(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7705481号
(P7705481)

(45)発行日　令和7年7月9日(2025.7.9)

(24)登録日　令和7年7月1日(2025.7.1)

(51)国際特許分類

Ｇ１０Ｌ	*15/22*	*(2006.01)*	Ｇ１０Ｌ	15/22	３００Ｚ
Ｇ１０Ｌ	*15/04*	*(2013.01)*	Ｇ１０Ｌ	15/04	３００Ｚ
Ｇ１０Ｌ	*15/10*	*(2006.01)*	Ｇ１０Ｌ	15/10	５００Ｔ
Ｇ１０Ｌ	*25/51*	*(2013.01)*	Ｇ１０Ｌ	15/04	３００Ａ
Ｇ１０Ｌ	*13/00*	*(2006.01)*	Ｇ１０Ｌ	15/10	２００Ｗ

ＦＩ

請求項の数　33　（全49頁）　最終頁に続く

(21)出願番号　特願2023-569701(P2023-569701)
(86)(22)出願日　令和3年11月29日(2021.11.29)
(65)公表番号　特表2024-528367(P2024-528367A)
(43)公表日　令和6年7月30日(2024.7.30)
(86)国際出願番号　PCT/US2021/060987
(87)国際公開番号　WO2023/022743
(87)国際公開日　令和5年2月23日(2023.2.23)
審査請求日　令和6年1月5日(2024.1.5)
(31)優先権主張番号　63/233,877
(32)優先日　令和3年8月17日(2021.8.17)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　17/532,819
(32)優先日　令和3年11月22日(2021.11.22)
最終頁に続く

(73)特許権者　502208397
グーグル エルエルシー
Ｇｏｏｇｌｅ ＬＬＣ
アメリカ合衆国 カリフォルニア州 ９４０４３ マウンテン ビュー アンフィシアター パークウェイ １６００
１６００ Ａｍｐｈｉｔｈｅａｔｒｅ Ｐａｒｋｗａｙ ９４０４３ Ｍｏｕｎｔａｉｎ Ｖｉｅｗ， ＣＡ Ｕ．Ｓ．Ａ．
(74)代理人　100108453
弁理士　村山 靖彦
(74)代理人　100110364
弁理士　実広 信哉
(74)代理人　100133400
弁理士　阿部 達彦
最終頁に続く

(54)【発明の名称】　自動アシスタントに関するソフトエンドポインティングを用いて自然な会話を可能にすること

(57)【特許請求の範囲】

【請求項１】

１つまたは複数のプロセッサによって実装される方法であって、前記方法が、

自動音声認識(ASR)モデルを使用して、ASR出力のストリームを生成するためにオーディオデータのストリームを処理するステップであって、前記オーディオデータのストリームがユーザのクライアントデバイスの1つまたは複数のマイクロフォンによって生成され、前記オーディオデータのストリームが、前記クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントに向けられた、前記ユーザによって提供された発話の一部をキャプチャする、ステップと、

自然言語理解(NLU)モデルを使用して、NLU出力のストリームを生成するために前記ASR出力のストリームを処理するステップと、

前記NLU出力の前記ストリームに基づいて、前記自動アシスタントが前記発話のフルフィルメントを開始できると判定するステップと、

前記オーディオデータのストリームを処理することに基づいて、前記発話の前記一部に関連付けられたオーディオベースの特徴を決定するステップと、

前記発話の前記一部に関連付けられた前記オーディオベースの特徴に基づいて、前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定するステップと、

前記ユーザが前記発話の提供を一時停止したと判定したことに応答して、および少なくとも前記NLU出力のストリームに基づいて前記自動アシスタントが前記発話のフルフィル

メントを開始することができると判定したことに応答して、
前記発話のフルフィルメントの開始を控えるステップと、
前記ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、前記ユーザへの可聴提示のために提供されるべき前記自然な会話出力が、前記自動アシスタントが、前記ユーザが前記発話の提供を完了するのを待っていることを示す、ステップと、
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップと
を含む、
方法。
【請求項２】
前記クライアントデバイスの前記1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップが、前記ユーザがしきい値持続時間の間に前記発話の提供を一時停止したと判定したことにさらに応答する、請求項1に記載の方法。
【請求項３】
前記発話の前記一部に関連付けられた前記オーディオベースの特徴に基づいて、前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定するステップが、
オーディオベースの分類機械学習(ML)モデルを使用して、出力を生成するために、前記発話の前記一部に関連付けられた前記オーディオベースの特徴を処理するステップと、
前記オーディオベースの分類MLモデルを使用して生成された前記出力に基づいて、前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定するステップと
を含む、
請求項1または2に記載の方法。
【請求項４】
前記NLU出力のストリームに基づいて、フルフィルメントデータのストリームを生成させるステップをさらに含み、
前記自動アシスタントが前記発話のフルフィルメントを開始することができると判定するステップが、前記フルフィルメントデータのストリームにさらに基づく、
請求項1から3のいずれか一項に記載の方法。
【請求項５】
前記ユーザが前記発話の提供を完了したと判定したことに応答して、前記自動アシスタントに、前記フルフィルメントデータのストリームに基づいて前記発話のフルフィルメントを開始させるステップをさらに含む、請求項4に記載の方法。
【請求項６】
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させる間、前記ASRモデルを利用する1つまたは複数の自動アシスタント構成要素をアクティブに保つステップをさらに含む、請求項4に記載の方法。
【請求項７】
前記ASR出力のストリームに基づいて、前記発話が特定の単語または句を含むかどうかを判定するステップと、
前記発話が前記特定の単語または句を含むと判定したことに応答して、
前記発話の前記一部に関連付けられた前記オーディオベースの特徴に基づいて、前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定することを控えるステップと、
前記自動アシスタントに、前記フルフィルメントデータのストリームに基づいて前記発話のフルフィルメントを開始させるステップと

をさらに含む、請求項4に記載の方法。
【請求項８】
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップの後に、前記ユーザがしきい値持続時間内に前記発話を提供し続けたかどうかを判定するステップと、
前記ユーザが前記しきい値持続時間内に前記1つまたは複数の発話を提供し続けなかったと判定したことに応答して、
NLUデータのストリームおよび/または前記フルフィルメントデータのストリームに基づいて、前記自動アシスタントが前記発話のフルフィルメントを開始することができるかどうかを判定するステップと、
前記自動アシスタントが前記フルフィルメントデータのストリームに基づいて前記発話のフルフィルメントを開始することができると判定したことに応答して、
前記自動アシスタントに、前記フルフィルメントデータのストリームに基づいて前記発話のフルフィルメントを開始させるステップと
をさらに含む、請求項4に記載の方法。
【請求項９】
前記クライアントデバイスの前記1つまたは複数のスピーカを介して、自然な会話出力を前記ユーザへの可聴提示のために提供させるステップの後に、前記ユーザがしきい値持続時間内に前記発話を提供し続けたかどうかを判定するステップと、
前記ユーザが前記発話を提供し続けなかったと判定したことに応答して、
前記ユーザへの可聴提示のために提供されるべき追加の自然な会話出力を決定するステップであって、前記ユーザへの可聴提示のために提供されるべき前記追加の自然な会話出力が、前記ユーザが前記発話の提供を完了することを要求する、ステップと、
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記追加の自然な会話出力を前記ユーザへの可聴提示のために提供させるステップと
をさらに含む、請求項1から8のいずれか一項に記載の方法。
【請求項１０】
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させる間、前記クライアントデバイスのディスプレイを介して、1つまたは複数のグラフィカル要素を前記ユーザへの視覚的提示のために提供させるステップであって、前記ユーザへの視覚的提示のために提供されるべき前記1つまたは複数のグラフィカル要素が、前記自動アシスタントが、前記ユーザが前記発話の提供を完了するのを待っていることを示す、ステップをさらに含む、請求項1から9のいずれか一項に記載の方法。
【請求項１１】
前記ASR出力が、前記オーディオデータのストリーム内にキャプチャされた前記発話の前記一部に対応するストリーミング転写を含み、
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させる間、前記クライアントデバイスの前記ディスプレイを介して、前記ストリーミング転写を前記ユーザへの視覚的提示のために提供させるステップであって、前記1つまたは複数のグラフィカル要素が、前記クライアントデバイスの前記ディスプレイを介して前記ユーザへの視覚的提示のために提供される前記ストリーミング転写の先頭に追加されるか、または末尾に追加される、ステップをさらに含む、
請求項10に記載の方法。
【請求項１２】
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させる間、前記クライアントデバイスの1つまたは複数の発光ダイオード(LED)を点灯させるステップであって、前記1つまたは複数のLEDが、前記自動アシスタントが、前記ユーザが前記発話の提供を完了するのを待っているこ

とを示すために点灯される、ステップをさらに含む、請求項1から11のいずれか一項に記載の方法。
【請求項１３】
前記発話の前記一部に関連付けられた前記オーディオベースの特徴が、イントネーション、トーン、ストレス、リズム、テンポ、ピッチ、休止、休止に関連する1つまたは複数の文法、および引き延ばされた音節のうちの1つまたは複数を含む、請求項1から12のいずれか一項に記載の方法。
【請求項１４】
前記ユーザへの可聴提示のために提供されるべき前記自然な会話出力を決定するステップが、
前記クライアントデバイスのオンデバイスメモリ内に自然な会話出力のセットを維持するステップと、
前記発話の前記一部に関連付けられた前記オーディオベースの特徴に基づいて、前記自然な会話出力のセットの中から前記自然な会話出力を選択するステップと
を含む、
請求項1から13のいずれか一項に記載の方法。
【請求項１５】
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップが、
前記ユーザへの可聴提示のために提供される他の出力よりも低い音量において、前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップを含む、
請求項1から14のいずれか一項に記載の方法。
【請求項１６】
前記クライアントデバイスの前記1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップが、
テキスト読み上げ(TTS)モデルを使用して、前記自然な会話出力を含む合成音声オーディオデータを生成するために前記自然な会話出力を処理するステップと、
前記クライアントデバイスの前記1つまたは複数のスピーカを介して、前記合成音声オーディオデータを前記ユーザへの可聴提示のために提供させるステップと
を含む、
請求項1から15のいずれか一項に記載の方法。
【請求項１７】
前記クライアントデバイスの前記1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップが、
前記クライアントデバイスのオンデバイスメモリから、前記自然な会話出力を含む合成音声オーディオデータを取得するステップと、
前記クライアントデバイスの前記1つまたは複数のスピーカを介して、前記合成音声オーディオデータを前記ユーザへの可聴提示のために提供させるステップと
を含む、
請求項1から16のいずれか一項に記載の方法。
【請求項１８】
前記1つまたは複数のプロセッサが、前記ユーザの前記クライアントデバイスによってローカルに実装される、請求項1から17のいずれか一項に記載の方法。
【請求項１９】
ユーザのクライアントデバイス上で、1つまたは複数のプロセッサによって実装される方法であって、前記方法が、
自動音声認識(ASR)モデルを使用して、ASR出力のストリームを生成するためにオーディオデータのストリームを処理するステップであって、前記オーディオデータのストリームが、前記クライアントデバイスの1つまたは複数のマイクロフォンによって生成され、

前記オーディオデータのストリームが、前記クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントに向けられた前記ユーザの発話の一部をキャプチャする、ステップと、
自然言語理解(NLU)モデルを使用して、NLU出力のストリームを生成するために前記ASR出力のストリームを処理するステップと、
前記NLU出力の前記ストリームに基づいて、前記自動アシスタントが前記発話のフルフィルメントを開始できると判定するステップと、
少なくとも前記NLU出力のストリームに基づいて、前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定するステップと、
前記ユーザが前記発話の提供を一時停止し、前記発話の提供を完了していないと判定したことに応答して、
前記発話のフルフィルメントの開始を控えるステップと、
前記ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、前記ユーザへの可聴提示のために提供されるべき前記自然な会話出力が、前記自動アシスタントが、前記ユーザが前記発話の提供を完了するのを待っていることを示す、ステップと、
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップと
を含む、
方法。
【請求項２０】
前記NLU出力のストリームに基づいて、前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定するステップが、前記NLU出力のストリームに基づいて、前記自動アシスタントが前記発話のフルフィルメントを開始することができるかどうかを判定するステップを含み、
前記ユーザが前記発話の提供を一時停止したと判定するステップが、前記自動アシスタントが前記NLU出力のストリームに基づいて前記発話のフルフィルメントを開始することができないと判定するステップを含む、
請求項19に記載の方法。
【請求項２１】
前記クライアントデバイスの前記1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップの後に、前記ユーザがしきい値持続時間内に前記発話を提供し続けたかどうかを判定するステップと、
前記ユーザが前記発話を提供し続けなかったと判定したことに応答して、
前記ユーザへの可聴提示のために提供されるべき追加の自然な会話出力を決定するステップであって、前記ユーザへの可聴提示のために提供されるべき前記追加の自然な会話出力が、前記ユーザが前記発話の提供を完了することを要求する、ステップと、
前記クライアントデバイスの1つまたは複数のスピーカを介して、前記追加の自然な会話出力を前記ユーザへの可聴提示のために提供させるステップと
をさらに含む、請求項20に記載の方法。
【請求項２２】
前記ユーザへの可聴提示のために提供されるべき前記追加の自然な会話出力が、前記発話の追加の部分がNLUデータのストリームに基づく特定のデータを含むことを要求する、請求項21に記載の方法。
【請求項２３】
ユーザのクライアントデバイス上で、1つまたは複数のプロセッサによって実装される方法であって、前記方法が、
自動音声認識(ASR)モデルを使用して、ASR出力のストリームを生成するためにオーディオデータのストリームを処理するステップであって、前記オーディオデータのストリームが、クライアントデバイスの1つまたは複数のマイクロフォンによって生成され、前記

オーディオデータのストリームが、前記クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントに向けられた前記ユーザの発話の一部をキャプチャする、ステップと、

自然言語理解(NLU)モデルを使用して、NLU出力のストリームを生成するために前記ASR出力のストリームを処理するステップと、

前記NLU出力の前記ストリームに基づいて、前記自動アシスタントが前記発話のフルフィルメントを開始できると判定するステップと、

前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定するステップと、

前記ユーザが前記発話の提供を一時停止し、前記発話の提供を完了していないと判定したことに応答して、及び前記NLU出力の前記ストリームに基づいて、前記自動アシスタントが前記発話のフルフィルメントを開始できると判定することに応答して、

前記発話のフルフィルメントの開始を控えるステップと、

前記ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、前記ユーザへの可聴提示のために提供されるべき前記自然な会話出力が、前記自動アシスタントが、前記ユーザが前記発話の提供を完了するのを待っていることを示す、ステップと、

前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップと、

前記クライアントデバイスの前記1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップの後に、前記ユーザがしきい値持続時間内に前記発話の提供を完了しなかったと判定したことに応答して、

前記少なくともNLUデータのストリームに基づいて、前記自動アシスタントが前記発話のフルフィルメントを開始することができるかどうかを判定するステップと、

前記NLUデータのストリームに基づいて、前記自動アシスタントが前記発話のフルフィルメントを開始することができると判定したことに応答して、

前記自動アシスタントに前記発話のフルフィルメントを開始させるステップと

を含む、

方法。

【請求項２４】

前記オーディオデータのストリームを処理することに基づいて、前記発話の前記一部に関連付けられたオーディオベースの特徴を決定するステップをさらに含み、

前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定するステップが、前記発話の前記一部に関連付けられた前記オーディオベースの特徴に基づく、

請求項23に記載の方法。

【請求項２５】

前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定するステップが、前記NLUデータのストリームに基づく、請求項23または24に記載の方法。

【請求項２６】

前記NLUデータのストリームに基づいて、前記自動アシスタントが前記発話のフルフィルメントを開始することができないと判定したことに応答して、

前記ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、前記ユーザへの可聴提示のために提供されるべき前記自然な会話出力が、前記ユーザが前記発話の提供を完了することを要求する、ステップと、

前記クライアントデバイスの1つまたは複数のスピーカを介して、追加の自然な会話出力を前記ユーザへの可聴提示のために提供させるステップと

をさらに含む、請求項23から25のいずれか一項に記載の方法。

【請求項２７】

前記ユーザへの可聴提示のために提供されるべき前記自然な会話出力が、前記発話の追加の部分が前記NLUデータのストリームに基づく特定のデータを含むことを要求する、請求項26に記載の方法。

【請求項２８】

前記自動アシスタントが前記発話のフルフィルメントを開始することができるかどうかを判定するステップが、前記発話のフルフィルメントに関連付けられた1つまたは複数の計算コストにさらに基づく、請求項23から27のいずれか一項に記載の方法。

【請求項２９】

前記発話のフルフィルメントに関連付けられた前記1つまたは複数の計算コストが、前記発話のフルフィルメントを実行することに関連する計算コスト、および前記発話の実行されたフルフィルメントを取り消すことに関連する計算コストのうちの1つまたは複数を含む、請求項28に記載の方法。

【請求項３０】

前記NLU出力のストリームに基づいて、フルフィルメントデータのストリームを生成させるステップをさらに含み、

前記自動アシスタントが前記発話のフルフィルメントを開始することができると判定するステップが、前記フルフィルメントデータのストリームにさらに基づく、
請求項23から29のいずれか一項に記載の方法。

【請求項３１】

1つまたは複数のプロセッサによって実装される方法であって、前記方法が、

オーディオデータのストリームを受信するステップであって、前記オーディオデータのストリームがユーザのクライアントデバイスの1つまたは複数のマイクロフォンによって生成され、前記オーディオデータのストリームが、前記クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントに向けられた、前記ユーザによって提供された発話の少なくとも一部をキャプチャする、ステップと、

前記オーディオデータのストリームを処理することに基づいて、前記発話の前記一部に関連付けられたオーディオベースの特徴を決定するステップと、

前記発話の前記一部に関連付けられた前記オーディオベースの特徴に基づいて、前記ユーザが前記発話の提供を一時停止したか、または前記発話の提供を完了したかを判定するステップと、

前記ユーザが前記発話の提供を一時停止し、前記発話の提供を完了していないと判定したことに応答して、および前記発話のフルフィルメントを開始することができると判定したことに応答して、

前記発話のフルフィルメントの開始を控えるステップと、

前記ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、前記ユーザへの可聴提示のために提供されるべき前記自然な会話出力が、前記自動アシスタントが、前記ユーザが前記発話の提供を完了するのを待っていることを示す、ステップと、

前記クライアントデバイスの1つまたは複数のスピーカを介して、前記自然な会話出力を前記ユーザへの可聴提示のために提供させるステップと

を含む、

方法。

【請求項３２】

少なくとも1つのプロセッサと、

実行されると、前記少なくとも1つのプロセッサに請求項1から31のいずれか一項に対応する動作を実行させる命令を記憶するメモリと

を備えるシステム。

【請求項３３】

実行されると、少なくとも1つのプロセッサに請求項1から31のいずれか一項に対応する動作を実行させるプログラムを記憶する非一時的なコンピュータ可読記憶媒体。

【発明の詳細な説明】
【背景技術】
【０００１】
人間は、本明細書では「自動アシスタント」と呼ばれる(「チャットボット」、「対話型パーソナルアシスタント」、「インテリジェントパーソナルアシスタント」、「パーソナル音声アシスタント」、「会話型エージェント」などとも呼ばれる)対話型ソフトウェアアプリケーションとの人間対コンピュータのダイアログに参加し得る。自動アシスタントは、典型的には、発話(またはタッチ/タイプ入力)の解釈および応答において、構成要素のパイプラインに依存する。例えば、自動音声認識(ASR)エンジンは、発話または発話に対応すると予測される音素の音声仮説(すなわち、用語および/または他のトークンのシーケンス)などのASR出力を生成するために、ユーザの発話に対応するオーディオデータを処理することができる。さらに、自然言語理解(NLU)エンジンは、発話(またはタッチ/タイプ入力)を提供する際のユーザの意図、およびオプションで意図に関連付けられたパラメータのスロット値などのNLU出力を生成するために、ASR出力(またはタッチ/タイプ入力)を処理することができる。さらに、NLU出力を処理し、発話に対する応答内容を取得するおよび/または発話に応答するアクションを実行するための構造化された要求などのフルフィルメント出力を生成するために、フルフィルメントエンジンが使用され得、フルフィルメント出力に基づいてフルフィルメントデータのストリームが生成され得る。
【０００２】
一般に、自動アシスタントとのダイアログセッションは、ユーザが発話を提供することによって開始され、自動アシスタントは、応答を生成するために前述の構成要素のパイプラインを使用して発話に応答することができる。ユーザは、追加の発話を提供することによってダイアログセッションを続けることができ、自動アシスタントは、追加応答を生成するために前述の構成要素のパイプラインを使用して追加の発話に応答することができる。別の言い方をすれば、これらのダイアログセッションは、一般に、ユーザが発話を提供するためにダイアログセッションを引き受け、ユーザが話すのを止めたときに自動アシスタントが発話に応答するためにダイアログセッションを引き受けるという点でターン制である。しかしながら、これらのターン制ダイアログセッションは、ユーザの観点からは、人間が実際に互いにどのように会話するかを反映していないので、自然ではない場合がある。
【０００３】
例えば、第1の人間は、単一の考えを第2の人間に伝えるために、複数の異なる発話を提供する場合があり、第2の人間は、第1の人間への応答を考案するために、複数の異なる発話の各々を考慮する可能性がある。場合によっては、第1の人間は、これらの複数の異なる発話の間に様々な時間量の間(または単一の発話を提供する様々な時間量の間)休止する場合がある。特に、第2の人間は、単に、複数の異なる発話のうちの第1の発話(またはその一部)に基づいて、または分離した複数の異なる発話の各々に基づいて、第1の人間への応答を完全には考案することができない場合がある。
【０００４】
同様に、これらのターン制ダイアログセッションにおいて、自動アシスタントは、複数の異なる発話に関する所与の発話のコンテキストを考慮することなく、またはユーザが所与の発話の提供の完了を待つことなく、ユーザの所与の発話(またはその一部)への応答を完全には考案することができない場合がある。結果として、これらのターン制ダイアログセッションは、ユーザがこれらのターン制ダイアログセッションの単一のターン中に単一の発話において自動アシスタントにユーザの考えを伝えようとするので、長引く可能性があり、それによって計算リソースを浪費する。さらに、ユーザが、これらのターン制ダイアログセッションの単一のターン中に複数の発話において自動アシスタントにユーザの考えを伝えようとする場合、自動アシスタントは、単純に失敗する場合があり、それによっても計算リソースを浪費する。例えば、自動アシスタントは、ユーザが発話を考案することを試みる際に長い一時停止をもたらすと、ユーザが話し終わったと早まって結論付け、

不完全な発話を処理し、不完全な発話によって意味のある意図が伝えられていないと(処理から)判定した結果として失敗する場合があり、または不完全な発話によって伝えられた誤った意図を(処理から)決定した結果として失敗する場合がある。それに加えて、ターン制ダイアログセッションは、アシスタント応答のレンダリング中に提供されたユーザの発話が意味あるように処理されるのを妨げる可能性ある。これは、発話を提供する前にアシスタント応答のレンダリングの完了を待つことをユーザに要求する可能性があり、それによってダイアログセッションを長引かせる。

【発明の概要】

【課題を解決するための手段】

【０００５】

本明細書で説明する実装形態は、ダイアログセッション中に自動アシスタントがユーザと自然な会話を実行することを可能にすることに向けられている。いくつかの実装形態は、ストリーミング自動音声認識(ASR)モデルを使用して、ASR出力のストリームを生成するため、にユーザのクライアントのマイクロフォンによって生成されたオーディオデータのストリームを処理することができる。オーディオデータのストリームは、クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントに向けられた、ユーザの発話の部分をキャプチャすることができる。さらに、ASR出力は、自然言語理解(NLU)出力のストリームを生成するために、NLUモデルを使用して処理されることが可能である。さらに、NLU出力は、フルフィルメントデータのストリームを生成するために、1つもしくは複数のフルフィルメントルールおよび/または1つもしくは複数のフルフィルメントモデルを使用して処理されることが可能である。それに加えて、オーディオデータのストリームを処理することに基づいて、発話のうちの1つまたは複数に関連付けられたオーディオベースの特徴が決定されることが可能である。発話の一部に関連付けられたオーディオベースの特徴は、例えば、イントネーション、トーン、ストレス、リズム、テンポ、ピッチ、引き延ばされた音節、休止、休止に関連する文法、および/またはオーディオデータのストリームの処理から導出され得る他のオーディオベースの特徴を含む。NLU出力のストリームおよび/またはオーディオベースの特徴に基づいて、自動アシスタントは、ユーザが発話の提供を一時停止したか、または発話の提供を完了したか(例えば、ソフトエンドポインティング)を判定することができる。

【０００６】

いくつかの実装形態において、ユーザが発話の提供を一時停止したと判定したことに応答して、自動アシスタントは、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示すために、(自動アシスタントが、様々な実装形態において発話のフルフィルメントが実行され得ると判定した場合であっても)ユーザに提示するための自然な会話出力を提供させることができる。いくつかの実装形態において、ユーザが発話の提供を完了したと判定したことに応答して、自動アシスタントは、フルフィルメント出力をユーザに提示するために提供させることができる。したがって、ユーザが発話の提供を一時停止したか、発話の提供を完了したかを判定することによって、自動アシスタントは、ターン制ダイアログセッションのようにユーザが発話の提供を一時停止した後に単にユーザに応答するのではなく、ユーザがなにを言ったのか、どのように言ったのかに基づいて、ユーザが自分の考えを完了するのを自然に待つことができる。

【０００７】

例えば、ユーザが自動アシスタントとのダイアログセッションに参加しており、「Arnolllldに電話をかけて(call Arnolllld's)」という発話を提供したと仮定する。ユーザが発話を提供すると、ASR出力のストリーム、NLU出力のストリーム、およびフルフィルメントデータのストリームは、発話をキャプチャするオーディオデータのストリームを処理することに基づいて生成されることが可能である。特に、この例において、発話が受信された時点において、ASR出力のストリームは、発話に対応する認識されたテキスト(例えば、「Arnoldに電話をかけて(call Arnold's)」)を含み得、NLU出力のストリームは、予測された「かけて(call)」または「電話をかけて(phone call)」という意図に関連する着信者

パラメータに関する「Arnold」のスロット値を有する予測された「かけて(call)」または「電話をかけて(phone call)」という意図を含み得、フルフィルメントデータのストリームは、フルフィルメント出力として実行されると、クライアントデバイスまたはクライアントデバイスと通信する追加のクライアントデバイスに、エンティティ参照「Arnold」に関連付けられたユーザの連絡先エントリとの通話を開始させるアシスタントコマンドを含むことができる。さらに、発話に関連付けられたオーディオベースの特徴は、オーディオデータのストリームを処理することに基づいて生成され得、例えば、ユーザが着信者パラメータについて正確になにを意図しているのかが不明であることを示す(例えば、「Arnolllldに電話をかけて(call Arnolllld's)」における「llll」によって示されるような)引き延ばされた音節を含む可能性がある。したがって、この例において、自動アシスタントがNLUデータのストリームに基づいて(例えば、クライアントデバイスまたは追加のクライアントデバイスに連絡先エントリ「Arnold」)との通話を開始させることによって)発話を履行することができる場合があるが、自動アシスタントは、オーディオベースの特徴に基づいて、ユーザが一時停止したと判定し、ユーザが発話を完了するための追加の時間を提供するために、発話を履行させることを控え得る。

【０００８】

むしろ、この例において、自動アシスタントは、ユーザに提示するための自然な会話出力を提供することを決定することができる。例えば、ユーザが発話の提供を一時停止したと判定したことに応答して(およびオプションで、ユーザがしきい値持続時間の間一時停止した後)、自動アシスタントは、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示すために、クライアントデバイスのスピーカを介するユーザへの可聴提示のために「うんうん(Mmhmm)」または「うん(Uh huhh)」(または他の音声相づち)などの自然な会話出力を提供させることができる。場合によっては、ユーザへの可聴提示のために提供される自然な会話出力の音量は、ユーザへの提示のために提供される他の可聴出力よりも低くされることが可能である。追加的または代替的に、クライアントデバイスがディスプレイを含む実装形態において、クライアントデバイスは、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示すために、バウンドする楕円とともに、発話のストリーミング転写などの1つまたは複数のグラフィカル要素をレンダリングすることができる。追加的または代替的に、クライアントデバイスが1つまたは複数の発光ダイオード(LED)を含む実装形態において、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示すために、クライアントデバイスはLEDのうちの1つまたは複数を点灯させることができる。特に、自然な会話出力がクライアントデバイスのユーザへの可聴提示のために提供されている間、1つまたは複数の自動アシスタント構成要素(例えば、ASR、NLU、フルフィルメント、および/または他の構成要素)が、オーディオデータのストリームの処理を継続するためにアクティブのままであることができる。

【０００９】

この例において、自然な会話出力が可聴提示のために提供されている間、または自然な会話出力が可聴提示のために提供された後、ユーザが前の発話の提供を完了するために「Arnoldのトラットリア(Arnold's Trattoria)」の発話を提供し、結果として「Arnoldのトラットリアに電話をかけて(call Arnold's Trattoria)」の発話を生じたとさらに仮定し、ここで、「Arnoldのトラットリア(Arnold's Trattoria)」は、架空のイタリア料理レストランである。したがって、ASR出力のストリーム、NLU出力のストリーム、およびフルフィルメントデータのストリームは、ユーザが発話を完了することに基づいて更新され得る。特に、NLU出力のストリームは、予測された「かけて(call)」または「電話をかけて(phone call)」の意図を依然として含む場合があるが、(例えば、連絡先エントリ「Arnold」ではなく)予測された「かけて(call)」または「電話をかけて(phone call)」の意図に関連付けられた着信者パラメータに関する「Arnoldのトラットリア(Arnold' Trattoria)」のスロット値を有し、フルフィルメントデータのストリームは、フルフィルメント出力として実行されると、クライアントデバイスまたはクライアントデバイスと通信する追

加のクライアントデバイスに、エンティティ参照「Arnoldのトラットリア(Arnold's Trattoria)」に関連付けられたレストランとの通話を開始させるアシスタントコマンドを含むことができる。さらに、自動アシスタントは、発話が完了したという判定に応答して、クライアントデバイスまたはクライアントデバイスと通信する追加のクライアントデバイスに通話を開始させることができる。

【００１０】

対照的に、自然な会話出力が可聴提示のために提供された後に(およびオプションで、自然な会話出力が可聴提示のために提供された後のしきい値持続時間の間)、ユーザが前の発話の提供を完了するためにいかなる発話も提供しなかったとさらに仮定する。この例において、自動アシスタントは、ユーザへの可聴提示のために提供されるべき追加の自然な会話出力がを決定することができる。しかしながら、追加の自然な会話は、クライアントデバイスのユーザが発話を完了することを明示的に要求することができ(例えば、「話していましたか(You were saying?)」、「私はなにか聞き逃しましたか(Did I miss something?)」など)、またはクライアントデバイスのユーザが予測された意図に関する特定のスロット値を提供することを明示的に要求することができる(例えば、「誰に電話をかけたかったですか(Who did you want to call?)」など)。いくつかの実装形態において、次いで、ユーザが前の発話の提供を完了するために「Arnoldのトラットリア(Arnold's Trattoria)」の発話を提供すると仮定すると、ASR出力のストリーム、NLU出力のストリーム、およびフルフィルメント出力のストリームは、更新され得、自動アシスタントは、上記で説明したように、(例えば、クライアントデバイスにエンティティ参照「Arnoldのトラットリア(Arnold's Trattoria)」に関連付けられたレストランとの通話を開始させることによって)発話を履行させることができる。

【００１１】

追加または代替の実装形態において、クライアントデバイスがディスプレイを含むと仮定すると、自動アシスタントは、ユーザへの視覚的提示のために複数の選択可能なグラフィカル要素を提供することができ、選択可能なグラフィカル要素の各々は、発話の1つまたは複数の部分の異なる解釈に関連付けられる。この例において、自動アシスタントは、選択されると、自動アシスタントにレストラン「Arnoldのトラットリア(Arnold's Trattoria)」との通話を開始させる第1の選択可能なグラフィカル要素と、選択されると自動アシスタントに連絡先エントリ「Arnold」との通話を開始させる第2の選択可能なグラフィカル要素とを提供することができる。次いで、自動アシスタントは、選択可能なグラフィカル要素のうちの所与の1つのユーザ選択の受信に基づいて通話を開始することができ、または1つまたは複数の選択可能なグラフィカル要素を提示させるしきい値持続時間内にユーザが選択可能なグラフィカル要素のうちの1つを選択しない場合、解釈に関連付けられたNLU測定値に基づいて通話を開始させることができる。例えば、この例において、自動アシスタントは、1つまたは複数の選択可能なグラフィカル要素がユーザへの提示のために提供された後、ユーザが、5秒、7秒、または任意の他のしきい値持続時間内に選択可能なグラフィカル要素のうちの1つまたは複数の選択を提供しない場合、レストラン「Arnoldのトラットリア(Arnold's Trattoria)」との通話を開始することができる。

【００１２】

別の例として、ユーザが自動アシスタントとのダイアログセッションに参加し、「私のカレンダーになにかありますかforrrr(what's on my calendar forrrr)」の発話を提供すると仮定する。ユーザが発話を提供すると、ASR出力のストリーム、NLU出力のストリーム、およびフルフィルメントデータのストリームは、発話をキャプチャするオーディオデータのストリームを処理することに基づいて生成され得る。特に、この例において、発話が受信された時点において、ASR出力のストリームは、発話に対応する認識されたテキスト(例えば、「私のカレンダーになにかありますか(what's on my calendar for)」)を含む場合があり、NLU出力のストリームは、予測された「カレンダー」または「カレンダー検索」の意図に関連付けられた日付パラメータに関する未知のスロット値を有する予測された「カレンダー」または「カレンダー検索」の意図を含む場合があり、フルフィルメ

ントデータのストリームは、フルフィルメント出力として実行されると、クライアントデバイスにユーザのカレンダー情報を検索させるアシスタントコマンドを含むことができる。同様に、発話に関連付けられたオーディオベースの特徴は、オーディオデータのストリームを処理することに基づいて生成され得、例えば、ユーザが日付パラメータについて確信がないことを示す(例えば、「私のカレンダーになにかありますかforrrr(what's on my calendar forrrr)」における「rrrr」によって示されるような)引き延ばされた音節を含む可能性がある。したがって、この例において、自動アシスタントは、(例えば、未知のスロット値に基づく)NLUデータのストリームおよび/または発話のオーディオベースの特徴に基づいて発話を履行することができない場合があり、自動アシスタントは、オーディオベースの特徴に基づいて、ユーザが一時停止したと判定し、ユーザが発話を完了するための追加の時間を提供するために、発話を履行させることを控え得る。

【００１３】

同様に、この例において、自動アシスタントは、ユーザに提示するための自然な会話出力を提供することを決定することができる。例えば、ユーザが発話の提供を一時停止したと判定したことに応答して(およびオプションで、ユーザがしきい値持続時間の間に一時停止した後)、自動アシスタントは、自動アシスタントが、ユーザが発話の提供を完了するのを待っていること、および/または自動アシスタントがユーザが発話の提供を完了するのを待っていることを示す他の指示を示すために、クライアントデバイスのスピーカを介してユーザに可聴提示するために「うんうん(Mmhmm)」または「うん(Uh huhh)」などの自然な会話出力を提供させることができる。しかしながら、自然な会話出力が可聴提示のために提供された後に(およびオプションで、自然な会話出力が可聴提示のために提供された後のしきい値持続時間の間)、ユーザが前の発話の提供を完了するためになにも提供しなかったとさらに仮定する。この例において、自動アシスタントは、予測された「カレンダー」または「カレンダー検索」の意図に関連付けられた未知の日付パラメータに関する現在の日付のスロット値を単純に推測し、ユーザが発話を完了していなかったとしても、自動アシスタントに、現在の日付に関するカレンダー情報を(例えば、可聴および/または視覚的に)ユーザに提供することによって発話を履行させ得る。追加または代替の実装形態において、自動アシスタントは、任意の発話の曖昧さを解消するため、任意の発話のフルフィルメントを確認するため、および/または任意のアシスタントコマンドを履行させる前に任意の他のアクションを実行するために、1つまたは複数の追加または代替の自動アシスタントの構成要素を利用することができる。

【００１４】

ユーザが最初に「私のカレンダーになにかありますかforrrr(what's on my calendar forrrr)」の発話を提供した後者の例などの様々な実装形態において、ユーザが最初に「Arnolllldに電話をかけて(call Arnolllld's)」の発話を提供した前者の例とは対照的に、自動アシスタントは、履行されるべき発話を履行すること、および/または発話が誤って履行された場合に発話のフルフィルメントを取り消すことに関連する1つまたは複数の計算コストを決定することができる。例えば、前者の例において、発話を履行することに関連する計算コストは、少なくとも、連絡先エントリ「Arnold」との通話を開始させることを含む可能性があり、発話のフルフィルメントを取り消すことに関連する計算コストは、少なくとも、「Arnold」に関連付けられた連絡先エントリとの通話を終了することと、ユーザとのダイアログセッションを再開始することと、追加の発話を処理することと、レストラン「Arnoldのトラットリア(Arnold's Trattoria)」との別の通話を開始させることとを含む可能性がある。さらに、前者の例において、ユーザが意図しない通話を開始することに関連する1つまたは複数のユーザコストは、比較的高い場合がある。また、例えば、後者の例において、発話を履行することに関連する計算コストは、少なくとも、現在の日付に関するカレンダー情報をユーザへの提示のために提供させることを含む可能性があり、発話のフルフィルメントを取り消すことに関連する計算コストは、ユーザによって指定された別の日付に関するカレンダー情報をユーザへの提示のために提供させることを含む可能性がある。さらに、後者の例において、誤ったカレンダー情報をユーザに提供すること

に関連する1つまたは複数のユーザコストは、比較的低い場合がある。別の言い方をすれば、前者の例におけるフルフィルメント(およびフルフィルメントを取り消すこと)に関連する計算コストは、後者の例におけるフルフィルメント(およびフルフィルメントを取り消すこと)に関連する計算コストよりも相対的に高い。したがって、自動アシスタントは、後者の例において、ダイアログセッションをより迅速かつより効率的に終了させる試みにおいて、後者の計算コストに基づいて推論された日付パラメータを用いて発話を履行することを決定し得るが、前者の例においては、前者の計算コストのためにそうではない。

【００１５】

本明細書で説明する技法を使用することによって、1つまたは複数の技法的利点が達成されることが可能である。1つの非限定的な例として、本明細書で説明する技法は、自動アシスタントがダイアログセッション中にユーザとの自然な会話に参加することを可能にする。例えば、自動アシスタントは、自動アシスタントがターン制ダイアログセッションに限定されないように、またはユーザに応答する前にユーザが話し終わったと判定することに依存しないように、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定し、それに応じてユーザへの提示のために提供される出力を適応させることができる。したがって、自動アシスタントは、ユーザがこれらの自然な会話に参加するときに、いつユーザに応答するか、およびユーザにどのように応答するかを決定することができる。これは、クライアントデバイスにおける計算リソースを節約し、ダイアログセッションをより迅速かつより効率的に終了させることができるという、様々な技法的利点を結果として生じる。例えば、自動アシスタントは、(自動アシスタントがフルフィルメントが実行されるべきであると予測する場合であっても)ユーザの代わりに任意のフルフィルメントを実行しようとする前に、ユーザからのより多くの情報を待つことができるので、自動アシスタントが失敗する回数が低減されることが可能である。また、例えば、ユーザが同じことを繰り返さなければならない回数、または自動アシスタントを再呼び出ししなければならない回数が低減されることが可能であるので、クライアントデバイスにおいて受信されるユーザ入力の量が低減されることが可能である。

【００１６】

本明細書で使用される「ダイアログセッション」は、ユーザと自動アシスタント(場合によっては、他の人間の参加者)との間の論理的に自己充足型の交換を含む場合がある。自動アシスタントは、セッション間の時間経過、セッション間のユーザコンテキスト(例えば、場所、予定された会議の前/最中/後など)の変更、ユーザと自動アシスタントとの間のダイアログ以外のユーザとクライアントデバイスとの間の1つまたは複数の介在する対話(例えば、ユーザがしばらくの間アプリケーションを切り替え、ユーザが、立ち去り、その後、スタンドアロンの音声起動製品に戻る)の検出、セッション間のクライアントデバイスのロック/スリープ、自動アシスタントと対話するために使用されるクライアントデバイスの変更など、様々な信号に基づいて、ユーザとの複数のダイアログセッション間を区別してもよい。

【００１７】

上記の説明は、本明細書で開示するいくつかの実装形態のみの概要として提供される。それらの実装形態および他の実装形態について、本明細書でさらに詳細に説明する。

【００１８】

本明細書で開示する技法は、クライアントデバイス上でローカルに、1つまたは複数のネットワークを介してクライアントデバイスに接続されたサーバによってリモートで、および/またはその両方で実装されることが可能であることが理解されるべきである。

【図面の簡単な説明】

【００１９】

【図１】本開示の様々な態様を実証し、本明細書で開示する実装形態が実装されることが可能な例示的な環境のブロック図である。

【図２】様々な実装形態による、図1の様々な構成要素を使用して本開示の様々な態様を実証する例示的なプロセスフローを示す図である。

【図３】様々な実装形態による、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定する例示的な方法を示すフローチャートである。
【図４】様々な実装形態による、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定する別の例示的な方法を示すフローチャートである。
【図５Ａ】様々な実装形態による、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定する様々な非限定な例を示す図である。
【図５Ｂ】様々な実装形態による、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定する様々な非限定な例を示す図である。
【図５Ｃ】様々な実装形態による、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定する様々な非限定な例を示す図である。
【図５Ｄ】様々な実装形態による、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定する様々な非限定な例を示す図である。
【図５Ｅ】様々な実装形態による、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定する様々な非限定な例を示す図である。
【図６】様々な実装形態による、コンピューティングデバイスの例示的なアーキテクチャを示す図である。
【発明を実施するための形態】
【００２０】
ここで図1に進むと、本開示の様々な態様を実証し、本明細書で開示する実装形態が実装されることが可能な例示的な環境のブロック図が示されている。例示的な環境は、クライアントデバイス110と自然会話システム180とを含む。いくつかの実装形態において、自然会話システム180は、クライアントデバイス110においてローカルに実装されることが可能である。追加または代替の実装形態において、自然会話システム180は、図1に示すようにクライアントデバイス110からリモートで(例えば、リモートサーバにおいて)実装されることが可能である。これらの実装形態において、クライアントデバイス110および自然会話システム180は、1つまたは複数の有線またはワイヤレスのローカルエリアネットワーク(Wi-Fi LANを含む「LAN」、メッシュネットワーク、Bluetooth、近距離無線通信など)またはワイドエリアネットワーク(インターネットを含む「WAN」)などの1つまたは複数のネットワーク199を介した互いに通信可能に結合され得る。
【００２１】
クライアントデバイス110は、例えば、デスクトップコンピュータ、ラップトップコンピュータ、タブレット、携帯電話、車両のコンピューティングデバイス(例えば、車載通信システム、車載娯楽システム、車載ナビゲーションシステム)、スタンドアロンの対話型スピーカ(オプションでディスプレイを有する)、スマートテレビなどのスマート家電、および/またはコンピューティングデバイスを含むユーザのウェアラブル装置(例えば、コンピューティングデバイスを有するユーザの腕時計、コンピューティングデバイスを有するユーザの眼鏡、仮想または拡張現実コンピューティングデバイス)のうちの1つまたは複数であり得る。追加のおよび/または代替のクライアントデバイスが提供され得る。
【００２２】
クライアントデバイス110は、自動アシスタントクライアント114を実行することができる。自動アシスタントクライアント114のインスタンスは、クライアント110のオペレーティングシステムとは別の(例えば、オペレーティングシステムの「上に」インストール

された)アプリケーションであることが可能であり、または代替的にはクライアントデバイス110のオペレーティングシステムによって直接実装されることが可能である。自動アシスタントクライアント114は、クライアントデバイス110においてローカルに実装された、または(例えば、リモートサーバにおいて)図1に示すようにネットワーク199のうちの1つまたは複数を介してクライアントデバイス110からリモートで実装された自然会話システム180と対話することができる。自動アシスタントクライアント114は(およびオプションでリモートサーバとの対話によって)、ユーザに視点から、ユーザが人間対コンピュータのダイアログに参加し得る自動アシスタント115の論理的インスタンスであるように見えるものを形成し得る。自動アシスタント115のインスタンスが図1に示されており、クライアントデバイス110の自動アシスタントクライアント114と自然会話システム180とを含む破線によって囲まれている。したがって、クライアントデバイス110上で実行される自動アシスタントクライアント114と関与するユーザは、実際には、自動アシスタント115のユーザ自身の論理インスタンス(または世帯もしくは他のユーザグループの間で共有される自動アシスタント115の論理インスタンス)と関与し得ることが理解されるべきである。簡潔さおよび単純さのために、本明細書で使用する自動アシスタント115は、クライアントデバイス110上でローカルに、および/またはクライアントデバイス110からリモートで(例えば、追加的または代替的に自然会話システム180のインスタンスを実装し得るリモートサーバにおいて)実行される自動アシスタントクライアント114を指す。

【００２３】

様々な実装形態において、クライアントデバイス110は、1つまたは複数のユーザインターフェース入力デバイスを使用してクライアントデバイス110のユーザによって提供されるユーザ入力を検出するように構成されたユーザ入力エンジン111を含み得る。例えば、クライアントデバイス110は、クライアントデバイス110のユーザの発話またはクライアントデバイス110の環境における他の音をキャプチャするオーディオデータなどのオーディオデータを生成する1つまたは複数のマイクロフォンを備え得る。追加的または代替的に、クライアントデバイス110は、視覚構成要素のうちの1つまたは複数の視野内で検出された画像および/または動き(例えば、ジェスチャ)をキャプチャする視覚データを生成するように構成された1つまたは複数の視覚構成要素を備え得る。追加的または代替的に、クライアントデバイス110は、クライアントデバイス110に向けられたタッチ入力をキャプチャする1つまたは複数の信号を生成するように構成された1つまたは複数のタッチ感知構成要素(例えば、キーボードおよびマウス、スタイラス、タッチスクリーン、タッチパネル、1つまたは複数のハードウェアボタンなど)を備え得る。

【００２４】

様々な実装形態において、クライアントデバイス110は、1つまたは複数のユーザインターフェース出力デバイスを使用してクライアントデバイス110のユーザに可聴および/または視覚的提示のためのコンテンツを提供するように構成されたレンダリングエンジン112を含み得る。例えば、クライアントデバイス110は、クライアントデバイス110の1つまたは複数のスピーカを介してクライアントデバイス110のユーザに可聴提示のためにコンテンツが提供されることを可能にする1つまたは複数のスピーカを備え得る。追加的または代替的に、クライアントデバイス110は、クライアントデバイス110のディスプレイまたはプロジェクタを介してクライアントデバイスのユーザに視覚的提示のためにコンテンツが提供されることを可能にするディスプレイまたはプロジェクタを備え得る。他の実装形態において、クライアントデバイス110は、(例えば、ネットワーク199のうちの1つまたは複数を介して)1つまたは複数の他のコンピューティングデバイスと通信し得、他のコンピューティングデバイスのうちの1つまたは複数のユーザインターフェース入力デバイスおよび/またはユーザインターフェース出力デバイスは、それぞれ、クライアントデバイス110のユーザによって提供されるユーザ入力を検出するため、ならびに/またはクライアントデバイス110のユーザへの可聴および/もしくは視覚的提示のためのコンテンツを提供するために利用され得る。追加的または代替的に、クライアントデバイス110は、自動アシスタント115がクライアントデバイス110のユーザからのユーザ入力を処理し、クライ

アントデバイス110のユーザがユーザ入力を提供し続けるのを待っていることの指示を提供するため、および/または自動アシスタント115が任意の他の機能を実行していることの指示を提供するために、1つまたは複数の色において照明され得る1つまたは複数の発光ダイオード(LED)を備え得る。

【００２５】

様々な実装形態において、クライアントデバイス110は、対応するユーザからの承認を得て、検出された存在、特に人間の存在を示す信号を提供するように構成された1つまたは複数の存在センサ113を含み得る。それらの実装形態のうちのいくつかにおいて、自動アシスタント115は、クライアントデバイス110における(またはクライアントデバイス110のユーザに関連付けられた他のコンピューティングデバイスにおける)ユーザの存在に少なくとも部分的に基づいて、発話を満たすクライアントデバイス110(またはクライアントデバイス110のユーザに関連付けられた他のコンピューティングデバイス)を特定することができる。発話は、クライアントデバイス110および/もしくはクライアントデバイス110のユーザに関連付けられた他のコンピューティングデバイスにおいて(例えば、レンダリングエンジン112を介して)応答コンテンツをレンダリングすることによって、クライアントデバイス110および/もしくはクライアントデバイス110のユーザに関連付けられた他のコンピューティングデバイスを制御させることによって、ならびに/またはクライアントデバイス110および/もしくはクライアントデバイス110のユーザに関連付けられた他のコンピューティングデバイスに、発話を満たすために任意の他のアクションを実行させることによって満たされることが可能である。本明細書で説明するように、自動アシスタント115は、ユーザが近くにいるかまたは最近近くにいた場所に基づいてクライアントデバイス110(または他のコンピューティングデバイス)を決定する際に存在センサ113に基づいて決定されたデータを活用し、対応するコマンドをクライアントデバイス110のみに(または他のコンピューティングデバイスに)提供することができる。いくつかの追加のまたは代替の実装形態において、自動アシスタント115は、任意のユーザ(任意のユーザまたは特定のユーザ)が現在クライアントデバイス110(または他のコンピューティングデバイス)に近接しているかどうかを判定する際に存在センサ113に基づいて決定されたデータを活用することができ、クライアントデバイス110(または他のコンピューティングデバイス)に近接しているユーザに基づいて、クライアントデバイス110(または他のコンピューティングデバイス)へのデータの提供および/またはクライアントデバイス110(または他のコンピューティングデバイス)からのデータの提供をオプションで抑制することができる。

【００２６】

存在センサ113は、様々な形態であり得る。例えば、クライアントデバイス110は、ユーザの存在を検出するために、ユーザ入力エンジン111に関して上記で説明したユーザインターフェース入力構成要素のうちの1つまたは複数(例えば、上記で説明したマイクロフォン、視覚構成要素、および/またはタッチ感知構成要素)を利用することができる。追加的または代替的に、クライアントデバイス110は、視野内の物体から放射される赤外(「IR」)光を測定する受動型赤外線(「PIR」)センサなどの、他のタイプの光ベースの存在センサ113を備え得る。

【００２７】

追加的または代替的に、いくつかの実装形態において、存在センサ113は、人間の存在またはデバイスの存在に関連付けられた他の現象を検出するように構成され得る。例えば、いくつかの実施形態において、クライアントデバイス110は、例えば、ユーザによって携帯/操作される他のコンピューティングデバイス(例えば、モバイルデバイス、ウェアラブルコンピューティングデバイスなど)および/または他のコンピューティングデバイスによって放射される様々なタイプのワイヤレス信号(例えば、無線、超音波、電磁気などの波)を検出する存在センサ113を備え得る。例えば、クライアントデバイス110は、(例えば、超音波対応マイクロフォンなどの超音波/赤外線受信機を介して)他のコンピューティングデバイスによって検出され得る、超音波または赤外線などの、人間には知覚できない波を放射するように構成され得る。

【００２８】
追加的または代替的に、クライアントデバイス110は、ユーザによって携帯/操作される他のコンピューティングデバイス(例えば、モバイルデバイス、ウェアラブルコンピューティングデバイスなど)によって検出され、ユーザの特定の位置を決定するために使用され得る無線波(例えば、Wi-Fi、Bluetooth、セルラなど)などの、他のタイプの人間が知覚できない波を放射し得る。いくつかの実装形態において、例えば、クライアントデバイス110への/からのGPSおよび/またはWi-Fi信号に基づいて、人の位置を検出するために、GPSおよび/またはWi-Fi三角測量が使用され得る。他の実装形態において、ユーザによって携帯/操作される他のコンピューティングデバイスによって放射される信号に基づいて特定の人の位置を決定するために、飛行時間、信号強度などの他のワイヤレス信号特徴が、クライアントデバイス110によって、単独または集合的に使用され得る。
【００２９】
追加的または代替的に、いくつかの実装形態において、クライアントデバイス110は、ユーザの声からユーザを認識するために、話者識別(SID)を実行し得る。いくつかの実装形態において、次いで、例えば、クライアントデバイス110の存在センサ113(およびオプションでクライアントデバイス110のGPSセンサ、Soliチップ、および/または加速度計)によって、話者の動きが決定され得る。いくつかの実装形態において、そのような検出された動きに基づいて、ユーザの位置が予測され得、この位置は、ユーザの位置に対するクライアントデバイス110および/または他のコンピューティングデバイスの近接性に少なくとも部分的に基づいて、クライアントデバイス110および/または他のコンピューティングデバイスにおいて任意のコンテンツがレンダリングされるときにユーザの位置であると想定され得る。いくつかの実装形態において、ユーザは、特に、最後の関与からあまり時間が経過していない場合、ユーザが自動アシスタント115と関与した最後の位置にいると単純に想定され得る。
【００３０】
さらに、クライアントデバイス110および/または自然会話システム180は、データ(例えば、ソフトウェアアプリケーション、1つまたは複数のファーストパーティ(1P)エージェント171、1つまたは複数のサードパーティエージェント(3P)172など)を記憶するための1つもしくは複数のメモリ、データにアクセスし、データを実行するための1つもしくは複数のプロセッサ、および/または1つもしくは複数のネットワークインターフェースなどの、ネットワーク199のうちの1つまたは複数を介する通信を容易にする他の構成要素を含み得る。いくつかの実装形態において、ソフトウェアアプリケーション、1Pエージェント171、および/または3Pエージェント172のうちの1つもしくは複数は、クライアントデバイス110においてローカルにインストールされることが可能であるが、他の実装形態において、ソフトウェアアプリケーション、1Pエージェント171、および/または3Pエージェント172のうちの1つもしくは複数は、(例えば、1つまたは複数のサーバによって)リモートでホストされることが可能であり、ネットワーク199のうちの1つまたは複数を介してクライアントデバイス110によってアクセス可能であることが可能である。クライアントデバイス110、他のコンピューティングデバイス、および/または自動アシスタント115によって実行される動作は、複数のコンピューティングデバイスにわたって分散され得る。自動アシスタント115は、例えば、ネットワーク(例えば、図1のネットワーク199のうちの1つまたは複数)を介して互いに結合された1つまたは複数の場所におけるクライアントデバイス110および/または1つもしくは複数のコンピュータ上で実行されるコンピュータプログラムとして実装され得る。
【００３１】
いくつかの実装形態において、自動アシスタント115によって実行される動作は、自動アシスタントクライアント114を介してクライアント110においてローカルに実装され得る。図1に示すように、自動アシスタントクライアント114は、自動音声認識(ASR)エンジン120A1と、自然言語理解(NLU)エンジン130A1と、フルフィルメントエンジン140A1と、テキスト読み上げ(TTS)エンジン150A1とを含み得る。いくつかの実装形態において

、自動アシスタント115によって実行される動作は、図1に示すように自然会話システム180がクライアントデバイス110からリモートで実装される場合などでは、複数のコンピュータシステムにわたって分散され得る。これらの実装形態において、自然会話システム180がクライアントデバイス110からリモートに(例えば、リモートサーバにおいて)実装される実装形態において、自動アシスタント115は、追加的または代替的に、自然会話システム180のASRエンジン120A2、NLUエンジン130A2、フルフィルメントエンジン140A2、およびTTSエンジン150A2を利用し得る。

【００３２】

図2を参照してより詳細に説明するように、これらのエンジンの各々は、1つまたは複数の機能を実行するように構成され得る。例えば、ASRエンジン120A1および/または120A2は、機械学習(ML)モデルデータベース115A内に記憶されたストリーミングASRモデル(例えば、リカレントニューラルネットワーク(RNN)モデル、トランスフォーマーモデル、および/またはASRを実行することができる任意の他のタイプのMLモデル)を使用して、ASR出力のストリームを生成するために、発話の少なくとも一部をキャプチャし、クライアントデバイス110のマイクロフォンによって生成されたオーディオデータのストリームを処理することができる。特に、ストリーミングASRモデルは、オーディオデータのストリームが生成されるときに、ASR出力のストリームを生成するために利用されることが可能である。さらに、NLUエンジン130A1および/または130A2は、MLモデルデータベース115A内に記憶されたNLUモデル(例えば、長短期記憶(LSTM)、ゲート付き回帰型ユニット(GRU)、および/またはNLUを実行することができる任意の他のタイプのRNNもしくは他のMLモデル)および/または文法ベースのNLUルールを使用して、NLU出力のストリームを生成するために、ASR出力のストリームを処理することができる。さらに、フルフィルメントエンジン140A1および/または140A2は、NLU出力のストリームに基づいて生成されたフルフィルメントデータのストリームに基づいて、フルフィルメント出力のセットを生成することができる。フルフィルメントデータのストリームは、例えば、ソフトウェアアプリケーション、1Pエージェント171、および/または3Pエージェント172のうちの1つもしくは複数を使用して生成されることが可能である。最後に、TTSエンジン150A1および/または150A2は、MLモデルデータベース115A内に記憶されたTTSモデルを使用して、テキストデータに対応するコンピュータで生成された合成音声を含む合成音声オーディオデータを生成するために、テキストデータ(例えば、自動アシスタント115によって考案されたテキスト)を処理することができる。特に、MLモデルデータベース115A内に記憶されたMLモデルは、クライアントデバイス110内にローカルに記憶されたオンデバイスMLモデル、またはクライアントデバイス110および/もしくは(例えば、自然会話システムがリモートサーバによって実装される実装形態において)他のシステムの両方にアクセス可能な共有MLモデルであることが可能である。

【００３３】

様々な実装形態において、ASR出力のストリームは、例えば、オーディオデータのストリーム内にキャプチャされたクライアントデバイス110のユーザの発話(またはその1つもしくは複数の部分)に対応すると予測される音声仮説(例えば、用語仮説および/または転写仮説)のストリーム、音声仮説の各々についての1つまたは複数の対応する予測値(例えば、確率、対数尤度、および/または他の値)、オーディオデータのストリーム内にキャプチャされたクライアントデバイス110のユーザの発話に対応すると予測される複数の音素、および/または他のASR出力を含むことができる。それらの実装形態のいくつかの変形例において、ASRエンジン120A1および/または120A2は、(例えば、対応する予測値に基づいて)発話に対応する認識されたテキストとして音声仮説のうちの1つまたは複数を選択することができる。

【００３４】

様々な実装形態において、NLU出力のストリームは、例えば、ASR出力のストリーム内に含まれる用語のうちの1つもしくは複数(例えば、すべて)に対する認識されたテキストの1つまたは複数の注釈を含む注釈付きの認識されたテキストのストリーム、ASR出力のス

トリーム内に含まれる用語のうちの1つもしくは複数(例えば、すべて)に対する認識されたテキストに基づいて決定された1つもしくは複数の予測された意図、ASR出力のストリーム内に含まれる用語のうちの1つもしくは複数(例えば、すべて)に対する認識されたテキストに基づいて決定された1つもしくは複数の予測された意図の各々に関連付けられた対応するパラメータに関する予測されたおよび/もしくは推定されたスロット値、ならびに/または他のNLU出力を含むことができる。例えば、NLUエンジン130A1および/または130A2は、用語にそれらの文法的な役割で注釈を付けるように構成された品詞タガー(図示せず)を含み得る。追加的または代替的に、NLUエンジン130A1および/または130A2は、人(例えば、文学の登場人物、有名人、公人などを含む)、組織、場所(実在および架空)への参照などの、認識されたテキストの1つまたは複数のセグメント内のエンティティ参照に注釈を付けるように構成されたエンティティタガー(図示せず)を含み得る。いくつかの実装形態において、エンティティに関するデータが、ナレッジグラフ(図示せず)などの1つまたは複数のデータベース内に記憶され得る。いくつかの実装形態において、ナレッジグラフは、既知のエンティティ(および場合によってはエンティティ属性)を表すノード、ならびにノードを接続してエンティティ間の関係を表すエッジを含み得る。エンティティタガーは、(例えば、人などのエンティティクラスへのすべての参照の識別を可能にするために)高レベルの粒度および/または(例えば、特定の人などの特定のエンティティへのすべての参照の識別を可能にするために)低レベルの粒度においてエンティティへの参照に注釈を付け得る。エンティティタガーは、特定のエンティティを解決するために自然言語入力のコンテンツに依存し得、および/または特定のエンティティを解決するためにナレッジグラフもしくは他のエンティティデータベースとオプションで通信し得る。

【００３５】

追加的または代替的に、NLUエンジン130A1および/または130A2は、1つまたは複数のコンテキストキューに基づいて、同じエンティティへの参照をグループ化または「クラスタ化」するように構成された共参照リゾルバ(図示せず)を含み得る。例えば、共参照リゾルバは、「それらを買う(buy them)」という入力を受信する直前にレンダリングされたクライアントデバイスの通知において「劇場のチケット(theatre tickets)」が言及されていることに基づいて、「それらを買う(buy them)」という自然言語入力において、「それら(them)」という用語を「劇場のチケットを買う(buy theatre tickets)」に解決するために利用され得る。いくつかの実装形態において、NLUエンジン130A1および/または130A2の1つまたは複数の構成要素は、NLUエンジン130A1および/または130A2の1つまたは複数の他の構成要素からの注釈に依存し得る。例えば、いくつかの実装形態において、エンティティタガーは、特定のエンティティへのすべての言及に注釈を付ける際に、共参照リゾルバからの注釈に依存し得る。また、例えば、いくつかの実装形態において、共参照リゾルバは、同じエンティティへの参照をクラスタ化する際に、エンティティタガーからの注釈に依存し得る。

【００３６】

様々な実装形態において、フルフィルメントデータのストリームは、例えば、ソフトウェアアプリケーション、1Pエージェント171、および/または3Pエージェント172のうちの1つもしくは複数によって生成された1つまたは複数のフルフィルメント出力を含むことができる。NLU出力のストリームに基づいて生成された1つまたは複数の構造化された要求は、ソフトウェアアプリケーション、1Pエージェント171、および/または3Pエージェント172のうちの1つもしくは複数に送信されることが可能であり、ソフトウェアアプリケーション、1Pエージェント171、および/または3Pエージェント172のうちの1つもしくは複数は、構造化された要求のうちの1つまたは複数を受信したことに応答して、発話を満たすと予想されるフルフィルメント出力を送信することができる。フルフィルメントエンジン140A1および/または140A2は、クライアントデバイス110において受信されたフルフィルメント出力を、フルフィルメントデータのストリームに対応するフルフィルメント出力のセット内に含めることができる。特に、フルフィルメントデータのストリームは、クライアントデバイス110のユーザが発話を提供するときに生成されることが可能で

ある。さらに、フルフィルメント出力エンジン164は、フルフィルメント出力のストリームから1つまたは複数のフルフィルメント出力を選択することができ、フルフィルメント出力のうちの選択された1つまたは複数は、発話を満たすためにクライアントデバイス110のユーザに提示するために提供されることが可能である。1つまたは複数のフルフィルメント出力は、例えば、発話に応答すると予測され、スピーカを介してクライアントデバイス110のユーザに提示するために聴覚的にレンダリングされることが可能な可聴コンテンツ、発話に応答すると予測され、ディスプレイを介してクライアントデバイス110のユーザに提示するために視覚的にレンダリングされることが可能な視覚コンテンツ、ならびに/または実行されると、クライアントデバイス110および/もしくは(例えば、ネットワーク199のうちの1つまたは複数を介して)クライアントデバイス110と通信する他のコンピューティングデバイスを発話に応答して制御させるアシスタントコマンドを含むことができる。

【００３７】

図1は、単一のユーザを有する単一のクライアントデバイスに関して説明されているが、それは、例のためのものであり、限定することを意味していないことが理解されるべきである。例えば、ユーザの1つまたは複数の追加のクライアントデバイスも、本明細書で説明する技法を実装することができる。例えば、クライアントデバイス110、1つもしくは複数の追加のクライアントデバイス、および/またはユーザの任意の他のコンピュータデバイスは、本明細書で説明する技法を用いることができるデバイスのエコシステムを形成することができる。これらの追加のクライアントデバイスおよび/またはコンピューティングデバイスは、(例えば、ネットワーク199のうちの1つまたは複数を介して)クライアントデバイス110と通信し得る。別の例として、所与のクライアントデバイスが、共有設定(例えば、ユーザのグループ、世帯、共有の居住空間など)において複数のユーザによって利用されることが可能である。

【００３８】

本明細書で説明するように、自動アシスタント115は、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定することができる。この決定を行う際、自動アシスタントは、自然会話エンジン160を利用することができる。様々な実装形態において、図1に示すように、自然会話エンジン160は、音響エンジン161と、一時停止エンジン162と、時間エンジン163と、自然会話出力エンジン164と、フルフィルメント出力エンジン165とを含むことができる。

【００３９】

いくつかの実装形態において、音響エンジン161は、オーディオデータのストリームを処理することに基づいて、オーディオベースの特徴を決定することができる。それらの実装形態のいくつかの変形例において、音響エンジン161は、MLモデルデータベース115A内に記憶されたオーディオベースのMLモデルを使用して、オーディオベースの特徴を決定するためにオーディオデータのストリームを処理することができる。追加または代替の実装形態において、音響エンジン161は、1つまたは複数のルールを使用して、オーディオベースの特徴を決定するためにオーディオデータのストリームを処理することができる。オーディオベースの特徴は、例えば、オーディオデータのストリーム内にキャプチャされた発話に関連付けられた韻律特性、および/または他のオーディオベースの特徴を含むことができる。韻律特性は、例えば、イントネーション、トーン、ストレス、リズム、テンポ、ピッチ、引き延ばされた音節、休止、休止に関連する文法、および/またはオーディオデータのストリームの処理から導出され得る他のオーディオベースの特徴を含む、音節およびより大きい音声単位の1つまたは複数の特性を含むことができる。さらに、韻律特性は、例えば、感情状態、形式(例えば、声明、質問、または命令)、皮肉、風刺、話す調子、および/または強調の指示を提供することができる。言い換えれば、韻律特性は、所与のユーザの個々の音声特徴に依存せず、個々の発話および/または複数の発話の組み合わせに基づいてダイアログセッション中に動的に決定されることが可能な音声の特徴である。

【００４０】
いくつかの実装形態において、一時停止エンジン162は、クライアントデバイス110のユーザが音声データのストリーム内にキャプチャされた発話の提供を一時停止したか、または発話の提供を完了したかを判定することができる。それらの実装形態のいくつかの変形例において、一時停止エンジン162は、音響エンジン161を使用して決定されたオーディオベースの特徴の処理に基づいて、クライアントデバイスの110のユーザが発話の提供を一時停止したと判定することができる。例えば、一時停止エンジン162は、MLモデルデータベース115A内に記憶されたオーディオベースの分類MLモデルを使用して、出力を生成するためにオーディオベースの特徴を処理し、オーディオベースの分類MLモデルを使用して生成された出力に基づいて、クライアントデバイス110のユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定することができる。出力は、例えば、クライアントデバイス110のユーザが発話の提供を一時停止したか、または発話の提供を完了したかを示す、1つまたは複数の予測測定値(例えば、バイナリ値、対数尤度、確率など)を含むことができる。例えば、クライアントデバイス110のユーザが「Arnolllldに電話をかけて(call Arnolllld's)」の発話を提供すると仮定し、ここで、「llll」は、発話内に含まれる引き延ばされた音節を示す。この例において、オーディオベースの特徴は、発話が引き延ばされた音節を含むことの指示を含む可能性があり、結果として、オーディオベースの分類MLモデルを使用して生成された出力は、ユーザが発話の提供を完了していないことを示す場合がある。

【００４１】
それらの実装形態の追加または代替の変形例において、一時停止エンジン162は、NLUエンジン130A1および/または130A2を使用して生成されたNLUデータのストリームに基づいて、クライアントデバイス110のユーザが発話の提供を一時停止したと判定することができる。例えば、一時停止エンジン162は、予測された意図、ならびに/または予測された意図に関連付けられた対応するパラメータに関する予測および/もしくは推定されたスロット値関する予測されたスロット値に基づいて、クライアントデバイス110のユーザが発話の提供を一時停止したか、または発話の提供を完了したかどうかにかかわらず、オーディオデータのストリームを処理することができる。例えば、クライアントデバイス110のユーザが「Arnolllldに電話をかけて(call Arnolllld's)」の発話を提供すると仮定し、ここで、「llll」は、発話内に含まれる引き延ばされた音節を示す。この例において、NLUデータのストリームは、「電話をかけて(call)」の予測された意図と、「Arnold」のエンティティパラメータに関するスロット値とを含むことができる。しかしながら、この例において、(発話が履行されることが可能であるように)自動アシスタント115がエンティティ「Arnold」に関連付けられた連絡先エントリにアクセスできる場合があっても、自動アシスタント115は、発話を処理することに基づいて決定されたオーディオベースの特徴内に含まれる引き延ばされた音節に基づいて、エンティティ「Arnold」への呼び出しを開始しない場合がある。対照的に、この例において、ユーザが引き延ばされた音節を有する「Arnolllld」を提供しなかった場合、および/またはユーザが自動アシスタント115に発話のフルフィルメントを開始させるための明示的なコマンド(例えば、「いますぐArnoldに電話をかけて(call Arnold now)」、「すぐにArnoldに電話をかけて(call Arnold immediately)」など)を提供した場合、一時停止エンジン162は、クライアントデバイス110のユーザが発話の提供を完了したと判定し得る。

【００４２】
いくつかの実装形態において、自然会話出力エンジン163は、ユーザが発話の提供を一時停止したと判定したことに応答して、クライアントデバイスのユーザに提示するために提供されるべき自然な会話出力を決定することができる。それらの実装形態のいくつかの変形例において、自然会話出力エンジン163は、自然な会話出力のセットを決定することができ、NLUデータのストリームに関連付けられたNLU測定値および/またはオーディオベースの特徴に基づいて、自然な会話出力のセットの中から、ユーザへの提示(例えば、クライアントデバイス110の1つまたは複数のスピーカを介する可聴提示)のために提供され

るべき自然な会話出力のうちの1つまたは複数を(例えば、ランダムに、または自然な会話出力のセットを通して循環的に)選択することができる。それらの実装形態のいくつかのさらなる変形例において、自然な会話出力のスーパーセットが、クライアントデバイス110によってアクセス可能な1つまたは複数のデータベース(図示せず)内に(例えば、(例えば、TTSエンジン150A1および/または150A2を使用して)合成音声オーディオデータに変換されたテキストデータとして、および/または合成音声オーディオデータとして)記憶されることが可能であり、自然な会話出力のセットは、NLUデータのストリームに関連付けられたNLU測定値、および/またはオーディオベースの特徴に基づいて、自然な会話出力のスーパーセットから生成されることが可能である。

【００４３】

これらの自然な会話出力は、発話中のダイアログセッションの促進のために実施されることが可能であるが、必ずしも発話のフルフィルメントとして実施される必要はない。例えば、自然な会話出力は、自動アシスタントの115との対話を継続するという要望の指示をユーザが確認することに対する要求(例えば、「まだそこにいますか(Are you still there?)」など)、ユーザと自動アシスタント115との間のダイアログセッションの促進のためにユーザが追加のユーザ入力を提供することに対する要求(例えば、「誰に電話をかけてほしいですか(Who did you want me to call?)」など)、フィラースピーチ(「うんうん(Mmmhmm)」、「うん(Uh huhh)」、「わかった(Alright)」など)を含むことができる。様々な実装形態において、自然会話エンジン163は、自然な会話出力のセットを生成する際に、MLモデルデータベース115A内に記憶された1つまたは複数の言語モデルを利用することができる。他の実装形態において、自然会話エンジン163は、リモートシステム(例えば、リモートサーバ)から自然な会話出力のセットを取得し、クライアントデバイス110のオンデバイスメモリ内に自然な会話出力のセットを記憶することができる。

【００４４】

いくつかの実装形態において、フルフィルメント出力エンジン164は、ユーザが発話の提供を完了したと判定したことに応答して、または(例えば、図5Cに関して説明するように)ユーザが発話の提供を完了していないが、それにもかかわらず発話が履行されるべきであるという判定に応答して、フルフィルメント出力のストリームから、クライアントデバイスのユーザへの提示のために提供されるべき1つまたは複数のフルフィルメント出力を選択することができる。1Pエージェント171および3Pエージェント172が図1においてネットワーク199のうちの1つまたは複数を介して実装されるものとして示されているが、それは、例のためのものであり、限定することを意味していないことが理解されるべきである。例えば、1Pエージェント171および/または3Pエージェント172のうちの1つまたは複数は、クライアントデバイス110においてローカルに実装されることが可能であり、NLU出力のストリームは、アプリケーションプログラミングインターフェース(API)を介して1Pエージェント171および/または3Pエージェント172のうちの1つまたは複数に送信されることが可能であり、1Pエージェント171および/または3Pエージェント172のうちの1つまたは複数からのフルフィルメント出力は、APIを介して取得され、フルフィルメントデータのストリームに組み込まれることが可能である。追加的または代替的に、1Pエージェント171および/または3Pエージェント172のうちの1つまたは複数は、(例えば、それぞれ、1Pサーバおよび/または3Pサーバにおいて)クライアントデバイス110からリモートで実装されることが可能であり、NLU出力のストリームは、ネットワーク199のうちの1つまたは複数を介して1Pエージェント171および/または3Pエージェント172のうちの1つまたは複数に送信されることが可能であり、1Pエージェント171および/または3Pエージェント172のうちの1つまたは複数からのフルフィルメント出力は、ネットワーク199のうちの1つまたは複数を介して取得され、フルフィルメントデータのストリームに組み込まれることが可能である。

【００４５】

例えば、フルフィルメント出力エンジン164は、NLUデータのストリームに関連付けられたNLU測定値および/またはフルフィルメントデータのストリームに関連付けられたフ

ルフィルメント測定値に基づいて、フルフィルメントデータのストリームから1つまたは複数のフルフィルメント出力を選択することができる。NLU測定値は、例えば、NLUエンジン130A1および/もしくは130A2が、予測された意図がオーディオデータのストリーム内にキャプチャされた発話を提供したユーザの実際の意図に対応することをどの程度確信しているか、ならびに/または予測された意図に関連付けられたパラメータの推定および/もしくは予測されたスロット値が予測された意図に関連付けられたパラメータの実際のスロット値に対応することをどの程度確信しているかを示す、確率、対数尤度、バイナリ値などであることが可能である。NLU測定値は、NLUエンジン130A1および/または130A2がNLU出力のストリームを生成するときに生成されることが可能であり、NLU出力のストリーム内に含められることが可能である。フルフィルメント測定値は、例えば、フルフィルメントエンジン140A1および/または140A2が、予測フルフィルメント出力がユーザの所望のフルフィルメントに対応することをどの程度確信しているかを示す、確率、対数尤度、バイナリ値などであることが可能である。フルフィルメント測定値は、ソフトウェアアプリケーション、1Pエージェント171、および/もしくは3Pエージェント172のうちの1つもしくは複数がフルフィルメント出力を生成するときに生成されることが可能であり、フルフィルメントデータのストリームに組み込まれることが可能であり、ならびに/またはフルフィルメントエンジン140A1および/もしくは140A2がソフトウェアアプリケーション、1Pエージェント171、および/もしくは3Pエージェント172のうちの1つもしくは複数から受信されたフルフィルメントデータを処理するときに生成されることが可能であり、フルフィルメントデータのストリームに組み込まれることが可能である。

【００４６】

いくつかの実装形態において、ユーザが発話の提供を一時停止したと判定したことに応答して、時間エンジン165は、発話の提供の一時停止の持続時間および/または任意の後続の一時停止の持続時間を決定することができる。自動アシスタント115は、クライアントデバイス110のユーザへの提示のために提供されるべき自然な会話出力を選択する際に、自然会話出力エンジン163にこれらの一時停止の持続時間のうちの1つまたは複数を活用させることができる。例えば、クライアントデバイス110のユーザが「Arnolllldに電話をかけて(call Arnolllld's)」の発話を提供すると仮定し、ここで、「llll」は、発話内に含まれる引き延ばされた音節を示す。さらに、ユーザが発話の提供を一時停止したと判定されると仮定する。いくつかの実装形態において、クライアントデバイス110のユーザが発話の提供を一時停止したと判定したことに応答して、(例えば、「うんうん(Mmmhmm)」などを聴覚的にレンダリングすることによって)ユーザへの提示のために自然な会話出力が提供され得る。しかしながら、他の実装形態において、自然な会話出力は、ユーザが最初に一時停止してからしきい値持続時間が経過したと時間エンジン165が判定したことに応答して、ユーザへの提示のために提供され得る。さらに、自然な会話出力が提示のために提供されることに応答して、クライアントデバイス110のユーザが発話の提供を継続しないとさらに仮定する。この例において、ユーザが最初に一時停止してから追加のしきい値持続時間が経過した(または自然な会話出力がユーザへの提示のために提供されてから追加のしきい値持続時間が経過した)と時間エンジン165が判定したことに応答して、ユーザへの提示のために追加の自然な会話出力が提供され得る。したがって、ユーザへの提示のために追加の自然な会話出力を提供する際に、自然会話出力エンジン163は、クライアントデバイス110のユーザが発話を完了することを要求する(例えば、「話していましたか(You were saying?)」、「私はなにか聞き逃しましたか(Did I miss something?)」など)、またはクライアントデバイス110のユーザが予測された意図に関する特定のスロット値を提供することを要求する(例えば、「誰に電話をかけたかったですか(Who did you want to call?)」、「予約は何人分でしたか(And how many people was the reservation for?)」など)異なる自然な会話出力を選択することができる。

【００４７】

様々な実装形態において、自動アシスタント115が、クライアントデバイス110のユーザが発話を完了するのを待っている間、自動アシスタント115は、オプションで、フルフ

ィルメント出力のセット内のフルフィルメント出力を部分的に履行させることができる。例えば、自動アシスタント115は、フルフィルメント出力のセット内に含まれる1つまたは複数のフルフィルメント出力に基づいて、ソフトウェアアプリケーション、1Pエージェント171、3Pエージェント172、および/またはクライアントデバイス110のユーザに関連付けられた他のクライアントデバイス、スマートネットワークデバイスなどの(例えば、ネットワーク199のうちの1つまたは複数を介して)クライアントデバイス110と通信する追加のコンピューティングデバイスのうちの1つまたは複数との接続を確立することができ、合成音声を含む合成音声オーディオデータを生成させる(しかし、聴覚的にレンダリングさせない)ことができ、グラフィカルコンテンツを生成させる(しかし、視覚的にレンダリングさせない)、および/またはフルフィルメント出力のうちの1つもしくは複数の任意の他の部分的なフルフィルメントを実行することができる。結果として、クライアントデバイス110のユーザへの提示のためにフルフィルメント出力を提供させる際の待ち時間が短縮されることが可能である。

【００４８】

ここで図2に進むと、図1の様々な構成要素を使用して本開示の様々な態様を実証する例示的なプロセスフローが示されている。ASRエンジン120A1および/または120A2は、ASR出力のストリーム220を生成するために、MLモデルデータベース115A内に記憶されたストリーミングASRモデルを使用して、オーディオデータのストリーム201Aを処理することができる。NLUエンジン130A1および/または130A2は、NLU出力のストリーム230を生成するために、MLモデルデータベース115A内に記憶されたNLUモデルを使用して、ASR出力のストリーム220を処理することができる。いくつかの実装形態において、NLUエンジン130A1および/または130A2は、追加的または代替的に、NLU出力のストリーム230を生成する際に、非オーディオデータのストリーム201Bを処理することができる。非オーディオデータのストリーム201Bは、クライアントデバイス110の視覚構成要素によって生成された視覚データ、クライアントデバイス110のタッチ感知構成要素を介してユーザによって提供されたタッチ入力のストリーム、クライアントデバイス110のタッチ感知構成要素もしくは周辺デバイス(例えば、マウスおよびキーボード)を介してユーザによって提供されたタイプ入力のストリーム、および/またはクライアントデバイス110の任意の他のユーザインターフェース入力デバイスによって生成された任意の他の非オーディオデータを含むことができる。いくつかの実装形態において、1Pエージェント171は、1Pフルフィルメントデータ240Aを生成するために、NLU出力のストリームを処理することができる。追加または代替の実装形態において、3Pエージェント172は、3Pフルフィルメントデータ240Bを生成するために、NLU出力のストリーム230を処理することができる。フルフィルメントエンジン140A1および/または140A2は、1Pフルフィルメントデータ240Aおよび/または3Pフルフィルメントデータ240B(およびオプションで、NLU出力のストリーム230を処理するクライアントデバイス110においてアクセス可能な1つまたは複数のソフトウェアアプリケーションに基づいて生成された他のフルフィルメントデータ)に基づいて、フルフィルメントデータのストリーム240を生成することができる。さらに、音響エンジン161は、オーディオデータのストリーム201A内に含まれる1つまたは複数の発話(またはその一部)のオーディオベースの特徴261などの、オーディオデータのストリーム201Aに関連付けられたオーディオベースの特徴261を生成するために、オーディオデータのストリーム201Aを処理することができる。

【００４９】

一時停止エンジン162は、ブロック262において示すように、クライアントデバイスのユーザがオーディオデータのストリーム201A内にキャプチャされた発話の提供を一時停止したか、またはオーディオデータのストリーム201A内にキャプチャされた発話の提供を完了したかを判定するために、NLU出力のストリーム230および/またはオーディオベースの特徴261を処理することができる。自動アシスタント115は、ブロック262が、ユーザが発話の提供を一時停止したことを示しているか、または発話の提供を完了したことを示しているかに基づいて、自然な会話出力を提供するか、またはフルフィルメント出力

を提供するかを決定することができる。例えば、自動アシスタント115が、ブロック262における指示に基づいて、ユーザが発話の提供を一時停止したと判定すると仮定する。この例において、自動アシスタント115は、自然会話出力エンジン163に自然な会話出力263を選択させることができ、自動アシスタント115は、自然な会話出力263を、クライアントデバイス110のユーザへの提示のために提供させることができる。対照的に、自動アシスタント115が、ブロック262における指示に基づいて、ユーザが発話の提供を完了したと判定すると仮定する。この例において、自動アシスタント115は、フルフィルメント出力エンジン164に1つまたは複数のフルフィルメント出力264を選択させることができ、自動アシスタント115は、1つまたは複数のフルフィルメント出力264をクライアントデバイス110のユーザへの提示のために提供させることができる。いくつかの実装形態において、自動アシスタント115は、自然な会話出力263をクライアントデバイス110のユーザへの提示のために提供させるか、または1つもしくは複数のフルフィルメント出力264をクライアントデバイス110のユーザへの提示のために提供させるかを決定する際に、時間エンジン165によって決定された1つまたは複数の一時停止の持続時間265を考慮することができる。これらの実装形態において、自然な会話出力263および/または1つもしくは複数のフルフィルメント出力264は、1つまたは複数の一時停止の持続時間に基づいて適応されることが可能である。特定の機能および実施形態について、図1および図2に関して説明したが、それは、例のためのものであり、限定することを意味していないことが理解されるべきである。例えば、追加の機能および実施形態について、図3、図4、図5A～図5E、および図6に関して以下で説明する。

【００５０】

ここで図3に進むと、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定する例示的な方法300を示すフローチャートが示されている。便宜上、方法300の動作について、動作を実行するシステムを参照して説明する。方法300のこのシステムは、コンピューティングデバイス(例えば、図1および図5A～図5Eのクライアントデバイス110、図6のコンピューティングデバイス610、1つもしくは複数のサーバ、ならびに/または他のコンピューティングデバイス)の1つもしくは複数のプロセッサ、メモリ、および/または他の構成要素を含む。さらに、方法300の動作が特定の順序において示されているが、これは、限定することを意味していない。1つまたは複数の動作が、並べ替えられ、省略され、および/または追加され得る。

【００５１】

ブロック352において、システムは、ストリーミングASRモデルを使用して、ASR出力のストリームを生成するために、ユーザの発話の一部を含み、自動アシスタントに向けられたオーディオデータのストリームを処理する。オーディオデータのストリームは、ユーザのクライアントデバイスのマイクロフォンによって、クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントとのダイアログセッション中に生成されることが可能である。いくつかの実装形態において、ユーザが、1つもしくは複数の特定の単語および/もしくは句(例えば、「ねえ、アシスタント(Hey Assistant)」、「アシスタント(Assistant)」などのホットワード)、1つもしくは複数のボタン(例えば、ソフトウェアボタンおよび/またはハードウェアボタン)の作動、検出されると自動アシスタントを呼び出すクライアントデバイスの視覚構成要素によってキャプチャされた、1つもしくは複数のジェスチャを介して、ならびに/または任意の他の手段によって自動アシスタントを呼び出したと判定することに応答して、システムは、オーディオデータのストリームを処理し得る。ブロック354において、システムは、NLU出力のストリームを生成するために、NLUモデルを使用して、ASR出力のストリームを処理する。ブロック356において、システムは、NLU出力のストリームに基づいて、フルフィルメントデータのストリームを生成させる。ブロック358において、システムは、オーディオデータのストリームを処理することに基づいて、オーディオデータ内にキャプチャされた発話の部分に関連付けられたオーディオベースの特徴を決定する。オーディオベースの特徴は、例えば、発話の部分に関

連付けられた1つもしくは複数の韻律特性(例えば、イントネーション、トーン、ストレス、リズム、テンポ、ピッチ、休止、および/または他の韻律特性)、および/またはオーディオデータのストリームを処理することに基づいて決定されることが可能な他のオーディオベースの特徴を含むことができる。ブロック352～358の動作について、(例えば、図1および図2に関して)ここでより詳細に説明する。

【００５２】

ブロック360において、システムは、NLU出力のストリームおよび/またはオーディオデータ内にキャプチャされた発話の部分に関連付けられたオーディオベースの特徴に基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定する。いくつかの実装形態において、システムは、オーディオベースの分類MLモデルを使用して、出力を生成するためにオーディオベースの特徴を処理することができ、システムは、オーディオベースの分類MLモデルを使用して生成された出力に基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定することができる。オーディオベースの分類MLモデルを使用して生成された出力は、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを示す、1つまたは複数の予測測定値(例えば、バイナリ値、確率、対数尤度、および/または他の測定値)を含むことができる。例えば、出力が、ユーザが発話の提供を一時停止したという予測に関連する0.8の第1の確率と、ユーザが発話の提供を完了したという予測に関連する0.6の第2の確率とを含むと仮定する。この例において、システムは、予測測定値に基づいて、ユーザが発話の提供を一時停止したと判定することができる。追加または代替の実装形態において、システムは、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定するために、NLU出力のストリームを処理または分析することができる。例えば、システムが、予測された意図に関連付けられたNLU測定値、ならびに/または予測された意図に関連付けられた対応するパラメータに関する推定および/もしくは予測されたスロット値がNLU測定値しきい値を満たさないと判定した場合、またはシステムが、予測された意図に関連付けられた対応するパラメータに関するスロット値が未知であると判定した場合、自動アシスタントは、ユーザが発話の提供を一時停止したと判定し得る。特に、様々な実装形態において、システムは、オーディオベースの特徴とNLUデータのストリームの両方に基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定することができる。例えば、システムが、NLUデータのストリームに基づいて、発話が履行されることが可能であるが、オーディオベースの特徴が、ユーザが発話の提供を一時停止したことを示すことを判定する場合、ユーザによって提供され得る発話の任意の追加部分が、ユーザがどのように発話が履行されることを望むかを変更する場合があるので、システムは、ユーザが発話の提供を一時停止したと判定し得る。

【００５３】

ブロック360の反復において、システムが、ユーザが発話の提供を完了したと判定した場合、システムは、ブロック362に進むことができる。ブロック362において、システムは、自動アシスタントに発話のフルフィルメントを開始させる。例えば、システムは、フルフィルメントデータのストリームから、発話を満たすと予測される1つまたは複数のフルフィルメント出力を選択し、1つまたは複数のフルフィルメント出力を、クライアントデバイスまたはクライアントデバイスと通信する追加のコンピューティングデバイスを介してユーザへの提示のために提供させることができる。図1に関して上述したように、1つまたは複数のフルフィルメント出力は、例えば、発話に応答すると予測され、スピーカを介してクライアントデバイスのユーザに提示するために聴覚的にレンダリングされることが可能な可聴コンテンツ、発話に応答すると予測され、ディスプレイを介してクライアントデバイスのユーザに提示するために視覚的にレンダリングされることが可能な視覚コンテンツ、ならびに/または実行されると、クライアントデバイスおよび/もしくはクライアントデバイスと通信する他のコンピューティングデバイスを発話に応答して制御させるアシスタントコマンドを含むことができる。システムは、ブロック352に戻り、図3の方法300の追加の反復を実行することができる。

【００５４】
ブロック360の反復において、システムが、ユーザが発話の提供を一時停止したと判定した場合、システムは、ブロック364に進むことができる。ブロック364において、システムは、ユーザへの可聴提示のために提供されるべき自然な会話出力を決定する。さらに、ブロック366において、システムは、ブロック366において、システムは、自然な会話出力をユーザへの可聴提示のために提供させることができる。自然な会話出力は、NLUデータのストリームに関連付けられたNLU測定値および/またはオーディオベースの特徴に基づいて、クライアントデバイスのオンデバイスメモリ内に記憶された自然な会話出力のセットの中から選択され得る。いくつかの実装形態において、自然な会話出力のセット内に含まれる自然な会話出力のうちの1つまたは複数は、テキストデータに対応することができる。これらの実装形態において、選択された自然な会話出力に関連付けられたテキストデータは、選択された自然な会話出力に対応する合成音声を含む合成音声オーディオデータを生成するために、TTSモデルを使用して処理されることが可能であり、合成音声オーディオデータは、クライアントデバイスまたは追加のコンピューティングデバイスのスピーカを介してユーザへの提示のために聴覚的にレンダリングされることが可能である。
【００５５】
追加または代替の実装形態において、自然な会話出力のセット内に含まれる自然な会話出力のうちの1つまたは複数は、選択された自然な会話出力に対応する合成音声を含む合成音声オーディオデータに対応することができ、合成音声オーディオデータは、クライアントデバイスまたは追加のコンピューティングデバイスのスピーカを介してユーザへの提示のために聴覚的にレンダリングされることが可能である。特に、様々な実装形態において、ユーザへの可聴提示のために自然な会話出力を提供する際、自然な会話出力がユーザに対して再生される音量は、ユーザへの提示のために聴覚的にレンダリングされる他の出力よりも低い音量にすることができる。さらに、様々な実装形態において、ユーザへの可聴提示のために自然な会話出力を提供する際、1つまたは複数の自動アシスタント構成要素は、自動アシスタントがオーディオデータのストリームを処理し続けることを可能にするために、自然な会話出力がユーザへの可聴提示のために提供されている間、アクティブのままであることができる(例えば、ASRエンジン120A1および/もしくは120A2、NLUエンジン130A1および/もしくは130A2、ならびに/またはフルフィルメントエンジン140A1および/もしくは140A2)。
【００５６】
ブロック368において、システムは、自然な会話出力をユーザへの可聴提示のために提供させることに続いて、発話を履行するかどうかを決定する。いくつかの実装形態において、システムは、自然な会話出力をユーザへの可聴提示のために提供させることに続いて、ユーザが発話の提供を完了したと判定したことに応答して、発話を履行することを決定することができる。これらの実装形態において、ASR出力のストリーム、NLU出力のストリーム、およびフルフィルメントデータのストリームは、ユーザが発話の提供を完了したことに基づいて更新されることが可能である。追加または代替の実装形態において、システムは、(例えば、図5Cに関してより詳細に説明するように)自動アシスタントに発話のフルフィルメントを開始させることに関連する1つまたは複数のコストに基づいて、ユーザが発話の提供を完了しなかった場合でも、発話の部分に基づいて発話が履行されることが可能であると判定したことに応答して、発話を履行することを決定することができる。
【００５７】
ブロック368の反復において、システムが、ユーザへの可聴提示のために自然な会話出力を提供させることに続いて、発話を履行すると決定した場合、システムは、上記で説明したように自動アシスタントに発話のフルフィルメントを開始させるために、ブロック362に進む。ブロック368の反復において、システムが、ユーザへの可聴提示のために自然な会話出力を提供させることに続いて、発話を履行しないと決定した場合、システムは、ブロック364に戻る。ブロック364のこの後続の反復において、システムは、ユーザへの可聴提示のために提供されるべき追加の自然な会話出力を決定することができる。特に、

ブロック364のこの後続の反復において選択されたユーザへの可聴提示のために提供されるべき追加の会話出力は、ブロック364の以前の反復において選択されたユーザへの可聴提示のために提供されるべき自然な会話出力とは異なり得る。例えば、ブロック364の以前の反復において選択されたユーザへの可聴提示のために提供されるべき自然な会話出力は、自動アシスタントがまだ聞いており、ユーザが発話の提供を完了するのを待っていることの指示としてユーザに提供され得る(例えば、「うんうん(Mmhmm)」、「わかりました(Okay)」、「うん(Uh huhhh)」など)。しかしながら、ブロック364のこの後続の反復において選択されたユーザへの可聴提示のために提供されるべき自然な会話出力も、自動アシスタントがまだ聞いており、ユーザが発話を完了するのを待っていることの指示であるが、また、ユーザが発話を完了するか、または特定の入力を提供することをより明示的に促す指示としてユーザに提供され得る(例えば、「まだそこにいますか(Are you still there?)」、「予約は何人分でしたか(And how many people was the reservation for?)」など)。システムがブロック368の反復において発話を履行することを決定し、システムが上記で説明したように自動アシスタントに発話のフルフィルメントを開始させるためにブロック362に進むまで、システムは、ブロック364～368の反復を実行し続けることができる。

【００５８】

様々な実装形態において、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを示す1つまたは複数の予測測定値は、ユーザへの可聴提示のために自然な会話出力を提供するかどうか、および/またはいつ提供するかを決定する際に利用されることが可能である。例えば、オーディオベースの分類MLモデルを使用して生成された出力が、ユーザが発話の提供を一時停止したという予測に関連する0.8の第1の確率と、ユーザが発話の提供を完了したという予測に関連する0.6の第2の確率とを含むと仮定する。さらに、0.8の第1の確率が、ユーザが発話の提供を一時停止したことをシステムが強く確信していることを示す一時停止しきい値を満たすと仮定する。したがって、ブロック364の第1の反復において、システムは、音声バックチャンネルを自然な会話出力として利用させることができる(例えば、「うん(uh huh)」)。さらに、ブロック364の第2の反復において、システムは、ユーザが発話の提供を一時停止したことを強く確信しているので、システムは、別の音声バックチャンネルを自然な会話出力として利用させることができる(例えば、「うんうん(Mmmhmm)」または「私はここにいます(I'm here)」)。対照的に、オーディオベースの分類MLモデルを使用して生成された出力が、ユーザが発話の提供を一時停止したという予測に関連する0.5の第1の確率と、ユーザが発話の提供を完了したという予測に関連する0.4の第2の確率とを含むと仮定する。さらに、0.5の第1の確率が、ユーザが発話の提供を一時停止したことをシステムが強く確信していることを示す一時停止しきい値を満たさないと仮定する。したがって、ブロック364の第1の反復において、システムは、音声バックチャンネルを自然な会話出力として利用させることができる(例えば、「うん(uh huh)」)。しかしながら、ブロック364の第2の反復において、別の音声バックチャンネルの訥弁を自然な会話出力として利用させるのではなく、システムは、発話の処理に基づいて予測される予測された意図をユーザが確認することを要求し得る(例えば、「誰かに電話をかけたかったですか(Did you want to call someone?)」)。特に、ユーザへの可聴提示のために提供されるべき自然な会話出力を決定する際に、システムは、自然な会話出力のセットの中から、ユーザへの可聴提示のために提供されるべき所与の自然な会話出力をランダムに選択することができ、ユーザへの可聴提示のために提供されるべき所与の自然な会話出力を選択する際に自然な会話出力のセットを通して循環させることができ、または任意の他の方法でユーザへの可聴提示のために提供されるべき自然な会話出力を決定することができる。

【００５９】

図3は、ユーザへの可聴提示のために自然な会話出力を提供させる際の任意の時間的側面を考慮せずに本明細書で説明されているが、それは、例のためのものであることが理解されるべきである。様々な実装形態において、図4に関して以下で説明するように、シス

テムは、時間の様々なしきい値に基づいて、自然な会話出力のインスタンスのみをユーザへの可聴提示のために提供させ得る。例えば、図3の方法300において、システムは、ユーザが発話の提供を一時停止してから第1のしきい値持続時間が経過したと判定したことに応答して、自然な会話出力の初期インスタンスをユーザへの可聴提示のために提供させる得る。さらに、図3の方法300において、システムは、自然な会話出力の初期インスタンスがユーザへの可聴提示のために提供されてから第2のしきい値持続時間が経過したと判定したことに応答して、自然な会話出力の後続のインスタンスをユーザへの可聴提示のために提供させ得る。この例において、第1のしきい値持続時間および第2のしきい値持続時間は、同じまたは異なる場合があり、任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)に対応し得る。

【００６０】

ここで、図4に進むと、ユーザが発話の提供を一時停止したと判定したことに応答して自然な会話出力をユーザへの提示のために提供させるかどうかを判定する、および/または発話をいつ履行するかを決定する別の例示的な方法400を示すフローチャートが示されている。便宜上、方法400の動作について、動作を実行するシステムを参照して説明する。方法400のこのシステムは、コンピューティングデバイス(例えば、図1および図5A～図5Eのクライアントデバイス110、図6のコンピューティングデバイス610、1つもしくは複数のサーバ、および/または他のコンピューティングデバイス)の1つもしくは複数のプロセッサ、メモリ、および/または他の構成要素を含む。さらに、方法400の動作が特定の順序において示されているが、これは、限定することを意味していない。1つまたは複数の動作が、並べ替えられ、省略され、および/または追加され得る。

【００６１】

ブロック452において、システムは、ユーザの発話の一部を含み、自動アシスタントに向けられたオーディオデータのストリームを受信する。オーディオデータのストリームは、ユーザのクライントデバイスのマイクロフォンによって、クライントデバイスにおいて少なくとも部分的に実装された自動アシスタントとのダイアログセッション中に生成されることが可能である。ブロック454において、システムは、オーディオデータのストリームを処理する。システムは、図3の方法300の動作ブロック352～358に関して上記で説明したのと同じまたは同様の方法においてオーディオデータのストリームを処理することができる。

【００６２】

ブロック456において、システムは、NLU出力のストリーム、および/またはブロック454において発話を処理することに基づいて決定されたオーディオデータ内にキャプチャされた発話の部分に関連付けられたオーディオベースの特徴に基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定する。システムは、図3の方法300のブロック360の動作に関して説明したのと同じまたは同様の方法においてこの判定を行うことができる。ブロック456の反復において、システムが、ユーザが発話の提供を完了したと判定した場合、システムは、ブロック458に進むことができる。ブロック458において、図3の方法300のブロック360の動作に関して説明したのと同じまたは同様の方法において、自動アシスタントに発話のフルフィルメントを開始させる。システムは、ブロック452に戻り、図4の方法400の追加の反復を実行する。ブロック456の反復において、システムが、ユーザが発話の提供を一時停止したと判定した場合、システムは、ブロック460に進むことができる。

【００６３】

ブロック460において、システムは、発話を提供する際のユーザの一時停止がNしきい値を満たすかどうかを判定し、ここで、Nは、任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。ブロック460の反復において、システムが、発話を提供する際のユーザの一時停止がNしきい値を満たさないと判定した場合、システムは、ブロック454に戻り、オーディオデータのストリームを処理し続ける。ブロック460の反復において、システムが、発話を提供する際のユーザの一時停止がNしきい値を満たすと

判定した場合、システムは、ブロック460に進む。ブロック462において、システムは、ユーザへの可聴提示のために提供されるべき自然言語会話出力を決定する。ブロック464において、システムは、自然な会話出力をユーザへの可聴提示のために提供させる。システムは、それぞれ、図3の方法300のブロック364および366の動作に関して上記で説明したのと同じまたは同様の方法において、ブロック462および464の動作を実行することができる。別の言い方をすれば、図4の方法400の1つまたは複数の態様を利用する実装形態において、図3の方法300とは対照的に、システムは、ユーザが発話の提供を最初に一時停止した後、自然な会話出力をユーザへの可聴提示のために提供させる前に、N秒待機し得る。

【００６４】

ブロック466において、システムは、自然な会話出力をユーザへの可聴提示のために提供させた後の、発話を提供する際のユーザの一時停止がMしきい値を満たすかどうかを判定し、ここで、Mは、任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。ブロック466の反復において、システムが、発話を提供する際のユーザの一時停止がMしきい値を満たすと判定した場合、システムは、ブロック462に戻る。図3に関する上記の説明と同様に、ブロック462のこの後続の反復において、システムは、ユーザへの可聴提示のために提供されるべき追加の自然な会話出力を決定することができ、ブロック462のこの後続の反復において選択されたユーザへの可聴提示のために提供されるべき追加の会話出力は、ブロック462の以前の反復において選択されたユーザへの可聴提示のために提供されるべき追加の自然な会話出力とは異なる場合がある。別の言い方をすれば、システムは、ユーザに発話の提供を完了するように促すために、ブロック462の以前の反復において選択されたユーザへの可聴提示のために提供されるべき自然な会話出力を決定することができ、一方、システムは、ユーザに発話の提供を完了するように明示的に要求するために、ブロック462の後続の反復において選択されたユーザへの可聴提示のために提示されるべき追加の自然な会話出力を決定することができる。ブロック466の反復において、システムは、発話を提供する際のユーザの一時停止がMしきい値を満たさないと判定した場合、ブロック468に進む。

【００６５】

ブロック468において、システムは、自然な会話出力をユーザへの可聴提示のために提供させた後、発話を履行するかどうかを判定する。いくつかの実装形態において、システムは、自然な会話出力(および/または任意の追加の自然な会話出力)をユーザへの可聴提示のために提供させた後に、ユーザが発話の提供を完了したと判定したことに応答して、発話を履行することを決定することができる。これらの実装形態において、ASR出力のストリーム、NLU出力のストリーム、およびフルフィルメントデータのストリームは、ユーザが発話の提供を完了したことに基づいて更新されることが可能である。追加または代替の実装形態において、システムは、(例えば、図5Cに関してより詳細に説明するように)自動アシスタントに発話のフルフィルメントを開始させることに関連する1つまたは複数のコストに基づいて、ユーザが発話の提供を完了しなかった場合でも、発話の部分に基づいて発話が履行されることが可能であると判定したことに応答して、発話を履行することを決定することができる。

【００６６】

ブロック468の反復において、システムが、ユーザへの可聴提示のために自然な会話出力を提供させることに続いて、発話を履行すると決定した場合、システムは、上記で説明したように自動アシスタントに発話のフルフィルメントを開始させるために、ブロック458に進む。ブロック468の反復において、システムが、ユーザへの可聴提示のために自然な会話出力(および/または任意の追加の自然な会話出力)を提供させることに続いて、発話を履行しないと決定した場合、システムは、ブロック462に戻る。ブロック462の後続の反復については、上記で説明した。システムがブロック468の反復において発話を履行することを決定し、システムが上記で説明したように自動アシスタントに発話のフルフィルメントを開始させるためにブロック458に進むまで、システムは、ブロック462～468の

反復を実行し続けることができる。

【００６７】

ここで図5A～図5Eに進むと、ユーザが発話の提供を一時停止したと判定すること、および/または発話をいつ履行するかを決定することに応答して、自然な会話出力をユーザへの提示のために提供させるかどうかを判定する様々な非限定な例が示されている。自動アシスタントは、クライントデバイス110(例えば、図1に関して説明した自動アシスタント115)において少なくとも部分的に実装されることが可能である。自動アシスタントは、自動アシスタントとクライントデバイス110のユーザ101との間のダイアログセッションの促進のために実装されるべき自然な会話出力および/またはフルフィルメント出力を決定するために、自然会話システム(例えば、図1に関して説明した自然会話システム180)を利用することができる。図5A～図5Eに示すクライントデバイス110は、例えば、発話および/または他の可聴入力に基づいてオーディオデータを生成するためのマイクロフォン、合成音声および/または他の可聴出力を聴覚的にレンダリングするためのスピーカ、ならびにタッチ入力を受信するためおよび/または転写および/もしくは他の視覚的出力を視覚的にレンダリングするためのディスプレイ190を含む、様々なユーザインターフェース構成要素を含み得る。図5A～図5Eに示すクライントデバイス110は、ディスプレイ190を有するスタンドアロンのインタラクティブスピーカであるが、それは、例のためのものであり、限定することを意味していないことが理解されるべきである。

【００６８】

例えば、特に図5Aを参照すると、クライントデバイス110のユーザ101が、「アシスタント、Arnolllldに電話をかけて(Assistant, call Arnolllld's)」の発話552A1を提供し、次いで、552A2によって示されるようにN秒間一時停止すると仮定し、ここで、Nは、任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。この例において、自動アシスタントは、ASRモデルを使用して、ASR出力のストリームを生成するために、発話552A1をキャプチャするオーディオデータのストリームと、552A2によって示される一時停止とを処理させることができる。さらに、自動アシスタントは、NLU出力のストリームを生成するために、NLUモデルを使用して、ASR出力のストリームを処理させることができる。さらに、自動アシスタントは、クライントデバイス110においてアクセス可能なソフトウェアアプリケーション、クライントデバイス110においてアクセス可能な1Pエージェント、および/またはクライントデバイス110においてアクセス可能な3Pエージェントを使用して、NLU出力のストリームに基づいて、フルフィルメントデータのストリームを生成させることができる。この例において、発話552A1を処理することに基づいて、ASR出力が、オーディオデータのストリーム内にキャプチャされた発話552A1に対応する認識されたテキスト(例えば、「Arnoldに電話をかけて(call Arnold's)」に対応する認識されたテキスト)を含み、NLUデータのストリームが、予測された「かけて(call)」または「電話をかけて(phone call)」の意図に関連する着信者エンティティパラメータに関する「Arnold」のスロット値を有する予測された「かけて(call)」または「電話をかけて(phone call)」という意図を含み、フルフィルメントデータのストリームは、クライアントデバイス110に、実行されると、「Arnold」という名前のユーザ101の友人に関連付けられた連絡先エントリとの通話を開始させるアシスタントコマンドを含むと仮定する。したがって、発話552A1を処理することに基づいて、かつ任意の追加の発話を処理することなく、自動アシスタントは、アシスタントコマンドを実行させることによって発話552A1が満たされることが可能であると判定し得る。しかしながら、自動アシスタントが、発話552A1が履行されることが可能であると判定し得るとしても、自動アシスタントは、発話のフルフィルメントを開始することを控え得る。

【００６９】

いくつかの実装形態において、自動アシスタントは、発話552A1に関連付けられたオーディオベースの特徴を決定するために、オーディオベースのMLモデルを使用して、オーディオデータのストリームを処理させることができる。さらに、自動アシスタントは、ユーザが発話552A1の提供を一時停止したか、または発話の提供を完了したかを示す出力を生

成するために、オーディオベースの分類MLモデルを使用して、オーディオベースの特徴を処理させることができる。図5Aの例において、オーディオベースの分類MLモデルを使用して生成された出力が、(例えば、ユーザが「Arnolllld(Arnolllld's)」内の引き延ばされた文節を提供することによって示されるように)ユーザ101が発話552A1の提供を一時停止したことを示すと仮定する。したがって、この例において、自動アシスタントは、発話552A1のオーディオベースの特徴に少なくとも部分的に基づいて、発話552A1の履行を開始することを控え得る。

【００７０】

追加または代替の実装形態において、自動アシスタントは、発話552A1の履行に関連する1つまたは複数の計算コストを決定することができる。1つまたは複数の計算コストは、例えば、発話552A1のフルフィルメントを実行することに関連する計算コスト、発話552A1の実行された履行を取り消すことに関連する計算コスト、および/または他の計算コストを含むことができる。図5Aの例において、発話552A1のフルフィルメントを実行することに関連する計算コストは、少なくとも、「Arnold」に関連付けられた連絡先エントリとの通話を開始すること、および/または他のコストを含むことができる。さらに、発話552A1の実行されたフルフィルメントを取り消すことに関連する計算コストは、少なくとも、「Arnold」に関連付けられた連絡先エントリとの通話を終了すること、ユーザ101とのダイアログセッションを再開始すること、追加の発話を処理すること、および/または他のコストを含むことができる。したがって、この例において、自動アシスタントは、少なくとも、発話552A1を早まって履行することに関連する計算コストが比較的高いことに基づいて、発話552A1のフルフィルメントを開始することを控え得る。

【００７１】

結果として、自動アシスタントは、クライントデバイス110のスピーカを介するユーザ101への可聴提示のために(および、オプションで、552A2によって示されるように、発話552A1を提供した後、ユーザ101がN秒間一時停止したと判定したことに応答して)、図5Aに示すように「うんうん(Mmhmm)」などの自然な会話出力554Aを提供することを決定し得る。自然な会話出力554Aは、自動アシスタントがまだ聞いており、ユーザ101が発話552A1の提供を完了するのを待っていることの指示を提供するために、ユーザ101への可聴提示のために提供されることが可能である。特に、様々な実装形態において、自動アシスタントがユーザ101への提示のために自然な会話出力554Aを提供する間、オーディオデータのストリームを処理する際に利用される自動アシスタント構成要素(例えば、ASRエンジン120A1および/もしくは120A2、NLUエンジン130A1および/もしくは130A2、フルフィルメントエンジン140A1および/もしくは140A2、ならびに/または図1の音響エンジン161などの図1の他の自動アシスタント構成要素)が、クライントデバイス110においてアクティブのままであることができる。さらに、様々な実装形態において、自然な会話出力554Aは、ユーザ101が発話552A1を完了するのを妨げることを回避し、実際の人間間のより自然な会話を反映するために、他の可聴出力よりも低い音量においてユーザ101への可聴提示のために提供されることが可能である。

【００７２】

図5Aの例において、ユーザ101が「Arnoldのトラットリアに電話をかけて(Call Arnold's Trattoria)」の発話556Aを提供することによって発話552A1を完了したとさらに仮定し、ここで、「Arnoldのトラットリア(Arnold's Trattoria)」は、架空のイタリア料理レストランである。ユーザ101が発話556Aを提供することによって発話552A1を完了したことに基づいて、自動アシスタントは、ASR出力のストリーム、NLU出力のストリーム、およびフルフィルメントデータのストリームを更新させることができる。特に、自動アシスタントは、NLUデータの更新されたストリームが、予測された「かけて(call)」または「電話をかけて(phone call)」の意図を依然として含むが、以前に予測されたように、予測された「かけて(call)」または「電話をかけて(phone call)」に関連付けられた着信者エンティティパラメータに関して、「Arnold」ではなく、「Arnoldのトラットリア(Arnold's Trattoria)」のスロット値を有すると判定することができる。したがって、自

動アシスタントは、ユーザ101が発話556Aを提供することによって発話552A1を完了したことに応答して、クライントデバイス101(または、クライントデバイス101と通信する追加のクライントデバイス(例えば、ユーザ101に関連付けられたモバイルデバイス))に、「Arnoldのトラットリア(Arnold's Trattoria)」との通話を開始させ、オプションで、「わかりました、Arnoldのトラットリアに電話をかけています(Okay, calling Arnold's Trattoria)」の合成音声558Aをユーザ101への可聴提示のために提供させることができる。これらおよび他の方法において、自動アシスタントは、(例えば、連絡先エントリ「Arnold」に電話をかけることによって)発話552A1に基づいて決定されたユーザ101の予測された意図を誤って早まって履行することを控えることができ、(例えば、架空のレストラン「Arnoldのトラットリア(Arnold's Trattoria)」に電話をかけることによって)ユーザ101が発話556Aを介して発話552A1を完了したことに基づいて決定されたユーザ101の予測された意図を正しく履行するために、ユーザ101が自分の思考を完了するのを待つことができる。

【００７３】

別の例として、特に図5Bを参照すると、クライントデバイス110のユーザ101が、「アシスタント、Arnolllldに電話をかけて(Assistant, call Arnolllld's)」の発話552B1を提供し、次いで、552B2によって示されるようにN秒間一時停止すると再び仮定し、ここで、Nは、任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。図5Aと同様に、自動アシスタントが、発話552B1が履行されることが可能であると判定し得るとしても、自動アシスタントは、発話552B1に関連付けられたオーディオベースの特徴に基づいて、ならびに/または発話552B1のフルフィルメントを実行することおよび/もしくは発話552B1のフルフィルメントを取り消すことに関連する1つもしくは複数の計算コストに基づいて、発話552B1のフルフィルメントを開始することを控え得る。自動アシスタントが、図5Bに示すように「うんうん(Mmhmm)」などの自然な会話出力554B1を提供することを決定し、自然な会話出力554B1をクライアントデバイス110のユーザ101への可聴提示のために提供させるとさらに仮定する。しかしながら、図5Bの例において、図5Aの例とは対照的に、クライアントデバイス110のユーザ101が、554B2によって示されるように、自然な会話出力554B1をユーザ101への可聴提示のために提供させてからM秒以内に発話554B1を完了しなかったと仮定し、ここで、Mは、552B2によって示されるようなN秒とは同じまたは異なる場合がある任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。

【００７４】

結果として、図5Bの例において、自動アシスタントは、クライアントデバイス110のユーザ101への可聴提示のために提供されるべき追加の自然な会話出力556Bを決定することができる。特に、自動アシスタントが、ユーザ101が発話552B1を完了するのを待っていることを示す自然な会話出力554B1と同様に、音声バックチャンネルをクライアントデバイス110のユーザ101への可聴提示のために提供させるのではなく、追加の自然な会話出力556Bは、自動アシスタントが、ユーザ101が発話552B1を完了するのを待っていることをより明示的に示すこと、および/または(例えば、図5Cに関して以下で説明するように)ユーザ101がダイアログセッションの促進のために特定の入力を提供することを要求することができる。図5Bの例において、追加の自然な会話556Bがユーザ101への可聴提示のために提供されたことに応答して、クライアントデバイス110のユーザ101が、発話552B1の提供を完了するために、「すいません、Arnoldのトラットリアに電話をかけてください(Sorry, call Arnold's Trattoria)」の発話558Bを提供するとさらに仮定する。したがって、自動アシスタントは、ユーザ101が発話558Bを提供することによって発話552B1の提供を完了したことに応答して、クライアントデバイス110(またはクライアントデバイス110と通信する追加のクライアントデバイス(例えば、ユーザ101のモバイルデバイス))に「Arnoldのトラットリア(Arnold's Trattoria)」との通話を開始させ、オプションで、「わかりました、Arnoldのトラットリアに電話をかけています(Okay, calling Arnold's Trattoria)」の合成音声560Bをユーザ101への可聴提示のために提供させることが

できる。図5Bと同様に、自動アシスタントは、(例えば、連絡先エントリ「Arnold」に電話をかけることによって)発話552B1に基づいて決定されたユーザ101の予測された意図を誤って早まって履行することを控えることができ、ユーザ101が図5Bの例におけるようにより長い持続時間の間一時停止し得る場合であっても、(例えば、架空のレストラン「Arnoldのトラットリア(Arnold's Trattoria)」に電話をかけることによって)発話558Bを介して発話552B1を完了する際のユーザ101の予測された意図を正しく履行するために、ユーザ101が自分の思考を完了するのを待つことができる。

【００７５】

さらに別の例として、特に図5Cを参照すると、クライアントデバイス110のユーザ101が、「アシスタント、今夜Arnoldのトラットリアを6人のために予約してください(Assistant, make a reservation tonight at Arnold's Trattoria for six people)」の発話552C1を提供し、次いで、552C2によって示されるように、N秒間一時停止すると仮定し、ここで、Nは、任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。この例において、発話552C1を処理したことに基づいて、ASR出力が、オーディオデータのストリーム内にキャプチャされた発話552C1に対応する認識されたテキスト(例えば、「今夜Arnoldのトラットリアを6人のために予約してください(make a reservation tonight at Arnold's Trattoria for six people)」に対応する認識されたテキスト)を含み、NLUデータのストリームが、予測された「予約(reservation)」または「レストランの予約(restaurant reservation)」の意図に関連付けられたレストランエンティティパラメータに関する予測された「Arnoldのトラットリア(Arnold's Trattoria)」のスロット値と、予測された「予約(reservation)」または「レストランの予約(restaurant reservation)」の意図に関連付けられた予約日パラメータに関する「[今日の日付]」のスロット値と、予測された「予約(reservation)」または「レストランの予約(restaurant reservation)」の意図に関連付けられた人数パラメータに関する「6」のスロット値とを有する予測された「予約(reservation)」または「レストランの予約(restaurant reservation)」の意図を含むと仮定する。特に、発話552C1を提供する際、クライアントデバイス110のユーザ101は、「予約(reservation)」または「レストランの予約(restaurant reservation)」の意図に関連付けられた時間パラメータに関するスロット値を提供しなかった。結果として、NLUデータのストリームに基づいて、自動アシスタントは、ユーザ101が発話552C1の提供を一時停止したと判定し得る。

【００７６】

さらに、フルフィルメントデータのストリームが、実行されると、クライアントデバイス110においてアクセス可能なレストラン予約ソフトウェアアプリケーションおよび/またはクライアントデバイス110においてアクセス可能なレストラン予約エージェント(例えば、図1の1Pエージェント171および/または3Pエージェントのうちの1つ)を使用してクライアントデバイス110にレストランの予約を行わせるアシスタントコマンドを含むと仮定する。図5Cの例において、図5Aおよび図5Bの例とは対照的に、発話552C1を処理することに基づいて、かつ任意の追加の発話を処理することなく、自動アシスタントは、アシスタントコマンドを実行させることによって発話552C1が満たされることが可能であると判定し得る。この例において、自動アシスタントは、ユーザ101がレストランの予約を行うことを意図しているが、単に「予約(reservation)」または「レストランの予約(restaurant reservation)」の意図に関連付けられた時間パラメータに関するスロット値を提供しなかったことを示すNLUデータのストリームに関連付けられたNLU測定値に基づいて、発話552C1のフルフィルメントを開始し得る。したがって、自動アシスタントは、クライアントデバイス110においてアクセス可能なレストラン予約ソフトウェアアプリケーションおよび/またはクライアントデバイス110においてアクセス可能なレストラン予約エージェント(例えば、図1の1Pエージェント171および/または3Pエージェントのうちの1つ)との接続を確立し、発話552C1のフルフィルメントが完全には実行されることが可能ではないにもかかわらず、予約を行うことを開始するためにスロット値の提供を開始することができる。

【００７７】
特に、自動アシスタントが発話552C1のフルフィルメントを開始するとき、自動アシスタントは、少なくともNLUデータのストリームに基づいて、ユーザ101が発話552C1の提供を一時停止したと判定したので、自動アシスタントは、依然として、図5Cに示すような「うん(Uh huhh)」などの自然な会話出力554C1を提供することを決定し、自然な会話出力554C1をクライアントデバイス110のユーザ110への可聴提示のために提供させることができる。しかしながら、図5Cの例において、図5Bと同様に、クライアントデバイス110のユーザ101が、554C2によって示されるように、自然な会話出力554C1をユーザ101への可聴提示のために提供してからM秒以内に発話552C1を完了しなかったと仮定し、ここで、Mは、552C2によって示されるようなN秒とは同じまたは異なる場合がある任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。
【００７８】
結果として、図5Cの例において、自動アシスタントは、クライアントデバイス110のユーザ101への可聴提示のために提供されるべき追加の自然な会話出力556Cを決定することができる。特に、自動アシスタントが、ユーザ101が発話552C1を完了するのを待っていることを示す自然な会話出力554C1と同様に、音声バックチャンネルをクライアントデバイス110のユーザ101への可聴提示のために提供させるのではなく、追加の自然な会話出力556Cは、ユーザ101が「予約(reservation)」または「レストランの予約(restaurant reservation)」の意図に関連付けられた時間パラメータに関するスロット値を提供しなかったことに基づいて、「何時にですか(For what time?)」など、ユーザ101がダイアログセッションの促進のために特定の入力を提供することを要求することができる。さらに、図5Cの例において、追加の自然な会話556Cがユーザへの可聴提示のために提供されたことに応答して、クライアントデバイス110のユーザ101は、発話552C1の提供を完了するために、「午後7時(7:00PM)」の発話558Cを提供すると仮定する。したがって、自動アシスタントは、ユーザ101が発話558Cを提供することによって発話552C1を完了したことに応答して、以前に未知であったスロット値を使用してアシスタントコマンドのフルフィルメントを完了し、ユーザ101に代わってレストランの予約を行うことができる。これらおよび他の方法において、自動アシスタントは、自然な会話出力554C1を提供することによってユーザ101が自分の思考を完了するのを待ち、その後、ユーザ101が自然な会話出力554C1の提供に応答して自分の思考を完了しない場合、自然な会話出力556Cを提供することによってユーザ101に思考を完了するように促すことができる。
【００７９】
さらに別の例として、特に図5Dを参照すると、クライアントデバイス110のユーザ101が、「アシスタント、私のカレンダーになにかありますかforrrr(Assistant, what's on my calendar forrrr)」の発話552D1を提供し、次いで、552D2によって示されるように、N秒間一時停止すると仮定し、ここで、Nは、任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。この例において、発話552C1を処理することに基づいて、ASR出力が、オーディオデータのストリーム内にキャプチャされた発話552D1に対応する認識されたテキスト(例えば、「私のカレンダーになにかありますか(what's on my calendar for)」に対応する認識されたテキスト)を含み、NLUデータのストリームが、データパラメータに関する未知のスロット値を有する予測された「カレンダー」または「カレンダー検索」の意図を含むと仮定する。この例において、自動アシスタントは、NLUデータのストリームに基づいて、ユーザ101がデータパラメータに関するスロット値を提供しなかったので、ユーザが発話552D1の提供を一時停止したと判定し得る。追加的または代替的に、この例において、自動アシスタントは、発話552D1のオーディオベースの特徴に基づいて、発話552D1内に含まれる引き延ばされた音節(例えば、発話552D1内の「forrrr」を提供する際の「rrrr」)によって示されるように、ユーザ101が発話552D1の提供を一時停止したと判定し得る。
【００８０】
さらに、フルフィルメントデータのストリームが、実行されると、クライアントデバイ

ス110に、クライアントデバイス110においてアクセス可能なカレンダーソフトウェアアプリケーションおよび/またはクライアントデバイス110においてアクセス可能なカレンダーエージェント(例えば、図1の1Pエージェント171および/または3Pエージェントのうちの1つ)を使用して、ユーザ101のカレンダー情報を検索させるアシスタントコマンドを含むと仮定する。図5Dの例において、図5A～図5Cの例とは対照的に、発話552D1を処理することに基づいて、かつ任意の追加の発話を処理することなく、自動アシスタントは、アシスタントコマンドを実行させることによって発話552D1が満たされることが可能であると判定し得る。この例において、自動アシスタントは、ユーザ101が1つまたは複数のカレンダーエントリを検索することを意図しているが、単に「カレンダー」または「カレンダー検索」の意図に関連付けられたデータパラメータに関するスロット値を提供しなかったことを示すNLUデータのストリームに関連付けられたNLU測定に基づいて、発話552D1のフルフィルメントを開始し得る。したがって、自動アシスタントは、クライアントデバイス110においてアクセス可能なカレンダーソフトウェアアプリケーションおよび/またはクライアントデバイス110においてアクセス可能なカレンダーエージェント(例えば、図1の1Pエージェント171および/または3Pエージェントのうちの1つ)との接続を確立することができる。

【００８１】

自動アシスタントが発話552D1のフルフィルメントを開始するとき、自動アシスタントは、NLUデータのストリームに基づいて、ユーザ101が発話552D1の提供を一時停止したと判定したので、自動アシスタントは、依然として、図5Dに示すような「うん(Uh huhh)」などの自然な会話出力554D1を提供することを決定し、自然な会話出力554D1をクライアントデバイス110のユーザ110への可聴提示のために提供させることができる。しかしながら、図5Dの例において、図5Bおよび図5Cと同様に、クライアントデバイス110のユーザ101が、554D2によって示されるように、自然な会話出力554D1をユーザ101への可聴提示のために提供してからM秒以内に発話552D1を完了しなかったと仮定し、ここで、Mは、552D2によって示されるようなN秒とは同じまたは異なる場合がある任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。

【００８２】

しかしながら、図5Dの例において、自動アシスタントは、ユーザ101が発話552D1を完了し得なかった場合であっても、発話552D1のフルフィルメントを実行させることを決定することができる。自動アシスタントは、フルフィルメントを実行させることおよび/または任意の実行されたフルフィルメントを取り消すことに関連する計算コストのうちの1つまたは複数に基づいてこの決定を行い得る。この例において、計算コストのうちの1つまたは複数は、発話552D1のフルフィルメントにおいて「あなたは、今日2つのカレンダーエントリを有しています...(You have two calendar entries for today...)」の合成音声556D1をユーザ101への可聴提示のために提供させることと、ユーザ101が別の日のカレンダー情報を希望した場合、他の合成音声スピーチをユーザ101への可聴提示のために提供させることとを含むことができる。したがって、そうすることの計算コストが比較的低いので、ダイアログセッションをより迅速に行う試みにおいて、自動アシスタントは、進行し、「カレンダー」または「カレンダー検索」の意図に関連付けられたデータパラメータに関する現在の日の推定スロット値を用いて発話552D1のフルフィルメントを実行させることを決定し得る。

【００８３】

特に、様々な実装形態において、自動アシスタントが合成音声556D1をユーザ101への提示のために提供する間、オーディオデータのストリームを処理する際に利用される自動アシスタント構成要素(例えば、ASRエンジン120A1および/もしくは120A2、NLUエンジン130A1および/もしくは130A2、フルフィルメントエンジン140A1および/もしくは140A2、ならびに/または図1の音響エンジン161などの図1の他の自動アシスタント構成要素)は、クライアントデバイス110においてアクティブのままであることができる。したがって、これらの実装形態において、ユーザ101が、推定された現在の日付以外の異なる

日付を要求する別の発話を提供することによって、合成音声556D1の可聴提示中に自動アシスタントを中断した場合、自動アシスタントは、ユーザ101によって提供された異なる日付に基づいて、発話552D1のフルフィルメントを迅速かつ効率的に適応させることができる。追加または代替の実装形態において、合成音声556D1をユーザ101への可聴提示のために提供させた後に、自動アシスタントは、発話552D1のフルフィルメントを修正する機会をユーザ101に積極的に提供するために、「待ってください、私は一秒前にあなたを遮りましたか(Wait, did I cut you off a second ago?)」などの追加の合成音声556D2を聴覚的にレンダリングさせることができる。これらおよび他の方法において、自動アシスタントは、自然な会話出力554D1を提供することによってユーザ101が自分の思考を完了するのを待つことと、そうすることの計算コストが比較的低い場合に発話552D1を履行することによってダイアログセッションをより迅速かつ効率的に終了することとのバランスをとることができる。

【００８４】

図5A～図5Dの例について、自然な会話出力をユーザ101への可聴提示のために提供させることに関して説明したが、それは、例のためのものであり、限定することを意味していないことが理解されるべきである。例えば、図5Eを簡単に参照すると、クライアントデバイス110のユーザ101が「アシスタント、Arnolllldに電話をかけて(Assistant, call Arnolllld's)」の発話を提供し、次いで、N秒間一時停止すると再び仮定し、ここで、Nは、任意の正の整数および/またはその小数(例えば、2秒、2.5秒、3秒など)である。図5Eの例において、発話のストリーミング転写552Eが、クライアントデバイス110のディスプレイ190を介してユーザへの視覚的表示のために提供されることが可能である。いくつかの実装形態において、クライアントデバイス110のディスプレイ190は、ディスプレイ190上で移動し得るストリーミング転写552Eに付加された楕円などの、自動アシスタントが、ユーザ101が発話を完了するのを待っていることを示す1つまたは複数のグラフィカル要素191を追加的または代替的に提供することができる。図5Eに示すグラフィカル要素191は、ストリーミング転写に付加された楕円であるが、それは、例のためのものであり、限定することを意味しておらず、自動アシスタントが、ユーザ101が発話の提供を完了するのを待っていることを示すために、任意の他のグラフィカル要素がユーザ101への視覚的提示のために提供されることが可能であることが理解されるべきである。追加または代替の実装形態において、(例えば、破線192によって示されるように)自動アシスタントが、ユーザ101が発話の提供を完了するのを待っていることを示すために、1つまたは複数のLEDが点灯されることが可能であり、これは、クライアントデバイス110がディスプレイ190を持たない場合に特に有利であり得る。さらに、図5A～図5Eの例は、例のために提供されるものであり、限定することを意味していないことが理解されるべきである。

【００８５】

さらに、ユーザ101のクライアントデバイス110がディスプレイ190を含む実装形態において、ユーザが発話を提供する際、発話の様々な解釈に関連付けられた1つまたは複数の選択可能なグラフィカル要素が、ユーザへの視覚的提示のために提供されることが可能である。自動アシスタントは、1つもしくは複数の選択可能なグラフィカル要素のうちの所与の1つのユーザ101からのユーザ選択を受け取ることに基づいて、および/またはしきい値持続時間内にユーザ101からのユーザ選択が受け取られなかったことに応答して、1つもしくは複数の選択可能なグラフィカル要素のうちの所与の1つに関連付けられたNLU測定値に基づいて、発話のフルフィルメントを開始することができる。例えば、図5Aの例において、「アシスタント、Arnolllldに電話をかけて(Assistant, call Arnolllld's)」の発話552A1を受け取った後、第1の選択可能なグラフィカル要素がディスプレイを介してユーザ101への提示のために提供されることが可能であり、第1の選択可能なグラフィカル要素は、選択されると、自動アシスタントに「Arnold」に関連付けられた連絡先エントリに電話をかけさせる。しかしながら、ユーザが「Arnoldのトラットリアに電話をかけて(call Arnold's Trattoria)」の発話556Aを提供し続けると、1つまたは複数の選択可能なグラフィカル要素は、第2の選択可能なグラフィカル要素を含むように更新されること

が可能であり、第2の選択可能なグラフィカル要素は、選択されると、自動アシスタントに「Arnoldのトラットリア(Arnold's Trattoria)」に関連付けられたレストランに電話をかけさせる。この例において、ユーザ101が(提示されている第1の選択可能なグラフィカル要素または提示されている第2の選択可能なグラフィカル要素に関して)しきい値持続時間内に第1の選択可能なグラフィカル要素または第2の選択可能なグラフィカル要素の任意のユーザ選択を提供しないと仮定すると、自動アシスタントは、レストラン「Arnoldのトラットリア(Arnold's Trattoria)」との通話を開始することに関連付けられたNLU測定値が、連絡先エントリ「Arnold」との通話を開始することに関連付けられたNLU測定値と比較してユーザ101の真の意図をよりよく示すことに基づいて、レストラン「Arnoldのトラットリア(Arnold's Trattoria)」との通話を開始することができる。

【００８６】

ここで図6に進むと、本明細書で説明する技法の1つまたは複数の態様を実行するためにオプションで利用され得る例示的なコンピューティングデバイス610のブロック図が示されている。いくつかの実装形態において、クライアントデバイス、クラウドベースの自動アシスタント構成要素、および/または他の構成要素のうちの1つまたは複数は、例示的なコンピューティングデバイス610の1つまたは複数の構成要素を備え得る。

【００８７】

コンピューティングデバイス610は、典型的には、バスサブシステム612を介していくつかの周辺デバイスと通信する少なくとも1つのプロセッサ614を含む。これらの周辺デバイスは、例えば、メモリサブシステム625およびファイル記憶サブシステム626を含む記憶サブシステム624と、ユーザインターフェース出力デバイス620と、ユーザインターフェース入力デバイス622と、ネットワークインターフェースサブシステム616とを含み得る。入力デバイスおよび出力デバイスは、ユーザがコンピューティングデバイス610と対話することを可能にする。ネットワークインターフェースサブシステム616は、外部ネットワークへのインターフェースを提供し、他のコンピューティングデバイス内の対応するインターフェースデバイスに結合される。

【００８８】

ユーザインターフェース入力デバイス622は、キーボード、マウス、トラックボール、タッチパッド、もしくはグラフィックタブレットなどのポインティングデバイス、スキャナ、ディスプレイに組み込まれたタッチスクリーン、音声認識システムなどのオーディオ入力デバイス、マイクロフォン、および/または他のタイプの入力デバイスを含み得る。一般に、「入力デバイス」という用語の使用は、コンピューティングデバイスまたは通信ネットワークに情報を入力するためのすべての可能なタイプのデバイスおよび方法を含むことを意図している。

【００８９】

ユーザインターフェース出力デバイス620は、ディスプレイサブシステム、プリンタ、ファックス機、またはオーディオ出力デバイスなどの非視覚的ディスプレイを含み得る。ディスプレイサブシステムは、陰極線管(CRT)、液晶ディスプレイ(LCD)などのフラットパネルデバイス、投影デバイス、または可視画像を作成するためのなにか他のメカニズムを含み得る。ディスプレイサブシステムはまた、オーディオ出力デバイスを介するなどして、非視覚的ディスプレイを提供し得る。一般に、「出力デバイス」という用語の使用は、コンピューティングデバイス610からユーザまたは別の機械もしくはコンピューティングデバイスに情報を出力するためのすべての可能なタイプのデバイスおよび方法を含むことを意図している。

【００９０】

記憶サブシステム624は、本明細書で説明するモジュールのうちのいくつかまたはすべての機能を提供するプログラミング構造およびデータ構造を記憶する。例えば、記憶サブシステム624は、本明細書で開示する方法の選択された態様を実行するため、ならびに図1および図2に示す様々な構成要素を実装するためのロジックを含み得る。

【００９１】

これらのソフトウェアモジュールは、一般に、プロセッサ614単独で、または他のプロセッサと組み合わせて実行される。記憶サブシステム624内で使用されるメモリ625は、プログラム実行中の命令およびデータの記憶のためのメインランダムアクセスメモリ(RAM)630と、固定命令が記憶される読み取り専用メモリ(ROM)632とを含むいくつかのメモリを含むことができる。ファイル記憶サブシステム626は、プログラムおよびデータファイルのための永続的なストレージを提供することができ、ハードディスクドライブ、関連付けられたリムーバブルメディアを伴うフロッピーディスクドライブ、CD-ROMドライブ、光学ドライブ、またはリムーバブルメディアカートリッジを含み得る。特定の実装形態の機能を実装するモジュールは、記憶サブシステム624内のファイル記憶サブシステム626によって、またはプロセッサ614によってアクセス可能な他の機械内に記憶され得る。

【００９２】

バスサブシステム612は、コンピューティングデバイス610の様々な構成要素およびサブシステムを意図されたように互いに通信させるためのメカニズムを提供する。バスサブシステム612は、単一のバスとして概略的に示されているが、バスサブシステム612の代替実装形態は、複数のバスを使用し得る。

【００９３】

コンピューティングデバイス610は、ワークステーション、サーバ、コンピューティングクラスタ、ブレードサーバ、サーバファーム、または任意の他のデータ処理システムもしくはコンピューティングデバイスを含む様々なタイプのものであることが可能である。コンピュータおよびネットワークの絶えず変化する性質のために、図6に示すコンピューティングデバイス610の説明は、いくつかの実装形態を説明する目的のための特定の例としてのみ意図されている。コンピューティングデバイス610の他の多くの構成は、図6に示すコンピューティングデバイスよりも多いまたは少ない構成要素を有することが可能である。

【００９４】

本明細書で説明するシステムがユーザに関する個人情報を収集もしくは他の方法で監視するか、または個人情報および/もしくは監視情報を利用する可能性がある状況では、ユーザは、プログラムまたは機能がユーザ情報(例えば、ユーザのソーシャルネットワーク、社会的行為もしくは活動、職業、ユーザの好み、またはユーザの現在の地理的位置に関する情報)を収集するかどうかを制御する機会、またはユーザにより関連性がある可能性があるコンテンツサーバからのコンテンツを受信するかどうかおよび/もしくはどのように受信するかを制御する機会を提供され得る。また、特定のデータは、個人を特定し得る情報が除去されるように、それが記憶または使用される前に1つまたは複数の方法で処理され得る。例えば、ユーザの識別情報は、個人を特定し得る情報がユーザに関して決定され得ないように処理され得、またはユーザの地理的位置は、ユーザの特定の地理的位置が決定され得ないように、地理的位置情報が取得される場所(都市、郵便番号、または州レベルなど)で一般化され得る。したがって、ユーザは、情報がユーザに関してどのように収集および/または使用されるかについて制御し得る。

【００９５】

いくつかの実装形態において、1つまたは複数のプロセッサによって実装される方法が提供され、自動音声認識(ASR)モデルを使用して、ASR出力のストリームを生成するためにオーディオデータのストリームを処理するステップであって、オーディオデータのストリームがユーザのクライアントデバイスの1つまたは複数のマイクロフォンによって生成され、オーディオデータのストリームが、クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントに向けられた、ユーザによって提供された発話の一部をキャプチャする、ステップと、自然言語理解(NLU)モデルを使用して、NLU出力のストリームを生成するためにASR出力のストリームを処理するステップと、オーディオデータのストリームを処理することに基づいて、発話の一部に関連付けられたオーディオベースの特徴を決定するステップと、発話の一部に関連付けられたオーディオベースの特徴に基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定する

ステップと、ユーザが発話の提供を一時停止したと判定したことに応答して、および自動アシスタントが少なくともNLU出力に基づいて発話のフルフィルメントを開始することができると判定したことに応答して、ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、ユーザへの可聴提示のために提供されるべき自然な会話出力が、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示す、ステップと、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップとを含む。

【００９６】

本明細書で開示する技術のこれらおよび他の実装形態は、オプションで、以下の特徴のうちの1つまたは複数を含むことができる。

【００９７】

いくつかの実装形態において、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップは、ユーザがしきい値持続時間の間に発話の提供を一時停止したと判定したことにさらに応答し得る。

【００９８】

いくつかの実装形態において、発話の一部に関連付けられたオーディオベースの特徴に基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定するステップは、オーディオベースの分類機械学習(ML)モデルを使用して、出力を生成するために、発話の一部に関連付けられたオーディオベースの特徴を処理するステップと、オーディオベースの分類MLモデルを使用して生成された出力に基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定するステップとを含み得る。

【００９９】

いくつかの実装形態において、方法は、NLU出力のストリームに基づいて、フルフィルメントデータのストリームを生成させるステップをさらに含み得る。自動アシスタントが発話のフルフィルメントを開始することができると判定するステップは、フルフィルメントデータのストリームにさらに基づき得る。それらの実装形態のいくつかの変形例において、方法は、ユーザが発話の提供を完了したと判定したことに応答して、自動アシスタントに、フルフィルメントデータのストリームに基づいて発話のフルフィルメントを開始させるステップをさらに含み得る。それらの実装形態の追加または代替の変形例において、方法は、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させる間、ASRモデルを利用する1つまたは複数の自動アシスタント構成要素をアクティブに保つステップをさらに含み得る。それらの実装形態の追加または代替の変形例において、方法は、ASR出力のストリームに基づいて、発話が特定の単語または句を含むかどうかを判定するステップと、発話が特定の単語または句を含むと判定したことに応答して、発話の一部に関連付けられたオーディオベースの特徴に基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定することを控えるステップと、自動アシスタントに、フルフィルメントデータのストリームに基づいて発話のフルフィルメントを開始させるステップとをさらに含み得る。それらの実装形態の追加または代替の変形例において、方法は、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップの後に、ユーザがしきい値持続時間内に発話を提供し続けたかどうかを判定するステップと、ユーザがしきい値持続時間内に1つまたは複数の発話を提供し続けなかったと判定したことに応答して、NLUデータのストリームおよび/またはフルフィルメントデータのストリームに基づいて、自動アシスタントが発話のフルフィルメントを開始することができるかどうかを判定するステップと、自動アシスタントがフルフィルメントデータのストリームに基づいて発話のフルフィルメントを開始することができると判定したことに応答して、自動アシスタントに、フルフィルメントデータのストリームに基づいて発話のフルフィルメントを開始させるステップとをさらに含み得る。

【０１００】

いくつかの実装形態において、方法は、クライアントデバイスの1つまたは複数のスピ

ーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップの後に、ユーザがしきい値持続時間内に発話を提供し続けたかどうかを判定するステップと、ユーザが発話を提供し続けなかったと判定したことに応答して、ユーザへの可聴提示のために提供されるべき追加の自然な会話出力を決定するステップであって、ユーザへの可聴提示のために提供されるべき追加の自然な会話出力が、ユーザが発話の提供を完了することを要求する、ステップと、クライアントデバイスの1つまたは複数のスピーカを介して、追加の自然な会話出力をユーザへの可聴提示のために提供させるステップとをさらに含み得る。

【０１０１】

いくつかの実装形態において、方法は、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させる間、クライアントデバイスのディスプレイを介して、1つまたは複数のグラフィカル要素をユーザへの視覚的提示のために提供させるステップであって、ユーザへの視覚的提示のために提供されるべき1つまたは複数のグラフィカル要素が、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示す、ステップをさらに含み得る。それらの実装形態のいくつかの変形例において、ASR出力は、オーディオデータのストリーム内にキャプチャされた発話の一部に対応するストリーミング転写を含み得、方法は、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させる間、クライアントデバイスのディスプレイを介して、ストリーミング転写をユーザへの視覚的提示のために提供させるステップであって、1つまたは複数のグラフィカル要素が、クライアントデバイスのディスプレイを介してユーザへの視覚的提示のために提供されるストリーミング転写の先頭に追加されるか、または末尾に追加される、ステップをさらに含み得る。

【０１０２】

いくつかの実装形態において、方法は、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させる間、クライアントデバイスの1つまたは複数の発光ダイオード(LED)を点灯させるステップであって、1つまたは複数のLEDが、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示すために点灯される、ステップをさらに含み得る。

【０１０３】

いくつかの実装形態において、発話の一部に関連付けられたオーディオベースの特徴は、イントネーション、トーン、ストレス、リズム、テンポ、ピッチ、休止、休止に関連する1つまたは複数の文法、および引き延ばされた音節のうちの1つまたは複数を含み得る。

【０１０４】

いくつかの実装形態において、ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップは、クライアントデバイスのオンデバイスメモリ内に自然な会話出力のセットを維持するステップと、発話の一部に関連付けられたオーディオベースの特徴に基づいて、自然な会話出力のセットの中から自然な会話出力を選択するステップとを含み得る。

【０１０５】

いくつかの実装形態において、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップは、ユーザへの可聴提示のために提供される他の出力よりも低い音量において、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップを含み得る。

【０１０６】

いくつかの実装形態において、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップは、テキスト読み上げ(TTS)モデルを使用して、自然な会話出力を含む合成音声オーディオデータを生成するために自然な会話出力を処理するステップと、クライアントデバイスの1つまたは複

数のスピーカを介して、合成音声オーディオデータをユーザへの可聴提示のために提供させるステップとを含み得る。

【０１０７】

いくつかの実装形態において、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップは、クライアントデバイスのオンデバイスメモリから、自然な会話出力を含む合成音声オーディオデータを取得するステップと、クライアントデバイスの1つまたは複数のスピーカを介して、合成音声オーディオデータをユーザへの可聴提示のために提供させるステップとを含み得る。

【０１０８】

いくつかの実装形態において、1つまたは複数のプロセッサは、ユーザのクライアントデバイスによってローカルに実装され得る。

【０１０９】

いくつかの実装形態において、1つまたは複数のプロセッサによって実装される方法が提供され、自動音声認識(ASR)モデルを使用して、ASR出力のストリームを生成するためにオーディオデータのストリームを処理するステップであって、オーディオデータのストリームが、クライアントデバイスの1つまたは複数のマイクロフォンによって生成され、オーディオデータのストリームが、クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントに向けられたユーザの発話の一部をキャプチャする、ステップと、自然言語理解(NLU)モデルを使用して、NLU出力のストリームを生成するためにASR出力のストリームを処理するステップと、少なくともNLU出力のストリームに基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定するステップと、ユーザが発話の提供を一時停止し、発話の提供を完了していないと判定したことに応答して、ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、ユーザへの可聴提示のために提供されるべき自然な会話出力が、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示す、ステップと、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップとを含む。

【０１１０】

本明細書で開示する技術のこれらおよび他の実装形態は、オプションで、以下の特徴のうちの1つまたは複数を含むことができる。

【０１１１】

いくつかの実装形態において、NLU出力のストリームに基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定するステップは、NLU出力のストリームに基づいて、自動アシスタントが発話のフルフィルメントを開始することができるかどうかを判定するステップを含み得る。ユーザが発話の提供を一時停止したと判定するステップは、自動アシスタントがNLU出力のストリームに基づいて発話のフルフィルメントを開始することができないと判定するステップを含み得る。それらの実装形態のいくつかの変形例において、方法は、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップの後に、ユーザがしきい値持続時間内に発話を提供し続けたかどうかを判定するステップと、ユーザが発話を提供し続けなかったと判定したことに応答して、ユーザへの可聴提示のために提供されるべき追加の自然な会話出力を決定するステップであって、ユーザへの可聴提示のために提供されるべき追加の自然な会話出力が、ユーザが発話の提供を完了することを要求する、ステップと、クライアントデバイスの1つまたは複数のスピーカを介して、追加の自然な会話出力をユーザへの可聴提示のために提供させるステップとをさらに含み得る。それらの実装形態のいくつかのさらなる変形例において、ユーザへの可聴提示のために提供されるべき追加の自然な会話出力は、発話の追加の部分がNLUデータのストリームに基づく特定のデータを含むことを要求し得る。

【０１１２】

いくつかの実装形態において、1つまたは複数のプロセッサによって実装される方法が

提供され、自動音声認識(ASR)モデルを使用して、ASR出力のストリームを生成するためにオーディオデータのストリームを処理するステップであって、オーディオデータのストリームが、クライアントデバイスの1つまたは複数のマイクロフォンによって生成され、オーディオデータのストリームが、クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントに向けられたユーザの発話の一部をキャプチャする、ステップと、自然言語理解(NLU)モデルを使用して、NLU出力のストリームを生成するためにASR出力のストリームを処理するステップと、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定するステップと、ユーザが発話の提供を一時停止し、発話の提供を完了していないと判定したことに応答して、ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、ユーザへの可聴提示のために提供されるべき自然な会話出力が、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示す、ステップと、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップと、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップの後に、ユーザがしきい値持続時間内に発話の提供を完了しなかったと判定したことに応答して、少なくともNLUデータのストリームに基づいて、自動アシスタントが発話のフルフィルメントを開始することができるかどうかを判定するステップと、NLUデータのストリームに基づいて、自動アシスタントが発話のフルフィルメントを開始することができると判定したことに応答して、自動アシスタントに発話のフルフィルメントを開始させるステップとを含む。

【０１１３】

本明細書で開示する技術のこれらおよび他の実装形態は、オプションで、以下の特徴のうちの1つまたは複数を含むことができる。

【０１１４】

いくつかの実装形態において、方法は、オーディオデータのストリームを処理することに基づいて、発話の一部に関連付けられたオーディオベースの特徴を決定するステップをさらに含み得る。ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定するステップは、発話の一部に関連付けられたオーディオベースの特徴に基づき得る。

【０１１５】

いくつかの実装形態において、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定するステップは、NLUデータのストリームに基づき得る。

【０１１６】

いくつかの実装形態において、方法は、NLUデータのストリームに基づいて、自動アシスタントが発話のフルフィルメントを開始することができないと判定したことに応答して、ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、ユーザへの可聴提示のために提供されるべき自然な会話出力が、ユーザが発話の提供を完了することを要求する、ステップと、クライアントデバイスの1つまたは複数のスピーカを介して、追加の自然な会話出力をユーザへの可聴提示のために提供させるステップとをさらに含み得る。それらの実装形態のいくつかの変形例において、ユーザへの可聴提示のために提供されるべき自然な会話出力は、発話の追加の部分がNLUデータのストリームに基づく特定のデータを含むことを要求し得る。

【０１１７】

いくつかの実装形態において、自動アシスタントが発話のフルフィルメントを開始することができるかどうかを判定するステップは、発話のフルフィルメントに関連付けられた1つまたは複数の計算コストにさらに基づき得る。それらの実装形態のいくつかの変形例において、発話のフルフィルメントに関連付けられた1つまたは複数の計算コストは、発話のフルフィルメントを実行することに関連する計算コスト、および発話の実行されたフルフィルメントを取り消すことに関連する計算コストのうちの1つまたは複数を含み得る。

【０１１８】

いくつかの実装形態において、方法は、NLU出力のストリームに基づいて、フルフィルメントデータのストリームを生成させるステップをさらに含み得る。自動アシスタントが発話のフルフィルメントを開始することができると判定するステップは、フルフィルメントデータのストリームにさらに基づき得る。

【０１１９】

いくつかの実装形態において、1つまたは複数のプロセッサによって実装される方法が提供され、オーディオデータのストリームを受信するステップであって、オーディオデータのストリームがユーザのクライアントデバイスの1つまたは複数のマイクロフォンによって生成され、オーディオデータのストリームが、クライアントデバイスにおいて少なくとも部分的に実装された自動アシスタントに向けられた、ユーザによって提供された発話の少なくとも一部をキャプチャする、ステップと、オーディオデータのストリームを処理することに基づいて、発話の一部に関連付けられたオーディオベースの特徴を決定するステップと、発話の一部に関連付けられたオーディオベースの特徴に基づいて、ユーザが発話の提供を一時停止したか、または発話の提供を完了したかを判定するステップと、ユーザが発話の提供を一時停止し、発話の提供を完了していないと判定したことに応答して、ユーザへの可聴提示のために提供されるべき自然な会話出力を決定するステップであって、ユーザへの可聴提示のために提供されるべき自然な会話出力が、自動アシスタントが、ユーザが発話の提供を完了するのを待っていることを示す、ステップと、クライアントデバイスの1つまたは複数のスピーカを介して、自然な会話出力をユーザへの可聴提示のために提供させるステップとを含む。

【０１２０】

それに加えて、いくつかの実装形態は、1つまたは複数のコンピューティングデバイスの1つまたは複数のプロセッサ(例えば、中央処理装置(CPU)、グラフィックス処理ユニット(GPU)、および/またはテンソル処理ユニット(TPU))を含み、1つまたは複数のプロセッサは、関連付けられたメモリ内に記憶された命令を実行するように動作可能であり、命令は、前述の方法のいずれかの実行を引き起こすように構成される。いくつかの実装形態は、前述の方法のいずれかを実行するために1つまたは複数のプロセッサによって実行可能なコンピュータ命令を記憶する1つまたは複数の非一時的なコンピュータ可読記憶媒体も含む。いくつかの実装形態は、前述の方法のいずれかを実行するために1つまたは複数のプロセッサによって実行可能な命令を含むコンピュータプログラム製品も含む。

【符号の説明】

【０１２１】

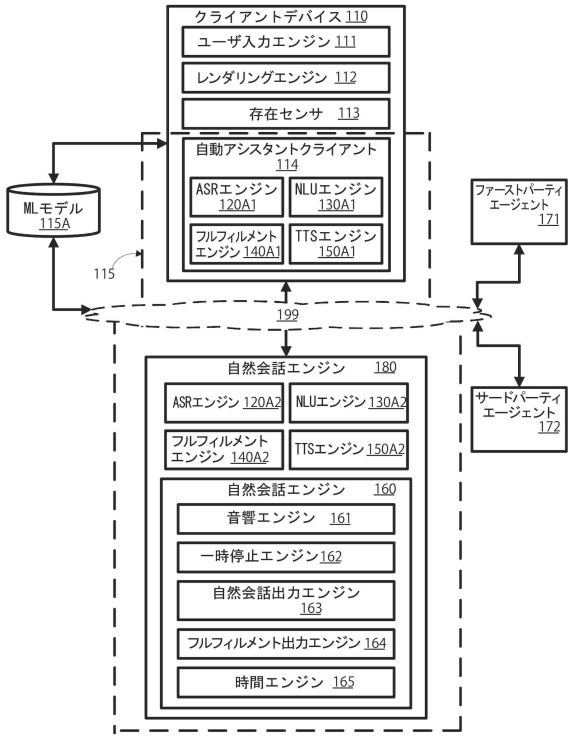
101 ユーザ
110 クライアントデバイス
111 ユーザ入力エンジン
112 レンダリングエンジン
113 存在センサ
114 自動アシスタントクライアント
115 自動アシスタント
115A 機械学習(ML)モデルデータベース、MLモデルデータベース
120A1 自動音声認識(ASR)エンジン、ASRエンジン
120A2 ASRエンジン
130A1 自然言語理解(NLU)エンジン、NLUエンジン
130A2 NLUエンジン
140A1 フルフィルメントエンジン
140A2 フルフィルメントエンジン
150A1 テキスト読み上げ(TTS)エンジン、TTSエンジン
150A2 TTSエンジン
160 自然会話エンジン
161 音響エンジン

162 一時停止エンジン
163 時間エンジン、自然会話出力エンジン、自然会話エンジン
164 自然会話出力エンジン、フルフィルメント出力エンジン
165 フルフィルメント出力エンジン、時間エンジン
171 ファーストパーティ(1P)エージェント、1Pエージェント
172 サードパーティ(3P)エージェント、3Pエージェント
180 自然会話システム
190 ディスプレイ
191 グラフィカル要素
192 破線
199 ネットワーク
201A オーディオデータのストリーム
201B 非オーディオデータのストリーム
220 ASR出力のストリーム
230 NLU出力のストリーム
240 フルフィルメントデータのストリーム
240A 1Pフルフィルメントデータ
240B 3Pフルフィルメントデータ
261 オーディオベースの特徴
262 ブロック
263 自然な会話出力
264 フルフィルメント出力
265 一時停止の持続時間
552A1 発話
552B1 発話
552C1 発話
552D1 発話
552E ストリーミング転写
554A 自然な会話出力
554B1 自然な会話出力
554C1 自然な会話出力
554D1 自然な会話出力
556A 発話
556B 追加の自然な会話出力、追加の自然な会話
556C 追加の自然な会話出力、追加の自然な会話、自然な会話出力
556D1 合成音声
556D2 追加の合成音声
558A 合成音声
558B 発話
558C 発話
560B 合成音声
610 コンピューティングデバイス
612 バスサブシステム
614 プロセッサ
616 ネットワークインターフェースサブシステム
620 ユーザインターフェース出力デバイス
622 ユーザインターフェース入力デバイス
624 記憶サブシステム
625 メモリサブシステム
626 ファイル記憶サブシステム

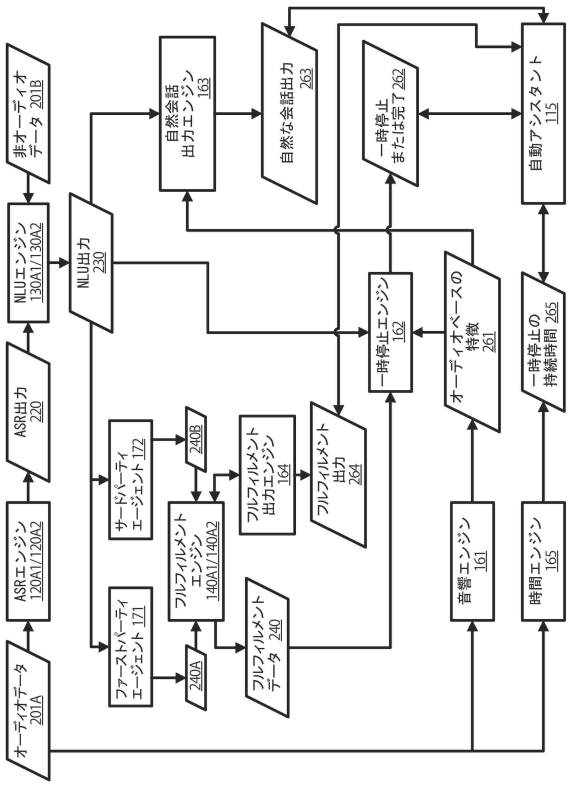
630 メインランダムアクセスメモリ(RAM)
632 読み取り専用メモリ(ROM)

【図面】

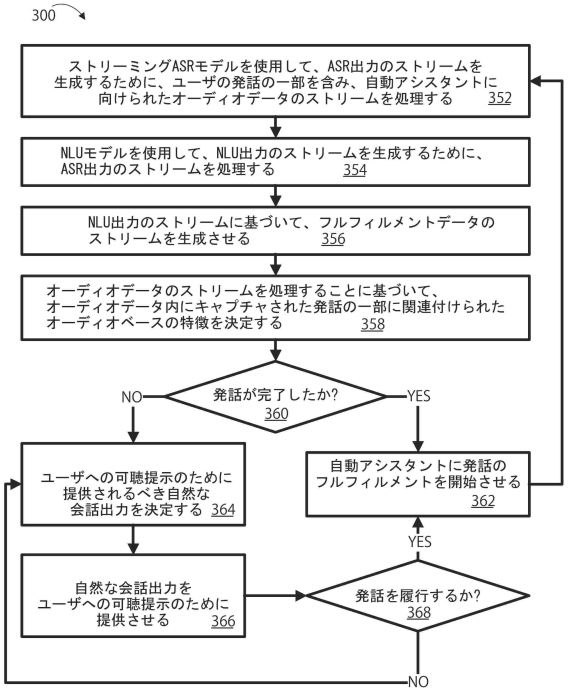
【図１】



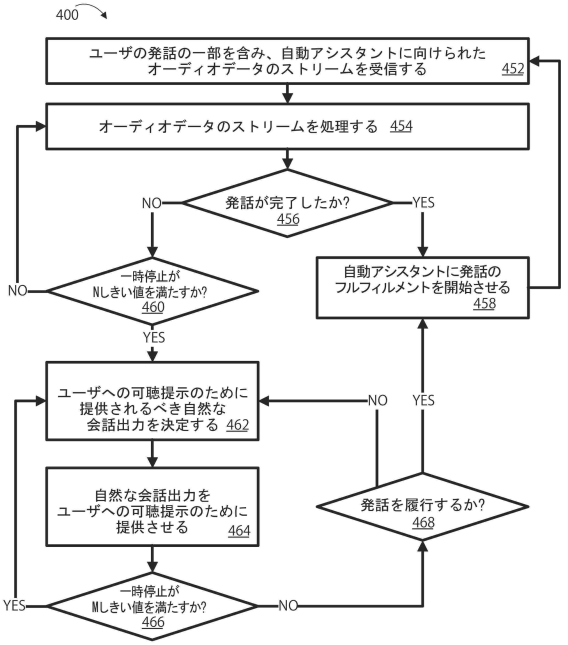
【図２】



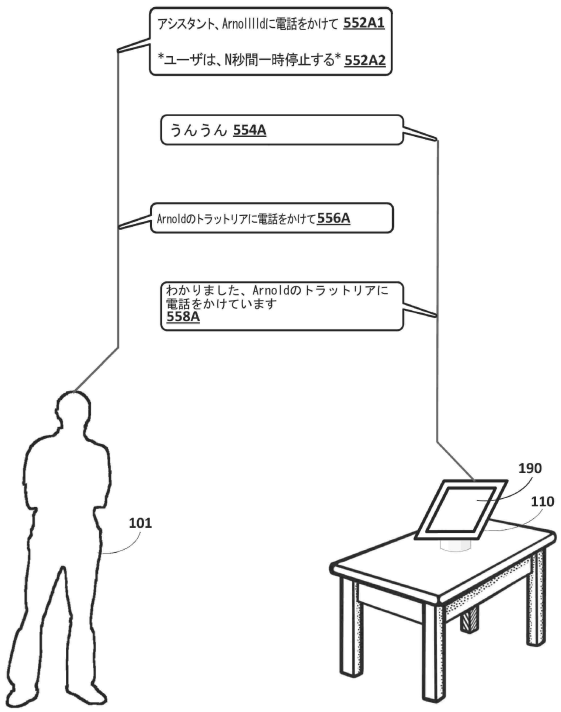
【図３】



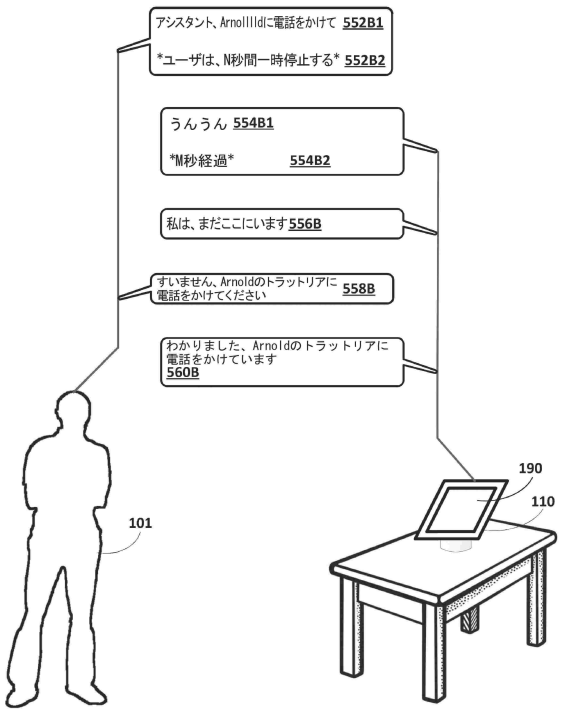
【図４】



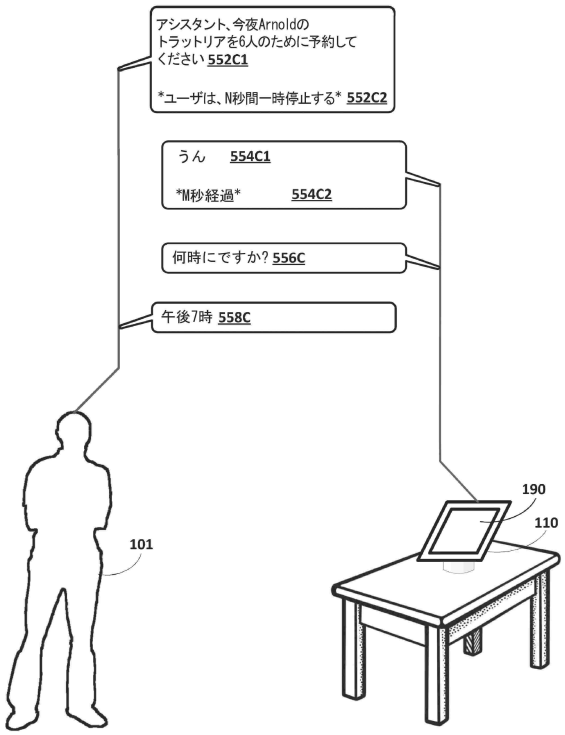
【図５Ａ】



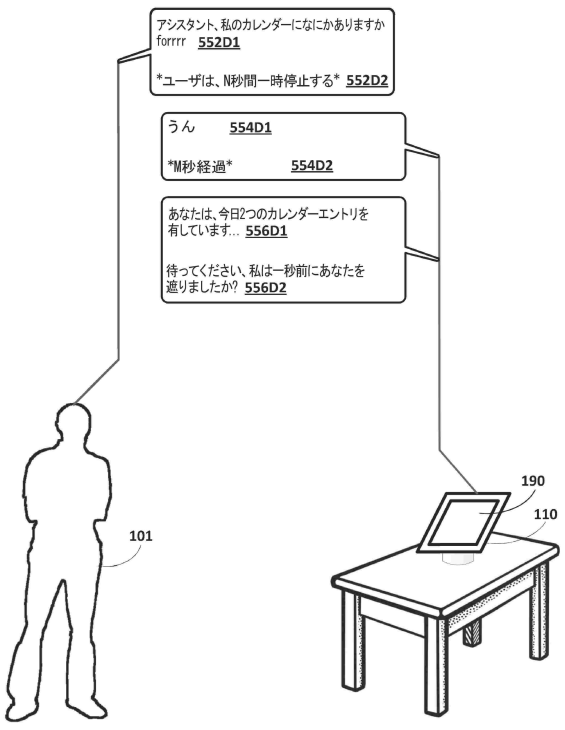
【図５Ｂ】



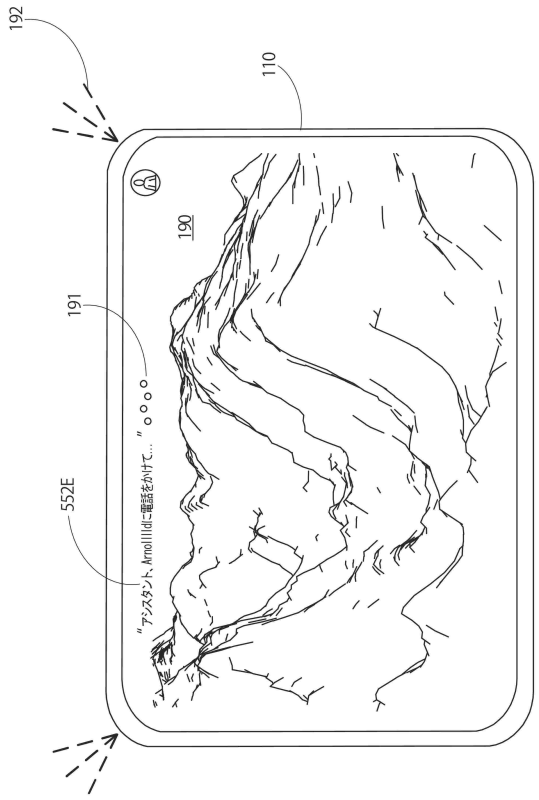
【図５Ｃ】



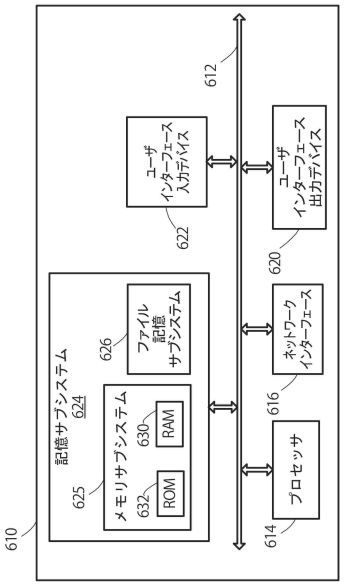
【図５Ｄ】



【図５Ｅ】


192
110
190
191
552E
"アシスタント、Arnoldに電話をかけて..."

【図６】


610
612
614
616
620
622
624
625
626
630
632
RAM
ROM
ユーザ
インターフェース
入力デバイス
ユーザ
インターフェース
出力デバイス
ネットワーク
インターフェース
プロセッサ
記憶サブシステム
メモリサブシステム
ファイル
記憶
サブシステム

フロントページの続き

(51)国際特許分類
G06F 3/16 (2006.01)

FI
G10L 25/51
G10L 13/00 100M
G06F 3/16 650
G06F 3/16 690
G06F 3/16 620

(33)優先権主張国・地域又は機関
米国(US)

(72)発明者 ジャクリン・コンツェルマン
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００
(72)発明者 トレヴァー・ストローマン
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００
(72)発明者 ジョナサン・ブルーム
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００
(72)発明者 ジョアン・ショークウィック
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００
(72)発明者 ジョセフ・スマール
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００
審査官 大野 弘

(56)参考文献
国際公開第２０２０／２２７５５７（WO，A１）
米国特許出願公開第２０２０／０３３５１２８（US，A１）
特表２００４－５１３４４５（JP，A）
特開２００８－２４１８９０（JP，A）
国際公開第２０１６／１５７６５０（WO，A１）
特開２００４－０８６００１（JP，A）
特開２０２１－１１７３７１（JP，A）
特表２０１８－５０４６２３（JP，A）
国際公開第２０１８／１９８８１２（WO，A１）

(58)調査した分野 (Int.Cl.，DB名)
G10L 15／22
G10L 15／04
G10L 15／10
G10L 25／51
G10L 13／00
G06F 3／16