態】
【００１８】
以下、図面に基づいて本発明の好適な実施形態について詳細に説明する。
【００１９】
図１は、本発明に係る音声認識装置１の構成を示す図である。音声認識装置１は、音声取得部２１、音声認識部２２、状況判定部２３、音声認識結果抽出部２４、出力部２５、及び音解析部２６を備える。また、これらの構成に加え、画像取得部２７、画像解析部２８、及び位置情報取得部２９を備えてもよい。
【００２０】
図２は、音声認識装置１として機能させるコンピュータ１０の構成例を示す図である。図２に示すように、コンピュータ１０は、制御部１１、記憶部１２、メディア入出力部１３、周辺機器Ｉ／Ｆ部１４、入力部１５、表示部１６、通信制御部１７、マイク１８等がバス１９を介して接続されて構成される。コンピュータ１０を音声認識装置１として機能させる場合、コンピュータ１０の制御部１１は、図１に示す音声認識装置１の各部（音声取得部２１、音声認識部２２、状況判定部２３、音声認識結果抽出部２４、出力部２５、音解析部２６、画像取得部２７、画像解析部２８、及び位置情報取得部２９）の機能を記述したプログラムを実行する。
【００２１】
制御部１１は、ＣＰＵ（Central Processing Unit）、ＲＯＭ（Read Only Memory）、ＲＡＭ（Random Access
Memory）等により構成される。
ＣＰＵは、記憶部１２、ＲＯＭ、記録媒体等に格納されるプログラムをＲＡＭ上のワークメモリ領域に呼び出して実行し、バス１９を介して接続された各部を駆動制御する。ＲＯＭは、コンピュータ１０のブートプログラムやＢＩＯＳ等のプログラム、データ等を恒久的に保持する。ＲＡＭは、ロードしたプログラムやデータを一時的に保持するとともに、制御部１１が各種処理を行うために使用するワークエリアを備える。制御部１１は、上記プログラムを読み出して実行することにより、図１に示す音声認識装置１の各部（音声取得部２１、音声認識部２２、状況判定部２３、音声認識結果抽出部２４、出力部２５、及び音解析部２６等）として機能する。すなわち、制御部１１はマイク１８または通信制御部１７等から入力された音声データを取得し、取得した音声データについて後述する音声認識処理（図３参照）を実行する。
【００２２】
記憶部１２は、例えば、ハードディスクドライブ等の記憶装置である。記憶部１２には制御部１１が実行するプログラムや、プログラム実行に必要なデータ、オペレーティングシステム等が格納されている。これらのプログラムコードは、制御部１１により必要に応じて読み出されてＲＡＭに移され、ＣＰＵに読み出されて実行される。
【００２３】
メディア入出力部１３は、例えば、ＣＤ、ＤＶＤ、ＭＯ等の各種記録媒体（メディア）のドライブ装置であり、メディアに対してデータの入出力（書込み／読み出し）を行う。
【００２４】
周辺機器Ｉ／Ｆ（インタフェース）部１４は、周辺機器を接続させるためのポートであり、周辺機器Ｉ／Ｆ部１４を介して周辺機器とのデータの送受信を行う。周辺機器Ｉ／Ｆ部１４は、ＵＳＢ等で構成されており、通常複数の周辺機器Ｉ／Ｆを有する。周辺機器との接続形態は有線、無線を問わない。
【００２５】
入力部１５は、例えば、キーボード、マウス等のポインティング・デバイス、テンキー等の入力装置であり、入力されたデータを制御部１１へ出力する。
表示部１６は、例えば液晶パネル、ＣＲＴモニタ等のディスプレイ装置と、ディスプレ

イ装置と連携して表示処理を実行するための論理回路（ビデオアダプタ等）で構成され、制御部１１の制御により入力された表示情報をディスプレイ装置上に表示させる。なお、入力部１５及び表示部１６は、表示画面にタッチパネル等の入力装置を一体的に設けたタッチパネルディスプレイとしてもよい。

【００２６】

通信制御部１７は、通信制御装置、通信ポート等を有し、ネットワーク３等との通信を制御する。

マイク１８は、音声を収集し、音声データとして制御部１１に入力する。

バス１９は、各装置間の制御信号、データ信号等の授受を媒介する経路である。

【００２７】

図１を参照して本発明に係る音声認識装置１の機能構成を説明する。

音声取得部２１は、会話の音声データを取得する。会話の音声データは、音声認識装置１がマイク１８を備えるものであれば、マイク１８から入力された音声データでもよいし、通信制御部１７及びネットワーク３を介して音声認識装置１と通信接続された機器とから入力されたものでもよい。

【００２８】

音声認識部２２は、取得した音声データのユーザの発話の内容を音声認識し、音声認識結果として１または複数の変換候補を求める。音声認識部２２は、発話の音声データと語とを対応付けた発話辞書や、音響モデル、言語モデル等の音声認識用データを有し、これらの音声認識用データを用いて、発話の音響や言語を解析し、発話の内容をテキストに変換する音声認識処理を行う。音声認識結果であるテキストは、状況判定部２３の会話内容判定部２３１に出力される。

【００２９】

状況判定部２３は、会話の状況を判定する。会話の状況とは、具体的には、会話の内容（ジャンル）、及び会話が行われている場所や目的等の周囲状況（シーン）である。図１に示すように、状況判定部２３は、会話内容を判定するための会話内容判定部２３１と、周囲状況を判定するための周囲状況判定部２３２とを有する。

【００３０】

会話内容判定部２３１は、会話の文に含まれる単語を解析することにより、会話の内容を求める。ここで求める会話の内容とは、話のジャンルまたは目的等である。ジャンルとは、「観光」、「ゲーム」、「飲食」、「映画」、「学校」、「医療」、…等のように、何についての会話であるかを示す分類である。目的とは、「接客」や「雑談」等のように会話がどのような目的で行われているかを示す分類である。会話内容判定部２３１は、例えば、単語とジャンルとを関連付けたデータをジャンルデータ５（図４参照）として記憶部１２に予め記憶しており、このジャンルデータ５を参照することにより会話の内容（ジャンル）を判定する。各単語は複数のジャンルに跨って含まれていてもよい。会話内容判定部２３１は、音声認識結果として得られる１または複数の文に含まれる単語から、会話内容の候補を求める。また、判定対象とする文だけでなく、それより前に入力された音声データから認識された文（音声認識結果）を判定対象に含むようにすることが望ましい。これにより、コンテキスト（文脈）を考慮して会話の内容を求めることができる。

【００３１】

周囲状況判定部２３２は、会話の音声データが入力されたときの周囲の音データの特徴に基づいて場所等の周囲状況を求める。音解析部２６は、音声取得部２１により取得した音声データから周囲の音データを抽出し、この周囲の音データの特徴を抽出し、周囲状況判定部２３２に出力する。周囲状況判定部２３２は、抽出した音データの特徴と状況とを関連付けたデータを音解析用データとして記憶部１２に予め記憶しており、この音解析用データに基づいて会話の周囲状況を判定する。例えば、「レストラン」の音解析用データには、食器等の音やテーブルでの会話、接客の音等の特徴が含まれる。また「アミューズメントパーク」の音解析用データには、歓声やアトラクションの音等の特徴が含まれる。周囲状況判定部２３２は、音データの特徴から１または複数の周囲状況の候補を求める。

【００３２】
なお、周囲状況判定部２３２は、音解析のみならず、画像や位置情報に基づいて周囲状況を判定してもよい。具体的には、図１に示すように画像取得部２７により会話中の様子や場所を撮影した画像（映像または静止画）等を解析する画像解析部２８を備え、画像解析部２８によって会話の場所や目的等、周囲状況を判定してもよい。また、ＧＰＳ（Global Positioning System）等の位置情報取得部２９を更に備え、周囲状況判定部２３２は、位置情報及び地図データ等に基づいて会話の場所（店舗や施設）等を求めることにより周囲状況を求めてもよい。
【００３３】
音声認識結果抽出部２４は、音声認識部２２により求めた複数の変換候補の中から状況判定部２３により判定した会話の状況に適した変換候補を抽出する。変換候補の抽出については後述する。
【００３４】
出力部２５は、音声認識結果抽出部２４により抽出した変換候補（テキスト）を出力する。出力は、表示部１６への表示や、制御部１１への通知、ネットワーク３を介した通信接続先への送信等、当該音声認識装置１に接続された各種機器に対する制御情報としての送信等も含むものとする。
【００３５】
次に、図３を参照して、音声認識装置１が実行する音声認識処理について説明する。
制御部１１は、記憶部１２から図３に示す音声認識処理に関するプログラム及びデータを読み出し、このプログラム及びデータに基づいて処理を実行する。
【００３６】
まず制御部１１（音声取得部２１）は、会話の音声データを取得する（ステップＳ１０１）。音声データは、マイク１８から入力されたものでもよいし、通信制御部１７及びネットワーク３を介して音声認識装置１と通信接続された機器から入力されたものでもよい。制御部１１（音声認識部２２）は、取得した音声データについて音声認識を行う（ステップＳ１０２）。ステップＳ１０２では、制御部１１（音声認識部２２）は、音声データに含まれる会話の音声を認識し、テキストに変換する処理を行う。制御部１１（音声認識部２２）は、音声認識処理の結果、１または複数の変換候補を得る。複数の変換候補がある場合に、ステップＳ１０３～ステップＳ１０４の処理により会話の状況を判定する。
【００３７】
制御部１１（状況判定部２３の会話内容判定部２３１）は、会話の状況として、会話の内容（ジャンル等）を判定する（ステップＳ１０３）。制御部１１（会話内容判定部２３１）は、ステップＳ１０２の音声認識の結果（変換候補）に含まれる語の意味を解析することにより、会話の内容を求める。ここで求める会話の内容とは、会話のジャンルまたは目的等である。会話内容判定部２３１は、例えば、記憶部１２に予め記憶されているジャンルデータ５を参照することにより会話の内容（ジャンル）を判定する。
【００３８】
ジャンルデータ５は、図４に示すように、単語の読み（音声認識結果）について１または複数の変換候補となる語と、その語のジャンルとを関連付けたデータである。例えば、音声認識結果「たいけん」の変換候補は、「大剣」と「体験」等があり、変換候補「大剣」のジャンルは「ＲＰＧ（ゲーム）」、変換候補「体験」のジャンルは「観光」である。このように、ひとつの音声認識結果について１または複数の変換候補と各変換候補に応じたジャンルが格納されている。各変換候補について複数のジャンルが関連づけられていてもよい。会話内容判定部２３１は、ジャンルデータ５を参照することにより、音声認識結果について、１または複数の会話内容の候補（ジャンル候補）を求める。例えば、音声認識結果「おすすめのたいけんはありますか」であれば、「たいけん」という語が含まれるため、ジャンル候補として、「ＲＰＧ（ゲーム）」と「観光」が求められる。
【００３９】
次に、制御部１１（状況判定部２３の周囲状況判定部２３２）は、周囲の状況を判定す

る（ステップＳ１０４）。制御部１１（周囲状況判定部２３２）は、会話の音声データが入力されたときの周囲の音の特徴を解析し、音の特徴に基づいて場所等の周囲状況を求める。例えば、「レストラン」で収録された音には、食器等の音や接客の音等の特徴が含まれている。周囲状況判定部２３２は、音の特徴と状況とを関連付けたデータを音特徴データとして記憶部１２に予め記憶しており（不図示）、この音特徴データに基づいて会話の周囲状況を判定するようにしてもよい。制御部１１（周囲状況判定部２３２）は、１または複数の周囲状況の候補を求める。なお、周囲状況は、場所に限定されず、「接客」、「授業」、「雑談」等のように、会話の目的等としてもよい。制御部１１は、音の特徴解析による周囲状況の判定結果として、例えば、「観光案内所」、「接客」等を得る。

【００４０】

なお、制御部１１（周囲状況判定部２３２）は、周囲の音の特徴のみならず、画像や位置情報に基づいて周囲状況を判定してもよい。具体的には、画像取得部２７（カメラ等）により会話中の様子を撮影した映像（画像）等を取得し、解析する画像解析部２８を備え、画像解析部２８によって会話の音声データが入力されたときの会話の場所や目的等、周囲状況を求めてもよい。また、ＧＰＳ等の位置情報を取得し、位置情報及び予め記憶されている地図データに基づいて会話の場所（店舗や施設）等を求めることにより周囲状況を求めてもよい。

【００４１】

制御部１１（音声認識結果抽出部２４）は、ステップＳ１０２で得た音声認識結果の複数の変換候補のうち、ステップＳ１０３及びステップＳ１０４において求めた会話の状況（会話内容（ジャンル）及び周囲状況）に適した変換候補を抽出する（ステップＳ１０５）。例えば、音声認識結果が「おすすめのたいけんはありますか」の場合、この文に含まれる「たいけん」の語には、「大剣」と「体験」の変換候補がある。ステップＳ１０３で会話の内容が「ＲＰＧ（ゲーム）」、「観光」、…と判定され、ステップＳ１０４で周囲の状況が「観光」、「接客」、…と判定された場合、制御部１１は会話内容のジャンルと周囲状況とをマッチングし、尤度の高いジャンルの語を抽出する。上記例では、会話の状況として「観光」が尤もらしいと判定されるため、変換候補「体験」を選択し、入力音声の音声認識結果として「おすすめの体験はありますか」を得る。

【００４２】

制御部１１は、ステップＳ１０３及びステップＳ１０４で判定した状況（上記例では、「観光」）を状況履歴データとして時間情報（音声データの入力時刻等）と関連付けて記憶部１２に保存する（ステップＳ１０６）。

【００４３】

制御部１１（出力部２５）は、ステップＳ１０５で抽出した音声認識結果を出力する。出力は、表示部１６への表示や、制御部１１への通知、ネットワーク３を介した通信接続先への送信等、当該音声認識装置１に接続された各種機器への制御信号の送信等も含むものとする。

【００４４】

音声認識結果を出力すると、入力された音声データに対する音声認識処理を終了する。

【００４５】

なお、上述の音声認識処理において、ステップＳ１０６で保存した履歴に基づき、会話内容を判定するようにしてもよい。すなわち、ステップＳ１０３において、前の文までの会話の状況の履歴が保存されている場合は、制御部１１は、前の文までの会話の状況から会話の内容（ジャンル）を絞り込んでもよい。

【００４６】

例えば、図５に示すように、１７時５０分に「いらっしゃいませ」、「何かお探しですか」という会話の音声が入力され、１７時５０分における音声（「いらっしゃいませ」、「何かお探しですか」）の状況として、「店」、「ファミレス」、「ドラッグストア」等の会話ジャンルが求められるものとする。会話ジャンルの各候補にはそれぞれ尤度が付与されているものとする。例えば、語「いらっしゃいませ」に対する会話ジャンル「店」の

尤度は「１．０」であり、「ファミレス」の尤度は「０．９」であり、「ドラッグストア」の尤度は「０．９」、…等である。尤度は、例えば語とジャンルとを対応付けたジャンルデータ５に予め付与されているものとする。また、１７時５０分における音解析による状況判定結果として、「観光案内所」と「受付」が尤度とともに求められる。例えば、「観光案内所」の尤度は「０．８」、「受付」の尤度は「０．５」のように求められるものとする。この音解析による状況判定結果の尤度は、入力された音と予め記憶されている音特徴データとの一致度等から付与するものとすればよい。

【００４７】

次に、１７時５６分に「○○レジャー施設はどこですか」、「ここから５分の場所にあります」という会話の音声が入力される。制御部１１は、１７時５６分における会話のジャンルを「遊園地」、「観光」等と判定する。それぞれの尤度は「遊園地」が「０．９」、「観光」が「０．８」とする。また１７時５６分における音解析による状況判定結果として、「観光案内所」、「店頭」、「接客」を得る。「観光案内所」の尤度は「０．８」、「店頭」の尤度は「０．８」、「接客」の尤度は「０．７」であるとする。

【００４８】

その後、１８時００分に、処理対象である「おすすめのたいけんはありますか」という音声が入力されるものとする。制御部１１は、１８時００分における会話のジャンルを「ＲＰＧ（ゲーム）」、「観光」等と判定する。それぞれの尤度として「ＲＰＧ（ゲーム）」は「０．５」、「観光」は「０．２」を得るものとする。また１８時００分における音解析による状況判定結果として、「観光案内所」、「接客」を得る。「観光案内所」の尤度は「０．８」、「接客」の尤度は「０．７」であるとする。

【００４９】

「たいけん」の各変換候補（「大剣」、「体験」）の会話内容に基づく尤度は、ＲＰＧ（ゲーム）は「０．５」、観光は「０．２」であるが、音による状況判定では、ＲＰＧ（ゲーム）という候補はなし（尤度「０」）、観光（観光案内所）は尤度「０．８」である。これらを併せると、

「ＲＰＧ（ゲーム）」の尤度＝会話内容「０．５」＋音判定「０」＝０．５

「観光」の尤度＝会話内容「０．２」＋音判定「０．８」＝１．０

となる。

【００５０】

したがって、状況としては「ＲＰＧ（ゲーム）」よりも「観光」の尤度が高く適切である。従って、制御部１１は、「たいけん」の変換結果（音声認識結果）としては「体験」が適していると判断する。このように、音声認識や音判定の履歴を遡って状況判定に利用すれば、コンテキスト（文脈、会話の流れ）を考慮した音声認識結果を得ることが可能となる。例えば最近の「ＲＰＧ（ゲーム）」「観光」のコンテキストの履歴を見ると「観光」が多く出現するので、この値（＝重み）を音声認識や音判定の尤度に掛け合わせて足したものを比較し判定することもできる。

「ＲＰＧ（ゲーム）」の尤度＝（会話内容「０．５」＋音判定「０」）×ゲームのコンテキストの重み「０」＝０

「観光」の尤度＝（会話内容「０．２」＋音判定「０．８」）×観光のコンテキストの重み「０．７」＝０．７

【００５１】

以上説明したように、本実施の形態の音声認識装置１は、会話の音声データを認識し、音声認識結果として複数の変換候補を求め、音声認識の結果のみならず周囲音等を考慮して会話の状況を判定することにより、複数の変換候補の中から会話の状況に適した候補を抽出する。これにより、会話の状況に適した音声認識結果を得ることが可能となり、音声認識精度を向上できる。

【００５２】

なお、本発明の音声認識装置１をスマートフォン２やタブレット等の通信機器や、インターネット等のネットワーク３に接続されたＰＣ（Personal Computer）７に適用する場

合において、図６に示す音声認識システム１００のように、スマートフォン２等からアクセス可能なサーバに本発明に係る音声認識装置１の各機能部（音声取得部２１、音声認識部２２、状況判定部２３、音声認識結果抽出部２４、出力部２５、音解析部２６等）を備える構成としてもよい。すなわち、スマートフォン２、ＰＣ７等はマイク１８から入力された会話の音声データをネットワーク３を介して音声認識装置１（サーバ）に送信すると、音声認識装置１は、図３に示す音声認識処理を実行し、音声認識結果を音声入力元のスマートフォン２等に返すものとしてもよい。

【００５３】

また、本発明に係る音声認識装置１は、対話型ロボット６に適用してもよい。この場合、本発明に係る音声認識装置１の各機能（音声取得部２１（マイク）、音声認識部２２、状況判定部２３、音声認識結果抽出部２４、出力部２５、音解析部２６、画像取得部２７（カメラ）、画像解析部２８、位置情報取得部２９）をロボット６が備える構成とする。或いは、対話型ロボット６の音声取得部２１（マイク）、画像取得部２７（カメラ）から入力された音声や画像を、サーバ（音声認識装置１）に送信し、サーバ（音声認識装置１）は、図３に示す音声認識処理を実行し、音声認識結果を音声入力元の対話型ロボット６に返すものとしてもよい。

【００５４】

その他、本発明に係る音声認識装置１は、カーナビゲーションシステム等の各種情報機器や家電等に適用することも可能である。

【００５５】

以上、添付図面を参照して、本発明に係る音声認識装置等の好適な実施形態について説明したが、本発明は係る例に限定されない。当業者であれば、本願で開示した技術的思想の範疇内において、各種の変更例または修正例に想到し得ることは明らかであり、それらについても当然に本発明の技術的範囲に属するものと了解される。

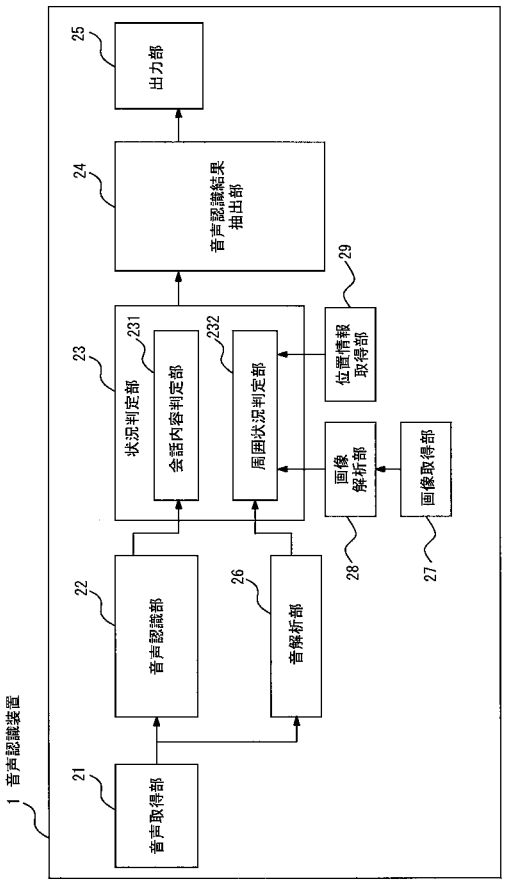
【符号の説明】

【００５６】

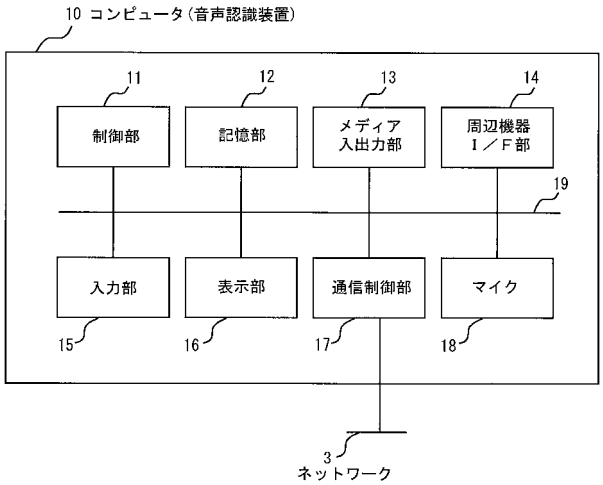
１……音声認識装置
１０……コンピュータ
１１……制御部
１２……記憶部
１３……メディア入出力部
１４……周辺機器Ｉ／Ｆ部
１５……入力部
１６……表示部
１７……通信制御部
１８……マイク
１９……バス
２１……音声取得部
２２……音声認識部
２３……状況判定部
２３１……会話内容判定部
２３２……周囲状況判定部
２４……音声認識結果抽出部
２５……出力部
２６……音解析部
２７……画像取得部
２８……画像解析部
２９……位置情報取得部
３……ネットワーク
５……ジャンルデータ

６……対話型ロボット
７……ＰＣ
１００……音声認識システム

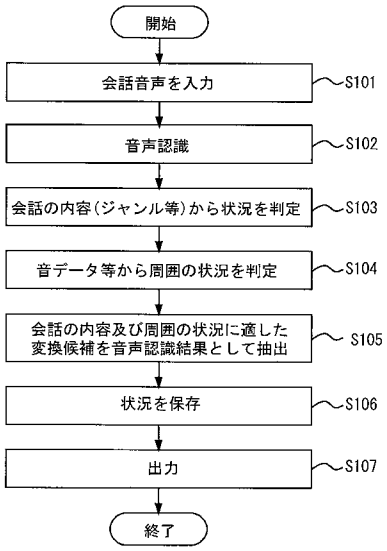
【図１】



【図２】



【図３】

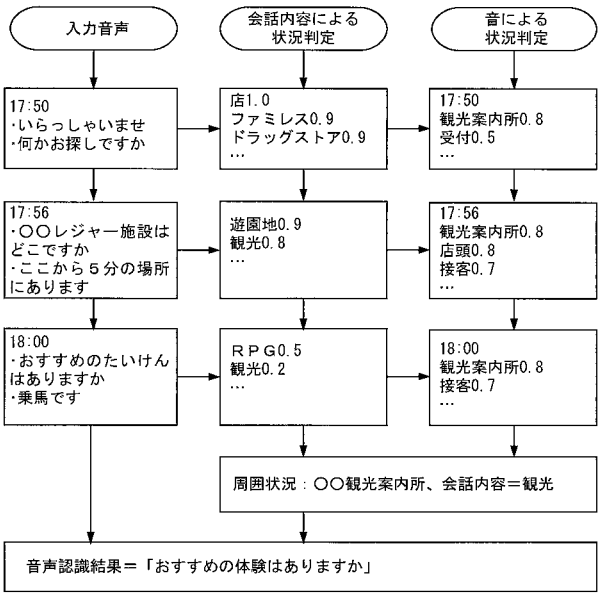


【図４】

5 ジャンルデータ

音声認識結果	変換候補	ジャンル
たいけん	大剣	ＲＰＧ（ゲーム）、…
	体験	観光、…
	…	
…		

【図５】



【図６】