(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7432556号
(P7432556)

(45)発行日　令和6年2月16日(2024.2.16)

(24)登録日　令和6年2月7日(2024.2.7)

(51)国際特許分類		FI		
G06T 13/40 (2011.01)		G06T	13/40	
G10L 15/00 (2013.01)		G10L	15/00	200Z
G10L 13/00 (2006.01)		G10L	13/00	100M
G06F 3/16 (2006.01)		G06F	3/16	650
G06F 3/01 (2006.01)		G06F	3/16	690

請求項の数　17　外国語出願　（全22頁）　最終頁に続く

(21)出願番号　特願2021-87333(P2021-87333)
(22)出願日　令和3年5月25日(2021.5.25)
(65)公開番号　特開2021-168139(P2021-168139A)
(43)公開日　令和3年10月21日(2021.10.21)
審査請求日　令和3年5月25日(2021.5.25)
(31)優先権主張番号　202011598915.9
(32)優先日　令和2年12月30日(2020.12.30)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　514322098
ベイジン バイドゥ ネットコム サイエンス テクノロジー カンパニー リミテッド
Ｂｅｉｊｉｎｇ Ｂａｉｄｕ Ｎｅｔｃｏｍ Ｓｃｉｅｎｃｅ Ｔｅｃｈｎｏｌｏｇｙ Ｃｏ．，Ｌｔｄ．
中華人民共和国 ペキン １０００８５，ハイディアン ディストリクト，シャンディ テンス ストリート，１０番，バイドゥ キャンパス ２階
２／Ｆ Ｂａｉｄｕ Ｃａｍｐｕｓ，Ｎｏ．１０，Ｓｈａｎｇｄｉ １０ｔｈ Ｓｔｒｅｅｔ，Ｈａｉｄｉａｎ Ｄｉｓｔｒｉｃｔ，Ｂｅｉｊｉｎｇ １０００
最終頁に続く

(54)【発明の名称】　マンマシンインタラクションのための方法、装置、機器および媒体

(57)【特許請求の範囲】

【請求項１】

受信した音声信号に基づいて、前記音声信号に対する回答の回答テキストを生成することと、

音声信号ユニットとテキストユニットとのマッピング関係に基づいて、１セットのテキストユニットを含む前記回答テキストに対応する回答音声信号を生成し、生成した前記回答音声信号は前記１セットのテキストユニットに対応する１セットの音声信号ユニットを含むことと、

前記回答テキストに基づいて、仮想オブジェクトによって表現される表情および／または動作の標識を確定することと、

前記回答音声信号、前記表情および／または動作の標識に基づいて、前記仮想オブジェクトを含む出力ビデオを生成し、前記出力ビデオは前記回答音声信号に基づいて確定された、前記仮想オブジェクトによって表現される唇形シーケンスを含むこととを含み、

前記回答音声信号を生成することは、

前記回答テキストを１セットのテキストユニットに分割することと、

音声信号ユニットとテキストユニットとのマッピング関係に基づいて、前記１セットのテキストユニットにおけるテキストユニットに対応する音声信号ユニットを取得することであって、前記１セットのテキストユニットから前記テキストユニットを選択することと、音声信号ユニットとテキストユニットとのマッピング関係に基づいて、音声ライブラリから前記テキストユニットに対応する前記音声信号ユニットを検索することとを含む、音

声信号ユニットを取得することと、
前記取得された音声信号ユニットに基づいて、前記回答音声信号を生成することとを含み、
前記音声ライブラリには音声信号ユニットとテキストユニットとの前記マッピング関係が記憶され、前記音声ライブラリにおける音声信号ユニットは、取得した、前記仮想オブジェクトに関する音声記録データを分割することで得られるものであり、前記音声ライブラリにおけるテキストユニットは、分割で得られた音声信号ユニットに基づいて確定されるものであり、
前記出力ビデオを生成することは、
前記回答音声信号を１セットの音声信号ユニットに分割することと、
前記１セットの音声信号ユニットに対応する前記仮想オブジェクトの唇形シーケンスを取得することと、
対応する前記表情および／または動作の標識に基づいて、前記仮想オブジェクトについての当該表情および／または動作のビデオセグメントを取得することと、
前記唇形シーケンスを前記ビデオセグメントに結合して前記出力ビデオを生成することとを含み、
前記仮想オブジェクトについての当該表情および／または動作の前記ビデオセグメントを取得することは、
表情および／または動作の標識とビデオセグメントとの間の事前に記憶されたマッピング関係を利用して、対応する前記表情および／または動作の標識に基づいて当該表情および／または動作の前記ビデオセグメントを取得することを含む、
マンマシンインタラクションのための方法。
【請求項２】
前記回答テキストを生成することは、
前記受信した音声信号を識別して入力テキストを生成することと、
前記入力テキストに基づいて、前記回答テキストを取得することとを含む、請求項１に記載の方法。
【請求項３】
前記入力テキストに基づいて、前記回答テキストを取得することは、
入力テキストと前記仮想オブジェクトの人格属性を用いて回答テキストを生成する機械学習モデルである対話モデルに、前記入力テキストと前記仮想オブジェクトの人格属性を入力して前記回答テキストを取得することを含む、請求項２に記載の方法。
【請求項４】
前記対話モデルは、前記仮想オブジェクトの人格属性と、入力テキストサンプルと回答テキストサンプルを含む対話サンプルトとを利用してトレーニングすることで得られるものである、請求項３に記載の方法。
【請求項５】
前記表情および／または動作の標識を確定することは、
テキストを用いて表情および／または動作の標識を確定する機械学習モデルである表情および動作識別モデルに、前記回答テキストを入力して、前記表情および／または動作の標識を取得することを含む、請求項１に記載の方法。
【請求項６】
前記唇形シーケンスを前記ビデオセグメントに結合して前記出力ビデオを生成することは、
前記ビデオセグメントにおける時間軸での所定の時間位置におけるビデオフレームを確定することと、
前記唇形シーケンスから前記所定の時間位置に対応する唇形を取得することと、
前記唇形を前記ビデオフレームに結合して前記出力ビデオを生成することとを含む、請求項１に記載の方法。
【請求項７】

前記回答音声信号と前記出力ビデオとを関連付けて出力することとをさらに含む、請求項１に記載の方法。
【請求項８】
受信した音声信号に基づいて、前記音声信号に対する回答の回答テキストを生成するように構成される回答テキスト生成モジュールと、
音声信号ユニットとテキストユニットとのマッピング関係に基づいて、１セットのテキストユニットを含む前記回答テキストに対応する回答音声信号を生成し、生成された前記回答音声信号は前記１セットのテキストユニットに対応する１セットの音声ユニットを含むように構成される第１回答音声信号生成モジュールと、
前記回答テキストに基づいて、仮想オブジェクトによって表現される表情および／または動作の標識を確定するように構成される標識確定モジュールと、
前記回答音声信号、前記表情および／または動作の標識に基づいて、前記仮想オブジェクトを含む出力ビデオを生成し、前記出力ビデオは、前記回答音声信号に基づいて確定された、前記仮想オブジェクトによって表現される唇形シーケンスを含むように構成される第１出力ビデオ生成モジュールとを含み、
前記第１回答音声信号生成モジュールは、
前記回答テキストを１セットのテキストユニットに分割するように構成されるテキストユニット分割モジュールと、
音声信号ユニットとテキストユニットとのマッピング関係に基づいて、前記１セットのテキストユニットにおけるテキストユニットに対応する音声信号ユニットを取得する音声信号ユニット取得モジュールであって、１セットのテキストユニットから前記テキストユニットを選択するように構成されるテキストユニット選択モジュールと、音声信号ユニットとテキストユニットとのマッピング関係に基づいて、音声ライブラリから前記テキストユニットに対応する前記音声信号ユニットを検索するように構成される検索モジュールとを含む、音声信号ユニット取得モジュールと、
前記取得された音声信号ユニットに基づいて、前記回答音声信号を生成するように構成される第２回答音声信号生成モジュールとを含み、
前記音声ライブラリには音声信号ユニットとテキストユニットとの前記マッピング関係が記憶され、前記音声ライブラリにおける音声信号ユニットは、取得された、前記仮想オブジェクトに関する音声記録データを分割することで得られるものであり、前記音声ライブラリにおけるテキストユニットは、分割で得られた音声信号ユニットに基づいて確定されるものであり、
前記第１出力ビデオ生成モジュールは、
前記回答音声信号を１セットの音声信号ユニットに分割するように構成される音声信号分割モジュールと、
前記１セットの音声信号ユニットに対応する前記仮想オブジェクトの唇形シーケンスを取得するように構成される唇形シーケンス取得モジュールと、
対応する前記表情および／または動作の標識に基づいて、前記仮想オブジェクトについての当該表情および／または動作のビデオセグメントを取得するように構成されるビデオセグメント取得モジュールと、
前記唇形シーケンスを前記ビデオセグメントに結合して前記出力ビデオを生成するように構成される第２出力ビデオ生成モジュールとを含み、
前記仮想オブジェクトについての当該表情および／または動作の前記ビデオセグメントを取得することは、
表情および／または動作の標識とビデオセグメントとの間の事前に記憶されたマッピング関係を利用して、対応する前記表情および／または動作の標識に基づいて当該表情および／または動作の前記ビデオセグメントを取得することを含む、
マンマシンインタラクションのための装置。
【請求項９】
前記回答テキスト生成モジュールは、

前記受信した音声信号を識別して入力テキストを生成するように構成される入力テキスト生成モジュールと、
前記入力テキストに基づいて、前記回答テキストを取得するように構成される回答テキスト取得モジュールとを含む、請求項８に記載の装置。
【請求項１０】
前記回答テキスト取得モジュールは、
入力テキストと前記仮想オブジェクトの人格属性を用いて回答テキストを生成する機械学習モデルである対話モデルに前記入力テキストと前記仮想オブジェクトの人格属性を入力して前記回答テキストを取得するように構成される、モデルに基づく回答テキスト取得モジュールを含む、請求項９に記載の装置。
【請求項１１】
前記対話モデルは、前記仮想オブジェクトの人格属性および入力テキストサンプルと回答テキストサンプルとを含む対話サンプルトを利用してトレーニングすることで得られるものである、請求項１０に記載の装置。
【請求項１２】
前記標識確定モジュールは、
テキストを用いて表情および／または動作の標識を確定する機械学習モデルである表情および動作識別モデルに前記回答テキストを入力して、前記表情および／または動作の標識を取得するように構成される表情動作標識取得モジュールを含む、請求項８に記載の装置。
【請求項１３】
前記第２出力ビデオ生成モジュールは、
前記ビデオセグメントにおける時間軸での所定の時間位置におけるビデオフレームを確定するように構成されるビデオフレーム確定モジュールと、
前記唇形シーケンスから前記所定の時間位置に対応する唇形を取得するように構成される唇形取得モジュールと、
前記唇形を前記ビデオフレームに結合して前記出力ビデオを生成するように構成される結合モジュールとを含む、請求項８に記載の装置。
【請求項１４】
前記回答音声信号と前記出力ビデオとを関連付けて出力するように構成される出力モジュールをさらに含む、請求項８に記載の装置。
【請求項１５】
少なくとも１つのプロセッサ、および
前記少なくとも１つのプロセッサに通信接続されたメモリを含み、
前記メモリには、前記少なくとも１つのプロセッサによって実行可能なコマンドが記憶され、前記コマンドは前記少なくとも１つのプロセッサによって実行されることにより、前記少なくとも１つのプロセッサが請求項１～７のいずれか一項に記載の方法を実行する、電子機器。
【請求項１６】
コンピュータに請求項１～７のいずれか一項に記載の方法を実行させるためのコンピュータコマンドが記憶された非一時的コンピュータ可読記憶媒体。
【請求項１７】
プロセッサによって実行されると、請求項１～７のいずれか一項に記載の方法を実現するコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、人工知能の分野に関し、特にディープラーニング、音声技術およびコンピュータビジョン分野におけるマンマシンインタラクションのための方法、装置、機器および媒体に関する。

【背景技術】
【０００２】
コンピュータ技術の急速な発展に伴って、人間と機械のインタラクションがますます多くなっている。ユーザの体験を向上させるために、マンマシンインタラクション技術が急速に発展している。ユーザが音声コマンドを出した後、計算機器は音声識別技術によってユーザの音声を識別する。識別を完了した後に、ユーザの音声コマンドに応じる操作を実行する。このような音声インタラクション方式はマンマシンインタラクションの体験を改善する。しかしながら、マンマシンインタラクションのプロセスにおいては、多くの解決する必要のある問題がまだ存在している。
【発明の概要】
【０００３】
本開示は、マンマシンインタラクションのための方法、装置、機器および媒体を提供する。
本開示の第１態様によれば、マンマシンインタラクションのための方法が提供される。この方法は、受信した音声信号に基づいて、音声信号に対する回答の回答テキストを生成することを含む。この方法は、音声信号ユニットとテキストユニットとのマッピング関係に基づいて、１セットのテキストユニットを含む回答テキストに対応する回答音声信号を生成することをさらに含む。この方法は、回答テキストに基づいて、仮想オブジェクトによって表現される表情および／または動作の標識を確定することをさらに含む。この方法は、回答音声信号、表情および／または動作の標識に基づいて、仮想オブジェクトを含む出力ビデオを生成することを含み、出力ビデオは、回答音声信号に基づいて確定された、仮想オブジェクトによって表現される唇形シーケンスを含む。
【０００４】
本開示の第２態様によれば、マンマシンインタラクションのための装置が提供される。この装置は、受信した音声信号に基づいて、音声信号に対する回答の回答テキストを生成するように構成される回答テキスト生成モジュールと、音声信号ユニットとテキストユニットとのマッピング関係に基づいて、１セットのテキストユニットを含む回答テキストに対応する回答音声信号を生成し、生成された回答音声信号は１セットのテキストユニットに対応する１セットの音声ユニットを含むように構成される第１回答音声信号生成モジュールと、回答テキストに基づいて、仮想オブジェクトによって表現される表情および／または動作の標識を確定する標識確定モジュールと、回答音声信号、表情および／または動作の標識に基づいて、仮想オブジェクトを含む出力ビデオを生成し、出力ビデオは回答音声信号に基づいて確定された、仮想オブジェクトによって表現される唇形シーケンスを含むように構成される第１出力ビデオ生成モジュールとを含む。
【０００５】
本開示の第３態様によれば、電子機器が提供される。この電子機器は、少なくとも１つのプロセッサ、および少なくとも１つのプロセッサに通信接続されるメモリを含み、ここで、メモリには、少なくとも１つのプロセッサによって実行可能なコマンドが記憶され、コマンドは少なくとも１つのプロセッサによって実行されることにより、少なくとも１つのプロセッサが本開示の第１態様の方法を実行することができる。
【０００６】
本開示の第４態様によれば、コンピュータに本開示の第１態様の方法を実行させるためのコンピュータコマンドが記憶された非一時的コンピュータ可読記憶媒体が提供される。
本開示の第５態様によれば、コンピュータプログラムを含むコンピュータプログラム製品が提供される。前記コンピュータプログラムはプロセッサによって実行されると、本開示の第１態様の方法を実現する。
【０００７】
理解できるように、この部分に説明される内容は、本開示の実施形態の肝心または重要な特徴を示すことを目的とせず、本開示の保護範囲を限定するためのものではないことである。本開示の他の特徴は、以下の明細書によって理解されやすくなる。

【図面の簡単な説明】
【０００８】
図面は、本発明をより良く理解するためのものであり、本開示に対する限定を構成していない。
【図１】本開示の複数の実施形態を実現することができる環境１００を示す概略図である。
【図２】本開示のいくつかの実施形態によるマンマシンインタラクションのためのプロセス２００を示すフローチャートである。
【図３】本開示のいくつかの実施形態によるマンマシンインタラクションのための方法３００を示すフローチャートである。
【図４】本開示のいくつかの実施形態による対話モデルをトレーニングするための方法４００を示すフローチャートである。
【図５Ａ】本開示のいくつかの実施形態による対話モデルネットワーク構造を示す例である。
【図５Ｂ】本開示のいくつかの実施形態によるマスクテーブルを示す例である。
【図６】本開示のいくつかの実施形態による回答音声信号を生成するための方法６００を示すフローチャートである。
【図７】本開示のいくつかの実施形態による表情および／または動作の説明例７００を示す概略図である。
【図８】本開示のいくつかの実施形態による表情および動作識別モデルを取得して使用するための方法８００を示すフローチャートである。
【図９】本開示のいくつかの実施形態による出力ビデオを生成するための方法９００を示すフローチャートである。
【図１０】本開示のいくつかの実施形態による出力ビデオを生成するための方法１０００を示すフローチャートである。
【図１１】本開示の実施形態によるマンマシンインタラクションのための装置１１００を示す概略的ブロック図である。
【図１２】本開示の複数の実施形態を実施することができる機器１２００を示すブロック図である。
【発明を実施するための形態】
【０００９】
以下、図面に合わせて本開示の例示的な実施形態を説明し、それに含まれる本開示の実施形態における様々な詳細が理解を助けるためので、それらは単なる例示的なものと考えられるべきである。したがって、当業者であれば、本開示の範囲および精神から逸脱することなく、本明細書で説明される実施形態に対して様々な変更および修正を行うことができることをを認識すべきである。同様に、明瞭と簡潔のために、以下の説明では公知の機能および構造についての説明を省略する。
【００１０】
本開示の実施形態の説明において、用語「含む」およびその類似用語はオープンな包含であり、すなわち「含むが、これらに限定されない」ことを理解されたい。用語「に基づいて」は、「少なくとも部分的に基づいて」ことを理解されたい。用語「一実施形態」または「該実施形態」は、「少なくとも１つの実施形態」ことを理解されたい。用語「第１」、「第２」などは異なるまたは同じオブジェクトを指すことができる。以下には他の明示的および暗示的な定義をさらに含む可能性もある。
【００１１】
機械を人間のように人間と対話させることは人工知能の重要な目標である。現在、機械と人間のインタラクションの形式がインターフェースによるインタラクションから言語によるインタラクションへと進化している。しかしながら、従来の技術案では、ただ内容が限られたインタラクションだけであり、または音声の出力しかい実行できない。例えばインタラクションの内容は主に、「天気を調べろ」、「音楽を再生しろ」、「アラームを設定しろ」など、限られた分野でのコマンド型のインタラクションに限られる。また、イン

タラクションのモードも単一で、音声またはテキストによるインタラクションのみがある。また、マンマシンインタラクションには人格属性を欠けて、机械は対話する人よりも、ツールのようなものである。

【００１２】

上述した問題を解決するために、本開示の実施形態によれば、改善案が提供される。この案において、計算機器は、受信した音声信号に基づいて、音声信号に対する回答の回答テキストを生成する。次に、計算機器は回答テキストに対応する回答音声信号を生成する。計算機器は、回答テキストに基づいて、仮想オブジェクトによって表現される表情および／または動作の標識を確定する。続いて、計算機器は、回答音声信号、表情および／または動作の標識に基づいて、仮想オブジェクトを含む出力ビデオを生成する。この方法により、インタラクションの内容の範囲を著しく増加させ、マンマシンインタラクションの品質とレベルを向上させ、ユーザ体験を向上させることができる。

【００１３】

図１は、本開示の複数の実施形態を実現することができる環境１００の概略図を示す。この例示的な環境は、マンマシンインタラクションを実現するために利用できる。この例示的な環境１００は、計算機器１０８および端末機器１０４を含む。

【００１４】

端末１０４における仮想人物などの仮想オブジェクト１１０は、ユーザ１０２と対話するために利用できる。インタラクションプロセスにおいて、ユーザ１０２は、端末１０４に問い合わせまたはチャット語句を送信することができる。端末１０４は、ユーザ１０２の音声信号を取得し、ユーザから入力された音声信号に対する回答を仮想オブジェクト１１０によって表現するために使用され、これによって人間と機械の対話を実現することができる。

【００１５】

端末１０４は任意のタイプの計算機器として実現されることができ、携帯電話（例えばスマートフォン）、ラップトップコンピュータ、ポータブルデジタルアシスタント（ＰＤＡ）、電子ブックリーダ、ポータブルゲームコンソール、ポータブルメディアプレイヤ、ゲームコンソール、セットトップボックス（ＳＴＢ）、スマートテレビ（ＴＶ）、パーソナルコンピュータ、車載コンピュータ（例えば、ナビゲーションユニット）、ロボットなどを含むがこれらに限定されない。

【００１６】

端末１０４は、取得された音声信号をネットワーク１０６を介して計算機器１０８に送信する。計算機器１０８は、端末１０４から取得された音声信号に基づいて、対応する出力ビデオと出力音声信号を生成して、端末１０４上における仮想オブジェクト１１０によって表現することができる。

【００１７】

図１は、計算機器１０８において、入力された音声信号に基づいて出力ビデオおよび出力音声信号を取得するプロセスを示しており、これは一例に過ぎず、本開示への具体的な限定ではない。このプロセスは、端末１０４上で実現されてもよく、または一部が計算機器１０８上で、他の一部が端末１０４上で実現されてもよい。いくつかの実施形態では、計算機器１０８と端末１０４は一体に統合されてもよい。図１は、計算機器１０８がネットワーク１０６を介して端末１０４に接続されていることを示す。これは一例に過ぎず、本開示への具体的な限定ではない。計算機器１０８は、他の方法で端末１０４と接続することもでき、例えば、ネットワークケーブルで直接的に接続される。上記の例は、本開示を説明するためのものに過ぎず、本開示への具体的な限定ではない。

【００１８】

計算機器１０８は任意のタイプの計算機器として実現されることができ、パーソナルコンピュータ、サーバコンピュータ、ハンドヘルドまたはラップトップラップトップ機器、携帯機器（例えば携帯電話、パーソナルデジタルアシスタント（ＰＤＡ）、メディアプレイヤなど）、マルチプロセッサシステム、消費者向け電子製品、小型コンピュータ、大型

コンピュータ、上記システムまたは機器のいずれかを含む分散式計算環境などを含むがこれらに限定されない。サーバは、クラウドサーバであってもよく、クラウド計算サーバまたはクラウドホストとも呼ばれ、クラウド計算サービスシステム中のホスト製品として、従来の物理ホストとＶＰＳサービス（「Ｖｉｒｔｕａｌ　Ｐｒｉｖａｔｅ　Ｓｅｒｖｅｒ」、または「ＶＰＳ」と略称される）における、管理の難度が高く、業務拡張性が弱いという欠陥を解決する。サーバは、分散式システムのサーバであってもよいし、ブロックチェーンと組み合せられたサーバであってもよい。

【００１９】

計算機器１０８は、端末１０４から取得された音声信号を処理することで、回答のための出力音声信号および出力ビデオを生成する。

この方法により、インタラクションの内容の範囲を著しく増加させ、マンマシンインタラクションの品質とレベルを向上させ、ユーザ体験を向上させることができる。

【００２０】

上記の図１は、本開示の複数の実施形態を実現することができる環境１００の概略図を示す。以下、図２によってマンマシンインタラクションのための方法２００の概略図を説明する。この方法２００は、図１における計算機器１０８または任意の適当な計算機器によって実現することができる。

【００２１】

図２に示すように、計算機器１０８は、受信した音声信号２０２を取得する。次に、計算機器１０８は、受信した音声信号を音声識別（ＡＳＲ）して入力テキスト２０４を生成する。ここでは、計算機器１０８は、任意の適当な音声識別アルゴリズムを用いて入力テキスト２０４を取得することができる。

【００２２】

計算機器１０８は、回答用の回答テキスト２０６を取得するために、取得された入力テキスト２０４を対話モデルに入力する。この対話モデルはトレーニングされた機械学習モデルであり、そのトレーニングプロセスはオフラインで行うことができる。代替的または付加的には、この対話モデルはニューラルネットワークモデルであり、以下、図４および図５Ａと図５Ｂに関連してこの対話モデルのレーニングプロセスを紹介する。

【００２３】

その後、計算機器１０８は、音声合成技術（ＴＴＳ）により回答テキスト２０６を利用して回答音声信号２０８を生成するとともに、回答テキスト２０６に基づいて、現在の回答に使用されている表情および／または動作の標識２１０をさらに識別することができる。いくつかの実施形態では、この標識は表情および／または動作ラベルであってもよい。いくつかの実施形態では、この標識は表情および／または動作のタイプである。上記の例は、本開示を説明するためのものに過ぎず、本開示への具体的な限定ではない。

【００２４】

計算機器１０８は取得された表情および／または動作の標識に基づいて、出力ビデオ２１２を生成する。次に、回答音声信号２０８と出力ビデオ２１２を、端末上で同期して再生されるように端末に送信する。

【００２５】

上記の図２は、本開示の複数の実施形態によるマンマシンインタラクションのためのプロセス２００の概略図を示す。以下、図３に関連して、本開示のいくつかの実施形態によるマンマシンインタラクションのための方法３００のローチャートを説明する。図３の方法３００は、図１の計算機器１０８または任意の適当な計算機器によって実行することができる。

【００２６】

ブロック３０２において、受信した音声信号に基づいて、音声信号に対する回答の回答テキストを生成する。例えば、図２に示すように、計算機器１０８は、受信した音声信号２０２に基づいて、受信した音声信号２０２に対する回答テキスト２０６を生成する。

【００２７】

いくつかの実施形態では、計算機器１０８は、受信した音声信号を識別して入力テキスト２０４を生成する。入力テキストを取得するために任意の適当な音声識別技術を採用して音声信号を処理することができる。続いて、計算機器１０８は、入力テキスト２０４に基づいて、回答テキスト２０６を取得する。この方法によって、ユーザから受信された音声の回答テキストを迅速かつ効率的に取得することができる。

【００２８】

いくつかの実施形態では、計算機器１０８は、回答テキスト２０６を取得するために、入力テキストと仮想オブジェクトの人格属性を用いて回答テキストを生成する機械学習モデルである対話モデルに入力テキスト２０４と仮想オブジェクトの人格属性を入力する。代替的または付加的には、この対話モデルはニューラルネットワークモデルである。いくつかの実施形態では、この対話モデルは任意の適当な機械学習モデルであってもよい。上記の例は、本開示を説明するためのものに過ぎず、本開示への具体的な限定ではない。この方法によって、回答テキストを迅速かつ正確に確定することができる。

【００２９】

いくつかの実施形態では、対話モデルは、仮想オブジェクトの人格属性および入力テキストサンプルと回答テキストサンプルとを含む対話サンプルトを利用してレーニングすることで得られる。この対話モデルは計算機器１０８によってオフラインでトレーニングすることで得られてもよい。計算機器１０８は、まず仮想オブジェクトの人格属性を取得し、人格属性は仮想オブジェクトの、性別、年齢、星座などの、人と関連する特性を説明する。次に、計算機器１０８は、人格属性および入力テキストサンプルと回答テキストサンプルとを含む対話サンプルに基づいて、対話モデルをトレーニングする。トレーニングするときに、人格属性と入力テキストサンプルを入力とし、回答テキストサンプルを出力としてトレーニングする。いくつかの実施形態では、対話モデルは他の計算機器によってオフラインでトレーニングしてもよい。上記の例は、本開示を説明するためのものに過ぎず、本開示への具体的な限定ではない。この方法によって、対話モデルを迅速的に取得することができる。

【００３０】

以下、図４と図５Ａおよび図５Ｂに関連してこの対話モデルのレーニングを紹介する。図４は、本開示のいくつかの実施形態による対話モデルをトレーニングするための方法４００のフローチャートを示す。図５Ａおよび図５Ｂは本開示のいくつかの実施形態による対話モデルネットワーク構造および用いられるマスクテーブルの一例を示す。

【００３１】

図４に示すように、プレトレーニング段階４０４において、例えば１０億レベルの人間対話コーパスなどのソーシャルプラットフォーム上で自動的にマイニングされたコーパス４０２を用いて、モデルが基礎的なオープンドメイン対話能力を備えるように、対話モデル４０６をトレーニングする。次に、例えば５万レベルの特定の人格属性を有する対話コーパスなどの手動ラベル付け対話コーパス４１０を取得し、人格適合段階４０８において、指定の人格属性を用いて対話する能力を備えるように、対話モデル４０６をさらにトレーニングする。この指定の人格属性は、マンマシンインタラクションで使用しようとする仮想人物の、性別、年齢、趣味、星座などの人格属性である。

【００３２】

図５Ａは対話モデルのモデル構造を示し、それは入力５０４、モデル５０２およびさらなる回答５１２を含む。このモデルはディープラーニングモデルにおけるＴｒａｎｓｆｏｒｍｅｒモデルを用いており、モデルを使用するたびに、回答中の１つの単語を生成する。このプロセスは、具体的には、人格情報５０６、入力テキスト５０８、および回答５１０に既に生成された部分（例えば、単語１＆２）をモデルに入力して、さらなる回答５１２の次の単語（３）を生成し、このように再帰して、完全な回答文を生成する。モデルトレーニング時に、効率を向上させるために図５Ｂにおけるマスクテーブル５１４を用いて、回答の生成にバッチ（Ｂａｔｃｈ）処理の操作を行う。

【００３３】

ここで、図３に戻り、ブロック３０４において、音声信号ユニットとテキストユニットとのマッピング関係に基づいて、１セットのテキストユニットを含む回答テキストに対応する回答音声信号を生成し、生成された回答音声信号は１セットのテキストユニットに対応する１セットの音声信号ユニットを含む。例えば、計算機器１０８は、予め記憶された音声信号ユニットとテキストユニットとのマッピング関係を利用して、１セットのテキストユニットを含む回答テキスト２０６に対応する回答音声信号２０８を生成し、生成した回答音声信号は該セットのテキストユニットに対応する１セットの音声信号ユニットを含む。

【００３４】

いくつかの実施形態では、計算機器１０８は、回答テキスト２０６を１セットのテキストユニットに分割する。次に、計算機器１０８は、音声信号ユニットとテキストユニットとのマッピング関係に基づいて、１セットのテキストユニットにおけるテキストユニットに対応する音声信号ユニットを取得する。計算機器１０８は、音声ユニットに基づいて、回答音声信号を生成する。この方法によって、回答テキストに対応する回答音声信号を迅速かつ効率的に生成することができる。

【００３５】

いくつかの実施形態では、計算機器１０８は、１セットのテキストユニットからテキストユニットを選択する。次に、計算機器は、音声信号ユニットとテキストユニットとのマッピング関係に基づいて、音声ライブラリからテキストユニットに対応する音声信号ユニットを検索する。この方式によって、音声信号ユニットを迅速に取得することができ、このプロセスにかかる時間を短縮し、効率を向上させる。

【００３６】

いくつかの実施形態では、音声ライブラリに音声信号ユニットとテキストユニットとのマッピング関係が記憶され、音声ライブラリにおける音声信号ユニットは、取得された、仮想オブジェクトに関する音声記録データを分割することで取得されるものであり、音声ライブラリにおけるテキストユニットは、分割で得られた音声信号ユニットに基づいて確定されるものである。音声ライブラリは以下の方式によって生成される。まず、仮想オブジェクトに関連する音声記録データを取得する。例えば、仮想オブジェクトに対応する人間の声を録音する。次に、音声記録データを複数の音声信号ユニットに分割する。音声信号ユニットに分割された後、複数の音声信号ユニットに対応する複数のテキストユニットを確定し、ここで、第１音声信号ユニットは１つのテキストユニットに対応する。次に、複数の音声信号ユニットにおける音声信号ユニットと複数のテキストユニットにおける対応するテキストユニットとを関連付けて音声ライブラリに記憶し、それにより音声ライブラリが生成される。この方法により、テキストの音声信号ユニットを取得する効率を高め、取得時間を節約することができる。

【００３７】

以下、図６に関連して、回答音声信号を生成するプロセスを具体的に説明する。ここで、図６は、本開示のいくつかの実施形態による回答音声信号を生成するための方法６００のフローチャートを示す。

【００３８】

図６に示すように、機械が人間のチャットをよりリアルにシミュレートするために、仮想キャラクタと一致する人間の声を用いて回答音声信号を生成する。このプロセス６００はオフラインとオンラインの２つの部分に分割される。オフライン部分では、ブロック６０２において、仮想キャラクタと一致する人間の録音録画データを収集する。次に、ブロック６０４の後に、録音された音声信号を音声ユニットに分割し、対応するテキストユニットとアライメントすることで、単語ごとに対応する音声信号を記憶している音声ライブラリ６０６を取得する。このオフラインプロセスは、計算機器１０８または任意の他の適切な装置で行われることができる。

【００３９】

オンライン部分では、回答テキスト中の単語シーケンスに基づいて音声ライブラリ６０

６から対応する音声信号を抽出して出力音声信号を合成する。まず、ブロック６０８において、計算機器１０８は回答テキストを取得する。次に、計算機器１０８は回答テキスト６０８を１セットのテキストユニットに分割する。その後、ブロック６１０において、音声ライブラリ６０６からテキストユニットに対応する音声ユニットの抜き取りおよびスプライスを行う。次に、ブロック６１２において、回答音声信号を生成する。したがって、音声ライブラリを利用して回答音声信号をオンラインで取得することができる。

【００４０】

次に、図３に戻って引き続き説明し、ブロック３０６において、回答テキストに基づいて、仮想オブジェクトによって表現される表情および／または動作の標識を確定する。例えば、計算機器１０８は、回答テキスト２０６に基づいて、仮想オブジェクト１１０によって表現される表情および／または動作の標識２１０を確定する。

【００４１】

いくつかの実施形態では、計算機器１０８は、テキストを用いて表情および／または動作の標識を確定する機械学習モデルである表情および動作識別モデルに回答テキストを入力して、表情および／または動作の標識を取得する。この方法によって、テキストを迅速かつ正確に利用して、使用しようとする表情と動作を確定することができる。

【００４２】

以下、図７と図８に関連して表情および／または動作の標識および表情および動作の記述を説明する。図７は、本開示のいくつかの実施形態による表情および／または動作の例７００の概略図を示す。図８は、本開示のいくつかの実施形態による表情および動作識別モデルを取得し使用するための方法８００のフローチャートを示す。

【００４３】

対話において、仮想オブジェクト１１０の表情と動作は対話内容によって決定され、仮想人物は「私はとても嬉しいです」と答える場合、楽しい表情を用いることができ、「こんにちは」と答える場合、手を振る動作を用いることができ、このため、表情と動作識別は対話モデルにおける回答テキストに基づいて仮想人物の表情と動作ラベルを識別するものである。このプロセスには表情および動作ラベルシステムの設定と識別の２つの部分が含まれる。

【００４４】

図７において、対話過程に関する高頻度の表情および／または動作に１１個のラベルが設定される。いくつかのシーンでは表情と動作が共同で働くので、システムにおいては、あるラベルが表情であるか動作であるかを厳密に区別していない。いくつかの実施形態では、表情と動作をそれぞれ設定してから、異なるラベルまたは標識を割り当てることができる。回答テキストを利用して表情および／または動作のラベルまたは標識を取得する場合、トレーニングされたモデルよって取得してもよいし、トレーニングされた、表情に対するモデルと動作に対するモデルによって対応する表情ラベルと動作ラベルをそれぞれ取得してもよい。上記の例は、本開示を説明するためのものに過ぎず、本開示への具体的な限定ではない。

【００４５】

表情および動作ラベルの識別プロセスは、図８に示すように、オフラインフローとオンラインフローに分けられる。オフラインフローは、ブロック８０２において、対話テキストの手動ラベル付け表情および動作コーパスを取得する。ブロック８０４において、ＢＥＲＴ分類モデルをトレーニングし、表情および動作識別モデル８０６を取得する。オンラインフローでは、ブロック８０８において回答テキストを取得し、次に回答テキストを表情および動作識別モデル８０６に入力して、ブロック８１０において表情および動作識別を行う。次に、ブロック８１２において、表情および／または動作の標識を出力する。いくつかの実施形態では、この表情および動作識別モデルは、様々な適当なニューラルネットワークモデルなどの任意の適当な機械学習モデルを用いることができる。

【００４６】

次に、図３に戻って説明を続け、ブロック３０８において、回答音声信号、表情および

／または動作の標識に基づいて、仮想オブジェクトを含む出力ビデオを生成し、出力ビデオは回答音声信号に基づいて確定された、仮想オブジェクトによって表現される唇形シーケンスを含む。例えば、計算機器１０８は、回答音声信号２０８、表情および／または動作の標識２１０に基づいて、仮想オブジェクト１１０を含む出力ビデオ２１２を生成する。出力ビデオには、回答音声信号に基づいて確定された、仮想オブジェクトによって表現される唇形シーケンスを含む。このプロセスは、以下、図９と図１０に関連して詳細に説明する。

【００４７】

いくつかの実施形態では、計算機器１０８は、回答音声信号２０８と出力ビデオ２１２とを関連付けて出力する。この方法によって、正確なマッチングした音声とビデオの情報を生成することができる。このプロセスでは、回答音声信号２０８と出力ビデオ２１２とを時間的に同期させることによって、ユーザとやり取りをする。

【００４８】

この方法により、インタラクションの内容の範囲を著しく増加させ、マンマシンインタラクションの品質とレベルを向上させ、ユーザ体験を向上させることができる。

以上、図３から図８に関連して、本開示のいくつかの実施形態によるマンマシンインタラクションのための方法３００のローチャートを説明する。以下、図９に関連して、回答音声信号、表情および／または動作の標識に基づいて出力ビデオを生成するプロセスについて詳細に説明する。図９は、本開示のいくつかの実施形態による出力ビデオを生成するための方法９００のフローチャートを示す。

【００４９】

ブロック９０２において、計算機器１０８は回答音声信号を１セットの音声信号ユニットに分割する。いくつかの実施形態では、計算機器１０８は、ワード単位で音声信号ユニットを分割する。いくつかの実施形態では、計算機器１０８は、音節単位で音声信号ユニットを分割する。上記の例は、本開示を説明するためのものに過ぎず、本開示への具体的な限定ではない。当業者は任意の適当な音声サイズで音声ユニットを分割することができる。

【００５０】

ブロック９０４において、計算機器１０８は、１セットの音声信号ユニットに対応する仮想オブジェクトの唇形シーケンスを取得する。計算機器１０８は、対応するデータベースから音声信号ごとに対応する唇形ビデオを検索することができる。音声信号ユニットと唇形の対応関係を生成する場合、まず、仮想オブジェクトに対応する人間の発声ビデオを録画し、次に、ビデオから音声信号ユニットに対応する唇形を抽出する。次に、唇形と音声信号ユニットとを関連付けてデータベースに記憶する。

【００５１】

ブロック９０６において、計算機器１０８は、表情および／または動作の標識に基づいて、仮想オブジェクトについての対応する表情および／または動作のビデオセグメントを取得する。データベースまたは記憶装置には、表情および／または動作の標識と、対応する表情および／または動作のビデオセグメントとのマッピング関係が事前に記憶される。例えば表情および／または動作のラベルまたはタイプなどの標識を取得した後に、表情および／または動作の標識と、ビデオセグメントとのマッピング関係を利用して、対応するビデオを検索することができる。

【００５２】

ブロック９０８において、計算機器１０８は、唇形シーケンスをビデオセグメントに結合して出力ビデオを生成する。計算機器は、時系列に、取得された、１セットの音声信号ユニットに対応する唇形シーケンスをビデオセグメントの各フレームに結合する。

【００５３】

いくつかの実施形態では、計算機器１０８は、ビデオセグメントにおける時間軸での所定の時間位置におけるビデオフレームを確定する。次に、計算機器１０８は、唇形シーケンスから所定の時間位置に対応する唇形を取得する。唇形を取得した後、計算機器１０８

は唇形をビデオフレームに結合して出力ビデオを生成する。この方式により、正確な唇形を含むビデオを迅速に取得することができる。

【００５４】

この方法によって、仮想人物の唇形を音声と動作により正確にマッチングすることができ、ユーザの体験を改善する。

以上、図９に関連して、本開示のいくつかの実施形態による出力ビデオを生成するための方法９００のフローチャートを示す。以下、図１０に関連して、出力ビデオを生成するプロセスについてさらに説明する。図１０は、本開示のいくつかの実施形態による出力ビデオを生成するための方法１０００のフローチャートを示す。

【００５５】

図１０においては、生成されたビデオは、回答音声信号と表情動作ラベルに基づいて仮想人物を合成するビデオセグメントを含む。このプロセスは図１０に示すように、唇形ビデオの取得、表情動作ビデオの取得およびビデオのレンダリングの三つの部分を含む。

【００５６】

唇形ビデオの取得プロセスは、オンラインフローとオフラインフローに分けられる。オフラインフローでは、ブロック１００２において、音声および対応する唇形の人間ビデオの撮影を実行する。次に、ブロック１００４において、人間の音声と唇形ビデオのアライメントを実行する。このプロセスでは、音声ユニットごとに対応する唇形ビデオを取得する。その後、取得された音声ユニットと唇形ビデオとを関連付けて音声唇形ライブラリ１００６に記憶する。オンラインフローでは、ブロック１００８において、計算機器１０８は回答音声信号を取得する。次に、ブロック１０１０において、計算機器１０８は回答音声信号を音声信号ユニットに分割し、その後、唇形データベース１００６から音声信号ユニットに基づいて対応する唇形を抽出する。

【００５７】

表情動作ビデオの取得プロセスもオンラインフローとオフラインフローに分けられる。オフラインフローでは、ブロック１０１４において、人間の表情動作ビデオを撮影する。次に、ブロック１０１６において、ビデオを分割して表情および／または動作標識ごとに対応するビデオを取得し、即ち、表情および／または動作をビデオユニットとアライメントする。その後、表情および／または動作ラベルとビデオとを関連付けて表情および／または動作ライブラリ１０１８に記憶する。いくつかの実施形態では、表情および／または動作ライブラリ１０１８には、表情および／または動作の標識と、対応するビデオとのマッピング関係を記憶する。いくつかの実施形態では、表情および／または動作ライブラリにおいて、表情および／または動作の標識を用いて、マルチレベルマッピングを利用して対応するビデオを見つける。上記の例は、本開示を説明するためのものに過ぎず、本開示への具体的な限定ではない。

【００５８】

オンライン段階のフローでは、ブロック１０１２において、計算機器１０８は、入力表情および／動作の標識を取得する。次に、ブロック１０２０において、表情および／または動作の標識に基づいてビデオセグメントを抽出する。

【００５９】

その後、ブロック１０２２において、唇形シーケンスをビデオセグメントに結合する。このプロセスにおいて、表情と動作ラベルに対応するビデオは時間軸でのビデオフレームによってスティッチングされてなり、唇形シーケンスに基づいて、それぞれの唇形を時間軸での同じ位置のビデオフレームにレンダリングし、最終的に組み合わされたビデオを出力する。次に、ブロック１０２４において、出力ビデオを生成する。

【００６０】

図１１は、本開示の実施形態によるマンマシンインタラクションのための装置１１００の概略的ブロック図を示す。図１１に示すように、装置１１００は、受信した音声信号に基づいて、音声信号に対する回答の回答テキストを生成するように構成される回答テキスト生成モジュール１１０２を含む。装置１１００は、音声信号ユニットとテキストユニッ

トとのマッピング関係に基づいて、１セットのテキストユニットを含む回答テキストに対応する回答音声信号を生成し、生成された回答音声信号は１セットのテキストユニットに対応する１セットの音声ユニットを含むように構成される第１回答音声信号生成モジュール１１０４をさらに含む。装置１１００は、回答テキストに基づいて、仮想オブジェクトによって表現される表情および／または動作の標識を確定するように構成される標識確定モジュール１１０６をさらに含む。装置１１００は、回答音声信号、表情および／または動作の標識に基づいて、仮想オブジェクトを含む出力ビデオを生成し、出力ビデオは回答音声信号に基づいて確定された、仮想オブジェクトによって表現される唇形シーケンスを含むように構成される第１出力ビデオ生成モジュール１１０８をさらに含む。

【００６１】

いくつかの実施形態では、回答テキスト生成モジュール１１０２は、受信した音声信号を識別して入力テキストを生成するように構成される入力テキスト生成モジュールと、入力テキストに基づいて、回答テキストを取得するように構成される回答テキスト取得モジュールを含む。

【００６２】

いくつかの実施形態では、回答テキスト生成モジュールは、回答テキストを取得するために、入力テキストと仮想オブジェクトの人格属性を用いて回答テキストを生成する機械学習モデルである対話モデルに入力テキストと仮想オブジェクトの人格属性を入力するように構成されるモデルに基づく回答テキスト取得モジュールを含む。

【００６３】

いくつかの実施形態では、対話モデルは、仮想オブジェクトの人格属性および入力テキストサンプルと回答テキストサンプルとを含む対話サンプルトを利用してレーニングすることで得られるものである。

【００６４】

いくつかの実施形態では、第１回答音声信号生成モジュールは、回答テキストを１セットのテキストユニットに分割するように構成されるテキストユニット分割モジュールと、音声信号ユニットとテキストユニットとのマッピング関係に基づいて、１セットのテキストユニットにおけるテキストユニットに対応する音声信号ユニットを取得するように構成される音声信号ユニット取得モジュールと、音声ユニットに基づいて回答音声信号を生成するように構成される第２回答音声信号生成モジュールとを含む。

【００６５】

いくつかの実施形態では、音声信号ユニット取得モジュールは、音声信号ユニットとテキストユニットとのマッピング関係に基づいて、１セットのテキストユニットからテキストユニットを選択するように構成されるテキストユニット選択モジュールと、音声ライブラリからテキストユニットに対応する音声信号ユニットを検索するように構成される検索モジュールとを含む。

【００６６】

いくつかの実施形態では、音声ライブラリには音声信号ユニットとテキストユニットとのマッピング関係が記憶され、音声ライブラリにおける音声信号ユニットは、取得された、前記仮想オブジェクトに関する音声記録データを分割することで取得されるものであり、音声ライブラリにおけるテキストユニットは、分割で得られた音声信号ユニットに基づいて確定されるものである。

【００６７】

いくつかの実施形態では、標識判定モジュール１１０６は、テキストを用いて表情および／または動作の標識を確定する機械学習モデルである表情および動作識別モデルに回答テキストを入力して、表情および／または動作の標識を取得するように構成される表情動作標識取得モジュールを含む。

【００６８】

いくつかの実施形態では、第１出力ビデオ生成モジュール１１０８は回答音声信号を１セットの音声信号ユニットに分割するように構成される音声信号分割モジュールと、１セ

ットの音声信号ユニットに対応する仮想オブジェクトの唇形シーケンスを取得するように構成される唇形シーケンス取得モジュールと、表情および／または動作の標識に基づいて、仮想オブジェクトについての対応する表情および／または動作のビデオセグメントを取得するように構成されるビデオセグメント取得モジュールと、唇形シーケンスをビデオセグメントに結合して出力ビデオを生成するように構成される第２出力ビデオ生成モジュールとを含む。

【００６９】

いくつかの実施形態では、第２出力ビデオ生成モジュールは、ビデオセグメントにおける時間軸での所定の時間位置におけるビデオフレームを確定するように構成されるビデオフレーム確定モジュールと、唇形シーケンスから所定の時間位置に対応する唇形を取得するように構成される唇形取得モジュールと、唇形をビデオフレームに結合して出力ビデオを生成するように構成される結合モジュールとを含む。

【００７０】

いくつかの実施形態では、装置１１００は回答音声信号と出力ビデオとを関連付けて出力するように構成される出力モジュールをさらに含む。

本開示の実施形態によれば、本公開は、電子機器、可読記憶媒体およびコンピュータプログラム製品をさらに提供する。

【００７１】

図１２は、本開示の実施形態を実施するための例示的な電子機器１２００の概略的ブロック図を示す。図１の端末１０４および計算機器１０８は、電子機器１２００によって実現することができる。電子機器は、ラップトップ型コンピュータ、デスクトップ型コンピュータ、ワークステーション、パーソナルデジタルアシスタント、サーバ、ブレードサーバ、大型コンピュータ、その他の好適なコンピュータなど、様々なディジタルコンピュータを指すことを意図している。電子機器は、例えば、パーソナルデジタル処理、携帯電話、スマートフォン、ウェアラブル機器、その他の類似装置などの様々なモバイル機器を指すこともできる。本明細書に示される部材、それらの接続関係、およびそれらの機能は、ただ一例に過ぎず、本明細書に記載および／または請求の本開示の実現を制限することを意図するものではない。

【００７２】

図１２に示すように、機器１２００は、計算ユニット１２０１を含み、それはリードオンリーメモリ（ＲＯＭ）１２０２に記憶されたプログラムまた記憶ユニット１２０８からランダムアクセスメモリ（ＲＡＭ）１２０３にロードされたプログラムによって、種々の適当な操作と処理を実行することができる。ＲＡＭ１２０３には、機器１２００の動作に必要な種々のプログラムとデータを記憶することもできる。計算ユニット１２０１、ＲＯＭ１２０２およびＲＡＭ１２０３はバス１２０４によって互いに接続される。入力／出力（Ｉ／Ｏ）インターフェース１２０５もバス１２０４に接続される。

【００７３】

機器９００における複数の部材はＩ／Ｏインターフェース１２０５に接続され、この複数の部材は、例えば、キーボード、マウスなどの入力ユニット１２０６と、例えば、様々なタイプのディスプレイ、スピーカーなどの出力ユニット１２０７と、例えば、磁気ディスク、光ディスクなどの記憶ユニット１２０８と、例えば、ネットワークカード、モデム、無線通信送受信機などの通信ユニット１２０９と、を含む。通信ユニット１２０９は、機器１２００が例えば、インターネットなどのコンピュータネットワーク及び／又は様々な電気通信ネットワークを介して他の機器と情報／データのやり取りをすることを可能にする。

【００７４】

計算ユニット１２０１は処理および計算能力を有する様々な汎用および／または専用の処理コンポーネントであってもよい。計算ユニット１２０１の例には、中央処理ユニット（ＣＰＵ）、グラフィックス処理ユニット（ＧＰＵ）、様々な専用人工知能（ＡＩ）計算チップ、様々な機械学習モデルアルゴリズムを実行する計算ユニット、デジタル信号プロ

セッサ（ＤＳＰ）、および任意の適当なプロセッサ、コントローラ、マイクロコントローラなどが含まれるがこれらに限定されない。計算ユニット１２０１は以上で説明される例えば方法２００、３００、４００、６００、８００、９００および１０００のような様々な方法および処理を実行する。例えば、いくつかの実施形態では、方法２００、３００、４００、６００、８００、９００および１０００をコンピュータソフトウェアプログラムとして実現することができ、それは記憶ユニット１２０８などの機械可読媒体に有形的に含まれる。いくつかの実施形態では、コンピュータプログラムの一部または全部は、ＲＯＭ１２０２および／または通信ユニット１２０９を介して機器１２００にロードされたりインストールされたりすることができる。コンピュータプログラムがＲＡＭ１２０３にロードされて計算ユニット１２０１によって実行される場合、以上で説明される方法２００、３００、４００、６００、８００、９００および１０００の１つまたは複数のステップを実行することできる。代替的に、他の実施形態において、計算ユニット１２０１は、他の任意の適当な方法で（例えば、ファームウェアを用いて）、方法２００、３００、４００、６００、８００、９００および１０００を実行するように構成される。

【００７５】

ここで述べるシステムおよび技術の様々な実施形態は、デジタル電子回路システム、集積回路システム、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）、専用集積回路（ＡＳＩＣ）、専用標準製品（ＡＳＳＰ）、チップ上システムのシステム（ＳＯＣ）、コンプレックスプログラマブルロジックデバイス（ＣＰＬＤ）、コンピュータハードウェア、ファームウェア、ソフトウェア、および／またはそれらの組み合わせで実現されてもよい。これら様々な実施形態は、１つまたは複数のコンピュータプログラムに実装され、この１つまたは複数のコンピュータプログラムは、少なくとも１つのプログラマブルプロセッサを含むプログラマブルシステム上で実行することおよび／または解釈することが可能であり、このプログラマブルプロセッサは、専用または汎用のプログラマブルプロセッサであってもよいし、記憶システム、少なくとも１つの入力装置、および少なくとも１つの出力装置からデータおよびコマンドを受信し、この記憶システム、この少なくとも１つの入力装置、およびこの少なくとも１つの出力装置にデータおよびコマンドを送信することが可能である。

【００７６】

本開示の方法を実施するためのプログラムコードは、１つまたは複数のプログラミング言語の任意の組み合わせを用いて作成することができる。これらのプログラムコードは、汎用コンピュータ、専用コンピュータ、または他のプログラマブルデータ処理装置のプロセッサまたはコントローラに提供することができ、これによって、プログラムコードがプロセッサまたはコントローラによって実行されると、フローチャートおよび／またはブロック図で規定された機能／操作が実行される。プログラムコードは完全に機械上で実行されても、部分的に機械で実行されても、独立ソフトウェアパッケージとして部分的に機械で実行されかつ部分的に遠隔機械上で実行されても、または、完全に遠隔機械またはサーバー上で実行されてもよい。

【００７７】

本開示のコンテストにおいて、機械可読媒体は、コマンド実行システム、装置、また機器が使用するプログラムまたはコマンド実行システム、装置または機器と組み合わせて使用されるプログラムを含むか記憶することができる有形の媒体であってもよい。機械可読媒体は、機械可読信号媒体または機械可読記憶媒体であってもよい。機械可読媒体は、電子的、磁気的、光学的、電磁的、赤外線的、または半導体システム、装置や機器、または上記の内容の任意の適当な組み合わせを含むことができるが、これらに限定されない。機械可読記憶媒体のより具体的な例は、１つまたは複数のワイヤに基づく電気接続、ポータブルコンピュータディスク、ハードディスク、ランダムアクセスメモリ（ＲＡＭ）、リードオンリーメモリ（ＲＯＭ）、消去可能プログラマブル読み取り専用メモリ（ＥＰＲＯＭまたフラッシュメモリ）、光ファイバ、ポータブルコンパクトディスク読み取り専用メモリ（ＣＤ－ＲＯＭ）、光学記憶機器、磁気記憶機器、また上記の内容の任意の適当な組み

合わせを含むことができる。

【００７８】

ユーザとのインタラクションを提供するために、ここで述べたシステムおよび技術をコンピュータ上で実行することができる。このコンピュータは、ユーザに情報を表示するための表示装置（例えば、ＣＲＴ（Ｃａｔｈｏｄｅ　Ｒａｙ　Ｔｕｂｅ、陰極線管）またはＬＣＤ（ＬｉｑｕｉｄＣｒｙｓｔａｌ　Ｃｒｙｓｔａｌ　Ｄｉｓｐｌａｙ、液晶表示装置）モニタ）と、キーボードやポインティング装置を有し、ユーザはこのキーボードやポインティング装置（例えば、マウスやトラックボール）によって入力をコンピュータに提供することができる。他の種類の装置は、さらに、ユーザとのインタラクションを提供するために利用することができる。例えば、ユーザに提供されるフィードバックは、任意の形のセンシングフィードバック（例えば、視覚フィードバック、聴覚フィードバック、または触覚フィードバック）であってもよい。しかも、ユーザからの入力を、任意の形（ボイス入力、音声入力、触覚入力を含む）で受け付けてもよい。

【００７９】

ここで述べたシステムや技術は、バックステージ部材を含む計算システム（例えば、データサーバとして）や、ミドルウェア部材を含む計算システム（例えば、アプリケーションサーバ）や、フロントエンド部材を含む計算システム（例えば、グラフィカルユーザインタフェースやウェブブラウザを有するユーザコンピュータ、ユーザが、そのグラフィカルユーザインタフェースやウェブブラウザを通じて、それらのシステムや技術の実施形態とのインタラクティブを実現できる）、あるいは、それらのバックステージ部材、ミドルウェア部材、あるいはフロントエンド部材の任意の組み合わせからなる計算システムには実施されてもよい。システムの部材は、任意の形式や媒体のデジタルデータ通信（例えば、通信ネットワーク）により相互に接続されてもよい。通信ネットワークとしては、例えば、ＬＡＮ（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、ＷＡＮ（Ｗｉｄｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、インターネットを含む。

【００８０】

コンピュータシステムは、クライアントとサーバとを含んでもよい。クライアントとサーバとは、一般に互いに離れ、通常、通信ネットワークを介してやりとりを行う。クライアントとサーバの関係は、対応するコンピュータ上で動作し、かつ、互いにクライアントとサーバの関係を有するコンピュータプログラムにより生成される。

【００８１】

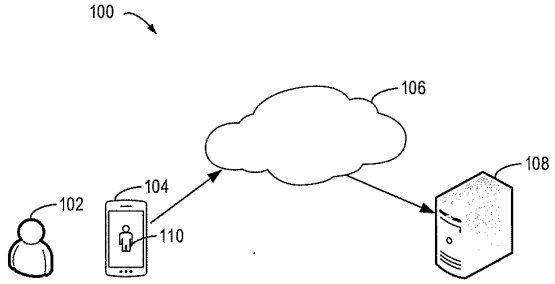
理解できるように、以上に示した様々な形式のフローを用いて、ステップを再び並び、増加または削除することができる。例えば、本開示に記載された各ステップは、並行して実行されてもよいし、順次実行されてもよいし、異なる順序で実行されてもよいし、本開示に開示された技術的解決手段が所望する結果を実現できれば、本明細書はここでは限定しない。

【００８２】

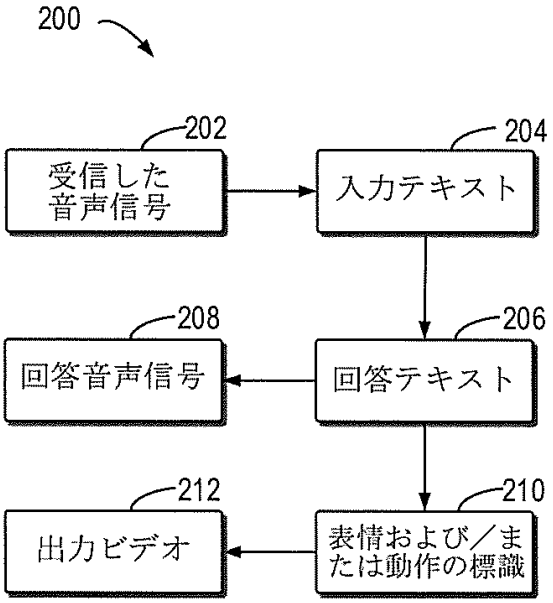
上述した具体的な実施形態は、本開示に係る保護範囲に対する制限を構成していない。当業者は、設計要件やその他の要因によって、種々の変更、組み合わせ、サブコンビネーション、代替が可能であることは明らかである。本開示における精神および原則から逸脱することなく行われるいかなる修正、同等物による置換や改良等などは、いずれも本開示の保護範囲に含まれるものである。

【図面】

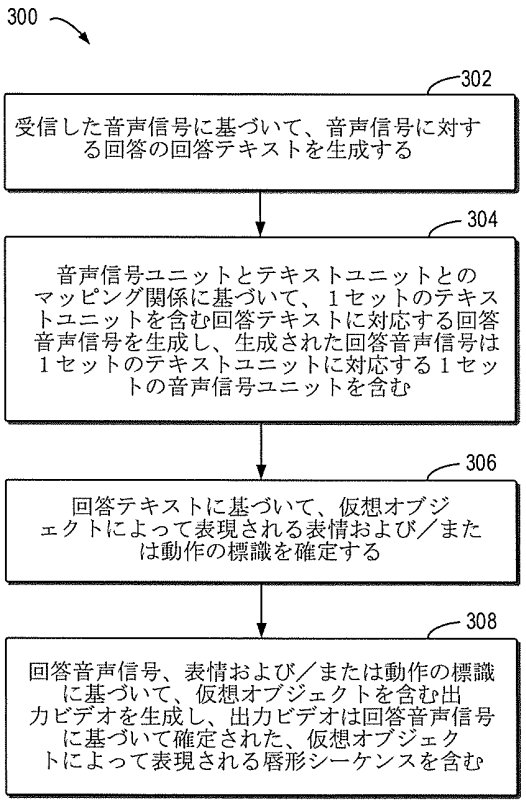
【図１】


100
106
108
104
102
110

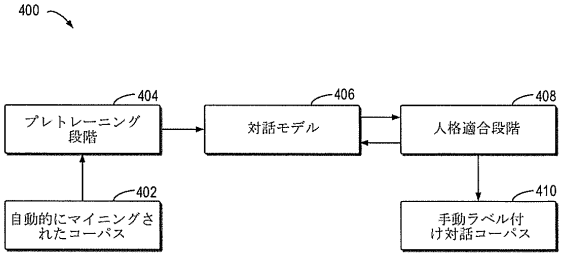
【図２】


200
202
受信した音声信号
204
入力テキスト
208
回答音声信号
206
回答テキスト
212
出力ビデオ
210
表情および／または動作の標識

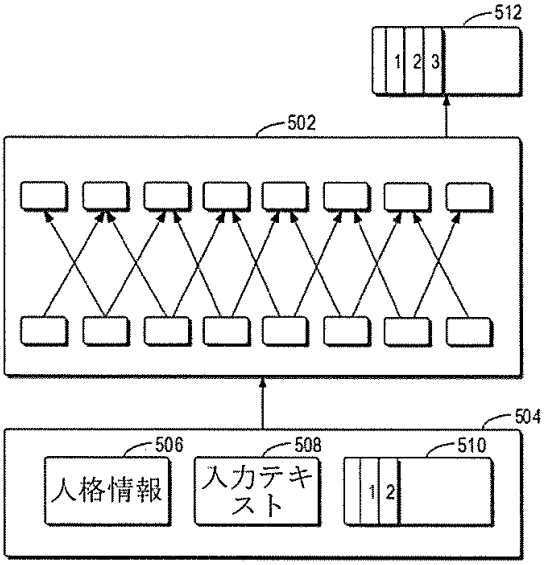
【図３】


300
302
受信した音声信号に基づいて、音声信号に対する回答の回答テキストを生成する
304
音声信号ユニットとテキストユニットとのマッピング関係に基づいて、１セットのテキストユニットを含む回答テキストに対応する回答音声信号を生成し、生成された回答音声信号は１セットのテキストユニットに対応する１セットの音声信号ユニットを含む
306
回答テキストに基づいて、仮想オブジェクトによって表現される表情および／または動作の標識を確定する
308
回答音声信号、表情および／または動作の標識に基づいて、仮想オブジェクトを含む出力ビデオを生成し、出力ビデオは回答音声信号に基づいて確定された、仮想オブジェクトによって表現される唇形シーケンスを含む

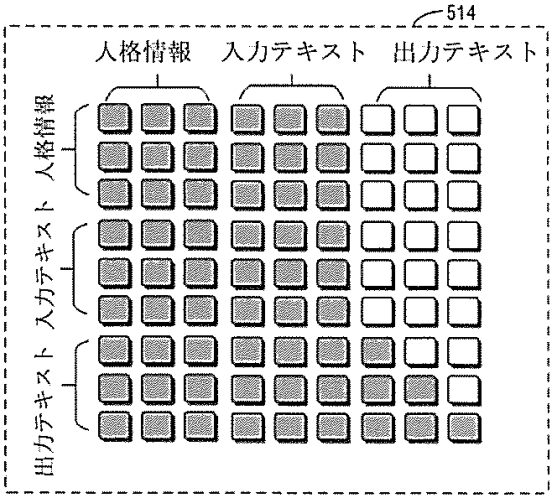
【図４】


400
404
プレトレーニング段階
406
対話モデル
408
人格適合段階
402
自動的にマイニングされたコーパス
410
手動ラベル付け対話コーパス

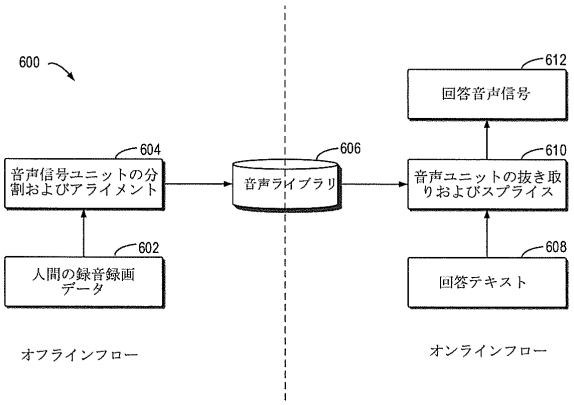
【図５Ａ】



【図５Ｂ】



【図６】

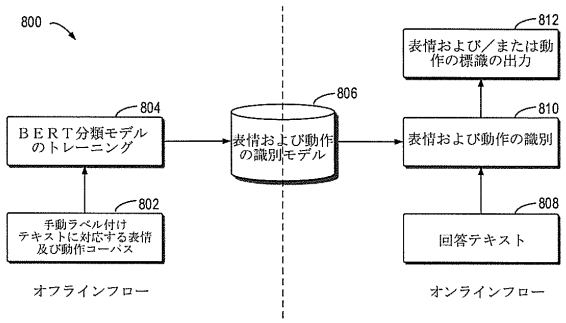


【図７】

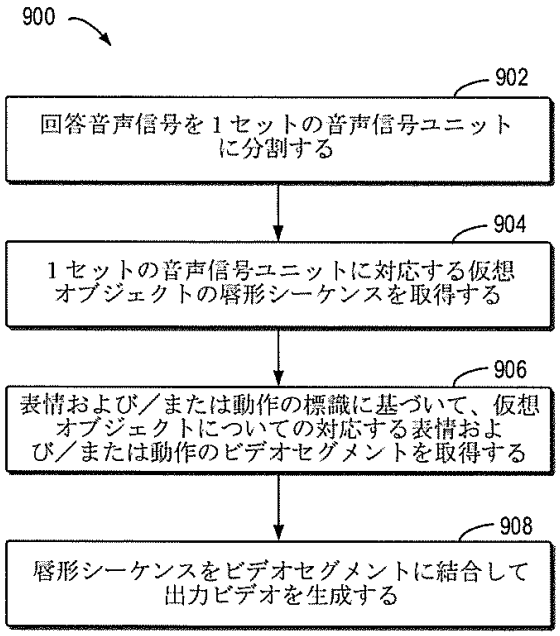
700

表情および／または動作ラベル	解説
自然な状態－耳を傾けること	両手を体の前に置き、顔の表情が自然で、まばたきやわずかなうなずきの動作
自然な状態－述べる	両手を体の前に置き、片手を外に向け、顔の表情が自然で、まばたきができること
微笑する	主に眉毛や顔全体で、微笑んでいるような表情を表現
口を開けて笑う	口を開けて、はっきりとした口の動きで数回笑う
しかめっ面	主に眉毛や顔全体で、しかめっ面をする
驚き	主に眉毛、目、顔全体で、驚きの表情を表現
ご挨拶とお別れ	右手を上げて手招きしている
自己紹介	右手を軽く上げ、襟元に向けて軽くタップする
確認－ジェスチャ	右手を上げてＯＫのジェスチャーをする
否定－頭部	頭を振って否定する（ショートアクション、話す前にアクションを完了させる）
思考と質問	少し首を傾げて、しかめっ面をしているような表情があること

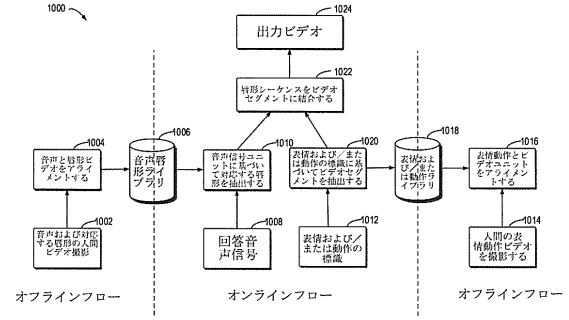
【図８】


800
812
表情および／または動作の標識の出力
804
ＢＥＲＴ分類モデルのトレーニング
806
表情および動作の識別モデル
810
表情および動作の識別
802
手動ラベル付けテキストに対応する表情及び動作コーパス
808
回答テキスト
オフラインフロー
オンラインフロー

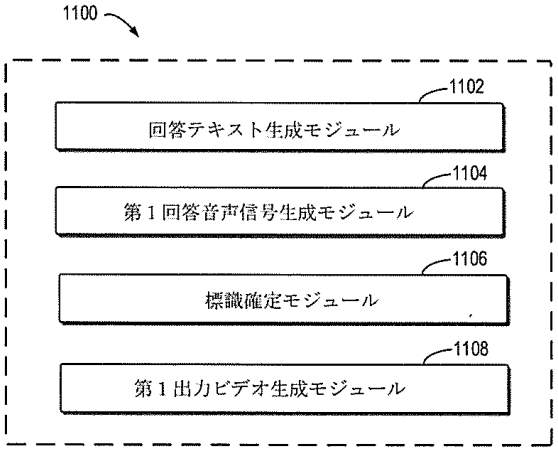
【図９】


900
902
回答音声信号を１セットの音声信号ユニットに分割する
904
１セットの音声信号ユニットに対応する仮想オブジェクトの唇形シーケンスを取得する
906
表情および／または動作の標識に基づいて、仮想オブジェクトについての対応する表情および／または動作のビデオセグメントを取得する
908
唇形シーケンスをビデオセグメントに結合して出力ビデオを生成する

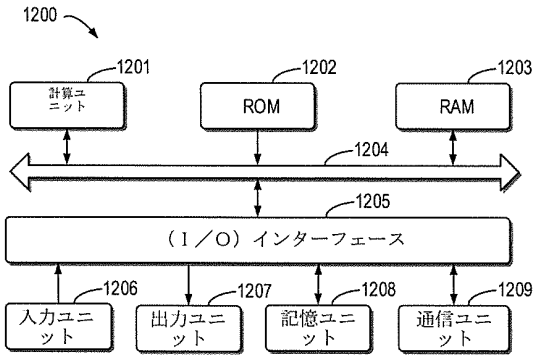
【図１０】


1000
1024
出力ビデオ
1022
唇形シーケンスをビデオセグメントに結合する
1004
音声と唇形ビデオをアライメントする
1006
音声唇形ライブラリ
1010
音声信号ユニットに基づいて対応する唇形を抽出する
1020
表情および／または動作の標識に基づいてビデオセグメントを抽出する
1018
表情および／または動作ライブラリ
1016
表情動作とビデオユニットをアライメントする
1002
音声および対応する唇形の人間ビデオ撮影
1008
回答音声信号
1012
表情および／または動作の標識
1014
人間の表情動作ビデオを撮影する
オフラインフロー
オンラインフロー
オフラインフロー

【図１１】


1100
1102
回答テキスト生成モジュール
1104
第１回答音声信号生成モジュール
1106
標識確定モジュール
1108
第１出力ビデオ生成モジュール

【図１２】


1200
1201
計算ユニット
1202
ROM
1203
RAM
1204
1205
（I／O）インターフェース
1206
入力ユニット
1207
出力ユニット
1208
記憶ユニット
1209
通信ユニット

フロントページの続き

(51)国際特許分類　　FI
　　G06F　3/01　510

　　85，China

(74)代理人　100118902
　　弁理士　山本 修
(74)代理人　100106208
　　弁理士　宮前 徹
(74)代理人　100196508
　　弁理士　松尾 淳一
(74)代理人　100138759
　　弁理士　大房 直樹
(72)発明者　ウエンチュエン・ウー
　　中華人民共和国 ペキン １０００８５， ハイディアン ディストリクト， シャンディ テンス ストリート， １０番， バイドゥ キャンパス ２階
(72)発明者　フア・ウー
　　中華人民共和国 ペキン １０００８５， ハイディアン ディストリクト， シャンディ テンス ストリート， １０番， バイドゥ キャンパス ２階
(72)発明者　ハイフオン・ワーン
　　中華人民共和国 ペキン １０００８５， ハイディアン ディストリクト， シャンディ テンス ストリート， １０番， バイドゥ キャンパス ２階

審査官　鈴木 圭一郎

(56)参考文献　特開平０９－０１６８００（ＪＰ，Ａ）
　　特開２００６－３３０４８４（ＪＰ，Ａ）
　　特開２００６－０９９１９４（ＪＰ，Ａ）
　　特開平１１－２３１８９９（ＪＰ，Ａ）
　　特開平１１－３３９０５８（ＪＰ，Ａ）
　　特開２０２０－１６０３４１（ＪＰ，Ａ）
　　特開２００４－３１００３４（ＪＰ，Ａ）
　　特開平９－０８１６３２（ＪＰ，Ａ）
　　高津 弘明、小林 哲則，対話エージェントのための性格モデル，言語処理学会第２１回年次大会 発表論文集 ［ｏｎｌｉｎｅ］，日本，言語処理学会，2015年03月09日，191～194

(58)調査した分野　(Int.Cl.，ＤＢ名)
　　Ｇ０６Ｔ　１３／４０
　　Ｇ１０Ｌ　１５／００
　　Ｇ１０Ｌ　１３／００
　　Ｇ０６Ｆ　３／１６
　　Ｇ０６Ｆ　３／０１