(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7522177号
(P7522177)

(45)発行日　令和6年7月24日(2024.7.24)

(24)登録日　令和6年7月16日(2024.7.16)

(51)国際特許分類

		FI		
G06F	*3/01 (2006.01)*	G06F	3/01	570
G06F	*3/16 (2006.01)*	G06F	3/16	650
		G06F	3/16	670
		G06F	3/01	510

請求項の数　15　（全36頁）

(21)出願番号　特願2022-500128(P2022-500128)
(86)(22)出願日　令和2年7月10日(2020.7.10)
(65)公表番号　特表2022-539794(P2022-539794A)
(43)公表日　令和4年9月13日(2022.9.13)
(86)国際出願番号　PCT/US2020/041499
(87)国際公開番号　WO2021/011331
(87)国際公開日　令和3年1月21日(2021.1.21)
審査請求日　令和5年6月12日(2023.6.12)
(31)優先権主張番号　62/873,775
(32)優先日　令和1年7月12日(2019.7.12)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　16/685,946
(32)優先日　令和1年11月15日(2019.11.15)
最終頁に続く

(73)特許権者　595020643
クゥアルコム・インコーポレイテッド
ＱＵＡＬＣＯＭＭ　ＩＮＣＯＲＰＯＲＡＴＥＤ
アメリカ合衆国、カリフォルニア州　９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ　５７７５
(74)代理人　110003708
弁理士法人鈴榮特許綜合事務所
(74)代理人　100108855
弁理士　蔵田 昌俊
(74)代理人　100158805
弁理士　井関 守三
(74)代理人　100112807
弁理士　岡田 貴志
最終頁に続く

(54)【発明の名称】　マルチモーダルユーザインターフェース

(57)【特許請求の範囲】
【請求項１】
マルチモーダルユーザ入力のためのデバイスであって、
第１の入力デバイスから受信された第１のデータを処理することと、前記第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、前記第１の入力は、コマンドに対応し、
前記第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることと、ここにおいて、前記フィードバックメッセージは、前記第１の入力モードとは異なる第２の入力モードに基づいて、前記第１の入力に関連するコマンドを識別する第２の入力を提供するように前記ユーザに命令する、
第２の入力デバイスから第２のデータを受信することと、前記第２のデータは、前記第２の入力を示し、
前記第１の入力のマッピングを、前記第２の入力に関連付けられた行為に対して更新することと、
を行うように構成された１つまたは複数のプロセッサを備える、デバイス。
【請求項２】
前記第１の入力モードは、スピーチモード、ジェスチャーモード、またはビデオモードのうちの１つであり、前記第２の入力モードは、前記スピーチモード、前記ジェスチャーモード、または前記ビデオモードのうちの異なる１つである、請求項１に記載のデバイス。
【請求項３】

前記フィードバックメッセージは、前記第１の入力をディスアンビギュエートするために前記第２の入力を提供するように前記ユーザに命令し、

オプションで、前記１つまたは複数のプロセッサは、前記第１の入力の認識処理に関連する確信度レベルが確信度しきい値を満たすことに失敗したことに応答して、前記フィードバックメッセージを送るようにさらに構成された、請求項１に記載のデバイス。

【請求項４】

前記更新されたマッピングは、前記第１の入力と前記第２の入力との組合せを前記コマンドに関連付ける、請求項１に記載のデバイス。

【請求項５】

前記１つまたは複数のプロセッサは、マルチモーダル認識エンジンを含み、前記マルチモーダル認識エンジンは、

組み合わされた埋め込みベクトルを生成するために、前記第１の入力モードに関連する第１の埋め込みネットワークと、前記第２の入力モードに関連する第２の埋め込みネットワークとの出力を組み合わせるように構成された融合埋め込みネットワークと、

前記組み合わされた埋め込みベクトルを特定のコマンドにマッピングするように構成された分類器と、

を含む、請求項１に記載のデバイス。

【請求項６】

前記ユーザに対応する第１の埋め込みネットワークデータおよび第１の重みデータと、

第２のユーザに対応する第２の埋め込みネットワークデータおよび第２の重みデータと、前記第１の埋め込みネットワークデータは、前記ユーザと前記第２のユーザとの間の入力コマンドの差に基づいて前記第２の埋め込みネットワークデータとは異なり、前記第１の重みデータは、前記ユーザと前記第２のユーザとの間の入力モード信頼性の差に基づいて前記第２の重みデータとは異なり、

を記憶するように構成されたメモリをさらに備える、請求項５に記載のデバイス。

【請求項７】

前記第１の入力モードは、ビデオモードに対応し、前記１つまたは複数のプロセッサは、照明しきい値を下回る値を有する周辺光メトリックに応答して前記フィードバックメッセージを送るように構成された、請求項１に記載のデバイス。

【請求項８】

前記第１の入力モードは、スピーチモードに対応し、前記１つまたは複数のプロセッサは、雑音しきい値を超える値を有する雑音メトリックに応答して前記フィードバックメッセージを送るように構成された、請求項１に記載のデバイス。

【請求項９】

グラフィカルユーザインターフェースを表すように構成されたディスプレイをさらに備える、請求項１に記載のデバイス。

【請求項１０】

１つまたは複数のキーワードまたは音声コマンドを含むオーディオ入力をキャプチャするように構成された１つまたは複数のマイクロフォンをさらに備える、請求項１に記載のデバイス。

【請求項１１】

１つまたは複数のジェスチャーまたは視覚的コマンドを含むビデオ入力をキャプチャするように構成された１つまたは複数のカメラをさらに備える、請求項１に記載のデバイス。

【請求項１２】

ジェスチャー入力を示すデータを受信するように構成された１つまたは複数のアンテナをさらに備える、請求項１に記載のデバイス。

【請求項１３】

前記フィードバックメッセージをレンダリングするかまたは前記ユーザにダイレクトするように構成された１つまたは複数のラウドスピーカーをさらに備える、請求項１に記載のデバイス。

【請求項１４】
マルチモーダルユーザ入力のための方法であって、
デバイスの１つまたは複数のプロセッサにおいて、第１の入力デバイスから受信された第１のデータを処理することと、前記第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、前記第１の入力は、コマンドに対応し、
前記１つまたは複数のプロセッサから、前記第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることと、ここにおいて、前記フィードバックメッセージは、前記第１の入力モードとは異なる第２の入力モードに基づいて、前記第１の入力に関連するコマンドを識別する第２の入力を提供するように前記ユーザに命令する、
前記１つまたは複数のプロセッサにおいて、第２の入力デバイスから第２のデータを受信することと、前記第２のデータは、前記第２の入力を示し、
前記第１の入力のマッピングを、前記第２の入力に関連付けられた行為に対して更新することと、
を備える、方法。
【請求項１５】
デバイスの１つまたは複数のプロセッサによって実行されたとき、前記１つまたは複数のプロセッサに、
第１の入力デバイスから受信された第１のデータを処理することと、前記第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、前記第１の入力が、コマンドに対応し、
前記第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることと、ここにおいて、前記フィードバックメッセージは、前記第１の入力モードとは異なる第２の入力モードに基づいて、前記第１の入力に関連するコマンドを識別する第２の入力を提供するように前記ユーザに命令する、
第２の入力デバイスから第２のデータを受信することと、前記第２のデータは、前記第２の入力を示し、
前記第１の入力のマッピングを、前記第２の入力に関連付けられた前記コマンドに対して更新することと、
を行わせる命令を備える非一時的コンピュータ可読媒体。
【発明の詳細な説明】
【優先権の主張】
【０００１】
[0001]本出願は、それらの各々の内容がそれらの全体として参照により本明細書に明確に組み込まれる、本願の譲受人が所有する２０１９年７月１２日に出願された米国仮特許出願第６２／８７３，７７５号と、２０１９年１１月１５日に出願された米国非仮特許出願第１６／６８５，９４６号との優先権の利益を主張する。
【技術分野】
【０００２】
[0002]本開示は、一般にユーザインターフェースに関係し、より詳細には、ユーザ入力の複数のモダリティ（modalities）をサポートするユーザインターフェースに関係する。
【背景技術】
【０００３】
[0003]多くのユーザインターフェースは、自動音声認識（ＡＳＲ）および自然言語処理（ＮＬＰ）に基づき、大規模カスタマーベース上で有用であるように多くの異なるコマンド、アクセント、および言語にわたってトレーニングされる。様々なユーザの間の広い適用可能性のためにそのようなユーザインターフェースをトレーニングすることは、広範なリソースを必要とし、ユーザインターフェースを大規模カスタマーベースのために一般的に適用可能にするためのトレーニングの大部分は、各個々のユーザが典型的には単一の言語、アクセント、およびサポートされるコマンドのサブセットのみを使用するので、ユー

ザごとのベースでは「浪費」になる。
【発明の概要】
【０００４】
[0004]本開示の一実装形態によれば、マルチモーダルユーザ入力のためのデバイスが、第１の入力デバイスから受信された第１のデータを処理するように構成された１つまたは複数のプロセッサを含む。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、第１の入力は、コマンドに対応する。１つまたは複数のプロセッサは、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送るように構成される。フィードバックメッセージは、第１の入力モードとは異なる第２の入力モードに基づいて、第１の入力に関連するコマンドを識別する第２の入力を提供するようにユーザに命令する。１つまたは複数のプロセッサは、第２の入力デバイスから第２のデータを受信することと、第２のデータが、第２の入力を示す、第２の入力によって識別されるコマンドに第１の入力を関連付けるようにマッピングを更新することとを行うように構成される。
【０００５】
[0005]本開示の別の実装形態によれば、マルチモーダルユーザ入力のための方法が、デバイスの１つまたは複数のプロセッサにおいて、第１の入力デバイスから受信された第１のデータを処理することを含む。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、第１の入力は、コマンドに対応する。本方法は、第１のデータを処理することに基づいて１つまたは複数のプロセッサから出力デバイスにフィードバックメッセージを送ることを含む。フィードバックメッセージは、第１の入力モードとは異なる第２の入力モードに基づいて、第１の入力に関連するコマンドを識別する第２の入力を提供するようにユーザに命令する。本方法は、１つまたは複数のプロセッサにおいて、第２の入力デバイスから第２のデータを受信することを含む。第２のデータは、第２の入力を示す。本方法はまた、１つまたは複数のプロセッサにおいて、第２の入力によって識別されるコマンドに第１の入力を関連付けるようにマッピングを更新することを含む。
【０００６】
[0006]本開示の別の実装形態によれば、マルチモーダルユーザ入力のための装置が、第１の入力デバイスから受信された第１のデータを処理するための手段を含む。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、第１の入力は、コマンドに対応する。本装置は、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送るための手段を含む。フィードバックメッセージは、第１の入力モードとは異なる第２の入力モードに基づいて、第１の入力に関連するコマンドを識別する第２の入力を提供するようにユーザに命令する。本装置は、第２の入力デバイスから第２のデータを受信するための手段を含み、第２のデータは、第２の入力を示す。本装置はまた、第２の入力によって識別されるコマンドに第１の入力を関連付けるようにマッピングを更新するための手段を含む。
【０００７】
[0007]本開示の別の実装形態によれば、非一時的コンピュータ可読媒体が、デバイスの１つまたは複数のプロセッサによって実行されたとき、１つまたは複数のプロセッサに、第１の入力デバイスから受信された第１のデータを処理させる命令を含む。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、第１の入力は、コマンドに対応する。命令は、１つまたは複数のプロセッサによって実行されたとき、１つまたは複数のプロセッサに、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送らせる。フィードバックメッセージは、第１の入力モードとは異なる第２の入力モードに基づいて、第１の入力に関連するコマンドを識別する第２の入力を提供するようにユーザに命令する。命令は、１つまたは複数のプロセッサによって実行されたとき、１つまたは複数のプロセッサに、第２の入力デバイスから第２のデータを受信させ、第２のデータは、第２の入力を示す。命令はまた、１つまたは複数のプロセッサによって実行されたとき、１つまたは複数のプロセッサに、第２の入力によって識別されるコ

マンドに第１の入力を関連付けるようにマッピングを更新させる。
【図面の簡単な説明】
【０００８】
【図１】[0008]本開示のいくつかの例による、マルチモーダルユーザ入力を処理するように動作可能なデバイスを含むシステムの特定の例示的な実装形態の図。
【図２】[0009]本開示のいくつかの例による、図１のデバイスの構成要素の特定の実装形態の図。
【図３】[0010]本開示のいくつかの例による、マルチモーダルユーザ入力を処理するように動作可能なデバイスを含むシステムの別の特定の実装形態の図。
【図４】[0011]本開示のいくつかの例による、マルチモーダルユーザ入力を処理するように動作可能なデバイスを含むシステムの別の特定の実装形態の一例の図。
【図５】[0012]本開示のいくつかの例による、マルチモーダルユーザ入力を処理するように動作可能なデバイスの別の実装形態の図。
【図６】[0013]本開示のいくつかの例による、図１のデバイスによって実施され得るマルチモーダルユーザ入力を処理する方法の実装形態の図。
【図７】[0014]本開示のいくつかの例による、図１のデバイスによって実施され得るマルチモーダルユーザ入力を処理する方法の別の実装形態の図。
【図８】[0015]本開示のいくつかの例による、図１のデバイスによって実施され得るマルチモーダルユーザ入力を処理する方法の別の実装形態の図。
【図９】[0016]本開示のいくつかの例による、図１のデバイスによって実施され得るマルチモーダルユーザ入力を処理する方法の別の実装形態の図。
【図１０】[0017]本開示のいくつかの例による、図１のデバイスによって実施され得るマルチモーダルユーザ入力を処理する方法の別の実装形態の図。
【図１１】[0018]本開示のいくつかの例による、マルチモーダルユーザ入力を処理するように動作可能な車両の図。
【図１２Ａ】[0019]本開示のいくつかの例による、マルチモーダルユーザ入力を処理するように動作可能な仮想現実または拡張現実ヘッドセットの図。
【図１２Ｂ】[0020]本開示のいくつかの例による、マルチモーダルユーザ入力を処理するように動作可能なウェアラブル電子デバイスの図。
【図１３】[0021]本開示のいくつかの例による、マルチモーダルユーザ入力を処理するように動作可能であるデバイスの特定の例示的な例のブロック図。
【発明を実施するための形態】
【０００９】
[0022]複数の入力モダリティを使用したユーザ対話を可能にするためのデバイスおよび方法について説明される。多くのユーザインターフェースは、自動音声認識（ＡＳＲ）および自然言語処理（ＮＬＰ）に基づき、大規模カスタマーベース上で有用であるように多くの異なるコマンド、アクセント、および言語にわたってトレーニングされる。様々なユーザの間の広い適用可能性のためにそのようなユーザインターフェースをトレーニングすることは、広範なリソースを必要とし、ユーザインターフェースを大規模カスタマーベースのために一般的に適用可能にするためのトレーニングの大部分は、各個々のユーザが典型的には単一の言語、アクセント、およびサポートされるコマンドのサブセットのみを使用するので、ユーザごとのベースでは「浪費」になる。
【００１０】
[0023]ユーザコマンドの解釈を個人化する能力とともに、マルチモーダルユーザ対話を可能にすることによって、本明細書で説明される技法は、マルチモーダルユーザインターフェースが、特定のユーザによる使用のためにトレーニングされることを可能にし、それにより、従来のユーザインターフェースの広い適用可能性のための広範なトレーニングを低減するかまたはなくす。いくつかの実装形態では、異なる埋め込（embedding）ネットワークは、異なる入力モダリティのために使用され（たとえば、スピーチ用の埋め込みネットワーク、視覚的入力用の埋め込みネットワーク、ジェスチャー入力用の埋め込みネッ

トワークなど）、それぞれのモダリティを使用して受信される異なるコマンド間で区別するように構成される。例示のために、「埋め込みネットワーク」は、埋め込みベクトルを生成するために、スピーチデータ（たとえば、時間領域スピーチデータまたは周波数領域スピーチデータ）などの入力データを処理するように構成された（たとえば、トレーニングされた）、１つまたは複数のニューラルネットワークレイヤを含むことができる。「埋め込みベクトル」は、入力データと比較して比較的低次元であり、入力データを表し、入力データの異なるインスタンス間で区別するために使用され得る、ベクトル（たとえば、複数の値のセット）である。異なる埋め込みネットワーク出力は、共通の埋め込み空間に変換され、組み合わされた埋め込みベクトルに融合される。たとえば、スピーチ入力のｎ次元のスピーチ埋め込みベクトルは、ｋ次元の第１の埋め込みベクトルに変換され得、ジェスチャー入力のｍ次元のジェスチャー埋め込みベクトルは、ｋ次元の第２の埋め込みベクトルに変換され得る（ここで、ｍ、ｎ、およびｋは、互いに等しいか、または異なり得る）。ｋ次元のベクトル空間（たとえば、共通の埋め込み空間）中で、ｋ次元の第１の埋め込みベクトルは、スピーチ入力を表し、ｋ次元の第２の埋め込みベクトルは、ジェスチャー入力を表す。ｋ次元の第１の埋め込みベクトルと、ｋ次元の第２の埋め込みベクトルとは、組み合わされた埋め込みベクトルを生成するために、ベクトル加算などによって組み合わされ得る。分類器は、出力を生成するために、組み合わされた埋め込みベクトルを解釈する。

【００１１】

[0024]埋め込みネットワークと分類器との各々は、様々なモダリティを介して受信されたユーザコマンドの認識を改善するように個々のユーザによって更新（たとえば、トレーニング）され得る。たとえば、高い確信度で解釈され得ない、話されたユーザコマンドが受信された場合、ユーザインターフェースは、話されたコマンドの意味に関してユーザに問い合わせることができ、ユーザは、ユーザインターフェースによって認識されるジェスチャー入力を実施することなどによって、異なるモダリティを使用して意味を入力することができる。

【００１２】

[0025]いくつかの実装形態では、ユーザインターフェースは、ユーザが入力モダリティを変更することを要求することができる。たとえば、再生ボリュームを上げるためのユーザの話されたコマンド「アップ」が、別のコマンド（たとえば、「オフ」）から確実に区別され得ない場合、ユーザインターフェースは、コマンドをより良く区別するためにユーザが別のモダリティを追加することを要求する（たとえば、話されたまた表示された）フィードバックメッセージを生成することができる。たとえば、ユーザは、「ボリュームを上げる」コマンドのために上方にポインティングすることなど、視覚的入力を追加することができる。ユーザインターフェースは、再生ボリュームを上げるためのマルチモーダルコマンドとして、話された入力「アップ」と、上方ポインティングの視覚的入力との組合せを認識するように更新され得る。したがって、コマンド認識精度を改善するために、（たとえば、シングルモーダルからマルチモーダルへの）個人化された更新が使用され得る。

【００１３】

[0026]いくつかの実装形態では、ユーザインターフェースは、ユーザ入力をより容易にディスアンビギュエートする（disambiguate）ために、ユーザが入力モダリティを変更することを要求する。たとえば、（たとえば、動いている車両中で）オーディオ雑音がユーザのスピーチの解釈を損なう実装形態では、ユーザインターフェースは、ユーザがモダリティを視覚的またはジェスチャーモダリティなどに変更することを要求するフィードバックメッセージを生成することができる。別の例として、低い光レベルがユーザの視覚的入力の解釈を損なう実装形態では、ユーザインターフェースは、ユーザがモダリティを、スピーチモダリティ、または手の移動および配向を検出するためにウェアラブル電子デバイス（たとえば、「スマートウォッチ」）の動き検出器を使用するジェスチャーモダリティなどに変更することを要求する、フィードバックメッセージを生成することができる。したがって、入力モダリティを変更するようにユーザに命令することは、コマンド認識精

度を改善するために使用され得る。

【００１４】

[0027]いくつかの実装形態では、ユーザインターフェースは、多因子認証プロセスの一部としてユーザが入力モダリティを変更することを要求する。たとえば、音声認証を実施するために、話されたユーザ入力を受信した後に、ユーザインターフェースは、次に、ユーザが視覚的またはジェスチャー入力を提供することを要求し得る。別の入力モダリティを使用して追加のユーザ入力を提供するようにとの要求は、スピーチ入力がユーザの記録されたスピーチの再生を示す特性を有するという検出など、前のユーザ入力における異常によってトリガされ得る。代替または追加として、要求は、ランダムに、または多因子認証プロセスのための確立された一連の認証入力の一部として生成され得る。入力モダリティを変更するようにユーザに命令することは、したがって、より高い精度、よりロバストなユーザ認証のために使用され得る。本明細書で使用されるとき、多因子認証プロセスのための認証入力に対応するユーザ入力は、ユーザコマンドに対応するユーザ入力とは別個である。例示のために、コマンドに対応するユーザ入力は、コマンド（たとえば、「ライトをオンにする」）に関連する行為または「スキル」を実施するための命令としてユーザインターフェースによって解釈される一方で、認証入力に対応するユーザ入力は、（たとえば、生体データまたは他のユーザ識別データの比較を介して）ユーザ入力が、記憶されたユーザプロファイルに関連付けられた同じユーザから生起するという尤度を決定するために、記憶されたユーザプロファイルのデータと比較される。

【００１５】

[0028]それの文脈によって明確に限定されない限り、「発生すること」という用語は、計算すること、生成すること、および／または提供することなど、それの通常の意味のいずれかを示すために使用される。それの文脈によって明確に限定されない限り、「提供すること」という用語は、計算すること、生成すること、および／または発生することなど、それの通常の意味のいずれかを示すために使用される。それの文脈によって明確に限定されない限り、「結合」されるという用語は、直接的または間接的な電気的接続または物理的接続を示すために使用される。接続が間接的である場合、「結合」されている構造の間に他のブロックまたは構成要素があり得る。たとえば、ラウドスピーカーは、ラウドスピーカーから壁への（またはその逆への）波（たとえば、音）の伝搬を可能にする介在する媒体（たとえば、空気）を介して近くの壁に音響的に結合され得る。

【００１６】

[0029]「構成」という用語は、それの特定の文脈によって示されるように、方法、装置、デバイス、システム、またはそれらの任意の組合せに関して使用され得る。「備える」という用語は、本明細書および特許請求の範囲において使用される場合、他の要素または動作を除外しない。（「ＡはＢに基づく」などにおけるような）「に基づく」という用語は、（ｉ）「に少なくとも基づく」（たとえば、「ＡはＢに少なくとも基づく」）、および特定の文脈において適切な場合、（ｉｉ）「に等しい」（たとえば、「ＡはＢに等しい」）という場合を含む、それの通常の意味のいずれかを示すために使用される。「ＡはＢに基づく」が「に少なくとも基づく」を含む場合（ｉ）、これは、ＡがＢに結合される構成を含み得る。同様に、「に応答して」という用語は、「に少なくとも応答して」を含む、それの通常の意味のいずれかを示すために使用される。「少なくとも１つ」という用語は、「１つまたは複数」を含む、それの通常の意味のいずれかを示すために使用される。「少なくとも２つ」という用語は、「２つ以上」を含む、それの通常の意味のいずれかを示すために使用される。

【００１７】

[0030]「装置」および「デバイス」という用語は、特定の文脈によって別段に規定されていない限り、総称的および互換的に使用される。別段に規定されていない限り、特定の特徴を有する装置の動作のいかなる開示も、類似の特徴を有する方法を開示すること（その逆も同様）をも明確に意図され、特定の構成による装置の動作のいかなる開示も、類似の構成による方法を開示すること（その逆も同様）をも明確に意図される。「方法」、「

プロセス」、「手順」、および「技法」という用語は、特定の文脈によって別段に規定されていない限り、総称的および互換的に使用される。「要素」および「モジュール」という用語は、より大きい構成の一部分を示すために使用され得る。「パケット」という用語は、ヘッダ部分とペイロード部分とを含むデータのユニットに対応し得る。文書の一部分の参照による任意の組込みはまた、その部分内で参照される用語または変数の定義が、文書内の他の場所、ならびに組み込まれた部分で参照される任意の図に現れる場合、そのような定義を組み込んでいると理解されたい。

【００１８】

[0031]本明細書で使用されるとき、「通信デバイス」という用語は、ワイヤレス通信ネットワークを介した音声および／またはデータ通信のために使用され得る電子デバイスを指す。通信デバイスの例は、スマートスピーカー、スピーカーバー、セルラーフォン、携帯情報端末（ＰＤＡ）、ハンドヘルドデバイス、ヘッドセット、ウェアラブルデバイス、ワイヤレスモデム、ラップトップコンピュータ、パーソナルコンピュータなどを含む。

【００１９】

[0032]図１は、ユーザ１０２がマルチモーダルユーザ入力のデバイス１１０と対話するシステム１００を示す。デバイス１１０は、第１の入力デバイス１１２と、第２の入力デバイス１１４と、場合によっては第３の入力デバイス１１６などの１つまたは複数の追加の入力デバイスと、出力デバイス１２０と、制御ユニット１０４とを含む。いくつかの実装形態では、デバイス１１０は、例示的および非限定的な例として、ポータブル通信デバイス（たとえば、「スマートフォン」）、ウェアラブルデバイス（たとえば、「スマートウォッチ」）、車両システム（たとえば、自動車エンターテインメントシステムとともに使用するための可動もしくはリムーバブルディスプレイ、ナビゲーションシステム、または自動運転制御システム）、あるいは仮想現実または拡張現実ヘッドセットを含むことができる。

【００２０】

[0033]第１の入力デバイス１１２は、第１の入力モードに基づく第１のユーザ入力を検出するように構成される。一例では、第１の入力デバイス１１２は、マイクロフォンを含み、第１の入力モードは、（たとえば、ＡＳＲ／ＮＬＰのための）スピーチモードを含む。例示のために、第１の入力デバイス１１２は、１つまたは複数のキーワードまたは音声コマンドを含むオーディオ入力をキャプチャするように構成された１つまたは複数のマイクロフォンを含むことができる。

【００２１】

[0034]第２の入力デバイス１１４は、第２の入力モードに基づく第２のユーザ入力を検出するように構成される。一例では、第２の入力デバイス１１４は、カメラを含み、第２の入力モードは、（たとえば、サムズアップ（thumbs-up）またはサムズダウン（thumbs-down）の手の位置、顔の表情など、ユーザ１０２の視覚的態様を検出するための）ビデオモードを含む。例示のために、第２の入力デバイス１１４は、１つまたは複数のジェスチャーまたは視覚的コマンドを含むビデオ入力をキャプチャするように構成された１つまたは複数のカメラを含むことができる。

【００２２】

[0035]第３の入力デバイス１１６は、第３の入力モードに基づく第３のユーザ入力を検出するように構成される。一例では、第３の入力デバイス１１６は、ジェスチャートラッカーを含み、第３の入力モードは、ジェスチャーモードを含む。第３の入力デバイス１１６は、ジェスチャー入力を示すデータ（たとえば、動きデータ）を受信するように構成された１つまたは複数のアンテナを含むことができる。例示のために、ユーザ１０２は、ユーザの手の移動を追跡する動きセンサー（たとえば、加速度計、ジャイロスコープなど）を含み、動きデータを第３の入力デバイス１１６に送信する、ブレスレットまたはウォッチを着用することができる。他の実装形態では、動き追跡電子デバイスは、人間ユーザ１０２中のサイバネティックインプラントなど、ユーザ１０２と一体化され得るか、またはユーザ１０２がロボットである実装形態では、ユーザ１０２の構成要素であり得る。

【００２３】
[0036]出力デバイス１２０は、ラウドスピーカーを使用した可聴出力、ディスプレイを使用した視覚的出力の生成を介して、１つまたは複数の他の出力モダリティ（たとえば、ハプティック）を介して、あるいはそれらの任意の組合せなどで、ユーザ１０２のために情報を出力するように構成される。たとえば、出力デバイス１２０は、以下でさらに説明されるように、制御ユニット１０４からメッセージデータ（たとえば、フィードバックメッセージ１４４）を受信することができ、ユーザ１０２への出力（たとえば、命令１４６）を生成することができる。特定の例では、出力デバイス１２０は、グラフィカルユーザインターフェースを表現するように構成されたディスプレイ、フィードバックメッセージ１４４をレンダリングするかまたはユーザ１０２にダイレクトするように構成された１つまたは複数のラウドスピーカー、あるいはそれらの組合せを含む。
【００２４】
[0037]制御ユニット１０４は、入力デバイス１１２～１１６からユーザ入力に対応するデータを受信し、出力デバイス１２０を介してユーザ１０２に提供されるべきフィードバックメッセージを生成するように構成される。制御ユニット１０４は、プロセッサ１０８と呼ばれる、１つまたは複数のプロセッサに結合されたメモリ１０６を含む。図２を参照しながらさらに説明されるように、メモリ１０６は、プロセッサ１０８による使用のためにアクセス可能な、１つまたは複数の埋め込みネットワークを表すデータと、組み合わされた埋め込み空間への埋め込みベクトルの１つまたは複数の変換を表すデータと、１つまたは複数の分類器を表すデータとを含むことができる。メモリ１０６はまた、マルチモーダル認識エンジン１３０、フィードバックメッセージ生成器１３２、またはそれらの両方を実装するためにプロセッサ１０８によって実行可能な命令を含むことができる。
【００２５】
[0038]プロセッサ１０８は、マルチモーダル認識エンジン１３０と、フィードバックメッセージ生成器１３２とを含む。いくつかの実装形態では、プロセッサ１０８は、マルチモーダル認識エンジン１３０とフィードバックメッセージ生成器１３２とを実装するための命令を実行するように構成された１つまたは複数の処理コアを含む。いくつかの実装形態では、プロセッサ１０８は、マルチモーダル認識エンジン１３０とフィードバックメッセージ生成器１３２との一方または両方を実装するように構成された専用回路を含む。一例では、プロセッサ１０８は、集積回路（ＩＣ）として実装される。
【００２６】
[0039]マルチモーダル認識エンジン１３０は、入力デバイス１１２～１１６のうちの１つまたは複数からデータを受信し、出力を生成するために受信データを処理するように構成される。たとえば、出力は、受信された入力に最も密接に一致するコマンドと、コマンドに関連する確信度（または尤度）インジケータとを含むことができる。いくつかの実装形態では、マルチモーダル認識エンジン１３０は、各入力モダリティについて、各入力モダリティの埋め込みベクトルを生成することなどによって、特定のトレーニングされたユーザ入力を他のトレーニングされたユーザ入力から区別するためのデータを生成するように構成される。マルチモーダル認識エンジン１３０は、ユニモーダルまたはマルチモーダルユーザ入力の一部として、入力デバイス１１２～１１６の各々を介して（もしあれば）どの認識されたユーザ入力が検出されたかを示す、組み合わされた埋め込みベクトルを生成するために、異なる入力モダリティに関連する埋め込みベクトルを組み合わせるように構成され得る。組み合わされた埋め込みベクトルは、組み合わされた埋め込みベクトルをコマンドにマッピングするようにトレーニングされた分類器を使用することなどによって、出力を決定するように処理される。マルチモーダル認識エンジン１３０において実装され得る構成要素の例示的な例については、図２に関して説明される。
【００２７】
[0040]フィードバックメッセージ生成器１３２は、出力デバイス１２０を介してユーザ１０２に出力されるべきフィードバックメッセージデータを生成するように構成される。たとえば、フィードバックメッセージ生成器１３２は、確信度レベルがしきい値を下回る

特定のコマンドであることが予測されるなど、適切に認識されなかったユーザ入力を繰り返すようにユーザ１０２に命令するために、フィードバックメッセージ１４４を出力デバイス１２０に送ることができる。他の例として、フィードバックメッセージ生成器１３２は、入力モダリティを変更するように、または１つの入力モダリティを使用して行われる入力を異なる入力モダリティを使用して行われる別の入力でオーグメントするようにユーザ１０２に命令するために、フィードバックメッセージ１４４を出力デバイス１２０に送ることができる。他の例は、ユーザ１０２がエミュレートするためのユーザ入力の記録サンプル、ユーザ１０２が識別するためのユーザの入力の記録サンプル、またはユーザ１０２がデバイス１１０を使用するのを支援するための他の情報を提供する、フィードバックメッセージデータを生成することを含む。例示的な例は、ユーザ１０２からの問合せを受信したことに応答して「アップ」に対応する動きを示すモーションビデオを表示すること、アップジェスチャーの動きに関連する最も類似している発話のオーディオ再生を生成すること、またはユーザ定義の動きにすでに密接に関連付けられている関係する発話のオーディオ再生を生成することなど、クロスモーダルサンプル取出しを含む。いくつかの例では、フィードバックメッセージ生成器１３２は、以下でより詳細に説明されるように、多因子認証プロセスに従って次の認証入力を提供するようにユーザ１０２に命令するために、フィードバックメッセージ１４４を生成するように構成される。

【００２８】

[0041]動作中に、ユーザ１０２は、第１の入力デバイス１１２によって検出された第１の入力モード（たとえば、バーバルコマンド）に基づいて、第１の入力１４０を提供する。第１の入力デバイス１１２は、第１の入力１４０を示す第１のデータ１４２を生成し、第１のデータ１４２を制御ユニット１０４に提供する。

【００２９】

[0042]プロセッサ１０８（たとえば、マルチモーダル認識エンジン１３０）は、第１の入力モード（たとえば、スピーチ）に基づくユーザ１０２からの第１の入力１４０を示す第１のデータ１４２を処理する。プロセッサ１０８（たとえば、フィードバックメッセージ生成器１３２）は、第１のデータ１４２の処理に基づいて出力デバイス１２０にフィードバックメッセージ１４４を送る。フィードバックメッセージ１４４は、たとえば、話された命令１４６のプレイアウトを介して、異なる入力モードを使用して第２の入力１４８を提供するようにユーザ１０２に命令する。第２の入力１４８は、第１の入力モードとは異なる第２の入力モード（たとえば、ビデオ）に基づき、マルチモーダル認識エンジン１３０が第１の入力１４０にどのように応答するかを更新するために使用され得る。本明細書で使用されるとき、異なる入力モードを使用することは、同じタイプの入力を使用するのではなく、異なるタイプの入力を使用することを意味する。各異なるタイプの入力は、様々な異なるセンサーを使用する。たとえば、スピーチ入力モードは、１つまたは複数のマイクロフォンを使用し得る。ジェスチャー入力モードは、動き検出を使用し得る。ビデオ入力モードは、カメラと、フレームのシーケンスとを使用し得る。概して、各入力モードは、その入力を提供するために使用され得る異なるタイプのセンサーを提供する。

【００３０】

[0043]いくつかの実装形態では、第１の入力１４０は、コマンドであり、フィードバックメッセージ１４４は、第１の入力１４０をディスアンビギュエートするための第２の入力１４８を提供するようにユーザ１０２に命令する。マルチモーダル認識エンジン１３０は、出力の不確実性（たとえば、話された入力が「アップ」を示すのか「オフ」を示すのかの不確実性）を示す、第１の入力１４０の認識処理に関連する確信度レベルが確信度しきい値を満たすのに失敗したことに応答して、フィードバックメッセージ１４４を送り得る。ユーザ１０２は、第２の入力１４８（たとえば、上方ポインティング）を提供し得、第２の入力１４８を示す第２のデータ１５０に基づいて、マルチモーダル認識エンジン１３０は、図２においてさらに詳細に説明されるように、第１の入力１４０（たとえば、スピーチ「アップ」）のマッピングを、第２の入力１４８に関連付けられた行為（たとえば、音楽ボリュームを上げる）に対して更新することができる。

【００３１】
[0044]別の実装形態では、マルチモーダル認識エンジン１３０は、第２の入力１４８と組み合わされた第１の入力１４０のマッピングを、第２の入力１４８に関連する行為に対して更新する。たとえば、雑音条件が、話された「アップ」コマンドの信頼できる認識を妨げるとき、マルチモーダル認識エンジン１３０は、ボリュームをアップするための単一のコマンドとして、ユーザの話された「アップ」コマンドと併せてユーザの「アップ」ビデオ入力（たとえば、上方ポインティング）を認識するように更新される。
【００３２】
[0045]したがって、いくつかの実装形態では、ユーザ１０２は、フィードバックメッセージ１４４のフィードバック機構と第２の入力１４８とを介して特定の行為を実施するためのコマンドとして特定の入力を認識するようにデバイス１１０を個人化することができる。例示のために、ユーザ１０２は、マルチモーダル認識エンジン１３０によって現在認識されないコマンド（第１の入力１４０）を話すことができ、フィードバックメッセージ１４４に応答して、ユーザ１０２は、認識されたコマンド（第２の入力１４８）を入力することによって、この認識されないコマンドにマッピングされるべき行為を識別することができる。同様に、デバイス１１０は、ユーザの選定されたモードが信頼できなくなったとき、入力モードを変更するようにユーザ１０２に命令することができる。たとえば、デバイス１１０が車両（たとえば、カーナビゲーションおよび／またはエンターテインメントシステム）中に実装されたとき、夜間運転中に、ユーザ１０２は、（低い照明条件により）ビデオの代わりにスピーチ入力またはジェスチャー入力を使用するように命令され得、ウィンドウが開いた状態で運転しているとき、ユーザ１０２は、（高い風雑音により）スピーチの代わりにジェスチャー入力またはビデオ入力を使用するように命令され得る。デバイス１１０が、仮想現実または拡張現実ヘッドセットなどのヘッドセット中に実装されたとき、入力モードを変更するようにユーザ１０２に命令するための同様の動作が実施され得る。
【００３３】
[0046]他の実装形態では、デバイス１１０は、多因子認証を実施するために使用される。たとえば、第１の入力１４０は、ユーザ１０２の第１の認証行為（たとえば、スピーカー検証のための話されたパスコード）に対応し得、フィードバックメッセージ１４４は、多因子認証手順の一部として、第２の認証行為として第２の入力１４８を提供する（たとえば、ユーザ１０２によって以前に選択された特定の手の構成を表示する）ようにユーザ１０２に命令する。デバイス１１０は、認証行為を実施するようにユーザ１０２に命令するための認証入力モードの数およびタイプをランダムにまたはアルゴリズム的に選択することができる。たとえば、デバイス１１０は、スピーチ入力（たとえば、第１の入力１４０）が、プレイアウトされている記録スピーチであり得るというインジケーションに応答して、命令１４６を生成することができ、カメラ（たとえば、第２の入力デバイス１１４）にウィンクするようにユーザ１０２に命令することなどによって、「ライブリネス」確認を要求し得る。
【００３４】
[0047]上記の例では、第２の入力１４８が第１の入力１４０とは異なるモードを使用することについて説明しているが、他の実装形態では、第２の入力１４８は、第１の入力１４０と同じモードを使用することができる。たとえば、第１の入力１４０の話されたコマンドは、解釈するのが困難であり得るが（たとえば、周辺雑音の存在下での「アップ」対「オフ」）、別の話されたコマンド（たとえば、「より大きく」）は、正しい行為（たとえば、ボリュームを上げる）を選択するために、他のマッピングされたコマンドとは十分に異なり得る。別の例として、トレーニングプロセス中に、ユーザ１０２は、トレーニングされていないスピーチコマンドとして「より大きく」発話し得、デバイス１１０は、「より大きく」という発話に関連付けられるべき行為を識別するように、命令１４６を介してユーザ１０２に命令し得る。ユーザ１０２は、ボリュームを上げるためのコマンドとしてデバイス１１０によって認識される第２の話された発話「アップ」を提供し得、マルチ

モーダル認識エンジン１３０は、「より大きく」を「ボリュームを上げる」行為にマッピングするように、ユーザ入力のマッピングを更新し得る。

【００３５】

[0048]図２は、特定の実装形態による、メモリ１０６と、マルチモーダル認識エンジン１３０と、プロセッサ１０８によって実行可能である１つまたは複数のアプリケーション２４０とを含む、制御ユニット１０４の構成要素の一例を示す。マルチモーダル認識エンジン１３０は、第１のユーザ入力（たとえば、スピーチ入力）を第１の埋め込みベクトル（たとえば、第１の埋め込みベクトル「Ｅ１」）にコンバートするように構成された第１の埋め込みネットワーク２０２を含む。第２の埋め込みネットワーク２０４は、第２のユーザ入力（たとえば、ジェスチャー入力）を第２の埋め込みベクトル（たとえば、第２の埋め込みベクトル「Ｅ２」）にコンバートするように構成される。マルチモーダル認識エンジン１３０は、第Ｎのユーザ入力（たとえば、ビデオ入力）を第Ｎの埋め込みベクトル（たとえば、第Ｎの埋め込みベクトル「Ｅｎ」）にコンバートするように構成された第Ｎの埋め込みネットワーク２０６を含む、１つまたは複数の追加の埋め込みネットワークを含み得る。マルチモーダル認識エンジン１３０は、本開示のいくつかの実施形態による任意の数の埋め込みネットワークを含み得る。

【００３６】

[0049]融合埋め込みネットワーク（fusion embedding network）２２０は、埋め込みネットワーク２０２～２０６の出力を組み合わせ、組み合わされた埋め込みベクトル「Ｃ」２２８など、組み合わされた埋め込みベクトルを生成するように構成される。たとえば、第１の変換２１２は、第１の共通の埋め込みベクトル２２２を生成するために、スピーチ埋め込みベクトルを「共通」の埋め込み空間にコンバートすることができる。第２の変換２１４は、第２の共通の埋め込みベクトル２２４を生成するために、ジェスチャー埋め込みベクトルを共通の埋め込み空間にコンバートすることができ、第Ｎの変換２１６は、第Ｎの共通の埋め込みベクトル２２６を生成するために、ビデオ埋め込みベクトルを共通の埋め込み空間にコンバートすることができる。共通の埋め込みベクトル２２２～２２６の各々は、それぞれ、対応する重みＷ１、Ｗ２、およびＷ３で重み付けされ、融合埋め込みネットワーク２２０において組み合わされ得る。マッピング２３０は、組み合わされた埋め込みベクトル２２８に対応する出力２３２と確信度レベル２３４とを選択するように構成される。たとえば、マッピング２３０は、組み合わされた埋め込みベクトルを特定の行為にマッピングするように構成された分類器２３１を含むことができる。例示のために、複数の埋め込みネットワーク２０２～２０６への組み合わされた入力から生じる出力２３２を決定するために、各モダリティ入力について個々の分類器を使用するのではなく、単一の分類器２３１が使用される。

【００３７】

[0050]マルチモーダル認識エンジン１３０によって使用される１つまたは複数のパラメータを示すデータは、メモリ１０６に記憶される。第１のユーザプロファイル２５０は、第１のユーザ（たとえば、ユーザ１０２）に関連付けられ、第１の埋め込みネットワークデータ２５２と、第１の重みデータ２５４と、第１の一時的調整データ２５６と、第１の履歴データ２５８とを含む。第１の埋め込みネットワークデータ２５２は、第１のユーザに対応すべき、第１の埋め込みネットワーク２０２と、第２の埋め込みネットワーク２０４と、第Ｎの埋め込みネットワーク２０６と、融合埋め込みネットワーク２２０とを含む、埋め込みネットワークを構成するためのデータ（たとえば、重みまたは他のパラメータもしくは値）を含む。第１の重みデータ２５４は、第１のユーザに対応すべき重み（たとえば、Ｗ１、Ｗ２、Ｗ３）を構成するための重み値を含む。第１の一時的調整データ２５６は、以下でさらに説明されるように、（たとえば、雑音の多い環境では重みＷ１を低減し、重みＷ２およびＷ３を増加させるための）一時的条件に基づいてマルチモーダル認識エンジン１３０の構成を調整するための値を含む。第１の履歴データ２５８は、第１のユーザに関連するヒストリカルデータを含み、マルチモーダル認識エンジン１３０によって処理される第１のユーザのマルチモーダル入力に対応する履歴傾向に基づいて、プロセッ

サ１０８が第１の埋め込みネットワークデータ２５２、第１の重みデータ２５４、またはそれらの両方を更新することを可能にする。

【００３８】

[0051]同様に、メモリ１０６は、第２のユーザのための第２の埋め込みネットワークデータ２６２と、第２の重みデータ２６４と、第２の一時的調整データ２６６と、第２の履歴データ２５８とを含む、第２のユーザに関連付けられた第２のユーザプロファイル２６０を含む。第１の埋め込みネットワークデータ２５２は、第１のユーザと第２のユーザとの間の入力コマンドの差に基づいて第２の埋め込みネットワークデータ２６２とは異なる。たとえば、第１のユーザと第２のユーザは、ビデオ入力を実施するときの異なるアクセント、異なるスタイルのジェスチャリング、異なる身体力学、またはそれらの任意の組合せを有し得る。第１の埋め込みネットワークデータ２５２は、第１のユーザのための埋め込みネットワークデータのデフォルトセットからのユーザ固有の変動を認識するように埋め込みネットワーク２０２～２０６および２２０をトレーニングした結果を表し得、第２の埋め込みネットワークデータ２６２は、第２のユーザのための埋め込みネットワークデータのデフォルトセットからのユーザ固有の変動を認識するように埋め込みネットワーク２０２～２０６および２２０をトレーニングした結果を表し得る。ただ２つのユーザプロファイル２５０、２６０が示されているが、デバイス１１０の複数のユーザのためのマルチモーダル認識エンジン１３０の動作をカスタマイズするために、任意の数のユーザプロファイルが含まれ得る。

【００３９】

[0052]異なるアクセント、ジェスチャースタイル、および身体力学など、個々のユーザ変動について調整することに加えて、第１の埋め込みネットワークデータ２５２はまた、第１のユーザによって決定されたユーザ入力の第１のカスタマイズされたセットを認識するように埋め込みネットワーク２０２～２０６および２２０をトレーニングした結果を表し得、第２の埋め込みネットワークデータ２６２はまた、第２のユーザによって決定されたユーザ入力の第２のカスタマイズされたセットを認識するように埋め込みネットワーク２０２～２０６および２２０をトレーニングした結果を表し得る。たとえば、第１のユーザは、オーディオ再生動作が進行中の間、スピーチコマンド「アップ」をボリュームを上げるためのコマンドとして認識するようにマルチモーダル認識エンジン１３０をカスタマイズ（たとえば、トレーニング）し得る。対照的に、第２のユーザは、オーディオ再生動作が進行中の間、スピーチコマンド「アップ」をプレイリスト上の前のオーディオトラックを選択するためのコマンドとして認識するようにマルチモーダル認識エンジン１３０をカスタマイズ（たとえば、トレーニング）し得る。

【００４０】

[0053]第１の重みデータ２５４は、第１のユーザと第２のユーザとの間の入力モード信頼性の差に基づいて第２の重みデータ２６４とは異なり得る。たとえば、プロセッサ１０８は、第１の履歴データ２５８などに基づいて、第１のユーザからのスピーチ入力が、第１のユーザからのジェスチャー入力と比較してあまり確実に解釈されないと決定し得る。その結果、第１のユーザからのスピーチ入力への依拠を低減し、ジェスチャー入力への依拠を増加させるために、第１の重みデータ２５４において、重みＷ１は、デフォルトＷ１値から低減され得、重みＷ２は、デフォルトＷ２値から増加され得る。対照的に、プロセッサ１０８は、第２の履歴データ２６８などに基づいて、第２のユーザからのスピーチ入力が、第２のユーザからのジェスチャー入力と比較してより確実であると決定し得る。その結果、第２のユーザからのジェスチャー入力への依拠を低減し、スピーチ入力への依拠を増加させるために、第２の重みデータ２６４において、重みＷ１は、デフォルトＷ１値から増加され得、重みＷ２は、デフォルトＷ２値から減少され得る。

【００４１】

[0054]アプリケーション２４０は、一時的調整器２９０と、データ調整器２９２とを含む。一時的調整器２９０は、一時的条件に基づいて、埋め込みネットワーク２０２、２０４、２０６、または２２０のうちの１つまたは複数の調整、重みＷ１～Ｗ３のうちの１つ

または複数の調整、あるいはそれらの組合せを決定するように構成される。たとえば、一時的調整器２９０は、検出された条件に基づいて、１つまたは複数の入力モダリティを強調するように、１つまたは複数の入力モダリティを強調しないように、あるいはそれらの組合せを行うように、重みＷ１～Ｗ３のうちの１つまたは複数を調整することができる。例示的および非限定的な例として、検出された条件は、以下でさらに詳細に説明されるように、周辺雑音データ２７２、周辺光データ２７４、ロケーションデータ２７６、またはユーザ選好２７８のうちの１つまたは複数によって示され得る。

【００４２】

[0055]データ調整器２９２は、一時的条件に基づかないと決定された変化を表すように埋め込みネットワークデータと重みデータとを更新するために、埋め込みネットワーク２０２、２０４、２０６、または２２０のうちの１つまたは複数の調整、重みＷ１～Ｗ３のうちの１つまたは複数の調整、あるいはそれらの組合せを決定するように構成される。いくつかの実装形態では、データ調整器２９２は、たとえば、マルチモーダル認識エンジン１３０がユーザ入力をより正確に認識する（たとえば、話されたコマンドのユーザの発音とデフォルトスピーチ認識モデルとの間の差に適応する）のを助けるユーザからのディスアンビギュエーションフィードバックを受信したことに応答して、または特定のコマンドへの入力のカスタムマッピングを示すユーザ入力（たとえば、ユーザが、以前不明であったビデオ入力として両手での「サムズアップ」ジェスチャーを入力し、このビデオ入力により、デバイス１１０がアラームをオフすべきであることを示す）に応答して、特定のコマンドへのユーザ入力の更新されたマッピングを示すために、埋め込みネットワーク２０２、２０４、２０６、または２２０のうちの１つまたは複数に対して更新トレーニングを実施するように構成される。

【００４３】

[0056]図１のシステム１００中に実装されたマルチモーダル認識エンジン１３０の動作の例示的な例では、ユーザ１０２は、顔認識、音声認識、または何らかの他の形態のユーザ認識などを介して、デバイス１１０へのマルチモーダル入力のソースとして識別される。ユーザ１０２からの入力を認識するようにマルチモーダル認識エンジン１３０を構成（たとえば、カスタマイズ）するために、埋め込みネットワーク２０２～２０６は、第１の埋め込みネットワークデータ２５２に基づいて更新され、重みＷ１、Ｗ２、およびＷ３は、第１の重みデータ２５４に基づいて更新され、いずれかの一時的調整は、第１の一時的調整データ２５６に基づいて適用される。

【００４４】

[0057]ユーザ１０２は、コマンドとして第１の入力１４０を提供する。第１の入力１４０は、十分な信頼性で何らかの特定のコマンドとして認識されず、フィードバックメッセージ１４４は、第１の入力１４０をディスアンビギュエートするための第２の入力１４８を提供するようにユーザ１０２に命令する。たとえば、フィードバックメッセージ１４４は、出力２３２の不確実性（たとえば、話された入力が「アップ」を示すのか「オフ」を示すのかの不確実性）を示す、第１の入力１４０の認識処理に関連する確信度レベル２３４が確信度しきい値２９４を満たすのに失敗したことに応答して送られ得る。他の実装形態では、フィードバックメッセージ１４４は、１つまたは複数の環境条件が検出されたことに応答して送られる。

【００４５】

[0058]たとえば、第１の入力１４０がビデオモードを介して受信される実装形態では、フィードバックメッセージ１４４は、照明しきい値２８６を下回る値を有する周辺光メトリック２８４に応答して送られる。たとえば、周辺光データ２７４は、デバイス１１０の１つまたは複数のセンサーを介して受信され、周辺光メトリック２８４を生成するために処理され得る。周辺光メトリック２８４は、周辺照明が、信頼できるビデオモード入力のためには薄暗すぎるかどうかを決定するために、照明しきい値２８６と比較され得る。フィードバックメッセージ１４４は、薄暗い照明がビデオ入力モードを信頼できないものにしていることをユーザに通知し得、別のモダリティ（たとえば、スピーチ）を使用して入

力を繰り返すようにユーザに命令し得る。
【００４６】
[0059]別の例として、第１の入力１４０がスピーチモードを介して受信される実装形態では、フィードバックメッセージ１４４は、雑音しきい値２８２を上回る値を有する雑音メトリック２８０（たとえば、信号対雑音比（ＳＮＲ）または周辺雑音測定値）に応答して送られる。たとえば、周辺雑音データ２７２は、デバイス１１０の１つまたは複数のセンサーを介して受信され（あるいはマイクロフォン入力信号の音声アクティビティ検出処理中に測定され）、雑音メトリック２８０を生成するために処理され得る。雑音メトリック２８０は、周辺雑音が、信頼できるスピーチモード入力のためには大きすぎるかどうかを決定するために、雑音しきい値２８２と比較され得る。フィードバックメッセージ１４４は、雑音環境がスピーチ入力モードを信頼できないものにしていることをユーザに通知し得、別のモダリティ（たとえば、ビデオ）を使用して入力を繰り返すようにユーザに命令し得る。
【００４７】
[0060]ユーザ１０２は、第２の入力１４８（たとえば、上方ポインティング）を提供し得、第２の入力１４８を示す第２のデータ１５０に基づいて、マルチモーダル認識エンジン１３０は、第１の入力１４０（たとえば、スピーチ「アップ」）のマッピングを、第２の入力１４８に関連付けられた行為（たとえば、音楽ボリュームを上げる）に対して更新することができる。例示のために、第１の埋め込みネットワーク２０２、第１の変換２１２、重みＷ１、融合埋め込みネットワーク２２０、またはマッピング２３０のうちの１つまたは複数は、マルチモーダル認識エンジン１３０が、音楽ボリュームを上げるためのコマンドとしてユーザの話された「アップ」をより正確に認識することを引き起こすように、データ調整器２９２によって調整され得る。
【００４８】
[0061]動作の例示的な例では、１つの入力モダリティが低精度条件を有すると決定された場合、（たとえば、一時的調整器２９０によって生成された一時的調整データに応答して）マルチモーダル認識エンジン１３０は、組み合わされた埋め込みベクトル２２８の生成のためにそのモダリティを使用する入力の影響を低減するかまたはなくすように１つまたは複数の設定を調整する。（たとえば、雑音しきい値２８２を超える雑音メトリック２８０により）スピーチモダリティが信頼できないと決定される、予測される、または推定される一方で、ジェスチャーおよびビデオモダリティが入力認識のために十分に信頼できると決定される例では、一時的調整器２９０は、スピーチ入力に関連する共通の埋め込みベクトル２２２に適用される重みＷ１を「０」値に設定し得る。ジェスチャー入力に関連する共通の埋め込みベクトル２２４に適用される重みＷ２と、ビデオ入力に関連する共通の埋め込みベクトル２２６に適用される重みＷ３とは、非０値に設定される（たとえば、ジェスチャー入力とビデオ入力が等しく信頼できるように扱われる実装形態では、Ｗ２＝Ｗ３＝０．５）。重みＷ１を「０」値に設定することにより、スピーチ入力が信頼できない状態である間、スピーチ入力が、得られた組み合わされた埋め込みベクトル２２８に影響を及ぼすのを防止する。
【００４９】
[0062]上記の例によれば、重みの初期設定は、各モダリティが入力認識について等しい重要性または信頼性を有することを示す、Ｗ１＝Ｗ２＝Ｗ３＝１／３を割り当て得る。スピーチモダリティは、（たとえば、雑音メトリック２８０が雑音しきい値２８２を超えるという検出、もしくは車両が動いている間に車両ウィンドウが開いているという検出を介した）大量の周辺雑音の検出により、またはスピーチ入力のしきい値数が所定の時間期間中に正確に認識されることに失敗することなどにより、信頼できないと後で決定または予測され得る。スピーチモダリティが信頼できないと決定または予測されたことに応答して、一時的調整器２９０は、入力認識に対するスピーチ入力の影響を除去するために、重みＷ１、Ｗ２、およびＷ３を、それぞれ０、１／２、および１／２に調整する。スピーチ入力モダリティがもはや信頼できなくはないという後続の決定（たとえば、風雑音が雑音し

きい値を下回るか、ウィンドウが閉じられるか、または車両が移動するのを止めた）に応答して、重みＷ１、Ｗ２、およびＷ３は、１／３のそれらの初期値にそれぞれ戻され得る。

【００５０】

[0063]別の例として、代わりに、ビデオモダリティが、周辺光の低い量の検出（たとえば、周辺光メトリック２８４が照明しきい値２８６を下回る）により、またはビデオ入力のしきい値数が所定の時間期間中に正確に認識されることに失敗することなどにより、信頼できないと決定または予測され得る。ビデオモダリティが信頼できないと決定または予測されたことに応答して、一時的調整器２９０は、入力認識に対するビデオ入力の影響を除去するために、重みＷ１、Ｗ２、およびＷ３を、それぞれ１／２、１／２、および０に調整する。ビデオ入力モダリティがもはや信頼できなくはないという後続の決定（たとえば、周辺光が照明しきい値を超えることを決定される）に応答して、重みＷ１、Ｗ２、およびＷ３は、１／３のそれらの初期値にそれぞれ戻され得る。

【００５１】

[0064]いくつかの実装形態では、複数の重みは、入力認識に対する複数の入力モダリティのインパクトを低減または除去するように調整される。たとえば、スピーチモダリティのみが使用されるべきであるという決定が行われる実装形態では、Ｗ１は「１」に設定され、Ｗ２とＷ３とは「０」に設定される。例示のために、デバイス１１０は、低い周辺照明条件を検出し得、また、アクティブなジェスチャー検出デバイスが検出されない（たとえば、ユーザのスマートウォッチが存在しないかまたは動きデータを送信していない）と決定し得る。別の例として、ユーザ１０２は、スピーチ入力のみを処理するように入力認識を制限するように、ユーザ選好２７８を入力することなどを介して、デバイス１１０に命令し得る。別の例として、１つまたは複数の入力モダリティを制限すべきかどうかを決定するために、ロケーションデータ２７６が使用され得る。たとえば、ユーザが車両を操作していることを示すロケーションデータ２７６に応答して、一時的調整器２９０は、ユーザの注意散漫を防ぐために、および車両の安全な操作を奨励するためになど、ジェスチャー入力とビデオ入力との認識を防止するようにユーザ入力モードを制限し得る。ユーザがもはや車両を操作しておらず、ユーザの自宅にいることを示すロケーションデータ２７６に応答して、一時的調整器２９０は、ジェスチャー入力とビデオ入力との認識を有効にするようにユーザ入力モードを復元し得る。

【００５２】

[0065]上記の例では重み値の例について説明されているが、そのような例示的な重み値は、例示的であり、限定的ではない。例示のために、重みを「０」に設定するのではなく、重みは、全体的な入力認識に対する関連する入力モダリティの影響を減少させるがなくしはしない、低減された値に設定され得る。別の例として、「信頼できる」入力モダリティは、入力モダリティの相対的信頼性を示し得る、等しくない重みを有し得る。例示のために、ジェスチャー入力が十分に信頼できると見なされ、ビデオ入力がジェスチャー入力よりも信頼できると見なされ、スピーチが信頼できないと決定された場合、重みは、Ｗ１＝０．１、Ｗ２＝０．４、およびＷ３＝０．５などの値に設定され得る。上記の例では、重みＷ１、Ｗ２、およびＷ３の和は１に等しいが、他の実装形態では、重みＷ１、Ｗ２、およびＷ３の和は、どんな特定の値にも制限されない。

【００５３】

[0066]信頼できないと決定された入力モダリティの影響を低減するかまたはなくすように１つまたは複数の重みを調整することの追加または代替として、いくつかの実装形態では、マルチモーダル認識エンジン１３０は、関連する埋め込みネットワークの出力を、利用可能なスキルの中から「なし」出力に強制するか、変換の出力を、「０」値を有する埋め込みベクトルへの「なし」カテゴリー入力のために共通の埋め込み空間に強制するか、またはそれらの組合せを行い得る。

【００５４】

[0067]いくつかの実装形態では、マルチモーダル認識エンジン１３０を含むデバイス１１０は、複数の入力モダリティの環境アウェア融合を実施する。たとえば、ユーザ１０２

が車を運転していると決定したことに応答して、ジェスチャー入力に関連する重みＷ２は、車を運転している間に安全でない手の動きを阻止するために、ユーザの手の動きがジェスチャー入力としてよりもむしろ車の操作に対応する可能性があることを示す、「０」に設定され得る。別の例として、ユーザ１０２が暗い部屋の中にいると決定したことに応答して、ビデオ入力に関連する重みＷ３は、「０」に設定され得る。別の例として、ユーザ１０２が雑音の多い環境の中にいると決定したことに応答して、スピーチ入力に関連する重みＷ１は、「０」に設定され得る。環境条件の決定は、デバイス１１０に組み込まれた１つまたは複数のセンサー（たとえば、周辺光センサー、周辺雑音センサー）、（たとえば、デバイス１１０と、ホームオートメーションシステム、モノのインターネットシステム、または別のシステムの１つまたは複数の構成要素との間の通信を介した）デバイス１１０の外部にある１つまたは複数のセンサー、あるいはそれらの任意の組合せに基づくことができる。

【００５５】

[0068]図３は、ヘッドセット３０２を着用しているユーザが、スマートフォンなどの別のデバイス、車などの車両システム、またはワイヤレスデジタルアシスタントアプリケーションを組み込んでいるスピーカーシステム（たとえば、「スマートスピーカー」）と通信している、マルチモーダルユーザ入力のためのシステム３００の一例を示す。ヘッドセット３０２は、図１のデバイス１１０に対応することができ、拡張現実（「ＡＲ」）、仮想現実（「ＶＲ」）、または複合現実（「ＭＲ」）オーディオおよびビデオ出力を着用者に提供するために、ディスプレイと、イヤバッド３０８または他のウェアラブル雑音生成デバイスなどのトランスデューサとを含むことができる。

【００５６】

[0069]ヘッドセット３０２は、ユーザ入力を検出するために、１つまたは複数のマイクロフォン、１つまたは複数のカメラなど、複数のセンサーを含むことができる。たとえば、１つまたは複数のマイクロフォンを介して受信されたオーディオ入力は、ヘッドセット３０２に組み込まれたかまたはそれに結合されたプロセッサにおいて１つまたは複数の動作３１０を実施するために使用され得る。たとえば、音環境分類を可能にするための機械学習、ヘッドセット３０２の着用者がいつ話しているかを決定するための自己音声の音声アクティビティ検出（ＶＡＤ）、音響イベント検出、およびモード制御（たとえば、シーケンスベースのユーザインターフェース）を使用することなど、オーディオ入力に対応するオーディオ信号を処理することが実施され得る。

【００５７】

[0070]１つまたは複数の動作３１０の結果は、１つまたは複数の行為３１２を生成するために使用され得る。たとえば、行為３１２は、アクティブ雑音消去（ＡＮＣ）フィルタをチューニングすること、１つまたは複数の支援的リスニング特徴を実装すること、マルチマイクロフォン音キャプチャのフィールドを調整すること（たとえば、「ＡｕｄｉｏＺｏｏｍ」）、あるいは拡張現実レンダリング、仮想現実レンダリング、または複合現実レンダリング（まとめて「ＸＲ」レンダリングと呼ばれる）を実施することを含むことができる。たとえば、結果は、空間透過モードでヘッドセット３０２にレンダリングされ得る。

【００５８】

[0071]ヘッドセット３０２において（たとえば、１つまたは複数のマイクロフォン、動き検出器、ジェスチャー検出器、カメラなどを介して）検出されたユーザ入力は、自動音声認識および自然言語処理、探索もしくは問合せ応答、またはそれらの両方など、１つまたは複数のスピーチベースの動作３０４の実施を開始するために使用され得る。１つまたは複数のスピーチベースの動作３０４は、ヘッドセット３０２と通信しているスマートフォンまたは他のポータブル通信デバイスなどにおいて、機械学習を使用して実施され得る。データ通信３０５（たとえば、ワイヤレスネットワーク通信、ワイヤライン通信、またはそれらの両方）は、外部処理リソース３０６（たとえば、機械学習を組み込んでいるクラウドベースＡＳＲ／ＮＬＰおよび探索サーバ）にオーディオスピーチデータを送ることを含み得る。探索および問合せ結果は、ヘッドセット３０２を介してユーザに返信され得

る。

【００５９】

[0072]図４は、例示的および非限定的な例では図３のヘッドセット３０２などによって実施され得る、マルチマイクロフォン音キャプチャのフィールドを調整すること（たとえば、「ＡｕｄｉｏＺｏｏｍ」）の例４００を示す。代表的なマイクロフォン４１２、４１４、および４１６など、複数のマイクロフォンが、ユーザの周りに配置される。ユーザは、極座標系の中心におり、０度角度方向を向くように配向されるものとして示されている。マイクロフォン４１２、４１４、および４１６は、指向性マイクロフォン、無指向性マイクロフォン、またはそれらの両方を含み、ユーザの周囲のオーディオ環境をキャプチャすることができる。第１の構成４０２では、マイクロフォン４１２～４１６からのオーディオの音処理は、ユーザ指示調整なしのオーディオ環境を表す、（たとえば、イヤフォンまたはイヤバッドを介した）ユーザへの可聴出力を生じる。

【００６０】

[0073]第２の構成４０４では、マルチモーダルインターフェース（たとえば、例示的な例として、ユーザジェスチャー、発話、ビデオ入力、またはそれらの組合せ）を介したユーザ入力に応答して、マイクロフォン４１２～４１６からのオーディオの音処理は、特定の空間領域４２０（たとえば、９０度角度方向の、またはユーザの左側の領域）から生起または到着する音を強調（たとえば、増幅）する一方で、空間領域４２０外のエリアから生起する音を減衰させるように調整される。第２の構成４０４に遷移することを生じるユーザ入力の例は、例示的および非限定的な例として、スピーチモダリティに基づく「左にズームする」スピーチシーケンス、ジェスチャーモダリティに基づく「手を左側にポインティングする」または「指を左側にポインティングする」ジェスチャーシーケンス、あるいはオーディオ（非スピーチ）モダリティに基づく「スナップ音を起こす」オーディオシーケンスを含むことができる。

【００６１】

[0074]いくつかの実装形態では、図１～図４を参照しながら上記で説明されたマルチモーダルインターフェースは、ユーザの近傍にあるロケーションまたはアクティビティ（たとえば、リビングルームでテレビジョンを見ること、またはキッチンで皿を洗うこと）などのコンテキストに応答する。たとえば、ウォッチまたはアームバンドベースの加速度計を使用してキャプチャされたジェスチャーは、検出されたコンテキストに基づいて解釈され得る。たとえば、手を振ることは、ターゲットコマンド「ライトをオンにする」として解釈され得、手を左側に反転させることは、「次の曲」または「次のチャンネル」として解釈され得、手を右側に反転させることは、「前の曲」、「前のチャンネル」、または「ドアオープン」として解釈され得る。例示的および非限定的な例として、閉じられた拳が形成される「グラブ」ジェスチャーは、「電話を取る」または「チャンネルを選択する」として解釈され得、長いグラブは、「曲を止める」、「アラームをキャンセルする」、または「ドアクローズ」として解釈され得、指を伸ばしている手の反時計回りの回転は、「ホームデバイスを発見する」として解釈され得る。コンテキストは、検出された音響イベント／環境シーケンスとの関連付けを介して決定され得る。たとえば、様々な音響イベントは、音響環境（たとえば、ユーザがどこにいるか）を推論するために、または適切なフィードバックタイミングを監視するために検出され得る。そのような検出可能な音響イベントの例は、ヘアドライヤー、掃除機、音楽、キッチンフード、料理、食事、皿の洗浄、屋内空調、電子レンジ、洗濯機、乾燥機、シャワー、およびテレビジョンを見ることを含む。

【００６２】

[0075]手のジェスチャー認識のためのデータセットは、手の移動を示す（たとえば、ｘ、ｙおよびｚ軸に沿った）３次元（３Ｄ）加速度計およびジャイロスコープセンサーデータを含むことができる。（たとえば、加速度計とジャイロスコープとからの）センサー信号の各成分は、３秒のウィンドウ（たとえば、１５０の読取り／ウィンドウ）など、固定幅のウィンドウであり得る。例示的および非限定的な例として、次、前、アップ／増加、

ダウン／減少、オン、オフ、および不明など、複数のジェスチャークラスが実装され得る。置換、時間ワーピング、スケーリング、大きさワーピング、ジッタ、およびクロッピングなど、１つまたは複数のデータオーグメンテーション技法が実装され得る。

【００６３】

[0076]手のジェスチャー認識のデータセットの統計的特徴などに基づく、特徴抽出が実施され得る。例示のために、抽出された特徴は、例示的および非限定的な例として、最小、最大、分散、平均、標準偏差、ＭＳＥ（最小２乗誤差）、ＡＣＦ（自己相関）、ＡＣＶ（自己共分散）、ゆがみ、尖度、平均交差率、ジッタ、または３分位数に対応することができる。

【００６４】

[0077]サポートベクターマシン（ＳＶＭ）、勾配ブースティング、分類器、積層長短期記憶リカレントニューラルネットワーク（ＬＳＴＭ－ＲＮＮ）、シーケンスツーシーケンスエンコーダデコーダモデルウィズアテンション、１つまたは複数の他のモデル、あるいはそれらの任意の組合せなど、１つまたは複数のモデルが手のジェスチャー認識のために使用され得る。

【００６５】

[0078]いくつかの態様では、マルチモーダル認識エンジン１３０は、ターゲット行為に直接マッピングされたシーケンス埋め込みベクトルを生成することを学習またはトレーニングすることができる。入力シーケンスの例は、（たとえば、ジェスチャー入力のための）加速度計もしくはジャイロスコープ時系列、スピーチコマンド時系列、またはオーディオ時系列を含む。エンコーダデコーダＬＳＴＭ－ＲＮＮウィズアテンションは、入力シーケンスに関連するターゲット行為クラスを示すためのソフトマックスレイヤへの出力を生成するためになど、可変長時系列信号を固定長および弁別ベクトルとして表す埋め込みベクトルを生成することを学習するために使用され得る。

【００６６】

[0079]いくつかの態様では、マルチモーダル認識エンジン１３０は、異なる行為クラスの登録と設計とのために埋め込みベクトルを使用することができる。たとえば、いくつかの異なる入力シーケンスが登録され得、１つまたは複数の分類器は、各ターゲット行為にマッピングされた埋め込みベクトルを使用して設計され得る。たとえば、埋め込みをターゲット行為にマッピングするために、ＳＶＭ、Ｋ平均、ｋ近傍法（ＫＮＮ）、コサイン（ｃｏｓ）距離、または他の設計が実装され得る。更新されたシステムの精度を検証するために、ユーザシーケンスのテストが実施され得る。

【００６７】

[0080]いくつかの態様では、登録およびＳＶＭ／Ｋ平均／ＫＮＮ設計の後に、分類器評価に関連するメトリックは、クラス間の分離があまりにあいまいであり、シーケンス整形が実施され得ることを示す。そのような場合、フィードバックメッセージ生成器１３２は、他のクラスとの混同を引き起こすいくつかの問題があるシーケンスをユーザに示すためのフィードバックを生成することができる。たとえば、混同されたクラスの動き、オーディオ、またはスピーチシーケンスは、出力デバイス１２０などを介して、ユーザに再生され得る。ユーザは、どのシーケンスが混同を引き起こすかを了解することができ、ターゲットクラス間の分離を改善しディスアンビギュエーションを提供するために新しいシーケンスを発話する／ジェスチャーで示すことができる。代替的に、混同を招く入力シーケンスは、入力シーケンス間のあいまいさが未決定にレンダリングされるように、ユーザによって、マルチモーダルユーザインターフェースを介して、同じ行為／クラスに一致させられ得る。ユーザフィードバックを受信した後に、マルチモーダル認識エンジン１３０は、ＳＶＭ／Ｋ平均／ＫＮＮ設計を再登録および修正することができ、フィードバックメッセージ生成器１３２は、混同がある場合に、シーケンスマッピングが互いに十分に別個になるまで、入力シーケンスを繰り返すようにユーザに再プロンプトすることができる。たとえば、「混同行列」は、異なるシーケンス間のあいまいさの量を表すことができ、トレーニングは、混同行列が準対角になるまで繰り返され得る。

【００６８】
[0081]いくつかの態様では、他の行為クラスとの混同を引き起こす「問題がある」入力シーケンスを検出したことに応答して、マルチモーダル入力を用いたシーケンス整形が実施され得る。デバイス１１０は、ユーザが、問題がある入力シーケンスの各々のためにマルチモーダル入力を使用することを望むかどうかをユーザに要求することができる。たとえば、「オフ」および「ボリュームダウン」のためのユーザの特定のジェスチャーが、マルチモーダル認識エンジン１３０にとって区別するのが困難である場合、出力デバイス１２０は、「あなたは、『オフ』カテゴリーのためにバーバルコマンド『オフにする』を使用したいですか？」という問合せをユーザに出力し得る。別の例として、出力デバイス１２０は、「あなたは、『ボリュームダウン』カテゴリーのためにバーバルコマンド『ボリュームを下げる』を使用したいですか？」という問合せを出力し得る。ユーザが（たとえば、ジェスチャー混同によりバーバルコマンドを追加するために）マルチモーダル入力を使用することを選択したことに応答して、マルチモーダルキューがアクティブにされ得、デバイス１１０は、マルチモーダル入力シーケンスを使用した混同の確率を含めるように混同行列を調整することができる。

【００６９】
[0082]いくつかの態様では、入力シーケンスをディスアンビギュエートするために、対話型連続検証が使用され得る。たとえば、ユーザは、どのカテゴリーがどのマルチモーダル入力に登録されたかを忘れることがある。ユーザとデバイス１１０との間でダイアログベースの対話が行われ得る。たとえば、ジェスチャー入力が「オフ」カテゴリーとして検出された場合、出力デバイス１２０は、「あなたは、『オフ』カテゴリーまたは『次』カテゴリーを意図していますか？」をユーザに問い合わせ得る。ユーザは、「オフ」と答えることがあり、マルチモーダル認識エンジン１３０は、「オフ」コマンドをアクティブにし得る。

【００７０】
[0083]図５は、図１３に関してさらに説明されるように、半導体チップまたはパッケージなどの個別構成要素に組み込まれたマルチモーダル認識エンジン１３０とフィードバックメッセージ生成器１３２とを含む、デバイス５０２の実装形態５００を示す。例示のために、デバイス５０２は、マルチモーダル認識エンジン１３０とフィードバックメッセージ生成器１３２とに関して説明される動作を実施するために、記憶された命令を実行するように構成された１つまたは複数のプロセッサ（たとえば、プロセッサ１０８）を含むことができる。デバイス５０２は、図１の入力デバイス１１２～１１６のうちの１つまたは複数からのデータなど、センサーデータ５０４がデバイス５０２の外部の１つまたは複数のセンサーから受信されることを可能にするために、第１のバスインターフェースなどのセンサーデータ入力５１０を含む。デバイス５０２はまた、（たとえば、出力デバイス１２０に）フィードバックメッセージ１４４を送ることを可能にするために、第２のバスインターフェースなどの出力５１２を含む。デバイス５０２は、図１１に示されている車両、図１２Ａに示されている仮想現実もしくは拡張現実ヘッドセット、図１２Ｂに示されているウェアラブル電子デバイス、または図１３に示されているワイヤレス通信デバイスなどの中に、複数のセンサーと出力デバイスとを含むシステム中の構成要素として、マルチモーダルユーザインターフェース処理の実装を可能にする。

【００７１】
[0084]図６を参照すると、例示的および非限定的な例として、図１のデバイス１１０もしくは制御ユニット１０４、図５のデバイス５０２、またはそれらの両方によって実施され得る、マルチモーダルユーザ入力を処理する方法６００の特定の実装形態が示されている。

【００７２】
[0085]方法６００は、６０２において、デバイスのプロセッサにおいて、第１の入力デバイスから受信された第１のデータを処理することを含む。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示す。たとえば、図１を参照すると、プロセッ

サ１０８は、第１の入力デバイス１１２から受信された第１のデータ１４２を処理する。第１のデータ１４２は、第１の入力モードに基づくユーザ１０２からの第１の入力１４０を示す。

【００７３】

[0086]方法６００はまた、６０４において、デバイスのプロセッサから、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることを含む。フィードバックメッセージは、第１の入力モードとは異なる第２の入力モードに基づく第２の入力を提供するようにユーザに命令する。たとえば、図１を参照すると、制御ユニット１０４は、第１のデータ１４２を処理することに基づいて出力デバイス１２０にフィードバックメッセージ１４４を送る。フィードバックメッセージ１４４は、第２の入力モードに基づく第２の入力１４８を提供するようにユーザ１０２に命令する。

【００７４】

[0087]方法６００は、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）デバイス、特定用途向け集積回路（ＡＳＩＣ）、中央処理ユニット（ＣＰＵ）などの処理ユニット、デジタル信号プロセッサ（ＤＳＰ）、コントローラ、別のハードウェアデバイス、ファームウェアデバイス、またはそれらの任意の組合せによって実装され得る。一例として、方法６００は、本明細書で説明されるように、命令を実行するプロセッサによって実施され得る。

【００７５】

[0088]図７を参照すると、例示的および非限定的な例として、図１の制御ユニット１０４、図５のデバイス５０２、またはそれらの両方によって実施され得る、マルチモーダルユーザ入力を処理する方法７００の特定の実装形態が示されている。

【００７６】

[0089]方法７００は、７０２において、第１の入力デバイスから受信された第１のデータを処理することを含む。第１のデータは、第１の入力モードに基づくユーザからのコマンドに対応する第１の入力を示す。たとえば、図１を参照すると、プロセッサ１０８は、第１の入力デバイス１１２から受信された第１のデータ１４２を処理する。第１のデータ１４２は、第１の入力モードに基づくユーザ１０２からのコマンドに対応する第１の入力１４０を示す。

【００７７】

[0090]方法７００はまた、７０４において、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることを含む。フィードバックメッセージは、第１の入力をディスアンビギュエートする（disambiguate）ために、第１の入力モードとは異なる第２の入力モードに基づく第２の入力を提供するようにユーザに命令する。たとえば、図１を参照すると、制御ユニット１０４は、第１のデータ１４２を処理することに基づいて出力デバイス１２０にフィードバックメッセージ１４４を送る。フィードバックメッセージ１４４は、第１の入力１４０をディスアンビギュエートするために、第１の入力モードとは異なる第２の入力モードに基づく第２の入力１４８を提供するようにユーザ１０２に命令する。

【００７８】

[0091]方法７００は、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）デバイス、特定用途向け集積回路（ＡＳＩＣ）、中央処理ユニット（ＣＰＵ）などの処理ユニット、ＤＳＰ、コントローラ、別のハードウェアデバイス、ファームウェアデバイス、またはそれらの任意の組合せによって実装され得る。一例として、方法７００は、本明細書で説明されるように、命令を実行するプロセッサによって実施され得る。

【００７９】

[0092]図８を参照すると、例示的および非限定的な例として、図１の制御ユニット１０４、図５のデバイス５０２、またはそれらの両方によって実施され得る、マルチモーダルユーザ入力を処理する方法８００の特定の実装形態が示されている。

【００８０】

[0093]方法８００は、８０２において、第１の入力デバイスから受信された第１のデータを処理することを含む。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、第１のデータは、ユーザの第１の認証行為（authentication action）に対応する。たとえば、図１を参照すると、プロセッサ１０８は、第１の入力デバイス１１２から受信された第１のデータ１４２を処理する。第１のデータ１４２は、第１の入力モードに基づくユーザ１０２からの第１の入力１４０を示し、第１のデータ１４２は、ユーザ１０２の第１の認証行為に対応する。

【００８１】

[0094]方法８００はまた、８０４において、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることを含む。フィードバックメッセージは、多因子認証手順（multi-factor authentication procedure）の一部として、第２の認証行為として、第１の入力モードとは異なる第２の入力モードに基づいて、第２の入力を提供するようにユーザに命令する。たとえば、図１を参照すると、制御ユニット１０４は、第１のデータ１４２を処理することに基づいて出力デバイス１２０にフィードバックメッセージ１４４を送る。フィードバックメッセージ１４４は、多因子認証手順の一部として、第２の認証行為として、異なるフォーム第１の入力モードである第２の入力モードに基づいて、第２の入力１４８を提供するようにユーザ１０２に命令する。

【００８２】

[0095]方法８００は、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）デバイス、特定用途向け集積回路（ＡＳＩＣ）、中央処理ユニット（ＣＰＵ）などの処理ユニット、ＤＳＰ、コントローラ、別のハードウェアデバイス、ファームウェアデバイス、またはそれらの任意の組合せによって実装され得る。一例として、方法８００は、本明細書で説明されるように、命令を実行するプロセッサによって実施され得る。

【００８３】

[0096]図９を参照すると、例示的および非限定的な例として、図１の制御ユニット１０４、図５のデバイス５０２、またはそれらの両方によって実施され得る、マルチモーダルユーザ入力を処理する方法９００の特定の実装形態が示されている。

【００８４】

[0097]方法９００は、９０２において、第１の入力モードに基づく第１のユーザ入力を検出することを含む。たとえば、図１を参照すると、第１の入力デバイス１１２は、第１の入力モードに基づく第１のユーザ入力１４０を検出する。

【００８５】

[0098]方法９００はまた、９０４において、第２の入力モードに基づく第２のユーザ入力を検出することを含む。たとえば、図１を参照すると、第２の入力デバイス１１４は、第２の入力モードに基づく第２のユーザ入力１４８を検出する。

【００８６】

[0099]方法９００はまた、９０６において、第１のユーザ入力を第１の埋め込みベクトルにコンバートするように構成された第１の埋め込みネットワークを使用して、第１の埋め込みベクトルを生成することを含む。たとえば、図２を参照すると、第１の埋め込みネットワーク２０２は、第１のユーザ入力を第１の埋め込みベクトルにコンバートすることによって、第１の埋め込みベクトルを生成する。

【００８７】

[0100]方法９００はまた、９０８において、第２のユーザ入力を第２の埋め込みベクトルにコンバートするように構成された第２の埋め込みネットワークを使用して、第２の埋め込みベクトルを生成することを含む。たとえば、図２を参照すると、第２の埋め込みネットワーク２０４は、第２のユーザ入力を第２の埋め込みベクトルにコンバートすることによって、第２の埋め込みベクトルを生成する。

【００８８】

[0101]方法９００はまた、９１０において、組み合わされた埋め込みベクトルを生成するために、第１の埋め込みネットワークと第２の埋め込みネットワークとの出力を組み合

わせるように構成された融合埋め込みネットワークを使用して、組み合わされた埋め込みベクトルを生成することを含む。たとえば、図２を参照すると、融合埋め込みネットワーク２２０は、組み合わされた埋め込みベクトルを生成するために、第１の埋め込みネットワーク２０２と第２の埋め込みネットワーク２０４との出力を組み合わせる。

【００８９】

[0102]方法９００はまた、９１２において、分類器を使用して、組み合わされた埋め込みベクトルを特定の行為にマッピングすることを含む。たとえば、図２を参照すると、マッピング２３０は、組み合わされた埋め込みベクトルを特定の行為にマッピングする。

【００９０】

[0103]方法９００は、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）デバイス、特定用途向け集積回路（ＡＳＩＣ）、中央処理ユニット（ＣＰＵ）などの処理ユニット、ＤＳＰ、コントローラ、別のハードウェアデバイス、ファームウェアデバイス、またはそれらの任意の組合せによって実装され得る。一例として、方法９００は、本明細書で説明されるように、命令を実行するプロセッサによって実施され得る。

【００９１】

[0104]図１０を参照すると、例示的および非限定的な例として、図１の制御ユニット１０４、図５のデバイス５０２、またはそれらの両方によって実施され得る、マルチモーダルユーザ入力を処理する方法１０００の特定の実装形態が示されている。

【００９２】

[0105]方法１０００は、１００２において、第１の入力デバイスから受信された第１のデータを処理することを含む。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、第１の入力は、コマンドに対応する。たとえば、図１を参照すると、プロセッサ１０８は、第１の入力デバイス１１２から受信された第１のデータ１４２を処理する。第１のデータ１４２は、第１の入力モードに基づくユーザ１０２からの第１の入力１４０を示す。

【００９３】

[0106]方法１０００はまた、１００４において、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることを含む。フィードバックメッセージは、第１の入力モードとは異なる第２の入力モードに基づいて、第１の入力に関連するコマンドを識別する第２の入力を提供するようにユーザに命令する。たとえば、図１を参照すると、制御ユニット１０４は、第１のデータ１４２を処理することに基づいて出力デバイス１２０にフィードバックメッセージ１４４を送る。フィードバックメッセージ１４４は、別の入力モードを使用して、第１の入力１４０に関連するコマンドを再び入力するようにユーザ１０２に命令する。一例では、第１の入力モードは、スピーチモード、ジェスチャーモード、またはビデオモードのうちの１つであり、第２の入力モードは、スピーチモード、ジェスチャーモード、またはビデオモードのうちの異なる１つである。

【００９４】

[0107]いくつかの実装形態では、フィードバックメッセージは、第１の入力をディスアンビギュエートするために第２の入力を提供するようにユーザに命令する。たとえば、フィードバックメッセージは、確信度レベル２３４が確信度しきい値２９４よりも小さいときなど、第１の入力の認識処理に関連する確信度レベルが確信度しきい値を満たすことに失敗したことに応答して送られ得る。いくつかの例では、第１の入力モードは、ビデオモードに対応し、フィードバックメッセージは、照明しきい値２８６よりも小さい値を有する周辺光メトリック２８４など、照明しきい値を下回る値を有する周辺光メトリックに応答して送られる。他の例では、第１の入力モードは、スピーチモードに対応し、フィードバックメッセージは、雑音しきい値２８２よりも大きい値を有する雑音メトリック２８０など、雑音しきい値を超える値を有する雑音メトリックに応答して送られる。

【００９５】

[0108]方法１０００はまた、１００６において、第２の入力デバイスから第２のデータを受信することを含み、第２のデータは、第２の入力を示す。たとえば、図１を参照する

と、制御ユニット１０４は、第１の入力１４０に関連する特定のコマンドを識別する、第２の入力１４８に関連する第２のデータ１５０を受信する。

【００９６】

[0109]方法１０００はまた、１００８において、第２の入力によって識別されるコマンドに第１の入力を関連付けるようにマッピングを更新することを含む。たとえば、図１を参照すると、制御ユニット１０４は、特定のコマンドに第１の入力１４０を関連付けるように、コマンドへのユーザ入力のマッピングを更新する。いくつかの実装形態では、更新されたマッピングは、コマンドが、第１の入力モードを介した第１の入力と第２の入力とモードを介した第２の入力とのコンカレントな（たとえば、少なくとも部分的に時間的に重複している）受信を介してより確実に認識されるように、ユーザのためにコマンドへの入力のマッピングをカスタマイズするためになど、第１の入力と第２の入力との組合せをコマンドに関連付ける。いくつかの実装形態では、マッピングを更新することは、ユーザに関連する埋め込みネットワークデータ（たとえば、第１の埋め込みネットワークデータ２５２）を更新すること、またはユーザに関連する重みデータ（たとえば、第１の重みデータ２５４）を更新することのうちの少なくとも１つを含む。

【００９７】

[0110]方法１０００は、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）デバイス、特定用途向け集積回路（ＡＳＩＣ）、中央処理ユニット（ＣＰＵ）などの処理ユニット、ＤＳＰ、コントローラ、別のハードウェアデバイス、ファームウェアデバイス、またはそれらの任意の組合せによって実装され得る。一例として、方法１０００は、本明細書で説明されるように、命令を実行するプロセッサによって実施され得る。

【００９８】

[0111]図１１は、車ダッシュボードデバイス１１０２などの車両ダッシュボードデバイスに組み込まれたデバイス１１０の実装形態１１００の一例を示す。複数のセンサー１１５０は、１つまたは複数のマイクロフォン、カメラ、または他のセンサーを含むことができ、図１の入力デバイス１１２～１１６に対応することができる。単一のロケーションに示されているが、他の実装形態では、車両操作者からおよび各搭乗者からのマルチモーダル入力を検出するために車両中の各シートに近接して位置する１つまたは複数のマイクロフォンおよび１つまたは複数のカメラのアレイなど、センサー１１５０のうちの１つまたは複数は、車両のキャビン内の様々なロケーションに分散されるなど、車両の他のロケーションに配置され得る。

【００９９】

[0112]ディスプレイ１１２０などの視覚的インターフェースデバイスは、出力デバイス１２０に対応することができ、車の運転者に見えるように車ダッシュボードデバイス１１０２内に取り付けられるかまたはその上に配置される（たとえば、車両ハンドセットマウントに着脱可能に固定される）。マルチモーダル認識エンジン１３０とフィードバックメッセージ生成器１３２とは、マルチモーダル認識エンジン１３０とフィードバックメッセージ生成器１３２とが車両の乗員に見えないことを示すために、破線の境界で示されている。マルチモーダル認識エンジン１３０とフィードバックメッセージ生成器１３２とは、図１のデバイス１１０中のように、ディスプレイ１１２０およびセンサー１１５０をも含むデバイス中に実装され得るか、または図５のデバイス５０２中のように、ディスプレイ１１２０およびセンサー１１５０とは別個であり、それらに結合され得る。

【０１００】

[0113]図１２Ａは、仮想現実、拡張現実、または複合現実ヘッドセットなど、ヘッドセット１２０２に組み込まれたマルチモーダル認識エンジン１３０とフィードバックメッセージ生成器１３２との一例を示す。ディスプレイ１２２０などの視覚的インターフェースデバイスは、出力デバイス１２０に対応することができ、ヘッドセット１２０２が着用されている間、ユーザへの拡張現実または仮想現実の画像またはシーンの表示を可能にするために、ユーザの眼の前に配置される。センサー１２５０は、１つまたは複数のマイクロフォン、カメラ、または他のセンサーを含むことができ、図１の入力デバイス１１２～１

１６に対応することができる。単一のロケーションに示されているが、他の実装形態では、マルチモーダル入力を検出するためにヘッドセット１２０２の周りに分散された１つまたは複数のマイクロフォンおよび１つまたは複数のカメラのアレイなど、センサー１２５０のうちの１つまたは複数は、ヘッドセット１２０２の他のロケーションに配置され得る。

【０１０１】

[0114]図１２Ｂは、ディスプレイ１２２０とセンサー１２５０とを含む、「スマートウォッチ」として示されている、ウェアラブル電子デバイス１２０４に組み込まれたマルチモーダル認識エンジン１３０とフィードバックメッセージ生成器１３２との一例を示す。センサー１２５０は、たとえば、ビデオ、スピーチ、およびジェスチャーなどのモダリティに基づくユーザ入力の検出を可能にする。また、単一のロケーションに示されているが、他の実装形態では、センサー１２５０のうちの１つまたは複数は、ウェアラブル電子デバイス１２０４の他のロケーションに配置され得る。

【０１０２】

[0115]図１３は、ワイヤレス通信デバイス実装形態（たとえば、スマートフォン）またはデジタルアシスタントデバイス実装形態などにおける、マルチモーダル認識エンジン１３０を含むデバイス１３００の特定の例示的な実装形態のブロック図を示す。様々な実装形態では、デバイス１３００は、図１３に示されているものよりも多いまたは少ない構成要素を有し得る。例示的な実装形態では、デバイス１３００は、デバイス１１０に対応し得る。例示的な実装形態では、デバイス１３００は、図１～図１２Ｂを参照しながら説明された１つまたは複数の動作を実施し得る。

【０１０３】

[0116]特定の実装形態では、デバイス１３００は、マルチモーダル認識エンジン１３０を含むプロセッサ１３０６（たとえば、プロセッサ１０８に対応する中央処理ユニット（ＣＰＵ））を含む。デバイス１３００は、１つまたは複数の追加のプロセッサ１３１０（たとえば、１つまたは複数のＤＳＰ）を含み得る。プロセッサ１３１０は、スピーチおよび音楽コーダデコーダ（コーデック）１３０８を含み得る。スピーチおよび音楽コーデック１３０８は、音声コーダ（「ボコーダ」）エンコーダ１３３６、ボコーダデコーダ１３３８、またはそれらの両方を含み得る。

【０１０４】

[0117]デバイス１３００は、メモリ１３８６と、コーデック１３３４とを含み得る。メモリ１３８６は、メモリ１０６に対応し得、マルチモーダル認識エンジン１３０、フィードバックメッセージ生成器１３２、アプリケーション２４０のうちの１つまたは複数、あるいはそれらの任意の組合せに関して説明された機能を実装するためにプロセッサ１３０６（あるいは１つまたは複数の追加のプロセッサ１３１０）によって実行可能である命令１３５６を含み得る。デバイス１３００は、トランシーバ１３５０を介して１つまたは複数のアンテナ１３５２に結合されたワイヤレスコントローラ１３４０を含み得る。いくつかの実装形態では、１つまたは複数のアンテナ１３５２は、ジェスチャー入力を示すデータを受信するように構成された１つまたは複数のアンテナを含む。

【０１０５】

[0118]デバイス１３００は、ディスプレイコントローラ１３２６に結合されたディスプレイ１３２８（たとえば、出力デバイス１２０）を含み得る。ディスプレイ１３２８は、フィードバックメッセージ１４４（たとえば、命令１４６）を出力するグラフィカルユーザインターフェースを表現するように構成され得る。コーデック１３３４は、デジタルアナログコンバータ（ＤＡＣ）１３０２と、アナログデジタルコンバータ（ＡＤＣ）１３０４とを含み得る。特定の実装形態では、コーデック１３３４は、１つまたは複数のマイクロフォン１３１２（たとえば、１つまたは複数のキーワードまたは音声コマンドを含むオーディオ入力をキャプチャするように構成された第１の入力デバイス１１２）からアナログ信号を受信し、アナログデジタル変換器１３０４を使用してアナログ信号をデジタル信号にコンバートし、デジタル信号をスピーチおよび音楽コーデック１３０８に提供し得る。スピーチおよび音楽コーデック１３０８は、デジタル信号を処理し得る。

【０１０６】
[0119]特定の実装形態では、スピーチおよび音楽コーデック１３０８は、オーディオ再生信号を表すデジタル信号をコーデック１３３４に提供し得る。コーデック１３３４は、デジタルアナログコンバータ１３０２を使用してデジタル信号をアナログ信号にコンバートし得、可聴信号を生成するために、アナログ信号を１つまたは複数のラウドスピーカー１３１４に提供し得る。１つまたは複数のラウドスピーカー１３１４は、出力デバイス１２０に対応することができ、図１のフィードバックメッセージ１４４をレンダリングするか、またはフィードバックメッセージ１４４をユーザにダイレクトするように構成され得る。
【０１０７】
[0120]特定の実装形態では、デバイス１３００は、１つまたは複数の入力デバイス１３３０を含む。入力デバイス１３３０は、図１の入力デバイス１１２～１１６のうちの１つまたは複数に対応することができる。たとえば、入力デバイス１３３０は、１つまたは複数のジェスチャーまたは視覚的コマンドを含むビデオ入力をキャプチャするように構成された１つまたは複数のカメラを含むことができる。
【０１０８】
[0121]特定の実装形態では、デバイス１３００は、システムインパッケージまたはシステムオンチップデバイス１３２２中に含まれ得る。特定の実装形態では、メモリ１３８６と、プロセッサ１３０６と、プロセッサ１３１０と、ディスプレイコントローラ１３２６と、コーデック１３３４と、ワイヤレスコントローラ１３４０とは、システムインパッケージまたはシステムオンチップデバイス１３２２中に含まれる。特定の実装形態では、入力デバイス１３３０（たとえば、図１の入力デバイス１１２～１１６のうちの１つまたは複数の）と、電源１３４４とは、システムインパッケージまたはシステムオンチップデバイス１３２２に結合される。その上、特定の実装形態では、図１３に示されているように、ディスプレイ１３２８と、入力デバイス１３３０と、マイクロフォン１３１２と、アンテナ１３５２と、電源１３４４とは、システムインパッケージまたはシステムオンチップデバイス１３２２の外部にある。特定の実装形態では、ディスプレイ１３２８と、入力デバイス１３３０と、マイクロフォン１３１２と、ラウドスピーカー１３１４と、アンテナ１３５２と、電源１３４４との各々は、インターフェースまたはコントローラなど、システムインパッケージまたはシステムオンチップデバイス１３２２の構成要素に結合され得る。
【０１０９】
[0122]デバイス１３００は、例示的および非限定的な例として、モバイル通信デバイス、スマートフォン、セルラーフォン、ラップトップコンピュータ、コンピュータ、タブレット、携帯情報端末、ディスプレイデバイス、テレビジョン、ゲーミングコンソール、音楽プレーヤ、ラジオ、デジタルビデオプレーヤ、デジタルビデオディスク（ＤＶＤ）またはＢｌｕ－ｒａｙ（登録商標）ディスクプレーヤ、チューナー、カメラ、ナビゲーションデバイス、仮想現実または拡張現実ヘッドセット、ウェアラブル電子デバイス、車両コンソールデバイス、あるいはそれらの任意の組合せを含み得る。
【０１１０】
[0123]説明される実装形態に関連して、マルチモーダルユーザ入力のためのデバイスは、第１の入力デバイスから受信された第１のデータを処理するマルチモーダル認識エンジンを含む。第１のデータは、第１の入力モード（たとえば、スピーチモード、ジェスチャーモード、またはビデオモード）に基づくユーザからの第１の入力を示す。フィードバックメッセージ生成器は、第１のデータを処理することに基づいて、第１の入力モードとは異なる第２の入力モードに基づく第２の入力を提供するようにユーザに命令するフィードバックメッセージを出力デバイスに送る。
【０１１１】
[0124]説明される実装形態に関連して、マルチモーダルユーザ入力のためのデバイスは、第１の入力デバイスから受信された第１のデータを処理するマルチモーダル認識エンジ

ンを含む。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示す。フィードバックメッセージ生成器は、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送る。フィードバックメッセージは、第１の入力に関連付けられるべき行為を識別するようにユーザに命令する。マルチモーダル認識エンジンは、第１の入力に関連付けられるべき特定の行為を識別する第２の入力を受信し、特定の行為に第１の入力を関連付けるように行為へのユーザ入力のマッピングを更新する。

【０１１２】

[0125]説明される実装形態に関連して、マルチモーダルユーザ入力のための装置は、第１の入力デバイスから受信された第１のデータを処理するための手段を含む。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、第１の入力は、コマンドに対応する。たとえば、第１のデータを処理するための手段は、プロセッサ１０８、マルチモーダル認識エンジン１３０、プロセッサ１３０６、１３１０によって実行可能な命令１３５６、１つまたは複数の他のデバイス、モジュール、回路、構成要素、あるいはそれらの組合せを含むことができる。

【０１１３】

[0126]本装置は、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送るための手段を含む。フィードバックメッセージは、第１の入力モードとは異なる第２の入力モードに基づいて、第１の入力に関連するコマンドを識別する第２の入力を提供するようにユーザに命令する。たとえば、送るための手段は、プロセッサ１０８、マルチモーダル認識エンジン１３０、フィードバックメッセージ生成器１３２、プロセッサ１３０６、１３１０によって実行可能な命令１３５６、１つまたは複数の他のデバイス、モジュール、回路、構成要素、あるいはそれらの組合せを含むことができる。

【０１１４】

[0127]本装置は、第２の入力デバイスから第２のデータを受信するための手段を含み、第２のデータは、第２の入力を示す。たとえば、第２のデータを受信するための手段は、プロセッサ１０８、マルチモーダル認識エンジン１３０、プロセッサ１３０６、１３１０によって実行可能な命令１３５６、１つまたは複数の他のデバイス、モジュール、回路、構成要素、あるいはそれらの組合せを含むことができる。

【０１１５】

[0128]本装置はまた、第２の入力によって識別されるコマンドに第１の入力を関連付けるようにマッピングを更新するための手段を含む。たとえば、更新するための手段は、プロセッサ１０８、マルチモーダル認識エンジン１３０、データ調整器２９２、プロセッサ１３０６、１３１０によって実行可能な命令１３５６、１つまたは複数の他のデバイス、モジュール、回路、構成要素、あるいはそれらの組合せを含むことができる。

【０１１６】

[0129]いくつかの実装形態では、非一時的コンピュータ可読媒体（たとえば、メモリ１０６、メモリ１３８６、またはそれらの任意の組合せ）は、デバイスの１つまたは複数のプロセッサ（たとえば、プロセッサ１０８、プロセッサ１３０６、プロセッサ１３１０、またはそれらの任意の組合せ）によって実行されたとき、図６～図１０の方法のうちの１つまたは複数の全部または一部に対応する動作を実施することなどによって、マルチモーダルユーザ入力を処理するための動作を１つまたは複数のプロセッサに実施させる命令（たとえば、命令１３５６）を含む。一例では、命令は、１つまたは複数のプロセッサによって実行されたとき、第１の入力デバイスから受信された第１のデータを１つまたは複数のプロセッサに処理させる。第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、第１の入力は、コマンドに対応する。命令は、１つまたは複数のプロセッサによって実行されたとき、１つまたは複数のプロセッサに、第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送らせる。フィードバックメッセージは、第１の入力モードとは異なる第２の入力モードに基づいて、第１の入力に関連するコマンドを識別する第２の入力を提供するようにユーザに命令する。命令は、１つまたは複数のプロセッサによって実行されたとき、１つまたは複数のプロセッサに、第２

の入力デバイスから第２のデータを受信させ、第２のデータは、第２の入力を示す。命令はまた、１つまたは複数のプロセッサによって実行されたとき、１つまたは複数のプロセッサに、第２の入力によって識別されるコマンドに第１の入力を関連付けるようにマッピングを更新させる。

【０１１７】

[0130]さらに、本明細書で開示される実装形態に関して説明される様々な例示的な論理ブロック、構成、モジュール、回路、およびアルゴリズムステップは、電子ハードウェア、プロセッサによって実行されるコンピュータソフトウェア、または両方の組合せとして実装され得ることを当業者は諒解されよう。様々な例示的な構成要素、ブロック、構成、モジュール、回路、およびステップについて、上記では概して、それらの機能に関して説明された。そのような機能がハードウェアとして実装されるか、プロセッサ実行可能命令として実装されるかは、特定の適用例および全体的なシステムに課された設計制約に依存する。当業者は、説明された機能を、特定の適用例ごとに様々な方法で実装し得、そのような実装の決定は、本開示の範囲からの逸脱を引き起こすと解釈されるべきではない。

【０１１８】

[0131]本明細書で開示される実装形態に関して説明された方法またはアルゴリズムのステップは、ハードウェアで直接実施されるか、プロセッサによって実行されるソフトウェアモジュールで実施されるか、またはその２つの組合せで実施され得る。ソフトウェアモジュールは、ランダムアクセスメモリ（ＲＡＭ）、フラッシュメモリ、読取り専用メモリ（ＲＯＭ）、プログラマブル読取り専用メモリ（ＰＲＯＭ）、消去可能プログラマブル読取り専用メモリ（ＥＰＲＯＭ）、電気的消去可能プログラマブル読取り専用メモリ（ＥＥＰＲＯＭ（登録商標））、レジスタ、ハードディスク、リムーバブルディスク、コンパクトディスク読取り専用メモリ（ＣＤ－ＲＯＭ）、または当技術分野で知られている任意の他の形態の非一時的記憶媒体中に常駐し得る。例示的な記憶媒体は、プロセッサが記憶媒体から情報を読み取り、記憶媒体に情報を書き込むことができるように、プロセッサに結合される。代替として、記憶媒体はプロセッサと一体であり得る。プロセッサと記憶媒体とは、特定用途向け集積回路（ＡＳＩＣ）中に存在し得る。ＡＳＩＣは、コンピューティングデバイスまたはユーザ端末中に存在し得る。代替として、プロセッサと記憶媒体とは、コンピューティングデバイスまたはユーザ端末中に個別構成要素として存在し得る。

【０１１９】

[0132]開示される実装形態の前の説明は、開示される実装形態を当業者が製作または使用することを可能にするために提供される。これらの実装形態への様々な変更は当業者には容易に明らかになり、本明細書で定義された原理は本開示の範囲から逸脱することなく他の実装形態に適用され得る。したがって、本開示は、本明細書に示された実装形態に限定されるものではなく、以下の特許請求の範囲によって定義されるような原理および新規の特徴に一致する可能な最も広い範囲を与えられるべきである。

以下に本願の出願当初の特許請求の範囲に記載された発明を付記する。

［Ｃ１］

マルチモーダルユーザ入力のためのデバイスであって、

第１の入力デバイスから受信された第１のデータを処理することと、前記第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、前記第１の入力は、コマンドに対応し、

前記第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることと、ここにおいて、前記フィードバックメッセージは、前記第１の入力モードとは異なる第２の入力モードに基づいて、前記第１の入力に関連するコマンドを識別する第２の入力を提供するように前記ユーザに命令する、

第２の入力デバイスから第２のデータを受信することと、前記第２のデータは、前記第２の入力を示し、

前記第２の入力によって識別される前記コマンドに前記第１の入力を関連付けるようにマッピングを更新することと、

を行うように構成された１つまたは複数のプロセッサを備える、デバイス。
［Ｃ２］
前記第１の入力モードは、スピーチモード、ジェスチャーモード、またはビデオモードのうちの１つであり、前記第２の入力モードは、前記スピーチモード、前記ジェスチャーモード、または前記ビデオモードのうちの異なる１つである、Ｃ１に記載のデバイス。
［Ｃ３］
前記フィードバックメッセージは、前記第１の入力をディスアンビギュエートするために前記第２の入力を提供するように前記ユーザに命令する、Ｃ１に記載のデバイス。
［Ｃ４］
前記１つまたは複数のプロセッサは、前記第１の入力の認識処理に関連する確信度レベルが確信度しきい値を満たすことに失敗したことに応答して、前記フィードバックメッセージを送るようにさらに構成された、Ｃ３に記載のデバイス。
［Ｃ５］
前記更新されたマッピングは、前記第１の入力と前記第２の入力との組合せを前記コマンドに関連付ける、Ｃ１に記載のデバイス。
［Ｃ６］
前記１つまたは複数のプロセッサは、マルチモーダル認識エンジンを含み、前記マルチモーダル認識エンジンは、
組み合わされた埋め込みベクトルを生成するために、前記第１の入力モードに関連する第１の埋め込みネットワークと、前記第２の入力モードに関連する第２の埋め込みネットワークとの出力を組み合わせるように構成された融合埋め込みネットワークと、
前記組み合わされた埋め込みベクトルを特定のコマンドにマッピングするように構成された分類器と、
を含む、Ｃ１に記載のデバイス。
［Ｃ７］
前記ユーザに対応する第１の埋め込みネットワークデータおよび第１の重みデータと、
第２のユーザに対応する第２の埋め込みネットワークデータおよび第２の重みデータと、前記第１の埋め込みネットワークデータは、前記ユーザと前記第２のユーザとの間の入力コマンドの差に基づいて前記第２の埋め込みネットワークデータとは異なり、前記第１の重みデータは、前記ユーザと前記第２のユーザとの間の入力モード信頼性の差に基づいて前記第２の重みデータとは異なり、
を記憶するように構成されたメモリをさらに備える、Ｃ６に記載のデバイス。
［Ｃ８］
前記第１の入力モードは、ビデオモードに対応し、前記１つまたは複数のプロセッサは、照明しきい値を下回る値を有する周辺光メトリックに応答して前記フィードバックメッセージを送るように構成された、Ｃ１に記載のデバイス。
［Ｃ９］
前記第１の入力モードは、スピーチモードに対応し、前記１つまたは複数のプロセッサは、雑音しきい値を超える値を有する雑音メトリックに応答して前記フィードバックメッセージを送るように構成された、Ｃ１に記載のデバイス。
［Ｃ１０］
グラフィカルユーザインターフェースを表すように構成されたディスプレイをさらに備える、Ｃ１に記載のデバイス。
［Ｃ１１］
１つまたは複数のキーワードまたは音声コマンドを含むオーディオ入力をキャプチャするように構成された１つまたは複数のマイクロフォンをさらに備える、Ｃ１に記載のデバイス。
［Ｃ１２］
１つまたは複数のジェスチャーまたは視覚的コマンドを含むビデオ入力をキャプチャするように構成された１つまたは複数のカメラをさらに備える、Ｃ１に記載のデバイス。

［Ｃ１３］
ジェスチャー入力を示すデータを受信するように構成された１つまたは複数のアンテナをさらに備える、Ｃ１に記載のデバイス。
［Ｃ１４］
前記フィードバックメッセージをレンダリングするかまたは前記ユーザにダイレクトするように構成された１つまたは複数のラウドスピーカーをさらに備える、Ｃ１に記載のデバイス。
［Ｃ１５］
前記ユーザは、ロボットまたは他の電子デバイスを含む、Ｃ１に記載のデバイス。
［Ｃ１６］
前記第１の入力デバイスと前記出力デバイスとは、仮想現実ヘッドセットまたは拡張現実ヘッドセットに組み込まれる、Ｃ１に記載のデバイス。
［Ｃ１７］
前記第１の入力デバイスと前記出力デバイスとは、車両に組み込まれる、Ｃ１に記載のデバイス。
［Ｃ１８］
マルチモーダルユーザ入力のための方法であって、
デバイスの１つまたは複数のプロセッサにおいて、第１の入力デバイスから受信された第１のデータを処理することと、前記第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、前記第１の入力は、コマンドに対応し、
前記１つまたは複数のプロセッサから、前記第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることと、ここにおいて、前記フィードバックメッセージは、前記第１の入力モードとは異なる第２の入力モードに基づいて、前記第１の入力に関連するコマンドを識別する第２の入力を提供するように前記ユーザに命令する、
前記１つまたは複数のプロセッサにおいて、第２の入力デバイスから第２のデータを受信することと、前記第２のデータは、前記第２の入力を示し、
前記１つまたは複数のプロセッサにおいて、前記第２の入力によって識別される前記コマンドに前記第１の入力を関連付けるようにマッピングを更新することと、
を備える、方法。
［Ｃ１９］
前記第１の入力モードは、スピーチモード、ジェスチャーモード、またはビデオモードのうちの１つであり、前記第２の入力モードは、前記スピーチモード、前記ジェスチャーモード、または前記ビデオモードのうちの異なる１つである、Ｃ１８に記載の方法。
［Ｃ２０］
前記フィードバックメッセージは、前記第１の入力をディスアンビギュエートするために前記第２の入力を提供するように前記ユーザに命令する、Ｃ１８に記載の方法。
［Ｃ２１］
前記フィードバックメッセージは、前記第１の入力の認識処理に関連する確信度レベルが確信度しきい値を満たすことに失敗したことに応答して送られる、Ｃ２０に記載の方法。
［Ｃ２２］
前記更新されたマッピングは、前記第１の入力と前記第２の入力との組合せを前記コマンドに関連付ける、Ｃ１８に記載の方法。
［Ｃ２３］
前記マッピングを更新することは、
前記ユーザに関連する埋め込みネットワークデータを更新すること、または
前記ユーザに関連する重みデータを更新すること、
のうちの少なくとも１つを含む、Ｃ１８に記載の方法。
［Ｃ２４］
前記第１の入力モードは、ビデオモードに対応し、前記フィードバックメッセージは、

照明しきい値を下回る値を有する周辺光メトリックに応答して送られる、Ｃ１８に記載の方法。
［Ｃ２５］
前記第１の入力モードは、スピーチモードに対応し、前記フィードバックメッセージは、雑音しきい値を超える値を有する雑音メトリックに応答して送られる、Ｃ１８に記載の方法。
［Ｃ２６］
マルチモーダルユーザ入力のための装置であって、
第１の入力デバイスから受信された第１のデータを処理するための手段と、前記第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、前記第１の入力は、コマンドに対応し、
前記第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送るための手段と、ここにおいて、前記フィードバックメッセージは、前記第１の入力モードとは異なる第２の入力モードに基づいて、前記第１の入力に関連するコマンドを識別する第２の入力を提供するように前記ユーザに命令する、
第２の入力デバイスから第２のデータを受信するための手段と、前記第２のデータは、前記第２の入力を示し、
前記第２の入力によって識別される前記コマンドに前記第１の入力を関連付けるようにマッピングを更新するための手段と、
を備える、装置。
［Ｃ２７］
前記更新されたマッピングは、前記第１の入力と前記第２の入力との組合せを前記コマンドに関連付ける、Ｃ２６に記載の装置。
［Ｃ２８］
デバイスの１つまたは複数のプロセッサによって実行されたとき、前記１つまたは複数のプロセッサに、
第１の入力デバイスから受信された第１のデータを処理することと、前記第１のデータは、第１の入力モードに基づくユーザからの第１の入力を示し、前記第１の入力が、コマンドに対応し、
前記第１のデータを処理することに基づいて出力デバイスにフィードバックメッセージを送ることと、ここにおいて、前記フィードバックメッセージは、前記第１の入力モードとは異なる第２の入力モードに基づいて、前記第１の入力に関連するコマンドを識別する第２の入力を提供するように前記ユーザに命令する、
第２の入力デバイスから第２のデータを受信することと、前記第２のデータは、前記第２の入力を示し、
前記第２の入力によって識別される前記コマンドに前記第１の入力を関連付けるようにマッピングを更新することと、
を行わせる命令を備える非一時的コンピュータ可読媒体。
［Ｃ２９］
前記第１の入力モードは、ビデオモードに対応し、前記フィードバックメッセージは、照明しきい値を下回る値を有する周辺光メトリックに応答して送られる、Ｃ２８に記載の非一時的コンピュータ可読媒体。
［Ｃ３０］
前記第１の入力モードは、スピーチモードに対応し、前記フィードバックメッセージは、雑音しきい値を超える値を有する雑音メトリックに応答して送られる、Ｃ２８に記載の非一時的コンピュータ可読媒体。

【図面】

【図１】

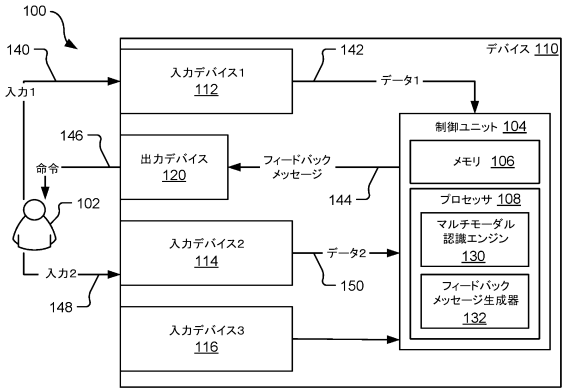

FIG. 1

【図２】

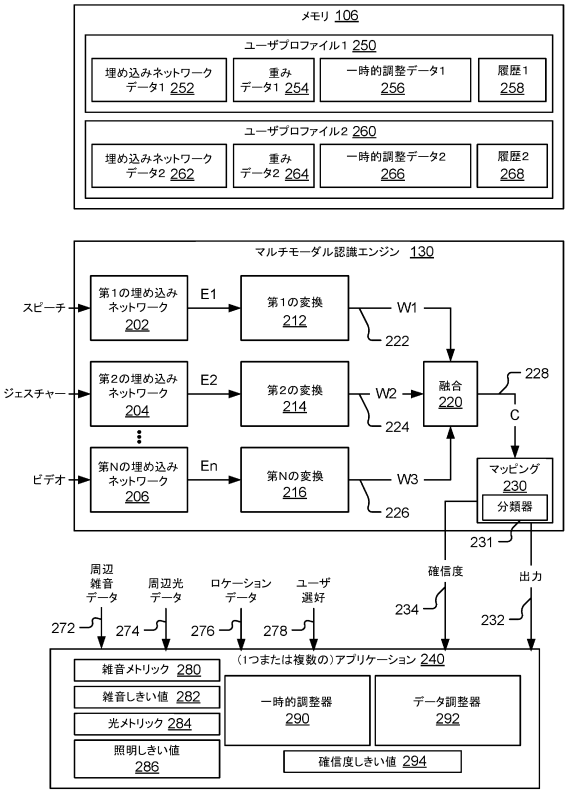

FIG. 2

【図３】

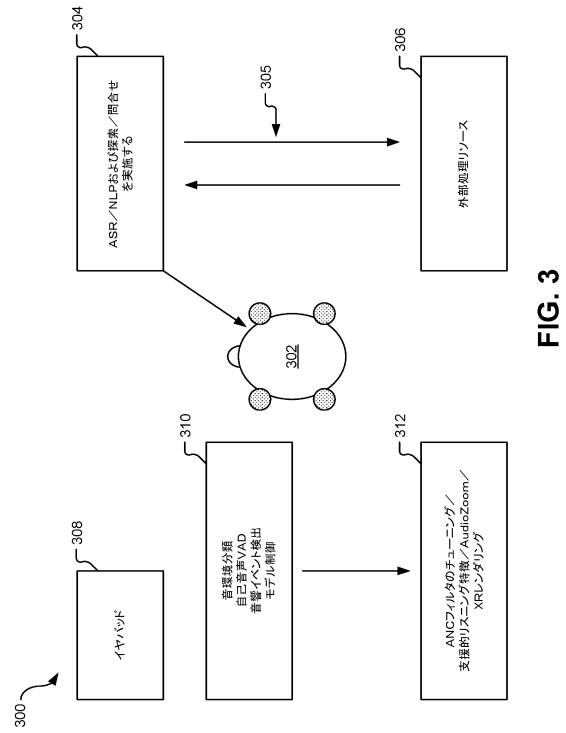

FIG. 3

【図４】

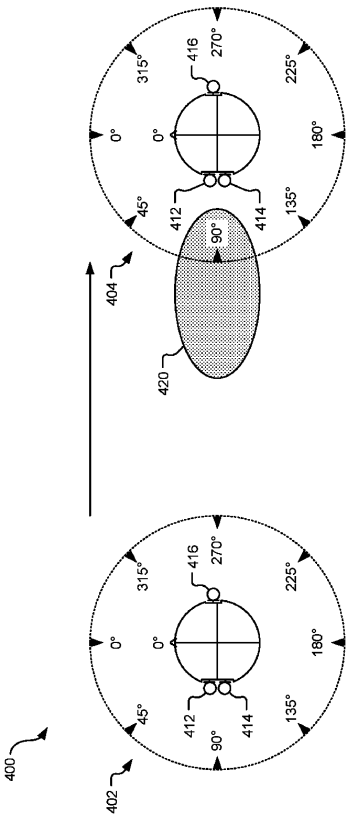

FIG. 4

【図５】

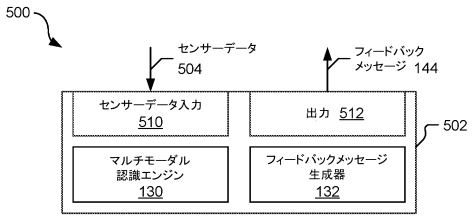

FIG. 5

【図６】

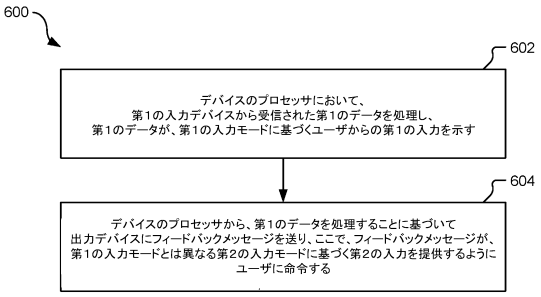

FIG. 6

【図７】

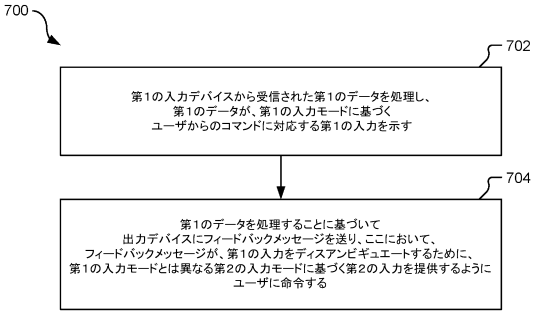

FIG. 7

【図８】

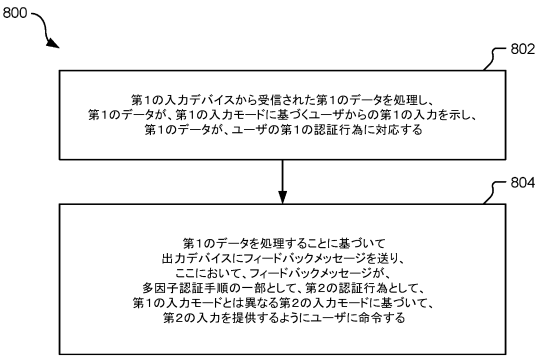

FIG. 8

【図９】

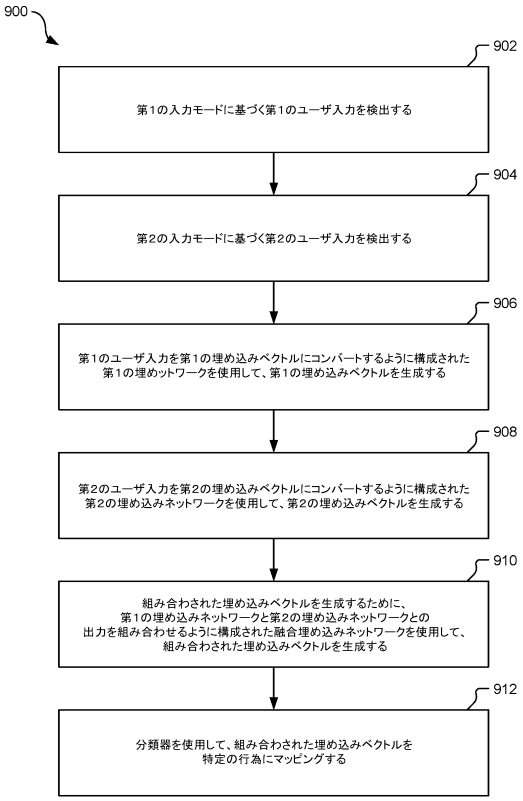

FIG. 9

【図１０】

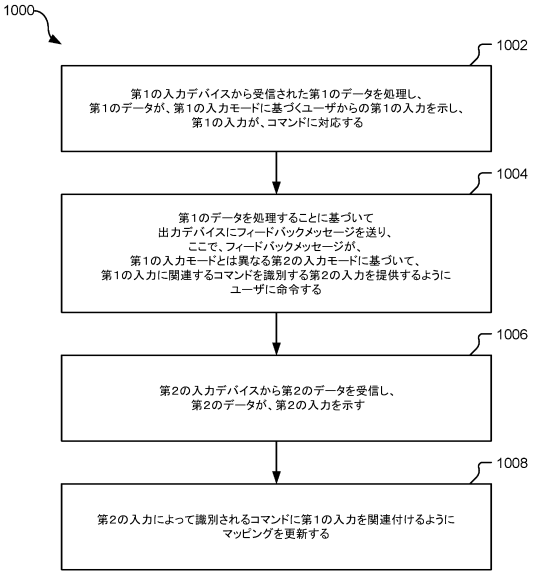

FIG. 10

【図１１】

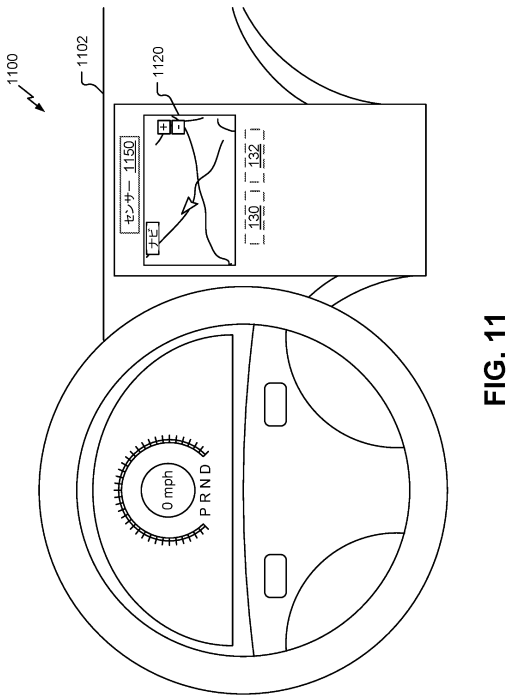

FIG. 11

【図１２Ａ】

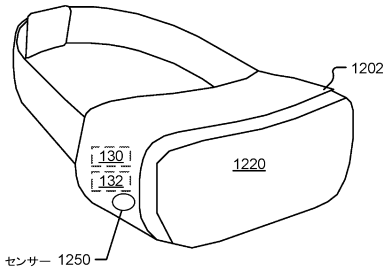

FIG. 12A

【図１２Ｂ】

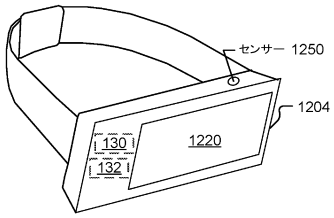

FIG. 12B

【図１３】

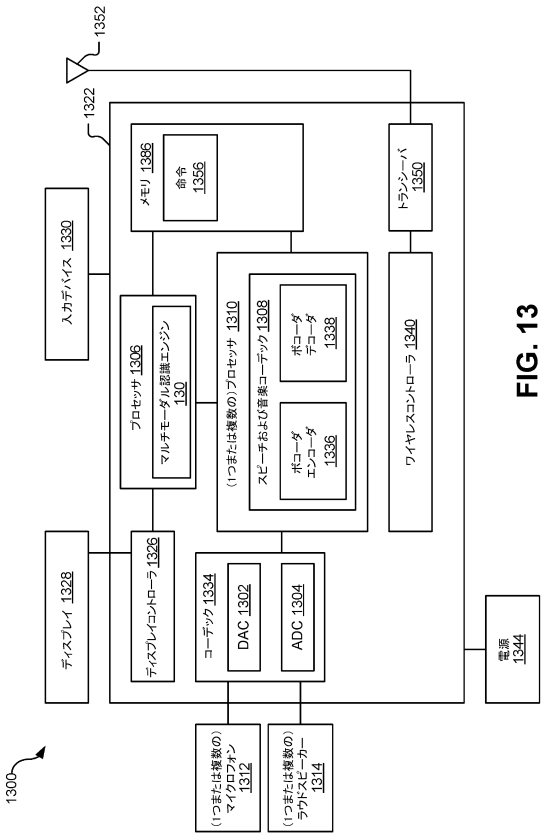

FIG. 13

フロントページの続き

(33)優先権主張国・地域又は機関
米国(US)
(72)発明者 チョウドハリー、ラビ
アメリカ合衆国、カリフォルニア州 ９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ ５７７５
(72)発明者 キム、レーフン
アメリカ合衆国、カリフォルニア州 ９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ ５７７５
(72)発明者 ムン、ソンクク
アメリカ合衆国、カリフォルニア州 ９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ ５７７５
(72)発明者 グオ、インイー
アメリカ合衆国、カリフォルニア州 ９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ ５７７５
(72)発明者 サキ、ファテメ
アメリカ合衆国、カリフォルニア州 ９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ ５７７５
(72)発明者 ビッサー、エリック
アメリカ合衆国、カリフォルニア州 ９２１２１－１７１４、サン・ディエゴ、モアハウス・ドライブ ５７７５
審査官 冨永 昌彦
(56)参考文献 米国特許出願公開第２０１８／０３２９６７７（ＵＳ，Ａ１）
国際公開第２０１４／０７０８７２（WO，Ａ２）
国際公開第２０１９／０２６６１７（WO，Ａ１）
特開２０１８－０３６９０２（ＪＰ，Ａ）
米国特許出願公開第２０２０／０１５２１９１（ＵＳ，Ａ１）
中国特許出願公開第１１０９９８７１８（ＣＮ，Ａ）
(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ３／０１
Ｇ０６Ｆ ３／０４８ － ０４８９５
Ｇ０６Ｆ ３／１６