(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7654792号
(P7654792)

(45)発行日　令和7年4月1日(2025.4.1)

(24)登録日　令和7年3月24日(2025.3.24)

(51)国際特許分類

Ｇ１０Ｌ　15/32　(2013.01)	Ｇ１０Ｌ	15/32	２２０Ｚ
Ｇ１０Ｌ　15/10　(2006.01)	Ｇ１０Ｌ	15/10	２００Ｗ

請求項の数　14　（全39頁）

(21)出願番号　特願2023-536526(P2023-536526)
(86)(22)出願日　令和3年12月14日(2021.12.14)
(65)公表番号　特表2024-505788(P2024-505788A)
(43)公表日　令和6年2月8日(2024.2.8)
(86)国際出願番号　PCT/US2021/063370
(87)国際公開番号　WO2022/191892
(87)国際公開日　令和4年9月15日(2022.9.15)
審査請求日　令和5年8月9日(2023.8.9)
(31)優先権主張番号　17/198,679
(32)優先日　令和3年3月11日(2021.3.11)
(33)優先権主張国・地域又は機関
米国(US)

(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】　自動音声認識のローカル実行のためのデバイス調停

(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行される方法であって、

クライアントデバイスにおいて、ユーザの口頭の発話をキャプチャするオーディオデータを検出するステップであって、

前記クライアントデバイスは、1つまたは複数の追加のクライアントデバイスを含む環境内にあり、ローカルネットワークを介して前記1つまたは複数の追加のクライアントデバイスとローカルで通信し、

前記1つまたは複数の追加のクライアントデバイスは、少なくとも第1の追加のクライアントデバイスを含む、ステップと、

前記クライアントデバイスにおいて、前記口頭の発話の候補テキスト表現を生成すべく、前記クライアントデバイスにおいてローカルで記憶された自動音声認識(「ASR」)モデルを使用して前記オーディオデータを処理するステップと、

前記クライアントデバイスにおいて、前記第1の追加のクライアントデバイスから前記ローカルネットワークを介して、前記口頭の発話の第1の追加の候補テキスト表現を受信するステップであって、

前記第1の追加のクライアントデバイスにおいてローカルで生成される前記口頭の発話の前記第1の追加の候補テキスト表現は、(a)前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータ、および(b)前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャするローカルで検

出されたオーディオデータに基づき、
前記口頭の発話の前記第1の追加の候補テキスト表現は、前記第1の追加のクライアントデバイスにおいてローカルで記憶された第1の追加のASRモデルを使用して、前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータおよび前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記ローカルで生成されたオーディオデータを処理することによって生成される、ステップと、
前記口頭の発話のテキスト表現を、前記口頭の発話の前記候補テキスト表現、および前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現に基づいて決定するステップと
を含む方法。
【請求項２】
前記1つまたは複数の追加のクライアントデバイスは、少なくとも前記第1の追加のクライアントデバイスと、第2の追加のクライアントデバイスとを含み、
前記クライアントデバイスにおいて、前記第1の追加のクライアントデバイスから前記ローカルネットワークを介して、前記第1の追加の候補テキスト表現を受信するステップは、
前記クライアントデバイスにおいて、前記第2の追加のクライアントデバイスから前記ローカルネットワークを介して、(a)前記オーディオデータ、および/または(b)前記第2の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする追加のローカルで検出されたオーディオデータに基づいて前記第2の追加のクライアントデバイスにおいてローカルで生成された前記口頭の発話の第2の追加の候補テキスト表現を受信するステップであって、前記口頭の発話の前記第2の追加の候補テキスト表現は、前記第2の追加のクライアントデバイスにおいてローカルで記憶された第2の追加のASRモデルを使用して前記オーディオデータおよび/または前記追加のローカルで生成されたオーディオデータを処理することによって生成される、ステップをさらに含み、
前記口頭の発話の前記候補テキスト表現、および前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現に基づいて、前記口頭の発話の前記テキスト表現を決定するステップは、
前記口頭の発話の前記候補テキスト表現、前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現、および前記第2の追加のクライアントデバイスによって生成された前記口頭の発話の前記第2の追加の候補テキスト表現に基づいて、前記口頭の発話の前記テキスト表現を決定するステップをさらに含む、請求項1に記載の方法。
【請求項３】
前記口頭の発話の前記候補テキスト表現、および前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現に基づいて、前記口頭の発話の前記テキスト表現を決定するステップは、
前記口頭の発話の前記候補テキスト表現、または前記口頭の発話の前記第1の追加の候補テキスト表現をランダムに選択するステップと、
前記ランダムな選択に基づいて前記口頭の発話の前記テキスト表現を決定するステップとを含む、請求項1に記載の方法。
【請求項４】
前記口頭の発話の前記候補テキスト表現、および前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現に基づいて、前記口頭の発話の前記テキスト表現を決定するステップは、
前記候補テキスト表現が前記テキスト表現である確率を示す前記候補テキスト表現の信頼度スコアを決定するステップであって、前記信頼度スコアは、前記クライアントデバイスの1つまたは複数のデバイスパラメータに基づく、ステップと、
前記追加の候補テキスト表現が前記テキスト表現である追加の確率を示す前記追加の候

補テキスト表現の追加の信頼度スコアを決定するステップであって、前記追加の信頼度スコアは、前記追加のクライアントデバイスの1つまたは複数の追加のデバイスパラメータに基づく、ステップと、
前記信頼度スコアと前記追加の信頼度スコアを比較するステップと、
前記比較に基づいて前記口頭の発話の前記テキスト表現を決定するステップとを含む、請求項1に記載の方法。
【請求項５】
前記口頭の発話の前記候補テキスト表現、および前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現に基づいて、前記口頭の発話の前記テキスト表現を決定するステップは、
前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータの品質を示すオーディオ品質値を決定するステップと、
前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする追加のオーディオデータの前記品質を示す追加のオーディオ品質値を決定するステップと、
前記オーディオ品質値と前記追加のオーディオ品質値を比較するステップと、
前記比較に基づいて前記口頭の発話の前記テキスト表現を決定するステップとを含む、請求項1または請求項4に記載の方法。
【請求項６】
前記口頭の発話の前記候補テキスト表現、および前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現に基づいて、前記口頭の発話の前記テキスト表現を決定するステップは、
前記クライアントデバイスにおいてローカルで記憶された前記ASRモデルの品質を示すASR品質値を決定するステップと、
前記追加のクライアントデバイスにおいてローカルで記憶された前記追加のASRモデルの前記品質を示す追加のASR品質値を決定するステップと、
前記ASR品質値と前記追加のASR品質値を比較するステップと、
前記比較に基づいて前記口頭の発話の前記テキスト表現を決定するステップとを含む、請求項1、請求項4、または請求項5に記載の方法。
【請求項７】
前記口頭の発話の前記第1の追加の候補テキスト表現は、複数の仮説を含み、前記口頭の発話の前記候補テキスト表現、および前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現に基づいて、前記口頭の発話の前記テキスト表現を決定するステップは、
前記クライアントデバイスを使用して前記複数の仮説を格付けし直すステップと、
前記口頭の発話の前記候補テキスト表現、および前記格付けし直された複数の仮説に基づいて、前記口頭の発話の前記テキスト表現を決定するステップとを含む、請求項1から6のいずれか一項に記載の方法。
【請求項８】
前記クライアントデバイスにおいて、前記第1の追加のクライアントデバイスから前記ローカルネットワークを介して、前記口頭の発話の前記第1の追加の候補テキスト表現を受信するステップに先立って、
(a)前記オーディオデータ、および/または(b)前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記ローカルで検出されたオーディオデータに基づいて、前記第1の追加のクライアントデバイスにおいてローカルで前記口頭の発話の前記第1の追加の候補表現を生成するかどうかを決定するステップをさらに含み、(a)前記オーディオデータ、および/または(b)前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記ローカルで検出されたオーディオデータに基づいて、前記第1の追加のクライアントデバイスにおいてローカルで前記口頭の発話の前記第1の追加の候補表現を生成するかどうかを決定するステップは、

前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータの品質を示すオーディオ品質値を決定するステップと、
前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記ローカルで検出されたオーディオデータの前記品質を示す追加のオーディオ品質値を決定するステップと、
前記オーディオ品質値と前記追加のオーディオ品質値を比較するステップと、
前記比較に基づいて、(a)前記オーディオデータ、および/または(b)前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記ローカルで検出されたオーディオデータに基づいて、前記第1の追加のクライアントデバイスにおいてローカルで前記口頭の発話の前記第1の追加の候補表現を生成するかどうかを決定するステップとを含む、請求項1から7のいずれか一項に記載の方法。
【請求項９】
前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータの前記品質を示す前記オーディオ品質値を決定するステップは、
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ホンを識別するステップと、
前記クライアントデバイスの前記1つまたは複数のマイクロホンに基づいて前記オーディオ品質値を決定するステップとを含み、
前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記ローカルで検出されたオーディオデータの前記品質を示す前記追加のオーディオ品質値を決定するステップは、
前記第1の追加のクライアントデバイスの1つまたは複数の第1の追加のマイクロホンを識別するステップと、
前記第1の追加のクライアントデバイスの前記1つまたは複数の第1の追加のマイクロホンに基づいて前記追加のオーディオ品質値を決定するステップとを含む、請求項8に記載の方法。
【請求項１０】
前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータの前記品質を示す前記オーディオ品質値を決定するステップは、
前記口頭の発話をキャプチャする前記オーディオデータを処理することに基づいて信号対ノイズ比値を生成するステップと、
前記信号対ノイズ比値に基づいて前記オーディオ品質値を決定するステップとを含み、
前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記ローカルで検出されたオーディオデータの前記品質を示す前記追加のオーディオ品質値を決定するステップは、
前記口頭の発話をキャプチャする前記オーディオデータを処理することに基づいて追加の信号対ノイズ比値を生成するステップと、
前記追加の信号対ノイズ比値に基づいて前記追加のオーディオ品質値を決定するステップとを含む、請求項8または請求項9に記載の方法。
【請求項１１】
前記クライアントデバイスにおいて、前記第1の追加のクライアントデバイスから前記ローカルネットワークを介して、前記口頭の発話の第1の追加の候補テキスト表現を受信するステップに先立って、前記口頭の発話の前記第1の追加の候補テキスト表現を求める要求を前記第1の追加のクライアントデバイスに送信するかどうかを決定するステップと、
前記口頭の発話の前記第1の追加の候補テキスト表現を求める前記要求を前記第1の追加のクライアントデバイスに送信することを決定することに応答して、前記口頭の発話の前記第1の追加の候補テキスト表現を求める前記要求を前記第1の追加のクライアントデバイスに送信するステップと
をさらに含む、請求項1から10のいずれか一項に記載の方法。
【請求項１２】
前記口頭の発話の前記第1の追加の候補テキスト表現を求める前記要求を前記第1の追加

のクライアントデバイスに送信するかどうかを決定するステップは、
ホットワードモデルを使用して前記ユーザの前記口頭の発話をキャプチャする前記オーディオデータの少なくとも一部分を処理することに基づいてホットワード信頼度スコアを決定するステップであって、前記ホットワード信頼度スコアは、前記オーディオデータの少なくとも前記一部分がホットワードを含むかどうかの確率を示す、ステップと、
前記ホットワード信頼度スコアが1つまたは複数の条件を満たすかどうかを判定するステップであって、前記ホットワード信頼度スコアが前記1つまたは複数の条件を満たすかどうかを判定することは、前記ホットワード信頼度スコアがしきい値を満たすかどうかを判定することを含む、ステップと、
前記ホットワード信頼度スコアがしきい値を満たすと判定することに応答して、前記ホットワード信頼度スコアが、前記オーディオデータの少なくとも前記一部分が前記ホットワードを含む弱い確率を示すかどうかを判定するステップと、
前記ホットワード信頼度スコアが、前記オーディオデータの少なくとも前記一部分が前記ホットワードを含む前記弱い確率を示すと判定したことに応答して、前記口頭の発話の前記第1の追加の候補テキスト表現を求める前記要求を前記第1の追加のクライアントデバイスに送信することを決定するステップとを含む、請求項11に記載の方法。
【請求項１３】
1つまたは複数のプロセッサによって実行されると、前記1つまたは複数のプロセッサに、
クライアントデバイスにおいて、ユーザの口頭の発話をキャプチャするオーディオデータを検出することであって、
前記クライアントデバイスは、1つまたは複数の追加のクライアントデバイスを含む環境内にあり、ローカルネットワークを介して前記1つまたは複数の追加のクライアントデバイスとローカルで通信し、
前記1つまたは複数の追加のクライアントデバイスは、少なくとも第1の追加のクライアントデバイスを含む、こと、
前記クライアントデバイスにおいて、前記口頭の発話の候補テキスト表現を生成すべく、前記クライアントデバイスにおいてローカルで記憶された自動音声認識(「ASR」)モデルを使用して前記オーディオデータを処理すること、
前記クライアントデバイスにおいて、前記第1の追加のクライアントデバイスから前記ローカルネットワークを介して、前記口頭の発話の第1の追加の候補テキスト表現を受信することであって、
前記第1の追加のクライアントデバイスにおいてローカルで生成される前記口頭の発話の前記第1の追加の候補テキスト表現は、(a)前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータ、および(b)前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャするローカルで検出されたオーディオデータに基づき、
前記口頭の発話の前記第1の追加の候補テキスト表現は、前記第1の追加のクライアントデバイスにおいてローカルで記憶された第1の追加のASRモデルを使用して、前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータおよび前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記ローカルで生成されたオーディオデータを処理することによって生成される、こと、ならびに
前記口頭の発話のテキスト表現を、前記口頭の発話の前記候補テキスト表現、および前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現に基づいて決定すること
を含む動作を実行させる命令を記憶するように構成されたコンピュータ可読記憶媒体。
【請求項１４】
1つまたは複数のプロセッサと、
1つまたは複数のプロセッサによって実行されると、前記1つまたは複数のプロセッサに

、
クライアントデバイスにおいて、ユーザの口頭の発話をキャプチャするオーディオデータを検出することであって、
前記クライアントデバイスは、1つまたは複数の追加のクライアントデバイスを含む環境内にあり、ローカルネットワークを介して前記1つまたは複数の追加のクライアントデバイスとローカルで通信し、
前記1つまたは複数の追加のクライアントデバイスは、少なくとも第1の追加のクライアントデバイスを含む、こと、
前記クライアントデバイスにおいて、前記口頭の発話の候補テキスト表現を生成すべく、前記クライアントデバイスにおいてローカルで記憶された自動音声認識(「ASR」)モデルを使用して前記オーディオデータを処理すること、
前記クライアントデバイスにおいて、前記第1の追加のクライアントデバイスから前記ローカルネットワークを介して、前記口頭の発話の第1の追加の候補テキスト表現を受信することであって、
前記第1の追加のクライアントデバイスにおいてローカルで生成される前記口頭の発話の前記第1の追加の候補テキスト表現は、(a)前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータ、および(b)前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャするローカルで検出されたオーディオデータに基づき、
前記口頭の発話の前記第1の追加の候補テキスト表現は、前記第1の追加のクライアントデバイスにおいてローカルで記憶された第1の追加のASRモデルを使用して、前記クライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記オーディオデータおよび前記第1の追加のクライアントデバイスにおいて検出された前記口頭の発話をキャプチャする前記ローカルで生成されたオーディオデータを処理することによって生成される、こと、ならびに
前記口頭の発話のテキスト表現を、前記口頭の発話の前記候補テキスト表現、および前記第1の追加のクライアントデバイスによって生成された前記口頭の発話の前記第1の追加の候補テキスト表現に基づいて決定すること
を含む動作を実行させる命令を記憶するように構成されたメモリと、
を含むシステム。

【発明の詳細な説明】
【技術分野】
【０００１】
本願は、自動音声認識のローカル実行のためのデバイス調停に関する。
【背景技術】
【０００２】
自動音声認識(ASR)技術は、口頭の自然言語入力をテキストに変換する。例えば、マイクロホンを使用してキャプチャされたオーディオデータが、テキストに変換されることが可能である。ASRシステムは、候補認識のセットを生成する際に使用するためのASRモデルを含むことが可能である。ASRシステムは、生成されたテキストを候補認識のセットから選択することができる。
【０００３】
人間は、「自動化されたアシスタント」と本明細書で呼ばれる(「デジタルエージェント」、「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」、「アシスタントアプリケーション」、「会話エージェント」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を用いて人間対コンピュータの対話をすることができる。例えば、人間(自動化されたアシスタントと対話するとき、「ユーザ」と呼ばれてよい)は、一部の事例において、テキストに変換され(例えば、ASR技術を使用してテキストに変換され)、その後、処理されることが可能である口頭の自然言語入力(すなわち、発話)を使用して自動化されたアシスタントにコマンドおよび/または要求を与え

ることができる。
【発明の概要】
【課題を解決するための手段】
【０００４】
本明細書において説明される実装形態は、所与のクライアントデバイスにおいて生成された口頭の発話の候補テキスト表現に基づいて、および/または口頭の発話の1つまたは複数の追加の候補テキスト表現に基づいて、口頭の発話のテキスト表現を生成することを対象とする。口頭の発話の追加の候補テキスト表現のそれぞれは、所与のクライアントデバイスを含むローカル環境にあり、所与のクライアントデバイスと同一の室内にあり、1つまたは複数のローカルネットワークを利用して所与のクライアントデバイスと通信する、所与のクライアントデバイスの定義された範囲内にあり、同一のユーザアカウントに対応し、追加の様態もしくは代替の様態で所与のクライアントデバイスを含む環境内にある、および/またはこれらの組合せである、1つまたは複数の追加のクライアントデバイスのうちの対応する1つにおいてローカルで生成される。口頭の発話の候補テキスト表現は、口頭の発話をキャプチャし、所与のクライアントデバイスにおいてキャプチャされたオーディオデータを処理することによって生成されることが可能である。候補テキスト表現は、所与のクライアントデバイスにおいてローカルで記憶された自動音声認識(ASR)モデルを使用して生成される。追加の候補テキスト表現は、追加のクライアントデバイスにおいて、追加のクライアントデバイスにおいてローカルで記憶されたASRモデルを使用して、オーディオデータを処理することによって、追加のクライアントデバイスによって生成されることが可能である。追加のクライアントデバイスにおいて処理されるオーディオデータは、所与のクライアントデバイスにおいてキャプチャされたオーディオデータであることが可能であり(例えば、オーディオデータは、所与のクライアントデバイスから追加のクライアントデバイスに送信されることが可能である)、またはオーディオデータは、追加のクライアントデバイスの(1つまたは複数の)マイクロホンを介してキャプチャされる追加のオーディオデータであることが可能である。
【０００５】
1つの実例として、「set the thermostat to 70 degrees(サーモスタットを21℃に設定する)」という口頭の発話をキャプチャするオーディオデータが、ユーザの携帯電話においてキャプチャされることが可能であり、口頭の発話の候補テキスト表現が、ユーザの携帯電話においてローカルで記憶されたASRモデルを使用してオーディオデータを処理することによって生成されることが可能である。一部の実装形態において、口頭の発話をキャプチャするオーディオデータは、ラップトップ、自動化されたアシスタントスマートスピーカ、および/または自動化されたアシスタントスマートディスプレイなどの、携帯電話を含む環境内にある、追加のクライアントデバイスに送信されることも可能である。それらの実装形態において、追加のクライアントデバイスのそれぞれが、対応するローカルで記憶された追加のASRモデルを使用してオーディオデータを処理することによって、対応する追加の候補テキスト表現を生成することが可能である。追加の候補テキスト表現は、次に、ユーザの携帯電話に送信されることが可能であり、携帯電話が、候補テキスト表現(携帯電話において生成された)、および受信された追加の(1つまたは複数の)候補テキスト表現(それぞれ、追加の(1つまたは複数の)クライアントデバイスのうちの対応する1つにおいて生成された)に基づいてテキスト表現を生成することが可能である。例えば、対応する追加のクライアントデバイスによってそれぞれが生成された、2つの追加の候補テキスト表現が、携帯電話において受信されることが可能である。
【０００６】
次に、携帯電話が、2つの追加の候補テキスト表現、および候補テキスト表現に基づいて最終的なテキスト表現を決定することが可能である。様々な技術が、最終的なテキスト表現を決定する際に利用されることが可能である。例えば、候補テキスト表現は、(1つまたは複数の)信頼測度(例えば、各語または他の断片に関する対応する測度)を伴って生成されることが可能であり、追加の候補表現はそれぞれ、対応する(1つまたは複数の)信頼測度

を伴って受信されることが可能であり、携帯電話は、最終的なテキスト表現を決定する際、その(1つまたは複数の)信頼測度を使用することができる。例えば、所与の追加の候補表現が、その表現が最も高信頼度を示す(1つまたは複数の)信頼測度を有することに基づいて、最終的なテキスト表現として生成されることが可能である。別の実例として、最終的なテキスト表現は、候補テキスト表現および追加の候補テキスト表現の中で最も頻出する(most common)(1つまたは複数の)言葉(word piece(s))を含むように生成されることが可能である。例えば、候補テキスト表現が「get the thermostat to 70 degrees」であると想定すると、第1の追加の候補テキスト表現が「set the thermostat to 7 degrees」であり、第2の追加の候補テキスト表現が「set the thermometer to 70 degrees」である。そのような事例において、「set the thermostat to 70 degrees」が、最終的なテキスト表現として生成されることが可能であり、この場合、2回出現する「set」が、1回出現する「get」に優先されて選択され、2回出現する「thermostat」が、1回出現する「thermometer」に優先されて選択され、2回出現する「70」が、1回出現する「7」に優先されて選択される。

【０００７】

前段の実例は、携帯電話が、ローカルでキャプチャされたオーディオデータを、ローカルASRを実行する際に追加の(1つまたは複数の)クライアントデバイスによって使用されるように追加の(1つまたは複数の)クライアントデバイスに送信することについて説明する。しかし、前述したとおり、一部の実装形態において、追加の(1つまたは複数の)クライアントデバイスのうちの1つまたは複数は、追加または代替として、対応する候補テキスト表現を生成する際、追加のクライアントデバイスのマイクロホンを介してローカルでキャプチャされたオーディオデータを利用することができる。それらの実装形態の一部において、所与のクライアントデバイスは、その所与のクライアントデバイスにおいてキャプチャされたオーディオデータを、オプションとして、追加の(1つまたは複数の)クライアントデバイスのいずれにも送信しないことが可能である。実例として、前の実例をつづけると、「Hey Assistant, set the thermostat to 70 degrees」という口頭の発話の追加の候補テキスト表現が、追加のクライアントデバイスにおいてローカルで記憶されたASRモデルを使用して追加のオーディオデータを処理することによって、追加のクライアントデバイスによって生成されることが可能である。追加のオーディオデータは、口頭の発話をキャプチャすることが可能であり、追加のクライアントデバイスの(1つまたは複数の)マイクロホンを介してキャプチャされることが可能である。

【０００８】

一部の実装形態において、オプションとして、追加の(1つまたは複数の)クライアントデバイスのそれぞれに関して、ローカルASRを実行する際に追加のクライアントデバイスによって使用されるように、所与のクライアントデバイスから追加のクライアントデバイスにオーディオデータを送信するかどうかについての決定が行われる。例えば、所与のクライアントデバイス(またはシステムの他の(1つまたは複数の)構成要素)が、追加のクライアントデバイスのハードウェア能力および/またはソフトウェア能力に基づいて、所与のクライアントデバイスを使用することによってキャプチャされたオーディオデータを追加のクライアントデバイスに送信するかどうかを決定することができる。追加のクライアントデバイスのハードウェア能力および/またはソフトウェア能力は、所与のクライアントデバイスにおいてローカルで記憶されたホームグラフもしくは他のデータから、および/または追加のクライアントデバイスによって所与のクライアントデバイスに送信されたデータに基づいて確かめられることが可能である。例えば、システムは、追加のクライアントデバイスが低品質のマイクロホンを有すると判定された場合、所与のクライアントデバイスにおいてキャプチャされたオーディオデータを追加のクライアントデバイスに送信することが可能である。例えば、システムは、スマートウォッチが低品質のマイクロホンを有するという知識に基づいて、携帯電話においてキャプチャされたオーディオデータをスマートウォッチに送信してよい。追加または代替として、システムは、所与のデバイスを使用してキャプチャされたオーディオデータの特徴(例えば、信号対ノイズ比)を判定することが可

能であり、その特徴に基づいて、および、オプションとして、追加のクライアントデバイスにおいてキャプチャされた追加のオーディオデータの特徴(例えば、信号対ノイズ比)に基づいて、追加のクライアントデバイスにオーディオデータを送信するかどうかを決定することが可能である。例えば、システムは、信号対ノイズ比が、キャプチャされたオーディオデータが劣悪な品質のものであることを示す場合、所与のクライアントデバイスにおいてキャプチャされたオーディオデータを送信しないことを決定することが可能である。別の実例として、システムは、追加のオーディオデータの(1つまたは複数の)特徴が、そのオーディオデータが高品質のものであることを示す場合、および/またはそのオーディオデータが、所与のクライアントデバイスにおいてキャプチャされたオーディオデータと比べて、より良好な品質のものであることを示す場合、所与のクライアントデバイスにおいてキャプチャされたオーディオデータを送信しないことを決定することが可能である。追加または代替として、システムは、所与のクライアントデバイスと追加の(1つまたは複数の)クライアントデバイスの間の通信リンク(例えば、デバイス間の配線接続、デバイス間の無線接続、その他)に基づいて、オーディオデータを送信しないことを決定することが可能である。例えば、システムは、所与のクライアントデバイスと追加のクライアントデバイスの間に低帯域幅の接続が存在する場合、および/または所与のクライアントデバイスと追加のクライアントデバイスの間の接続に高い潜時が存在する場合、オーディオデータを送信しないことを決定することが可能である。

【０００９】

さらに別の実例として、システムは、所与のクライアントデバイスにおいてキャプチャされたオーディオデータを追加のクライアントデバイスに送信するかどうかを、所与のクライアントデバイスおよび/または追加のクライアントデバイスにおけるオーディオデータの履歴上のインスタンスに基づいて決定することが可能である。例えば、システムは、所与のクライアントデバイスにおいてキャプチャされたオーディオデータのインスタンスが、履歴上、低品質のものである場合、および/または追加のクライアントデバイスにおいてキャプチャされたオーディオデータのインスタンスが、履歴上、高品質のものである場合、オーディオデータを送信しないことを決定することが可能である。同様に、システムは、所与のクライアントデバイスにおいてキャプチャされたオーディオデータのインスタンスが、履歴上、高品質のものである場合、および/または追加のクライアントデバイスにおいてキャプチャされたオーディオデータのインスタンスが、履歴上、低品質のものである場合、オーディオデータを送信することを決定することが可能である。さらなる実例として、システムは、所与のクライアントデバイスにおいてキャプチャされたオーディオデータを追加のクライアントデバイスに送信するかどうかを、追加のクライアントデバイスが所与のクライアントデバイスに物理的に近接しているかどうか(例えば、現在の近接性を判定すべく、記憶されたホームグラフおよび/または能動的な技術を使用して判定された)に基づいて決定することが可能である。例えば、システムは、追加のクライアントデバイスが所与のクライアントデバイスと同一の室内にない場合(例えば、ホームグラフに基づいて判定されて)、および/または所与のクライアントデバイスからしきい値距離を超えている場合(例えば、所与のクライアントデバイスと追加のクライアントデバイスの間の距離を判定する能動的技術に基づいて判定されて)に限って、オーディオデータを送信することを決定することが可能である。追加の実例として、システムは、所与のクライアントデバイスにおいてキャプチャされたオーディオデータを追加のクライアントデバイスに送信するかどうかを、追加のクライアントデバイスが音声活動をローカルで検出したかどうか(例えば、ローカル音声活動検出器を使用して)に基づいて決定することが可能である。例えば、システムは、追加のクライアントデバイスが音声活動をローカルで検出しない場合に限って、オーディオデータを送信することを決定することが可能である。

【００１０】

追加のクライアントデバイスが所与のクライアントデバイスからオーディオデータを受信する一部の追加の実装形態または代替の実装形態において、追加のクライアントデバイスは、ローカルASRを実行する際、オーディオデータを利用するか、または、代わりに、

ローカルでキャプチャされた追加のオーディオデータを利用するかを決定することが可能である。それらの実装形態のいくつかにおいて、追加のクライアントデバイスは、オーディオデータを利用するか、または追加のオーディオデータを利用するかを決定する際、オーディオデータを送信するかどうかを決定することに関して前述した(1つまたは複数の)考慮のうちの1つまたは複数を利用することが可能である。例えば、追加のクライアントデバイスは、オーディオデータの信号対ノイズ比と追加のオーディオデータの信号対ノイズ比を比較して、より高い信号対ノイズ比を有するオーディオデータを利用することが可能である。

【００１１】

前述したとおり、一部の実装形態において、所与のクライアントデバイスは、1つまたは複数の追加のクライアントデバイスを含む環境内にあることが可能である。例えば、携帯電話である所与のクライアントデバイスが、ユーザのスマートウォッチ、スタンドアローンの対話型スピーカ、およびスマートカメラを含む環境内にあることが可能である。それらの実装形態のいくつかにおいて、システムは、口頭の発話の1つまたは複数の追加の候補テキスト表現を生成する際に使用すべく1つまたは複数の追加のクライアントデバイスのうちの1つまたは複数を選択することが可能である。例えば、システムは、1つまたは複数のクライアントデバイスとの履歴上の対話、1つまたは複数の追加のクライアントデバイスのハードウェア能力および/またはソフトウェア能力、その他に基づいて、追加のクライアントデバイスのうちの1つまたは複数を選択することが可能である。例えば、システムは、データが、追加のクライアントデバイスが、所与のクライアントデバイスのローカルASRモデルと比べて、より堅牢である、より正確である、および/またはより新しい、追加のクライアントデバイスによってASRにおいて使用される、ローカルで記憶されたASRモデルを含むことを示すことに基づいて、追加のクライアントデバイスを選択することが可能である。追加または代替として、システムは、ユーザと追加のクライアントデバイスの間の以前の対話に基づいて、追加のクライアントデバイスを選択することが可能である。例えば、システムは、追加のクライアントデバイスが、ユーザからより多くのクエリを受け取っていること(したがって、ユーザがASRモデルにフィードバックを与える機会がより多い)に基づいて、追加のクライアントデバイスを選択することが可能である。それらの実装形態のいくつかにおいて、ユーザによってより頻繁に使用される追加のクライアントデバイスにおけるASRモデルは、ユーザの音声によりうまく合わせられることが可能であり、口頭の発話のより正確な候補テキスト表現を生成し得る。

【００１２】

前述したとおり、一部の実装形態において、口頭の発話のテキスト表現が、所与のクライアントデバイスにおいて生成された口頭の発話の候補テキスト表現に基づいて、および1つまたは複数の対応する追加のクライアントデバイスにおいて生成された口頭の発話の1つまたは複数の追加の候補テキスト表現に基づいて、生成されることが可能である。例えば、システムは、口頭の発話の候補テキスト表現のうちの1つまたは複数を口頭の発話のテキスト表現としてランダムに(または疑似ランダムに)選択することが可能であり、システムは、口頭の発話のテキスト表現を、所与のクライアントデバイスと1つまたは複数の追加のクライアントデバイスの間の履歴上の対話に基づいて選択することが可能であり、システムは、口頭の発話のテキスト表現を、所与のクライアントデバイスおよび/または1つまたは複数の追加のクライアントデバイスのハードウェア構成および/またはソフトウェア構成に基づいて選択することが可能であり、システムは、テキスト表現を、追加の条件もしくは代替の条件が満たされるかどうかに基づいて選択することが可能であり、システムは、口頭の発話のテキスト表現を、候補テキスト表現の中で最も頻度が高い言葉および/または最高信頼度に基づいて選択することが可能であり、システムは、口頭の発話のテキスト表現を、最高信頼度の(1つまたは複数の)候補テキスト表現に基づいて選択することが可能であり、および/またはシステムは、これらの組合せ行うことが可能である。

【００１３】

例えば、システムは、所与のクライアントデバイスと第1の追加のクライアントデバイ

スの間の履歴上の対話が、第1の追加のクライアントデバイスが正確である候補テキスト表現をより頻繁に生成することを示すことに基づいて、第1の追加のクライアントデバイスを使用して生成された第1の追加の候補テキスト表現を口頭の発話のテキスト表現として選択することが可能である。追加または代替として、システムは、第2の追加のクライアントデバイスにローカルであり、第2の追加の候補テキスト表現を生成する際に利用されるASRモデルに関連付けられた品質メトリックおよび/または他の(1つまたは複数の)メトリックに基づいて、第2の追加のクライアントデバイスを使用して生成された第2の追加の候補テキスト表現を口頭の発話のテキスト表現として選択することが可能である。

【００１４】

したがって、様々な実装形態が、環境における多数のクライアントデバイスのうちの対応する1つによってそれぞれが実行されるローカル音声認識のインスタンスに基づいて口頭の発話のテキスト表現を生成するための技術を提示する。デバイス調停技術を使用して、ユーザを含む環境内の単一のクライアントデバイスが、ユーザによって話される口頭の発話のテキスト表現を生成すべく選択されることが可能である。しかし、環境における1つまたは複数の追加のクライアントデバイスが、口頭の発話のより正確なテキスト表現を生成することが可能である。例えば、第1の追加のクライアントデバイスが、選択されたクライアントデバイスと比べて、ASRモデルのより新しく、および/またはより堅牢な、および/またはより正確なバージョンを有することが可能であり、第2の追加のクライアントデバイスが、選択されたクライアントデバイスによってキャプチャされたオーディオデータのインスタンスと比べて、より少ないノイズを包含するオーディオデータのインスタンスにおいて口頭の発話をキャプチャすることが可能である、といった具合である。このため、本明細書において開示される実装形態は、ローカル音声認識を実行する際に追加の(1つまたは複数の)クライアントデバイスを少なくとも選択的に活用することが可能であり、口頭の発話の最終的なテキスト表現を生成する際、(1つまたは複数の)ローカル音声認識から生成された(1つまたは複数の)追加の候補テキスト表現の少なくとも一部を少なくとも選択的に利用することが可能である。これら、およびその他の実装形態は、より正確な音声認識および/またはより堅牢な音声認識が行われることをもたらすことができる。このことは、音声認識が正確である尤度がより高く、その認識に依拠する下流の(1つまたは複数の)プロセス(例えば、自然言語理解)が、より正確な音声認識に鑑みて、より正確に実行され得るので、より効率的な人間/コンピュータ対話を可能にする。したがって、音声認識の失敗により、ユーザが口頭の発話を繰り返す必要性が生じることが低減される。このことは、人間/コンピュータ対話の全体的な持続時間を短縮し、その結果、さもなければ、長引いた対話に要求されることになるネットワークリソースおよび/または計算リソースを低減する。

【００１５】

本明細書において開示される様々な実装形態は、口頭の発話の1つまたは複数の対応する追加の候補テキスト表現を生成するために所与のクライアントデバイスを含む環境内にある1つまたは複数の追加のクライアントデバイスを選択的に選択することを対象とし、口頭の発話のテキスト表現は、所与のクライアントデバイスを使用して生成された候補テキスト表現、および対応する1つまたは複数の追加のクライアントデバイスを使用して生成された1つまたは複数の候補テキスト表現に基づいて生成されることが可能である。言い換えると、本明細書において開示される一部の実装形態は、追加の(1つまたは複数の)候補テキスト表現を生成するために追加の(1つまたは複数の)クライアントデバイスを常に活用することはせず、および/または追加の(1つまたは複数の)候補テキスト表現を生成するために利用可能なすべての追加の(1つまたは複数の)クライアントデバイスを常に活用することもしない。そうではなく、一部の実装形態は、追加の(1つまたは複数の)候補テキスト表現を生成するために任意の追加の(1つまたは複数の)クライアントデバイスを選択的にだけ利用してよく、および/または追加の(1つまたは複数の)候補テキスト表現を生成するためにいくつかの追加の(1つまたは複数の)クライアントデバイスだけを選択的に利用してよい。それらの実装形態は、代わりに、1つまたは複数の基準に基づいて、追加の(1つまた

は複数の)クライアントデバイスを利用するかどうか、および/またはいずれの追加の(1つまたは複数の)クライアントデバイスを利用するかを決定することができる。そのような基準の考慮は、より正確な音声認識(および、もたらされる計算リソース節約、ネットワークリソース節約、および/またはシステム潜時)を求める要望と、より正確な音声認識に要求される計算リソースおよび/またはネットワークリソースの使用量とのバランスをとるように努めることが可能である。これら、およびその他の様態で、計算リソース(例えば、バッテリ電力、電源、プロセッササイクル、メモリ、その他)が、口頭の発話の1つまたは複数の追加の候補テキスト表現を生成することを選択的にだけ決定することによって節約することが可能である。

【００１６】

1つの実例として、所与のクライアントデバイスが、口頭の発話がホットワードを含む確率を示すホットワード信頼度スコアを決定することが可能であり、音声認識のための追加の(1つまたは複数の)クライアントデバイスを利用するかどうか、および/またはいくつの追加の(1つまたは複数の)クライアントデバイスを利用するかを決定する際にホットワード信頼度スコアを利用することが可能である。例えば、所与のクライアントデバイスは、ホットワード信頼度スコアが自動化されたアシスタントを呼び出すのに必要なしきい値を満たすが、ホットワード信頼度スコアは、第2のしきい値を満たすことができない(例えば、5%未満しかしきい値を超えてない)と判定することが可能である。このことは、劣悪な品質のオーディオデータストリームが口頭の発話をキャプチャしていることを潜在的に示すことが可能である。それらの実装形態のいくつかにおいて、システムは、ホットワードにおける識別されたより弱い信頼度に基づいて、1つまたは複数の対応する追加のクライアントデバイスを使用して、口頭の発話の1つまたは複数の追加の候補テキスト表現を生成することを決定することが可能である。口頭の発話の追加の候補テキスト表現を利用することは、口頭の発話のテキスト表現の精度を高めることが可能である。一部の事例において、このことは、システムが口頭の発話の誤ったテキスト表現を生成するのを防止することが可能であり、すると、そのことが、ユーザが口頭の発話を繰り返さなければならないことを防止することが可能である。

【００１７】

別の実例として、システムは、ホットワード信頼度スコアがしきい値を満たすと判定することに加えて、所与のクライアントデバイスが、ホットワード信頼度スコアがホットワードにおける非常に強い信頼度を示す(例えば、信頼度が10%以上、しきい値を超える)ことを示すと判定することが可能であると判定することが可能である。例えば、所与のクライアントデバイスが、ホットワード信頼度スコアがしきい値を余裕で満たすと判定することが可能であり、そのことが、良好な品質のオーディオデータストリームが口頭の発話をキャプチャすることを示してよい。それらの実装形態のいくつかにおいて、システムは、口頭の発話の1つまたは複数の対応する追加の候補テキスト表現を生成するのに追加のクライアントデバイスのいずれも利用しないことが可能である。口頭の発話の1つまたは複数の追加の対応する候補テキスト表現を生成する1つまたは複数の追加のクライアントデバイスのこの選択的な使用は、追加または代替として、システムが口頭の発話をキャプチャするオーディオデータストリームの品質を信頼している状況において、口頭の発話の1つまたは複数の追加の候補テキスト表現を生成するのに必要な計算リソースを節約することが可能である。

【００１８】

本明細書において説明される技術は、ASRモデルを使用して口頭の発話のテキスト表現を生成することに関する。しかし、このことは、限定することを意図していない。一部の実装形態において、本明細書において説明される技術は、追加または代替として、(1つまたは複数の)ローカル自然言語理解(NLU)モデルを使用して口頭の発話のテキスト表現を処理することに基づいて、口頭の発話の意図を特定するのに、および/または意図に関する(1つまたは複数の)パラメータを特定するのに使用されることが可能である。

【００１９】

前段の説明は、本明細書において開示される一部の実装形態の概観としてのみ与えられる。本技術のこれら、およびその他の実装形態が、後段でさらに詳細に開示される。

【００２０】

前述の概念と、本明細書においてさらに詳細に説明される追加の概念のすべての組合せが、本明細書において開示される主題の一部であることが企図されることを理解されたい。例えば、本開示の終わりに記載される請求される対象のすべての組合せが、本明細書において開示される主題の一部であることが企図される。

【図面の簡単な説明】

【００２１】

【図１】本明細書において開示される様々な実装形態による複数のクライアントデバイスを含む環境内にいるユーザの例を示す図である。

【図２】本明細書において開示される様々な実装形態による、クライアントデバイス、第1の追加のクライアントデバイス、および第2の追加のクライアントデバイスを使用して口頭の発話のテキスト表現を生成することの実例を示す図である。

【図３】本明細書において開示される様々な実装形態が実装されてよい例示的な環境を示す図である。

【図４】本明細書において開示される様々な実装形態による、口頭の発話のテキスト表現を生成する例示的なプロセスを示すフローチャートである。

【図５】本明細書において開示される様々な実装形態による、1つまたは複数の追加のクライアントデバイスのサブセットを選択する例示的なプロセスを示すフローチャートである。

【図６】本明細書において開示される様々な実装形態による、口頭の発話の追加の候補テキスト表現を生成する例示的なプロセスを示すフローチャートである。

【図７】本明細書において開示される様々な実装形態による、口頭の発話のテキスト表現を生成する別の例示的なプロセスを示すフローチャートである。

【図８】本明細書において開示される様々な実装形態が実装されてよい別の例示的な環境を示す図である。

【図９】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】

【００２２】

図1は、複数のクライアントデバイスを含む例示的な環境100内にいるユーザを示す。図示される実例において、ユーザ102は、携帯電話104、スマートウォッチ106、ディスプレイを有する自動化されたアシスタント108、Wi-Fiアクセスポイント110、スマートカメラ112、およびラップトップコンピュータ114を有する環境100内にいる。環境100におけるクライアントデバイスは、単に例示的であり、ユーザは、1つまたは複数の追加のおよび/または代替のクライアントデバイスを含む環境内にいることが可能である。例えば、環境は、デスクトップコンピュータ、ラップトップコンピュータ、タブレットコンピューティングデバイス、携帯電話、スマートウォッチ、1つまたは複数の追加の、または代替のウェアラブルコンピューティングデバイス、スタンドアローンの対話型スピーカ、組み込まれたディスプレイを有する自動化されたアシスタント、Wi-Fiアクセスポイント、スマートサーモスタット、スマートオーブン、スマートカメラ、1つまたは複数の追加の、または代替のスマートコンピューティングデバイス、1つまたは複数の追加の、または代替のコンピューティングデバイス、および/またはこれらの組合せのうちの1つまたは複数を含むことが可能である。

【００２３】

一部の実装形態において、ユーザを含む環境内にあるクライアントデバイスは、自動化されたアシスタントクライアントのインスタンスを実行することが可能である。例えば、スマートウォッチ106が、自動化されたアシスタントクライアントのインスタンスを実行することが可能であり、携帯電話104が、自動化されたアシスタントクライアントのインスタンスを実行することが可能であり、ディスプレイ108を有する自動化されたアシスタ

ントが、自動化されたアシスタントクライアントのインスタンスを実行することが可能であり、Wi-Fiアクセスポイント110が、自動化されたアシスタントクライアントのインスタンスを実行することが可能であり、スマートカメラ112が、自動化されたアシスタントクライアントのインスタンスを実行することが可能であり、および/またはラップトップコンピュータ114が、自動化されたアシスタントクライアントのインスタンスを実行することが可能である。

【００２４】

一部の実装形態において、異なるクライアントデバイスはそれぞれ、異なるハードウェア構成および/または異なるソフトウェア構成を含むことが可能である。例えば、携帯電話104のマイクロホンが、スマートウォッチ106のマイクロホンと比べて、より良好であることがあり得る。このことは、スマートウォッチ106を使用してキャプチャされた追加のオーディオデータストリームと比べて、携帯電話104がより高い品質のオーディオデータストリームをキャプチャすることにつながる可能性がある。追加または代替として、ラップトップコンピュータ114のASRモデルが、スマートカメラ112のASRモデルと比べて、より正確な候補テキスト予測を生成する可能性がある。

【００２５】

例示的な実例として、ユーザ102が、「Hey Assistant, turn on all the lights」という口頭の発話を話すことが可能である。環境100におけるクライアントデバイスのうちの1つまたは複数が、口頭の発話をキャプチャするオーディオデータをキャプチャすることが可能である。様々な要因が、1つまたは複数のクライアントデバイスのそれぞれにおいてキャプチャされるオーディオデータの品質に影響を及ぼす可能性がある。一部の実装形態において、クライアントデバイスに対する環境内におけるユーザの姿勢(例えば、ユーザの位置および/または向き)が、クライアントデバイスのうちの1つまたは複数においてキャプチャされるオーディオデータの品質に影響を及ぼす可能性がある。例えば、ユーザの前方にあるクライアントデバイスが、ユーザの背後にあるクライアントデバイスと比べて、口頭の発話のより高い品質のオーディオデータストリームをキャプチャすることがあり得る。

【００２６】

追加または代替として、環境におけるノイズの源(例えば、吠える犬、ホワイトノイズマシン、テレビからのオーディオデータ、1名または複数名の追加のユーザが話していること、ノイズの1つまたは複数の追加の、または代替の源、および/またはこれらの組合せ)が、クライアントデバイスにおいてキャプチャされるオーディオデータストリームの品質に影響を及ぼすことが可能である。例えば、ユーザが口頭の発話を話している間、環境内で犬が吠えていることが可能である。クライアントデバイスに対する環境内における犬の姿勢(例えば、犬の位置および/または向き)が、クライアントデバイスのうちの1つまたは複数においてキャプチャされるオーディオデータの品質に影響を及ぼすことが可能である。例えば、犬に最も近いクライアントデバイスが、犬から遠く離れているクライアントデバイスと比べて、より低い品質のオーディオデータストリームをキャプチャする可能性がある。言い換えると、犬に最も近いデバイスによってキャプチャされるオーディオデータストリームは、環境におけるその他のクライアントデバイスのうちの1つまたは複数と比べて、より高いパーセンテージの吠える犬、およびより低いパーセンテージの口頭の発話をキャプチャする可能性がある。追加の要因および/または代替の要因が、環境におけるクライアントデバイスにおいてキャプチャされるオーディオデータストリームの品質に影響を及ぼすことが可能である。

【００２７】

一部の実装形態において、システムは、環境におけるクライアントデバイスから所与のクライアントデバイスを決定することが可能である。例えば、システムは、携帯電話104を所与のクライアントデバイスとして選択することが可能であり、携帯電話104にローカルのASRモデルを使用して口頭の発話をキャプチャするオーディオデータを処理することによって、口頭の発話の候補テキスト表現を生成することが可能である。追加または代替

として、システムは、口頭の発話の対応する追加の候補テキスト表現を生成する環境における追加のクライアントデバイスのサブセットを選択することが可能である。一部の実装形態において、システムは、本明細書において説明される図5のプロセス404により1つまたは複数の追加のクライアントデバイスを選択することが可能である。例えば、システムは、ディスプレイを有する自動化されたアシスタント108、スマートカメラ112、およびラップトップコンピュータ114のサブセットを選択することが可能である。

【００２８】

一部の実装形態において、システムは、所与のクライアントデバイスにおいてキャプチャされた口頭の発話をキャプチャするオーディオデータを、追加のクライアントデバイスの選択されたサブセットに送信するかどうかを決定することが可能である。一部の実装形態において、システムは、所与のクライアントデバイスにおいてキャプチャされたオーディオデータを、追加のクライアントデバイスのサブセットのうちの1つまたは複数に送信するかどうかを決定することが可能である。追加または代替として、システムは、所与のクライアントデバイスにおいてキャプチャされた口頭の発話をキャプチャするオーディオデータを、様々な様態で1つまたは複数の追加のクライアントデバイスに送信することが可能である。例えば、システムは、オーディオデータの圧縮されたバージョン(例えば、非可逆オーディオ圧縮および/または可逆オーディオ圧縮を使用してオーディオデータを処理することによって生成された)を送信することが可能であり、オーディオデータの暗号化されたバージョンを送信することが可能であり、ストリーミングの様態で(例えば、潜時を最小限に抑えるべく発話が話されるにつれ、リアルタイムで、またはほぼリアルタイムで)、オーディオデータの処理されていないバージョンを、および/またはこれらの組合せでオーディオデータを送信することが可能である。

【００２９】

一部の実装形態において、システムは、口頭の発話の1つまたは複数の追加の候補テキスト表現を生成することが可能である。クライアントデバイスのサブセットの中のそれぞれの追加のクライアントデバイスに関して、クライアントデバイスは、所与のクライアントデバイスにおいてキャプチャされたオーディオデータ、および/または対応する追加のクライアントデバイスにおいてキャプチャされたオーディオデータに基づいて、対応する追加の候補テキスト表現を生成するかどうかを決定することが可能である。一部の実装形態において、追加のクライアントデバイスは、対応する追加のクライアントデバイスにローカルのASRモデルを使用して、選択されたオーディオデータを処理することによって発話の対応する追加の候補テキスト表現を生成することが可能である。一部の実装形態において、システムは、本明細書において説明される図6のプロセス408により、口頭の発話の1つまたは複数の追加の候補テキスト表現を生成することが可能である。例えば、システムは、口頭の発話の第1の追加の候補テキスト表現を、ディスプレイを有する自動化されたアシスタント108にローカルのASRモデルにおいてオーディオデータを処理することによって生成することが可能であり、口頭の発話の第2の追加の候補テキスト表現を、スマートカメラ112にローカルのASRモデルにおいてオーディオデータを処理することによって生成することが可能であり、口頭の発話の第3の候補テキスト表現を、ラップトップコンピュータ114にローカルのASRモデルにおいてオーディオデータを処理することによって生成することが可能である。

【００３０】

一部の実装形態において、所与のクライアントデバイスは、口頭の発話の候補テキスト表現に基づいて口頭の発話のテキスト表現を生成することが可能である。一部の実装形態において、システムは、本明細書において説明される図7のプロセス412により、口頭の発話のテキスト表現を生成することが可能である。例えば、システムは、携帯電話104を使用して生成された候補テキスト表現、ディスプレイを有する自動化されたアシスタント108を使用して生成された第1の追加の候補テキスト表現、スマートカメラ112を使用して生成された第2の追加の候補テキスト表現、および/またはラップトップコンピュータ114を使用して生成された第3の追加の候補テキスト表現に基づいて、口頭の発話のテキスト

表現を生成することが可能である。

【００３１】

図2は、様々な実装形態による口頭の発話の候補テキスト表現を生成することの実例200を示す。図示される実例200は、ユーザを含む環境においてクライアントデバイス202と、第1の追加のクライアントデバイス204と、第2の追加のクライアントデバイス206とを含む。例えば、クライアントデバイスは、ユーザの携帯電話であることが可能であり、第1の追加のクライアントデバイスは、ディスプレイを有する自動化されたアシスタントであることが可能であり、第2の追加のクライアントデバイスは、スマートカメラであることが可能である。一部の実装形態において、クライアントデバイス202、第1の追加のクライアントデバイス204、および/または第2の追加のクライアントデバイス206はそれぞれ、自動化されたアシスタントクライアントのインスタンスを実行することが可能である。

【００３２】

ポイント208において、クライアントデバイス202が、口頭の発話をキャプチャするオーディオデータをキャプチャすることが可能である。例えば、クライアントデバイス202は、「set the temperature to 72 degrees」という口頭の発話をキャプチャすることが可能である。一部の実装形態において、ポイント210において、第1の追加のクライアントデバイス204が、口頭の発話をキャプチャするオーディオデータの第1の追加のインスタンスをキャプチャすることが可能である。例えば、第1の追加のクライアントデバイスが、「set the temperature to 72 degrees」という口頭の発話の第1の追加のインスタンスをキャプチャすることが可能である。追加または代替として、ポイント212において、第2の追加のクライアントデバイス206が、口頭の発話をキャプチャするオーディオデータの第2の追加のインスタンスをキャプチャすることが可能である。例えば、第2の追加のクライアントデバイスは、「set the temperature to 72 degrees」という口頭の発話の第2の追加のインスタンスをキャプチャすることが可能である。

【００３３】

一部の実装形態において、異なる品質のオーディオデータが、クライアントデバイス、第1の追加のクライアントデバイス、および/または第2の追加のクライアントデバイスにおいてキャプチャされる。例えば、クライアントデバイスのうちの1つが、より良好な品質の(1つまたは複数の)マイクロホンを有して、それゆえ、対応するクライアントデバイスがより高い品質のオーディオデータストリームをキャプチャすることを可能にすることがあり得る。追加または代替として、背景ノイズ(例えば、追加のユーザが話していること、犬が吠えていること、電子デバイスによって生成されたノイズ、幼児が泣いていること、テレビからのオーディオ、ノイズの追加の、または代替の源、および/またはこれらの組合せ)が、オーディオデータストリームのうちの1つまたは複数においてキャプチャされることがあり得る。一部の実装形態において、1つのクライアントデバイスにおいて、別のクライアントデバイスと比べて、より多くの背景ノイズがキャプチャされることが可能である。例えば、犬が、第1の追加のクライアントデバイスに対して、第2の追加のクライアントデバイスに対するのと比べて、より近いことが可能であり、口頭の発話をキャプチャするオーディオデータの第1の追加のインスタンスが、口頭の発話をキャプチャするオーディオデータの第2の追加のインスタンスと比べて、犬が吠えているのをより多くキャプチャすることが可能である。一部の実装形態において、クライアントデバイスのうちの1つまたは複数が、オーディオデータをキャプチャするのに必要なユーザインターフェース入力能力を有さない可能性があり(例えば、クライアントデバイスが、マイクロホンを有さない)、したがって、(1つまたは複数の)クライアントデバイスが、ポイント208、ポイント210、および/またはポイント212において対応するオーディオデータをキャプチャしない可能性がある。

【００３４】

一部の実装形態において、ポイント214において、クライアントデバイス202が、口頭の発話をキャプチャするオーディオデータ(すなわち、ポイント208においてクライアント

デバイス202を使用してキャプチャされたオーディオデータ)を、第1の追加のクライアントデバイス204および/または第2の追加のクライアントデバイス206に送信することが可能である。他の一部の実装形態において、クライアントデバイス202が、第1の追加のクライアントデバイス204および/または第2の追加のクライアントデバイス206(図示せず)にオーディオデータを送信することが可能である。例えば、クライアントデバイス202は、オーディオデータが劣悪な品質であるという指標に基づいて、口頭の発話をキャプチャするオーディオデータを送信しないことが可能である。

【００３５】

ポイント216において、第1の追加のクライアントデバイス204が、クライアントデバイス202においてキャプチャされたオーディオデータおよび/またはポイント212においてキャプチャされたオーディオデータの第1の追加のインスタンスを処理するかどうかを決定することが可能である。一部の実装形態において、第1の追加のクライアントデバイス204が、本明細書において説明される図6のプロセス408により、オーディオデータおよび/またはオーディオデータの第1の追加のインスタンスを処理するかどうかを決定することが可能である。同様に、ポイント218において、第2の追加のクライアントデバイス206が、クライアントデバイス202においてキャプチャされたオーディオデータ、および/またはポイント212においてキャプチャされたオーディオの第2の追加のインスタンスを処理するかどうかを決定することが可能である。一部の実装形態において、第2の追加のクライアントデバイス206が、本明細書において説明される図6のプロセス408により、オーディオデータ、および/またはオーディオデータの第2の追加のインスタンスを処理するかどうかを決定することが可能である。

【００３６】

ポイント220において、クライアントデバイス202が、口頭の発話の候補テキスト表現を生成することが可能である。一部の実装形態において、クライアントデバイス202が、クライアントデバイス202においてローカルで記憶されたASRモデルを使用して、口頭の発話をキャプチャするキャプチャされたオーディオデータを処理することによって、口頭の発話の候補テキスト表現を生成することが可能である。一部の実装形態において、第1の追加のクライアントデバイス204が、ポイント222において口頭の発話の第1の追加の候補テキスト表現を生成することが可能である。一部の実装形態において、口頭の発話の第1の追加の候補テキスト表現が、第1の追加のクライアントデバイスにおいてローカルで記憶されたASRモデルを使用してオーディオデータ、および/またはオーディオデータの第1の追加のインスタンスを処理することによって生成されることが可能である。一部の実装形態において、口頭の発話の第1の追加の候補テキスト表現が、本明細書において説明される図6のプロセス408により生成されることが可能である。同様に、ポイント224において、口頭の発話の第2の追加の候補テキスト表現が、第2の追加のクライアントデバイス206を使用して生成されることが可能である。一部の実装形態において、口頭の発話の第2の追加の候補テキスト表現が、第2の追加のクライアントデバイスにおいてローカルで記憶されたASRモデルを使用してオーディオデータ、および/またはオーディオデータの第2の追加のインスタンスを処理することによって生成されることが可能である。

【００３７】

ポイント226において、第1の追加のクライアントデバイス204が、口頭の発話の第1の追加の候補テキスト表現をクライアントデバイス202に送信することが可能である。同様に、ポイント228において、第2の追加のクライアントデバイス206が、口頭の発話の第2の追加の候補テキスト表現をクライアントデバイス202に送信することが可能である。

【００３８】

ポイント230において、クライアントデバイス202が、口頭の発話のテキスト表現を生成することが可能である。一部の実装形態において、クライアントデバイス202が、口頭の発話の候補テキスト表現、口頭の発話の第1の追加の候補テキスト表現、および/または口頭の発話の第2の追加の候補テキスト表現に基づいて、口頭の発話のテキスト表現を生成することが可能である。一部の実装形態において、クライアントデバイス202は、本明

細書において説明される図7のプロセス412により、口頭の発話のテキスト表現を生成することが可能である。

【００３９】

図3は、本明細書において開示される実装形態が実装されてよい例示的な環境300のブロック図を示す。例示的な環境300は、クライアントデバイス302と、追加のクライアントデバイス314とを含む。クライアントデバイス302は、(1つまたは複数の)ユーザインターフェース入出力デバイス304、候補テキスト表現エンジン306、テキスト表現エンジン308、追加のデバイスエンジン310、(1つまたは複数の)追加のまたは代替のエンジン(図示せず)、ASRモデル312、および/または(1つまたは複数の)追加のまたは代替のモデル(図示せず)を含むことが可能である。追加のクライアントデバイス314は、(1つまたは複数の)追加のユーザインターフェース入出力デバイス316、オーディオ源エンジン318、追加の候補テキスト表現エンジン320、(1つまたは複数の)追加のまたは代替のエンジン(図示せず)、追加のASRモデル322、および/または(1つまたは複数の)追加のまたは代替のモデル(図示せず)を含むことが可能である。

【００４０】

一部の実装形態において、クライアントデバイス302および/または追加のクライアントデバイス314は、ユーザインターフェース入出力デバイスを含んでよく、含んでよくユーザインターフェース入出力デバイスは、例えば、物理キーボード、タッチスクリーン(例えば、仮想キーボードもしくは他のテキスト入力機構を実装する)、マイクロホン、カメラ、ディスプレイ画面、および/または(1つまたは複数の)スピーカを含んでよい。例えば、ユーザの携帯電話が、ユーザインターフェース入出力デバイスを含んでよく、スタンドアローンのデジタルアシスタントハードウェアデバイスが、ユーザインターフェース入出力デバイスを含んでよく、第1のコンピューティングデバイスが、(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含んでよく、別個のコンピューティングデバイスが、(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含んでよい、といった具合である。一部の実装形態において、クライアントデバイス302および/または追加のクライアントデバイス314のすべて、または態様が、ユーザインターフェース入出力デバイスも包含するコンピューティングシステム上に実装されてよい。

【００４１】

クライアントデバイス302および/または追加のクライアントデバイス314の一部の非限定的な実例が、デスクトップコンピューティングデバイス、ラップトップコンピューティングデバイス、自動化されたアシスタントに少なくとも一部が専用であるスタンドアローンのハードウェアデバイス、タブレットコンピューティングデバイス、携帯電話コンピューティングデバイス、車両のコンピューティングデバイス(例えば、車両内通信システム、および車両内エンターテインメントシステム、車両内ナビゲーションシステム、車両内ナビゲーションシステム)、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、仮想現実コンピューティングデバイスもしくは拡張現実コンピューティングデバイス)のうちの1つまたは複数を含む。追加のコンピューティングシステムおよび/または代替のコンピューティングシステムが、備えられてよい。クライアントデバイス302および/または追加のクライアントデバイス314は、データおよびソフトウェアアプリケーションを記憶するための1つまたは複数のメモリと、データにアクセスするため、およびアプリケーションを実行するための1つまたは複数のプロセッサと、ネットワークを介した通信を円滑にするための他の構成要素とを含むことが可能である。クライアントデバイス302および/または追加のクライアントデバイス314によって実行される動作は、多数のコンピューティングデバイスにわたって分散されてよい。例えば、1つまたは複数のロケーションにおいて1つまたは複数のコンピュータ上で実行されるコンピューティングプログラムが、ネットワークを介して互いに結合されることが可能である。

【００４２】

一部の実装形態において、クライアントデバイス302が、(1つまたは複数の)ユーザイン

ターフェース入出力デバイス304を含んでよく、追加のクライアントデバイス314が、(1つまたは複数の)追加のユーザインターフェース入出力デバイス316を含むことが可能であり、デバイス316は、例えば、物理キーボード、タッチスクリーン(例えば、仮想キーボードもしくは他のテキスト入力機構を実装する)、マイクロホン、カメラ、ディスプレイ画面、および/または(1つまたは複数の)スピーカを含んでよい。一部の実装形態において、クライアントデバイス302および/または追加のクライアントデバイス314が、自動化されたアシスタント(図示せず)を含んでよく、自動化されたアシスタントのすべて、または態様が、ユーザインターフェース入出力デバイスを包含するクライアントデバイス(例えば、すべて、または態様が、「クラウドにおいて」実装されてよい)とは別個であり、遠隔である(1つまたは複数の)コンピューティングデバイス上に実装されてよい。それらの実装形態のいくつかにおいて、自動化されたアシスタントのそれらの態様は、ローカルエリアネットワーク(LAN)および/またはワイドエリアネットワーク(WAN)(例えば、インターネット)などの1つまたは複数のネットワークを介してコンピューティングデバイスと通信してよい。

【００４３】

一部の実装形態において、(1つまたは複数の)ユーザインターフェース入出力デバイス304が、ユーザによって話された口頭の発話をキャプチャするオーディオデータをキャプチャすることが可能である。例えば、クライアントデバイス304の1つまたは複数のマイクロホンが、「Hey Assistant, set an alarm for 8am」という口頭の発話をキャプチャするオーディオデータをキャプチャすることが可能である。一部の実装形態において、候補テキスト表現エンジン306が、ASRモデル312を使用して、口頭の発話をキャプチャするオーディオデータを処理して、口頭の発話の候補テキスト表現を生成することが可能である。

【００４４】

追加または代替として、追加のデバイスエンジン310が、環境300における1つまたは複数の追加のクライアントデバイスのサブセットを選択するのに使用されることが可能であり、クライアントデバイス302においてキャプチャされたオーディオデータを1つまたは複数の選択された追加のクライアントデバイスに送信するかどうかを決定するのに使用されることが可能であり、および/または口頭の発話をキャプチャするオーディオデータを1つまたは複数の選択された追加のクライアントデバイスに送信するのに使用されることが可能である。一部の実装形態において、追加のデバイスエンジン310が、本明細書において説明される図5のプロセス404により1つまたは複数の追加のクライアントデバイスのサブセットを選択することが可能である。例えば、追加のデバイスエンジン310は、追加のクライアントデバイス314を選択するのに使用されることが可能である。追加または代替として、追加のデバイスエンジン310は、クライアントデバイス302においてキャプチャされたオーディオデータを1つまたは複数の追加のクライアントデバイスに送信するかどうかを決定することが可能である。

【００４５】

一部の実装形態において、テキスト表現エンジン308が、クライアントデバイス302を使用して生成された口頭の発話の候補テキスト表現、および/または1つまたは複数の対応する追加のクライアントデバイスを使用して生成された口頭の発話の1つまたは複数の追加の候補テキスト表現に基づいて、口頭の発話のテキスト表現を生成するのに使用されることが可能である。例えば、テキスト表現エンジン308が、クライアントデバイス302を使用して生成された口頭の発話の候補テキスト表現、および/または追加のクライアントデバイス314を使用して生成された口頭の発話の追加の候補テキスト表現に基づいて、口頭の発話のテキスト表現を生成することが可能である。一部の実装形態において、テキスト表現エンジン308が、本明細書において説明される図7のプロセス412により口頭の発話のテキスト表現を生成することが可能である。

【００４６】

一部の実装形態において、追加のクライアントデバイス314が、(1つまたは複数の)追加のユーザインターフェース入出力デバイス316を使用して、口頭の発話をキャプチャする

オーディオデータの追加のインスタンスをキャプチャすることが可能である。例えば、追加のクライアントデバイス314は、追加のクライアントデバイスの1つまたは複数の追加のマイクロホンを使用して、「Hey Assistant, set an alarm for 8am」という口頭の発話の追加のインスタンスをキャプチャすることが可能である。一部の実装形態において、追加のクライアントデバイス314が、口頭の発話の追加の候補テキスト表現を生成するように、クライアントデバイス302を使用して生成された口頭の発話をキャプチャするオーディオデータ、および/または追加のクライアントデバイス314を使用して生成された口頭の発話をキャプチャする追加のオーディオデータを処理するかどうかを決定するのにオーディオ源エンジン318を使用することが可能である。一部の実装形態において、追加のクライアントデバイス314は、追加のASRモデル322を使用するオーディオ源エンジン318を使用して選択されたオーディオデータを処理することによって、口頭の発話の追加の候補テキスト表現を生成するように追加の候補テキスト表現エンジン320を使用することが可能である。一部の実装形態において、追加の候補テキスト表現エンジン320は、本明細書において説明される図6のプロセス408により口頭の発話の追加の候補テキスト表現を生成することが可能である。

【００４７】

図4は、本明細書において開示される様々な実装形態による、口頭の発話の候補テキスト表現を生成する例示的なプロセス400を示すフローチャートである。便宜のため、フローチャートの動作は、動作を実行するシステムに関連して説明される。このシステムは、クライアントデバイス302、追加のクライアントデバイス314、クライアントデバイス802、および/またはコンピューティングシステム810のうちの1つまたは複数の構成要素などの、様々なコンピュータシステムの様々な構成要素を含んでよい。さらに、プロセス400の動作は、特定の順序で示されるが、このことは、限定することは意図していない。1つまたは複数の動作が、並べ替えられること、省かれること、および/または追加されることが可能である。

【００４８】

ブロック402において、システムが、クライアントデバイスにおいて口頭の発話のオーディオデータをキャプチャし、クライアントデバイスは、1つまたは複数の追加のクライアントデバイスを含む環境内にある。一部の実装形態において、クライアントデバイスおよび/または追加のクライアントデバイスが、自動化されたアシスタントクライアントの対応するインスタンスを実行することが可能である。一部の実装形態において、ユーザが、携帯電話、ラップトップコンピュータ、スタンドアローンの自動化されたアシスタント、その他などの、いくつかのクライアントデバイスを有する室内にいることが可能である。一部の実装形態において、2つ以上のクライアントデバイスが、ユーザによって話された口頭の発話をキャプチャすることができる場合、従来のデバイス調停技術が、口頭の発話を処理するのに使用される所与のクライアントデバイスを決定するのに使用されることが可能である。例えば、口頭の発話をキャプチャするオーディオデータは、スタンドアローンの対話型スピーカの所与のクライアントデバイスにおいてキャプチャされることが可能であり、スタンドアローンの対話型スピーカは、携帯電話の第1の追加のクライアントデバイス、およびスマートカメラの第2の追加のクライアントデバイスを含む環境内にあることが可能である。

【００４９】

ブロック404において、システムは、1つまたは複数の追加のクライアントデバイスのサブセットを選択する。一部の実装形態において、システムは、図5に示されるプロセス404により1つまたは複数の追加のクライアントデバイスのサブセットを選択することが可能である。例えば、システムは、携帯電話の第1の追加のクライアントデバイス、スマートカメラの第2の追加のクライアントデバイス、または携帯電話の第1の追加のクライアントデバイスとスマートカメラの第2の追加のクライアントデバイスを選択することが可能である。

【００５０】

ブロック406において、システムは、ローカルASRモデルを使用して、キャプチャされたオーディオデータを処理することによって、口頭の発話の候補テキスト表現を生成する。一部の実装形態において、口頭の発話の候補テキスト表現は、ASRモデルを使用して生成された最高の格付けの仮説であることが可能である。追加または代替として、口頭の発話の候補テキスト表現は、ASRモデルを使用して生成された多数の仮説を含むことが可能である。

【００５１】

ブロック408において、システムは(オプションとして)、1つまたは複数の追加のクライアントデバイスにおいて口頭の発話の1つまたは複数の追加の候補テキスト表現を生成する。一部の実装形態において、システムは、図6に示されるプロセス408により、1つまたは複数の追加のクライアントデバイスにおいて1つまたは複数の追加の候補テキスト表現を生成することが可能である。例えば、システムは、第1の追加のクライアントデバイスにおいてローカルで記憶された第1の追加のASRモデルを使用して口頭の発話の第1の追加の候補テキスト表現を生成することが可能であり、および/またはシステムは、第2の追加のクライアントデバイスにおいてローカルで記憶された第2の追加のASRモデルを使用して、口頭の発話の第2の追加の候補テキスト表現を生成することが可能である。

【００５２】

ブロック410において、システムは、1つまたは複数の追加のクライアントデバイスの選択されたサブセットから口頭の発話の1つまたは複数の追加の候補テキスト表現を受信する。例えば、システムが、ブロック404において、第1の追加のクライアントデバイスおよび第2の追加のクライアントデバイスを選択した場合、システムは、第1の追加のクライアントデバイスにおいて生成された(例えば、図6のプロセス408により生成された)第1の追加の候補テキスト表現、および第2の追加のクライアントデバイスにおいて生成された(例えば、図6のプロセス408により生成された)第2の追加の候補テキスト表現を受信することが可能である。

【００５３】

ブロック412において、システムは、口頭の発話の候補テキスト表現、および口頭の発話の1つまたは複数の追加の候補テキスト表現に基づいて、口頭の発話のテキスト表現を生成する。一部の実装形態において、システムは、図7のプロセス412により、候補テキスト表現、および1つまたは複数の追加の候補テキスト表現に基づいて、口頭の発話のテキスト表現を生成することが可能である。

【００５４】

図5は、本明細書において開示される様々な実装形態による、1つまたは複数の追加のクライアントデバイスのサブセットを選択する例示的なプロセス404を示すフローチャートである。便宜のため、フローチャートの動作は、動作を実行するシステムに関連して説明される。このシステムは、クライアントデバイス302、追加のクライアントデバイス314、クライアントデバイス802、および/またはコンピューティングシステム810のうちの1つまたは複数の構成要素などの、様々なコンピュータシステムの様々な構成要素を含んでよい。さらに、プロセス404の動作は、特定の順序で示されるが、このことは、限定することは意図していない。1つまたは複数の動作が、並べ替えられること、省かれること、および/または追加されることが可能である。

【００５５】

ブロック502において、システムは、1つまたは複数の追加のクライアントデバイスのうちの追加のクライアントデバイスを選択し、1つまたは複数の追加のクライアントデバイスは、所与のクライアントデバイスを含む環境内にある。例えば、所与のクライアントデバイスが、第1の追加のクライアントデバイス、第2の追加のクライアントデバイス、および第3の追加のクライアントデバイスを含む環境内にあることが可能である。

【００５６】

ブロック504において、システムは、1つまたは複数のクライアントデバイスパラメータに基づいて、追加のクライアントデバイスを選択するかどうかを決定する。一部の実装

形態において、1つまたは複数のクライアントデバイスパラメータは、クライアントデバイスの電源、クライアントデバイスのハードウェア(例えば、クライアントデバイスが(1つまたは複数の)マイクロホン、プロセッサ、利用可能なメモリ、その他を有するかどうか)、クライアントデバイスのソフトウェア(例えば、ASRモデルバージョン、ASRモデルサイズ、ASRモデル容量、1つまたは複数の追加のモデルバージョンまたは代替のモデルバージョン、その他)、1つまたは複数の追加のデバイスパラメータまたは代替のデバイスパラメータ、および/またはこれらの組合せを含むことが可能である。例えば、一部の実装形態において、システムは、サブセットの中の1つまたは複数の追加のクライアントデバイスの各デバイスを含むことが可能である。

【００５７】

一部の実装形態において、システムは、電源コンセントに差し込まれることによって給電される1つまたは複数の追加のクライアントデバイスの各デバイス(例えば、交流で動作する各クライアントデバイス)を選択することが可能である。言い換えると、システムは、電力費用が無視できるようなものである場合、(1つまたは複数の)追加のクライアントデバイスを選択することが可能である。一部の実装形態において、システムは、クライアントデバイスのバッテリ電源が1つまたは複数の条件を満たす場合、追加のクライアントデバイスを選択することが可能である。例えば、システムは、残っているバッテリ電源がしきい値を超える(例えば、バッテリに25%を超える電力が残っている)場合、バッテリの容量がしきい値を超える(例えば、バッテリ容量が1000mAhを超える)場合、バッテリが、現在、充電中である場合、追加の(1つまたは複数の)条件または代替の(1つまたは複数の)条件が満たされる場合、および/またはこれらの組合せで、追加のクライアントデバイスを選択することが可能である。一部の実装形態において、システムは、追加のクライアントデバイスのハードウェアに基づいて、追加のクライアントデバイスを選択することが可能である。例えば、システムは、1つまたは複数のクライアントデバイスのサブセットを選択すべく機械学習モデルを使用して1つまたは複数の追加のクライアントデバイスの各デバイスのハードウェアを処理することが可能である。

【００５８】

一部の実装形態において、システムは、追加のクライアントデバイスが、プロセスの先行する繰り返しの中で以前に選択されているかどうかに基づいて、追加のクライアントデバイスを選択することが可能である。例えば、システムは、以前の口頭の発話を処理するときに、第1の追加のクライアントデバイスが選択されており、第2の追加のクライアントデバイスが選択されていないとシステムが判定した場合、第1の追加のクライアントデバイスを選択し、第2の追加のクライアントデバイスを選択しないことが可能である。

【００５９】

システムは、クライアントデバイスにおいてASRモデルを使用して生成された候補テキスト表現の信頼度を示す信頼度値を決定することが可能である。一部の実装形態において、システムは、信頼度値がしきい値を満たすかどうかなど、信頼度値が1つまたは複数の条件を満たすかどうかを判定することが可能である。システムは、信頼度値が候補テキスト表現における低い信頼度を示す場合、1つまたは複数の追加のクライアントデバイスを選択することが可能である。例えば、システムは、信頼度値がしきい値を下回る場合、1つまたは複数の追加のクライアントデバイスを選択することが可能である。

【００６０】

ブロック506において、システムは、さらなる追加のクライアントデバイスを選択するかどうかを決定する。一部の実装形態において、システムは、使用されていない追加のクライアントデバイスが残っているかどうか、しきい値数の追加のクライアントデバイスが選択されているかどうか、1つまたは複数の追加の条件または代替の条件が満たされるかどうか、および/またはこれらの組合せに基づいて、さらなる追加のクライアントデバイスを選択するかどうかを決定することが可能である。肯定的である場合、システムは、ブロック502に戻り、さらなる追加のクライアントデバイスを選択し、さらなる追加のクライアントデバイスに基づいてブロック504に進む。否定的である場合、プロセスは、終了す

る。

【００６１】

図6は、本明細書において開示される様々な実装形態による口頭の発話の追加の候補テキスト表現を生成する例示的なプロセス408を示すフローチャートである。便宜のため、フローチャートの動作は、動作を実行するシステムに関連して説明される。このシステムは、クライアントデバイス302、追加のクライアントデバイス314、クライアントデバイス802、および/またはコンピューティングシステム810の1つまたは複数の構成要素などの、様々なコンピュータシステムの様々な構成要素を含んでよい。さらに、プロセス408の動作は、特定の順序で示されるが、このことは、限定することは意図していない。1つまたは複数の動作が、並べ替えられること、省かれること、および/または追加されることが可能である。

【００６２】

ブロック602において、システムは、追加のクライアントデバイスにおいて、口頭の発話をキャプチャするオーディオデータの追加のインスタンスをキャプチャする。例えば、追加のクライアントデバイスは、「Hey Assistant, what is the temperature on Tuesday」という口頭の発話をキャプチャすることが可能である。

【００６３】

ブロック604において、システムは、追加のクライアントデバイスにおいて、所与のクライアントデバイスにおいてキャプチャされた口頭の発話をキャプチャするオーディオデータのインスタンスを受信し、所与のクライアントデバイスは、追加のクライアントデバイスを含む環境内にある。例えば、追加のクライアントデバイスは、所与のクライアントデバイスにおいてキャプチャされた「Hey Assistant, what is the temperature on Tuesday」という口頭の発話をキャプチャするオーディオデータを受信することが可能である。

【００６４】

ブロック606において、システムは、オーディオデータの追加のインスタンスとオーディオデータの受信されたインスタンスを比較する。

【００６５】

ブロック608において、システムは、その比較に基づいて、オーディオデータの追加のインスタンス、および/またはオーディオデータの受信されたインスタンスを処理するかどうかを決定する。一部の実装形態において、システムは、処理のためにオーディオデータのインスタンス、またはオーディオデータの追加のインスタンスをランダムに(または疑似ランダムに)選択することが可能である。一部の実装形態において、システムは、オーディオデータのインスタンスとオーディオデータの追加のインスタンスをともに選択することが可能である。一部の実装形態において、システムは、オーディオデータの品質に基づいて、処理のためにオーディオデータを選択することが可能である。例えば、システムは、追加のクライアントデバイスのマイクロホン、および/または所与のクライアントデバイスのマイクロホンに基づいて、オーディオデータの追加のインスタンス、またはオーディオデータのインスタンスを選択することが可能である。例えば、システムは、追加のクライアントデバイスのマイクロホンが所与のクライアントデバイスのマイクロホンと比べて、より良好な品質のオーディオデータをキャプチャする場合、オーディオデータの追加のインスタンスを選択することが可能である。

【００６６】

追加または代替として、システムは、オーディオデータのインスタンスに関する信号対ノイズ比、およびオーディオデータの追加のインスタンスに関する追加の信号対ノイズ比を決定することが可能である。システムは、より良好な品質のオーディオデータストリームを示す信号対ノイズ比を有するオーディオデータのインスタンスを選択することが可能である。追加の、または代替の知覚的品質メトリックが、より良好な品質のオーディオデータストリームを決定する際に利用されることが可能である。例えば、オーディオデータストリームの品質レベルを予測すべく訓練されている機械学習モデルが、オーディオデー

タストリームを選択する際に利用されることが可能である。

【００６７】

ブロック610において、システムは、口頭の発話の追加の候補テキスト表現を生成すべく、追加のクライアントデバイスにおいてローカルで記憶された追加のASRモデルを使用して、決定されたオーディオデータを処理する。例えば、オーディオデータの追加のインスタンスが処理のために選択された場合、システムは、追加のクライアントデバイスにおいてローカルで記憶された追加のASRモデルを使用して、オーディオデータの追加のインスタンスを処理することによって、口頭の発話の追加の候補テキスト表現を生成することが可能である。さらなる実例として、オーディオデータのインスタンスが処理のために選択された場合、システムは、追加のクライアントデバイスにおいてローカルで記憶された追加のASRモデルを使用してオーディオデータのインスタンスを処理することによって、口頭の発話の追加の候補テキスト表現を生成することが可能である。

【００６８】

ブロック612において、システムは、口頭の発話の追加の候補テキスト表現を所与のクライアントデバイスに送信する。

【００６９】

図7は、本明細書において開示される様々な実装形態による口頭の発話のテキスト表現を生成する例示的なプロセス412を示すフローチャートである。便宜のため、フローチャートの動作は、動作を実行するシステムに関連して説明される。このシステムは、クライアントデバイス302、追加のクライアントデバイス314、クライアントデバイス802、および/またはコンピューティングシステム810の1つまたは複数の構成要素などの、様々なコンピュータシステムの様々な構成要素を含んでよい。さらに、プロセス412の動作は、特定の順序で示されるが、このことは、限定することは意図していない。1つまたは複数の動作が、並べ替えられること、省かれること、および/または追加されることが可能である。

【００７０】

ブロック702において、システムは、クライアントデバイスにおいて口頭の発話のオーディオデータをキャプチャし、クライアントデバイスは、1つまたは複数の追加のクライアントデバイスを含む環境内にある。例えば、スタンドアローンの対話型スピーカが、「Hey Assistant, turn off the living room lights」という口頭の発話をキャプチャするオーディオデータをキャプチャすることが可能であり、スタンドアローンの対話型スピーカは、携帯電話およびスマートテレビを含む環境内にある。

【００７１】

ブロック704において、システムは、ローカルASRモデルを使用してクライアントデバイスにおいてオーディオデータを処理することによって、口頭の発話の候補テキスト表現を生成する。例えば、システムは、口頭の発話の候補テキスト表現を生成すべくスタンドアローンの対話型スピーカにローカルのASRモデルを使用して、「Hey Assistant, turn off the living room lights」という口頭の発話をキャプチャするオーディオデータを処理することが可能である。一部の実装形態において、システムは、ローカルASRモデルを使用して口頭の発話の候補テキスト表現を生成することが可能である。追加または代替として、システムは、ローカルASRモデルを使用して、口頭の発話のテキスト表現の多数の仮説を生成することが可能である。

【００７２】

ブロック706において、システムは、1つまたは複数の追加のクライアントデバイスから口頭の発話の1つまたは複数の候補テキスト表現を受信する。例えば、システムは、携帯電話から「Hey Assistant, turn off the living room lights」という口頭の発話の第1の追加の候補テキスト表現を、スマートテレビから「Hey Assistant, turn off the living room lights」という口頭の発話の第2の追加の候補テキスト表現を受信することが可能である。一部の実装形態において、1つまたは複数の追加の候補テキスト表現が、本明細書において説明される図6のプロセス408により1つまたは複数の追加のクライアント

デバイスを使用して生成されることが可能である。一部の実装形態において、システムは、追加のクライアントデバイスにローカルである対応するローカルASRモデルを使用して、1つまたは複数の追加のクライアントデバイスの各デバイスから口頭の発話の追加の候補テキスト表現を受信することが可能である。他の一部の実装形態において、システムは、追加のクライアントデバイスにローカルの対応するローカルASRモデルを使用して生成された、1つまたは複数の追加のクライアントデバイスの各デバイスからの口頭の発話の多数の候補テキスト表現を受信することが可能である。

【００７３】

ブロック708において、システムは、口頭の発話の候補テキスト表現を口頭の発話の1つまたは複数の追加の候補テキスト表現と比較する。

【００７４】

ブロック710において、システムは、その比較に基づいて、口頭の発話のテキスト表現を生成する。一部の実装形態において、システムは、口頭の発話の候補テキスト表現のうちの1つを口頭の発話のテキスト表現としてランダムに(または疑似ランダムに)選択することが可能である。例えば、システムは、第1の追加のクライアントデバイスを使用して生成された口頭の発話の候補テキスト表現を口頭の発話のテキスト表現としてランダムに(または疑似ランダムに)選択することが可能である。追加または代替として、システムは、所与のクライアントデバイスを使用して生成された口頭の発話の候補テキスト表現を口頭の発話のテキスト表現としてランダムに(または疑似ランダムに)選択することが可能である。

【００７５】

一部の実装形態において、システムは、口頭の発話の候補テキスト表現を格付けすることが可能であり、最も多くの「票」を有する口頭の発話の候補テキスト表現が、口頭の発話のテキスト表現として選択されることが可能である。例えば、システムは、所与のクライアントデバイスを使用して生成された「Hey Assistant, turn off the living room lights」という口頭の発話の候補テキスト表現と、第1の追加のクライアントデバイスを使用して生成された「Hey Assistant, turn on the living room lights」という口頭の発話の第1の追加の候補テキスト表現と、第2の追加のクライアントデバイスを使用して生成された「Hey Assistant, turn off the living room lights」という口頭の発話の第2の追加の候補テキスト表現とを比較することが可能である。言い換えると、クライアントデバイスのうちの2つ(例えば、所与のクライアントデバイスおよび第2の追加のクライアントデバイス)が、「Hey Assistant, turn off the living room lights」という口頭の発話の候補テキスト表現を生成した一方で、クライアントデバイスのうちの1つだけ(例えば、第1の追加のクライアントデバイス)が、「Hey assistant, turn on the living room lights」という口頭の発話の候補候補テキスト表現を生成した。一部の実装形態において、口頭の発話の候補テキスト表現は、一様に重み付けされることが可能である。例えば、システムは、3つのクライアントデバイスのうちの2つが「Hey Assistant, turn off the living room lights」を口頭の発話の候補テキスト表現として生成することに基づいて、「Hey Assistant, turn off the living room lights」を口頭の発話のテキスト表現として選択することが可能である。

【００７６】

他の一部の実装形態において、口頭の発話の候補テキスト表現が、口頭の発話を生成する際に使用されたクライアントデバイスに基づいて、重み付けされることが可能である。例えば、口頭の発話の候補テキスト表現が、候補テキスト表現を生成する際に使用されたASRモデルのバージョンに基づいて重み付けされること(例えば、システムが、口頭の発話の候補テキスト表現がより高い品質のASRモデルを使用して生成された場合に、その候補テキスト表現により大きい重み付けをすることが可能である)、対応するクライアントデバイスのハードウェアに基づいて重み付けされること(例えば、システムが、対応するクライアントデバイスがより高い品質のオーディオデータストリームをキャプチャする場合に、口頭の発話の候補テキスト表現により大きい重み付けをすることが可能である)、1つまたは複数の追加の、または代替の条件に基づいて重み付けされること、および/またはこれら

の組合せで重み付けされることが可能である。例えば、携帯電話が、より高い品質のオーディオデータをキャプチャするより良好なマイクロホンなどのより良好なハードウェアを有することが可能であり、より高い品質のバージョンのASRモデルを有することが可能である。一部の実装形態において、システムは、携帯電話(より高い品質のマイクロホンと、より高い品質のASRモデルとを有する)を使用して生成された口頭の発話の第1の追加の候補テキスト表現に、口頭の発話のその他の候補テキスト表現と比べて、より大きい重み付けをしてよい。一部の実装形態において、システムは、携帯電話を使用して生成された「Hey Assistant, turn on the living room lights(やあ、アシスタント、居間の照明をオンにして)」という候補テキスト表現を、口頭の発話の他の2つの候補表現が、居間の照明をオフにすることを示すにもかかわらず、口頭の発話のテキスト表現として選択することが可能である。

【００７７】

一部の実装形態において、システムは、口頭の発話の候補テキスト表現の複数の部分を選択的に組み合わせることが可能である。一部の実装形態において、システムは、仮説の上位N番までのリストを協力して生成すべく、所与のクライアントデバイスを使用して生成された1つまたは複数の候補テキスト表現、および1つまたは複数の追加のクライアントデバイスを使用して生成された1つまたは複数の候補テキスト表現を使用することが可能である。例えば、システムは、様々なデバイスからの仮説のリストをマージすることが可能である。

【００７８】

一部の実装形態において、システムは、候補テキスト表現が口頭の発話をキャプチャする確率を示す信頼度スコアを決定することが可能である。例えば、システムは、口頭の発話の候補テキスト表現の確率を示す信頼度スコア、第1の追加の候補テキスト表現が口頭の発話をキャプチャする確率を示す第1の追加の信頼度スコア、および第2の追加の候補テキスト表現が口頭の発話をキャプチャする確率を示す第2の追加の候補テキスト表現を生成することが可能である。一部の実装形態において、システムは、最高の信頼度スコアを有する口頭の発話の候補テキスト表現に基づいて、口頭の発話のテキスト表現を決定することが可能である。

【００７９】

追加または代替として、システムは、口頭の発話の候補テキスト表現の1つまたは複数の部分に基づいて信頼度スコアを生成することが可能である。一部の実装形態において、システムは、口頭の発話がホットワードをキャプチャする確率に基づいてホットワード信頼度スコアを生成することが可能である。例えば、システムは、口頭の発話の候補テキスト表現が「Hey Assistant」というホットワードを含む確率を示すホットワード信頼度スコアを生成することが可能である。

【００８０】

一部の実装形態において、システムは、所与のクライアントデバイスを使用して複数の候補テキスト表現を生成すること、第1の追加のクライアントデバイスを使用して口頭の発話の複数の第1の追加の候補テキスト表現を生成すること、および/または第2の追加のクライアントデバイスを使用して口頭の発話の複数の第2の追加の候補テキスト表現を生成することが可能である。一部の実装形態において、システムは、本明細書において説明される技術により、口頭の発話の複数の候補テキスト表現、口頭の発話の複数の第1の追加の候補テキスト表現、および/または口頭の発話の複数の第2の追加の候補テキスト表現に基づいて、口頭の発話のテキスト表現を決定することが可能である。

【００８１】

一部の実装形態において、システムは、口頭の発話の複数の候補テキスト表現のうちの1つまたは複数にバイアスをかけることが可能である。例えば、携帯電話が、より良好なASRモデルを有することが可能であるが、バイアスをかけるための連絡先のリストが、スタンドアローンの対話型スピーカを介してアクセス可能である(またはスタンドアローンの対話型スピーカを介してのみアクセス可能である)ことがあり得る。一部の実装形態において

、携帯電話(すなわち、「より良好な」ASRモデルを有するデバイス)を使用して生成された複数の第1の追加の候補テキスト表現が、スタンドアローンの対話型スピーカにおいて記憶された連絡先のリストを使用してバイアスをかけられることが可能である。それらの実装形態のいくつかにおいて、システムは、バイアスをかけることに基づいて口頭の発話のテキスト表現を決定することが可能である。

【００８２】

次に図8を参照すると、様々な実装形態が実行され得る例示的な環境が示される。図8について最初に説明され、図8は、自動化されたアシスタントクライアント804のインスタンスを実行するクライアントコンピューティングデバイス802を含む。1つまたは複数のクラウドベースの自動化されたアシスタント構成要素810が、808において全般的に示される1つまたは複数のローカルエリアネットワークおよび/またはワイドエリアネットワーク(例えば、インターネット)を介してクライアントデバイス802に通信可能に結合された1つまたは複数のコンピューティングシステム(ひとまとめにして「クラウド」コンピューティングシステムと呼ばれる)上に実装されることが可能である。

【００８３】

自動化されたアシスタントクライアント804のインスタンスが、1つまたは複数のクラウドベースの自動化されたアシスタント構成要素810との対話により、ユーザの見地から、ユーザが人間対コンピュータの対話をすることが可能である自動化されたアシスタント800の論理インスタンスに見えるものを形成してよい。そのような自動化されたアシスタント800のインスタンスが、図8に示される。それゆえ、一部の実装形態において、クライアントデバイス802上で実行される自動化されたアシスタントクライアント804と関係を結ぶユーザは、実際には、自動化されたアシスタント800のユーザ自らの論理インスタンスと関係を結ぶことが可能であることを理解されたい。簡明のため、本明細書において特定のユーザに「使われている」ものとして使用される「自動化されたアシスタント」という術語は、しばしば、ユーザによって操作されるクライアントデバイス802上で実行される自動化されたアシスタントクライアント804と1つまたは複数のクラウドベースの自動化されたアシスタント構成要素810(多数のクライアントコンピューティングデバイスの多数の自動化されたアシスタントクライアントの間で共有されてよい)の組合せを指す。また、一部の実装形態において、自動化されたアシスタント800は、自動化されたアシスタント800のその特定のインスタンスが実際にユーザに「使われている」かどうかにかかわらず、任意のユーザからの要求に応答してよいことも理解されたい。

【００８４】

クライアントコンピューティングデバイス802は、例えば、デスクトップコンピューティングデバイス、ラップトップコンピューティングデバイス、タブレットコンピューティングデバイス、携帯電話コンピューティングデバイス、ユーザの車両のコンピューティングデバイス(例えば、車両内通信システム、車両内エンターテインメントシステム、車両内ナビゲーションシステム)、スタンドアローンの対話型スピーカ、スマートテレビなどのスマート機器、および/またはコンピューティングデバイス(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、仮想現実コンピューティングデバイスもしくは拡張現実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であってよい。追加のクライアントコンピューティングデバイスおよび/または代替のクライアントコンピューティングデバイスが、備えられてよい。様々な実装形態において、クライアントコンピューティングデバイス802は、オプションとして、自動化されたアシスタントクライアント804に加えて、メッセージ交換クライアント(例えば、SMS、MMS、オンラインチャット)、ブラウザ、その他などである、1つまたは複数の他のアプリケーションを動作させてよい。それらの様々な実装形態のいくつかにおいて、その他のアプリケーションのうちの1つまたは複数が、オプションとして、自動化されたアシスタント800とインターフェースをとること(例えば、アプリケーションプログラミングインターフェースを介して)、または自動化されたアシスタントアプリケーションの独自のインスタンス((1つまたは複数の)クラウドベースの自動化されたアシスタント

構成要素810とインターフェースをとることも可能である)を含むことが可能である。
【００８５】
自動化されたアシスタント800は、クライアントデバイス802のユーザインターフェース入力デバイスおよびユーザインターフェース出力デバイスを介してユーザと人間対コンピュータの対話セッションを行う。ユーザプライバシーを保護すべく、および/またはリソースを節約すべく、多くの状況において、ユーザは、自動化されたアシスタントが口頭の発話を完全に処理するより前に、しばしば、自動化されたアシスタント800を明示的に呼び出さなければならない。自動化されたアシスタント800の明示的な呼出しは、クライアントデバイス802において受け取られた或るユーザインターフェース入力に応答して行われることが可能である。例えば、クライアントデバイス802を介して自動化されたアシスタント800を呼び出すことができるユーザインターフェース入力は、オプションとして、クライアントデバイス802のハードウェアおよび/または仮想ボタンの作動を含むことが可能である。さらに、自動化されたアシスタントクライアントは、1つまたは複数の口頭の呼出し句の存在を検出するように動作可能である呼出しエンジンなどの1つまたは複数のローカルエンジン806を含むことが可能である。呼出しエンジンは、口頭の呼出し句のうちの1つを検出することに応答して、自動化されたアシスタント800を呼び出すことが可能である。例えば、呼出しエンジンは、「Hey Assistant」、「OK Assistant」、および/または「Assistant」などの口頭の呼出し句を検出することに応答して、自動化されたアシスタント800を呼び出すことが可能である。呼出しエンジンは、口頭の呼出し句の出現を監視すべく、クライアントデバイス802の1つまたは複数のマイクロホンからの出力に基づくオーディオデータフレームのストリームを継続的に処理することが可能である(例えば、「不活性」モードに入っていない場合)。口頭の呼出し句の出現を監視している間、呼出しエンジンは、口頭の呼出し句を含まないオーディオデータフレームを破棄する(例えば、バッファに一時的に記憶した後)。しかし、呼出しエンジンが、処理されるオーディオデータフレーム内で口頭の呼出し句の出現を検出した場合、呼出しエンジンは、自動化されたアシスタント800を呼び出すことが可能である。本明細書において使用される、自動化されたアシスタント800を「呼び出すこと」は、自動化されたアシスタント800の1つまたは複数のそれまでに不活性の機能が活性化されるようにすることを含むことが可能である。例えば、自動化されたアシスタント800を呼び出すことは、1つまたは複数のローカルエンジン806および/またはクラウドベースの自動化されたアシスタント構成要素810が、呼出し句がそれに基づいて検出されたオーディオデータフレーム、および/または1つまたは複数の後続のオーディオデータフレームをさらに処理するようにさせることを含むことが可能である(呼び出す前には、オーディオデータフレームの追加の処理は、行われていなかったのに対して)。
【００８６】
自動化されたアシスタント800の1つまたは複数のローカルエンジン806は、オプションであり、例えば、前述の呼出しエンジン、ローカル音声テキスト化(voice-to-text)(「STT」)エンジン(キャプチャされたオーディオをテキストに変換する)、ローカルテキスト音声化(text-to-speech)(「TTS」)エンジン(テキストを音声に変換する)、ローカル自然言語プロセッサ(オーディオの意味論的な意味および/またはオーディオから変換されたテキストを決定する)、および/または他のローカル構成要素を含むことが可能である。クライアントデバイス802は、計算リソース(例えば、プロセッササイクル、メモリ、バッテリ、その他)の点で比較的制約されているため、ローカルエンジン806は、クラウドベースの自動化されたアシスタント構成要素810に含まれるそれに対応するものと比べて限られた機能を有する可能性がある。
【００８７】
クラウドベースの自動化されたアシスタント構成要素810は、(1つまたは複数の)ローカルエンジン806に対応するものと比べて、オーディオデータ、および/または他のユーザインターフェース入力のより堅牢な、および/またはより正確な処理を実行すべくクラウドの仮想的に無限のリソースを活用する。この場合も、様々な実装形態において、クライアン

トデバイス802は、呼出しエンジンが口頭の呼出し句を検出すること、または自動化されたアシスタント800の他の何らかの明示的な呼出しを検出することに応答して、オーディオデータおよび/または他のデータをクラウドベースの自動化されたアシスタント構成要素810に提供することが可能である。

【００８８】

例示されるクラウドベースの自動化されたアシスタント構成要素810は、クラウドベースのTTSモジュール812、クラウドベースのSTTモジュール814、自然言語プロセッサ816、対話状態トラッカ818、および対話マネージャ820を含む。一部の実装形態において、自動化されたアシスタント800のエンジンおよび/またはモジュールのうちの1つまたは複数は、省かれること、組み合わされること、および/または自動化されたアシスタント800とは別個の構成要素において実装されることが可能である。さらに、一部の実装形態において、自動化されたアシスタント800は、追加のエンジンおよび/またはモジュール、および/または代替のエンジンおよび/またはモジュールを含むことが可能である。クラウドベースのSTTモジュール814は、オーディオデータをテキストに変換することが可能であり、そのテキストが、次に、自然言語プロセッサ816に与えられてよい。

【００８９】

クラウドベースのTTSモジュール812は、テキストデータ(例えば、自動化されたアシスタント800によって作成された自然言語応答)を、コンピュータによって生成された音声出力に変換することが可能である。一部の実装形態において、TTSモジュール812は、コンピュータによって生成された音声出力を、例えば、1つまたは複数のスピーカを使用して、直接に出力されるようにクライアントデバイス802に供給してよい。他の実装形態において、自動化されたアシスタント800によって生成されたテキストデータ(例えば、自然言語応答)が、(1つまたは複数の)ローカルエンジン806のうちの1つに供給されてよく、そのローカルエンジン806が、次に、テキストデータを、ローカルで出力されるコンピュータによって生成された音声に変換してよい。

【００９０】

自動化されたアシスタント800の自然言語プロセッサ816が、自由形式の自然言語入力を処理し、自然言語入力に基づいて、自動化されたアシスタント800の1つまたは複数の他の構成要素によって使用されるように注釈付きの出力を生成する。例えば、自然言語プロセッサ816は、クライアントデバイス802を介してユーザによって提供されたオーディオデータの、STTモジュール814による変換であるテキスト入力である、自然言語自由形式入力を処理することが可能である。生成された注釈付きの出力は、自然言語入力の1つまたは複数の注釈を含んでよく、オプションとして、自然言語入力の項のうちの1つまたは複数(例えば、すべて)を含んでよい。

【００９１】

一部の実装形態において、自然言語プロセッサ816は、自然言語入力における様々な種類の文法情報を識別すること、およびそれらに注釈を付けることを行うように構成される。一部の実装形態において、自然言語プロセッサ816は、人々に対する言及(例えば、文学の登場人物、有名人、著名人、その他を含む)、組織、場所(現実および空想上の)、その他などの、1つまたは複数のセグメントにおけるエンティティ言及に注釈を付けるように構成されたエンティティタガ(entity tagger)(図示せず)をさらに、および/または代替として含んでよい。一部の実装形態において、自然言語プロセッサ816は、1つまたは複数のコンテキストキューに基づいて、同一のエンティティに対する言及をグループ化する、または「クラスタ化する」ように構成されたコアリファレンスリゾルバ(coreference resolver)(図示せず)をさらに、および/または代替として含んでよい。例えば、コアリファレンスリゾルバは、「there」という項を「I liked Hypothetical Cafe last time we ate there.」という自然言語入力における「Hypothetical Cafe」に解決するのに利用されることが可能である。一部の実装形態において、自然言語プロセッサ816の1つまたは複数の構成要素は、自然言語プロセッサ816の1つまたは複数の他の構成要素からの注釈に依拠してよい。一部の実装形態において、特定の自然言語入力を処理する際、自然言語プロ

セッサ816の1つまたは複数の構成要素は、1つまたは複数の注釈を決定すべく、関連する以前の入力および/または特定の自然言語入力の外部の関連する他のデータを使用してよい。

【００９２】

一部の実装形態において、対話状態トラッカ818が、例えば、人間対コンピュータ対話セッションの過程で、および/または多数の対話セッションをまたぐ1名または複数名のユーザの目標(または「意図」)の信念状態を含む「対話状態」を常に把握しているように構成されてよい。対話状態を判定する際、いくつかの対話状態トラッカが、対話セッションにおけるユーザ発話およびシステム発話に基づいて、対話においてインスタンス生成される(1つまたは複数の)スロットに関して最も尤度の高い(1つまたは複数の)値を決定しようと求めてよい。一部の技術は、スロットのセット、およびそれらのスロットに関連する値のセットを定義する固定のオントロジを利用する。一部の技術は、追加または代替として、個々のスロットおよび/または個々のドメインに合わせられてよい。例えば、一部の技術は、各ドメインにおける各スロットタイプに関してモデルを訓練することを要求してよい。

【００９３】

対話マネージャ820が、例えば、対話状態トラッカ818によって与えられる現在の対話状態を、複数の候補応答アクションのうちの1つまたは複数の「応答アクション」にマップするように構成されてよく、その応答アクションが、次に、自動化されたアシスタント800によって実行される。応答アクションは、現在の対話状態に依存して、様々な形態であってよい。例えば、最後の回(例えば、究極的なユーザによって所望されるタスクが実行されたときの)に先立って行われた対話セッションの回に対応する初期の対話状態および途中の対話状態が、自動化されたアシスタント800が追加の自然言語対話を出力することを含む様々な応答アクションにマップされてよい。この応答対話は、例えば、ユーザが実行することを意図するものと対話状態トラッカ818が信じる何らかのアクションに関するパラメータをユーザが提供する(すなわち、スロットを埋める)ことを求める要求を含んでよい。一部の実装形態において、応答アクションは、「要求する」(例えば、スロットを埋めるためのパラメータを求める)、「提案する(offer)」(例えば、ユーザのためのアクションもしくは一連のアクション(course of action)を示唆する)、「選択する」、「知らせる」(例えば、要求された情報をユーザに提供する)、「合致なし」(例えば、ユーザの前回の入力が理解されなかったことをユーザに通知する)、周辺デバイスに対するコマンド(例えば、電球をオフにする)、その他などのアクションを含んでよい。

【００９４】

図9は、本明細書において説明される技術の1つまたは複数の態様を実行するのにオプションとして利用されてよい例示的なコンピューティングデバイス910のブロック図である。一部の実装形態において、クライアントコンピューティングデバイスのうちの1つまたは複数、および/または他の(1つまたは複数の)構成要素が、例示的なコンピューティングデバイス910の1つまたは複数の構成要素を備えてよい。

【００９５】

コンピューティングデバイス910は、通常、バスサブシステム912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914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925およびファイルストレージサブシステム926、ユーザインターフェース出力デバイス920、ユーザインターフェース入力デバイス922、ならびにネットワークインターフェースサブシステム916を含む、ストレージサブシステム924を含んでよい。入力デバイスおよび出力デバイスは、コンピューティングデバイス910とのユーザ対話を可能にする。ネットワークインターフェースサブシステム916は、外部ネットワークに対するインターフェースをもたらし、他のコンピューティングデバイスにおける対応するインターフェースデバイスに結合される。

【００９６】

ユーザインターフェース入力デバイス922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ス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、マイクロホン、および/または他の種類の入力デバイスを含んでよい。一般に、「入力デバイス」という術語の使用は、情報をコンピューティングデバイス910に、または通信ネットワーク上に入力する可能なすべての種類のデバイスおよび方法を含むことを意図している。

【００９７】

ユーザインターフェース出力デバイス920は、ディスプレイサブシステム、プリンタ、ファックス装置、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んでよい。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、または視覚的画像を生成するための他の何らかの機構を含んでよい。また、ディスプレイサブシステムは、オーディオ出力デバイスを介するなどして非視覚的表示を提供してもよい。一般に、「出力デバイス」という術語の使用は、コンピューティングデバイス910からユーザに、または別の機械もしくは別のコンピューティングデバイスに情報を出力する可能なすべての種類のデバイスおよび方法を含むことを意図している。

【００９８】

ストレージサブシステム924は、本明細書において説明されるモジュールのいくつか、またはすべてのモジュールの機能をもたらすプログラミングおよびデータ構造を記憶する。例えば、ストレージサブシステム924は、図4、図5、図6、および/または図7の処理のうちの1つまたは複数の処理の選択された態様を実行するとともに、図3および/または図8に示される様々な構成要素を実装するロジックを含んでよい。

【００９９】

これらのソフトウェアモジュールは、一般に、プロセッサ914だけによって、または他のプロセッサとの組合せによって実行される。ストレージサブシステム924において使用されるメモリ925は、プログラム実行中に命令およびデータを記憶するためのメインランダムアクセスメモリ(「RAM」)930と、固定の命令が記憶された読取り専用メモリ(「ROM」)932とを含むいくつかのメモリを含むことが可能である。ファイルストレージサブシステム626が、プログラムおよびデータファイルのための永久ストレージを提供することが可能であり、ハードディスクドライブ、関連するリムーバブルメディアと一緒になったフロッピーディスクドライブ、CD-ROMドライブ、光ドライブ、またはリムーバブルメディアカートリッジを含んでよい。いくつかの実装形態の機能を実装するモジュールは、ストレージサブシステム924におけるファイルストレージサブシステム926によって、または(1つまたは複数の)プロセッサ914によってアクセス可能な他の機械に記憶されてよい。

【０１００】

バスサブシステム912が、コンピューティングデバイス910の様々な構成要素およびサブシステムに、意図どおりに互いに通信させるための機構を提供する。バスサブシステム912は、単一のバスとして概略で示されるものの、バスサブシステムの代替の実装形態は、多数のバスを使用してよい。

【０１０１】

コンピューティングデバイス910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、または他の任意のデータ処理システムもしくはデータ処理コンピューティングデバイスを含む様々な種類のものであり得る。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、図9に示されるコンピューティングデバイス910の説明は、一部の実装形態を例示する目的での特定の実例として意図されるに過ぎない。図9に示されるコンピューティングデバイスと比べて、より多くの、またはより少ない構成要素を有するコンピューティングデバイス910の他の多くの構成が、可能である。

【０１０２】

本明細書において説明されるシステムがユーザ(または、本明細書において、しばしば、「参加者」と呼ばれる)についての個人情報を収集する、または個人情報を使用してよい状

況において、ユーザは、プログラムもしくはフィーチャがユーザ情報(例えば、ユーザのソーシャルネットワーク、社会的行動もしくは社会的活動、職業、ユーザの選好、またはユーザの現在の地理的ロケーションについての情報)を収集するかどうかを管理する、またはユーザとより関係があり得るコンテンツをコンテンツサーバから受け取るかどうか、および/またはどのように受け取るかを管理する機会を与えられてよい。また、或るデータは、個人識別可能な情報が除去されるように、記憶される前、または使用される前に、1つまたは複数の方法で処置されてよい。例えば、ユーザの身元が、ユーザに関して個人識別可能な情報が特定され得ないように処置されてよく、またはユーザの地理的ロケーションが、ユーザの特定の地理的ロケーションが特定され得ないように、地理的ロケーション情報が獲得されるところに一般化されてよい(都市、郵便番号、または州レベルなどに)。このため、ユーザは、ユーザについて情報がどのように収集されるか、および/またはどのように使用されるかを管理してよい。

【０１０３】

一部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、方法は、クライアントデバイスにおいて、ユーザの口頭の発話をキャプチャするオーディオデータを検出することを含み、クライアントデバイスは、1つまたは複数の追加のクライアントデバイスを含む環境内にあり、ローカルネットワークを介して1つまたは複数の追加のクライアントデバイスとローカルで通信しており、1つまたは複数の追加のクライアントデバイスは、少なくとも第1の追加のクライアントデバイスを含む。方法は、クライアントデバイスにおいて、口頭の発話の候補テキスト表現を生成すべくクライアントデバイスにおいてローカルで記憶された自動音声認識(「ASR」)モデルを使用してオーディオデータを処理することをさらに含む。方法は、クライアントデバイスにおいて、第1の追加のクライアントデバイスからローカルネットワークを介して、口頭の発話の第1の追加の候補テキスト表現を受信することをさらに含み、第1の追加のクライアントデバイスにおいてローカルで生成される口頭の発話の第1の追加の候補テキスト表現は、(a)オーディオデータ、および/または(b)第1の追加のクライアントデバイスにおいて検出された口頭の発話をキャプチャするローカルで検出されたオーディオデータに基づき、口頭の発話の第1の追加の候補テキスト表現は、オーディオデータおよび/または第1の追加のクライアントデバイスにおいてローカルで記憶された第1の追加のASRモデルを使用してローカルで生成されたオーディオデータを処理することによって生成される。方法は、口頭の発話の候補テキスト表現、および第1の追加のクライアントデバイスによって生成された口頭の発話の第1の追加の候補テキスト表現に基づいて、口頭の発話のテキスト表現を決定することをさらに含む。

【０１０４】

本技術のこれら、およびその他の実装形態は、以下の特徴のうちの1つまたは複数を含むことが可能である。

【０１０５】

一部の実装形態において、1つまたは複数の追加のクライアントデバイスが、少なくとも第1の追加のクライアントデバイスと、第2の追加のクライアントデバイスとを含む。一部の実装形態において、クライアントデバイスにおいて、第1の追加のクライアントデバイスからローカルネットワークを介して、第1の追加の候補テキスト表現を受信することは、クライアントデバイスにおいて、第2の追加のクライアントデバイスからローカルネットワークを介して、(a)オーディオデータ、および/または(b)第2の追加のクライアントデバイスにおいて検出された口頭の発話をキャプチャする追加のローカルで検出されたオーディオデータに基づいて、第2の追加のクライアントデバイスにおいてローカルで生成された口頭の発話の第2の追加の候補テキスト表現を受信することをさらに含み、口頭の発話の第2の追加の候補テキスト表現は、オーディオデータ、および/または第2の追加のクライアントデバイスにおいてローカルで記憶された第2の追加のASRモデルを使用してローカルで生成された追加のオーディオデータを処理することによって生成される。一部の実装形態において、口頭の発話の候補テキスト表現、および第1の追加のクライアント

デバイスによって生成された口頭の発話の第1の追加の候補テキスト表現に基づいて口頭の発話のテキスト表現を決定することは、口頭の発話の候補テキスト表現、第1の追加のクライアントデバイスによって生成された口頭の発話の第1の追加の候補テキスト表現、および第2の追加のクライアントデバイスによって生成された口頭の発話の第2の追加の候補テキスト表現に基づいて口頭の発話のテキスト表現を決定することをさらに含む。

【０１０６】

一部の実装形態において、口頭の発話の候補テキスト表現、および第1の追加のクライアントデバイスによって生成された口頭の発話の第1の追加の候補テキスト表現に基づいて口頭の発話のテキスト表現を決定することは、口頭の発話の候補テキスト表現、または口頭の発話の第1の追加の候補テキスト表現をランダムに選択することを含む。一部の実装形態において、方法は、ランダムな選択に基づいて口頭の発話のテキスト表現を決定することをさらに含む。

【０１０７】

一部の実装形態において、口頭の発話の候補テキスト表現、および第1の追加のクライアントデバイスによって生成された口頭の発話の第1の追加の候補テキスト表現に基づいて口頭の発話のテキスト表現を決定することは、候補テキスト表現がテキスト表現である確率を示す候補テキスト表現の信頼度スコアを決定することを含み、信頼度スコアは、クライアントデバイスの1つまたは複数のデバイスパラメータに基づく。一部の実装形態において、方法は、追加の候補テキスト表現がテキスト表現である追加の確率を示す追加の候補テキスト表現の追加の信頼度スコアを決定することをさらに含み、追加の信頼度スコアは、追加のクライアントデバイスの1つまたは複数の追加のデバイスパラメータに基づく。一部の実装形態において、方法は、信頼度スコアと追加の信頼度スコアを比較することをさらに含む。一部の実装形態において、方法は、その比較に基づいて、口頭の発話のテキスト表現を決定することをさらに含む。

【０１０８】

一部の実装形態において、口頭の発話の候補テキスト表現、および第1の追加のクライアントデバイスによって生成された口頭の発話の第1の追加の候補テキスト表現に基づいて口頭の発話のテキスト表現を決定することは、クライアントデバイスにおいて検出された口頭の発話をキャプチャするオーディオデータの品質を示すオーディオ品質値を決定することを含む。一部の実装形態において、方法は、第1の追加のクライアントデバイスにおいて検出された口頭の発話をキャプチャする追加のオーディオデータの品質を示す追加のオーディオ品質値を決定することをさらに含む。一部の実装形態において、方法は、オーディオ品質値と追加のオーディオ品質値を比較することをさらに含む。一部の実装形態において、方法は、その比較に基づいて、口頭の発話のテキスト表現を決定することをさらに含む。

【０１０９】

一部の実装形態において、口頭の発話の候補テキスト表現、および第1の追加のクライアントデバイスによって生成された口頭の発話の第1の追加の候補テキスト表現に基づいて口頭の発話のテキスト表現を決定することは、クライアントデバイスにおいてローカルで記憶されたASRモデルの品質を示すASR品質値を決定することを含む。一部の実装形態において、方法は、追加のクライアントデバイスにおいてローカルで記憶された追加のASRモデルの品質を示す追加のASR品質値を決定することをさらに含む。一部の実装形態において、方法は、ASR品質値と追加のASR品質値を比較することをさらに含む。一部の実装形態において、方法は、その比較に基づいて口頭の発話のテキスト表現を決定することをさらに含む。

【０１１０】

一部の実装形態において、口頭の発話の第1の追加の候補テキスト表現は、複数の仮説を含み、口頭の発話の候補テキスト表現、および第1の追加のクライアントデバイスによって生成された口頭の発話の第1の追加の候補テキスト表現に基づいて口頭の発話のテキスト表現を決定することは、クライアントデバイスを使用して複数の仮説を格付けし直す

ことを含む。一部の実装形態において、方法は、口頭の発話の候補テキスト表現、および格付けし直された複数の仮説に基づいて口頭の発話のテキスト表現を決定することをさらに含む。

【０１１１】

一部の実装形態において、クライアントデバイスにおいて、第1の追加のクライアントデバイスからローカルネットワークを介して、口頭の発話の第1の追加の候補テキスト表現を受信することに先立って、(a)オーディオデータ、および/または(b)第1の追加のクライアントデバイスにおいて検出された口頭の発話をキャプチャするローカルで検出されたオーディオデータに基づいて、第1の追加のクライアントデバイスにおいてローカルで口頭の発話の第1の追加の候補表現を生成するかどうかを決定することをさらに含み、(a)オーディオデータ、および/または(b)第1の追加のクライアントデバイスにおいて検出された口頭の発話をキャプチャするローカルで検出されたオーディオデータに基づいて、第1の追加のクライアントデバイスにおいてローカルで口頭の発話の第1の追加の候補表現を生成するかどうかを決定することは、クライアントデバイスにおいて検出された口頭の発話をキャプチャするオーディオデータの品質を示すオーディオ品質値を決定することを含む。一部の実装形態において、方法は、第1の追加のクライアントデバイスにおいて検出された口頭の発話をキャプチャするローカルで検出されたオーディオデータの品質を示す追加のオーディオ品質値を決定することをさらに含む。一部の実装形態において、方法は、オーディオ品質値と追加のオーディオ品質値を比較することをさらに含む。一部の実装形態において、方法は、その比較に基づいて、(a)オーディオデータ、および/または(b)第1の追加のクライアントデバイスにおいて検出された口頭の発話をキャプチャするローカルで検出されたオーディオデータに基づいて、第1の追加のクライアントデバイスにおいてローカルで口頭の発話の第1の追加の候補表現を生成するかどうかを決定することをさらに含む。それらの実装形態の一部のバージョンにおいて、クライアントデバイスにおいて検出された口頭の発話をキャプチャするオーディオデータの品質を示すオーディオ品質値を決定することは、クライアントデバイスの1つまたは複数のマイクロホンを識別することを含む。それらの実装形態の一部のバージョンにおいて、方法は、クライアントデバイスの1つまたは複数のマイクロホンに基づいてオーディオ品質値を決定することをさらに含む。それらの実装形態の一部のバージョンにおいて、第1の追加のクライアントデバイスにおいて検出された口頭の発話をキャプチャするローカルで検出されたオーディオデータの品質を示す追加のオーディオ品質値を決定することは、第1の追加のクライアントデバイスの1つまたは複数の第1の追加のマイクロホンを識別することを含む。それらの実装形態の一部のバージョンにおいて、方法は、第1の追加のクライアントデバイスの1つまたは複数の第1の追加のマイクロホンに基づいて追加のオーディオ品質値を決定することをさらに含む。それらの実装形態の一部のバージョンにおいて、クライアントデバイスにおいて検出された口頭の発話をキャプチャするオーディオデータの品質を示すオーディオ品質値を決定することは、口頭の発話をキャプチャするオーディオデータを処理することに基づいて信号対ノイズ比値を生成することを含む。それらの実装形態の一部のバージョンにおいて、方法は、信号対ノイズ比値に基づいてオーディオ品質値を決定することをさらに含む。それらの実装形態の一部のバージョンにおいて、第1の追加のクライアントデバイスにおいて検出された口頭の発話をキャプチャするローカルで検出されたオーディオデータの品質を示す追加のオーディオ品質値を決定することは、口頭の発話をキャプチャするオーディオデータを処理することに基づいて、追加の信号対ノイズ比値を生成することを含む。それらの実装形態の一部のバージョンにおいて、方法は、追加の信号対ノイズ比値に基づいて、追加のオーディオ品質値を決定することをさらに含む。

【０１１２】

一部の実装形態において、クライアントデバイスにおいて、第1の追加のクライアントデバイスからローカルネットワークを介して、口頭の発話の第1の追加の候補テキスト表現を受信することに先立って、方法は、口頭の発話の第1の追加の候補テキスト表現を求める要求を第1の追加のクライアントデバイスに送信するかどうかを決定することをさら

に含む。一部の実装形態において、口頭の発話の第1の追加の候補テキスト表現を求める要求を第1の追加のクライアントデバイスに送信することを決定することに応答して、方法は、口頭の発話の第1の追加の候補テキスト表現を求める要求を第1の追加のクライアントデバイスに送信することをさらに含む。それらの実装形態の一部のバージョンにおいて、口頭の発話の第1の追加の候補テキスト表現を求める要求を第1の追加のクライアントデバイスに送信するかどうかを決定することは、ホットワードモデルを使用してユーザの口頭の発話をキャプチャするオーディオデータの少なくとも一部分を処理することに基づいて、ホットワード信頼度スコアを決定することを含み、ホットワード信頼度スコアは、オーディオデータの少なくとも一部分がホットワードを含むかどうかの確率を示す。それらの実装形態の一部のバージョンにおいて、方法は、ホットワード信頼度スコアが1つまたは複数の条件を満たすかどうかを判定することをさらに含み、ホットワード信頼度スコアが1つまたは複数の条件を満たすかどうかを判定することは、ホットワード信頼度スコアがしきい値を満たすかどうかを判定することを含む。それらの実装形態の一部のバージョンにおいて、ホットワード信頼度スコアがしきい値を満たすと判定したことに応答して、方法は、ホットワード信頼度スコアがオーディオデータの少なくとも一部分がホットワードを含む弱い確率を示すかどうかを判定することをさらに含む。それらの実装形態の一部のバージョンにおいて、ホットワード信頼度スコアがオーディオデータの少なくとも一部分がホットワードを含む弱い確率を示すと判定したことに応答して、方法は、口頭の発話の第1の追加の候補テキスト表現を求める要求を第1の追加のクライアントデバイスに送信することを決定することをさらに含む。

【０１１３】

さらに、一部の実装形態は、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、(1つまたは複数の)中央処理装置((1つまたは複数の)CPU))、(1つまたは複数の)グラフィックスプロセシングユニット(1つまたは複数の)GPU、および/または(1つまたは複数の)テンソルプロセシングユニット((1つまたは複数の)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、本明細書において説明される方法のいずれかの実行をもたらすように構成される。また、一部の実装形態は、本明細書において説明される方法のいずれかを実行すべく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つまたは複数の一過性の、または非一過性のコンピュータ可読記憶媒体も含む。

【符号の説明】

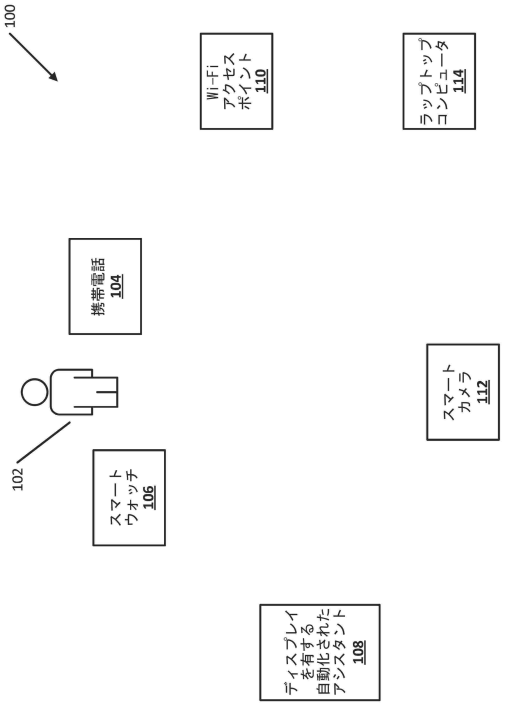
【０１１４】

100 環境
102 ユーザ
104 携帯電話
106 スマートウォッチ
108 ディスプレイを有する自動化されたアシスタント
110 Wi-Fiアクセスポイント
112 スマートカメラ
114 ラップトップコンピュータ
300 環境
302、314 クライアントデバイス
304、316 ユーザインターフェース入出力デバイス
306、320 候補テキスト表現エンジン
308 テキスト表現エンジン
310 追加のデバイスエンジン
312、322 ASRモデル
318 オーディオ源エンジン
800 自動化されたアシスタント
802 クライアントデバイス

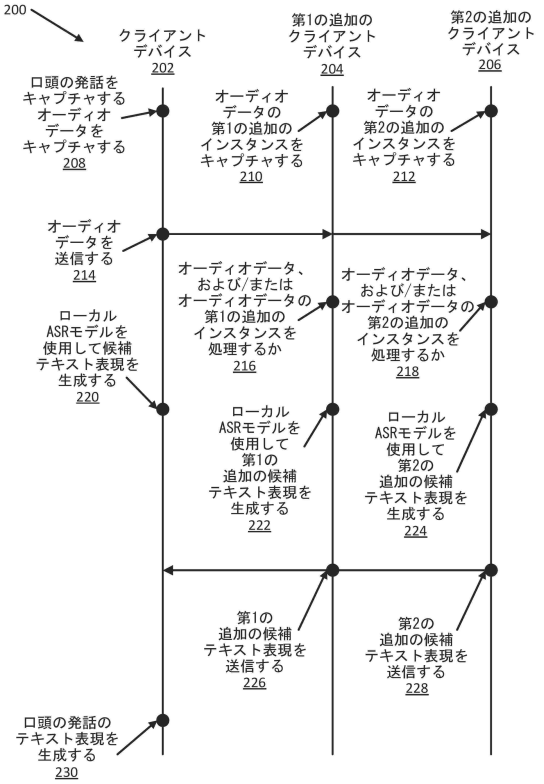
804 自動化されたアシスタントクライアント
806 ローカルエンジン
810 クラウドベースの自動化されたアシスタント構成要素
812 TTSモジュール
814 STTモジュール
816 自然言語プロセッサ
818 対話状態トラッカ
820 対話マネージャ
910 コンピューティングデバイス
912 バスサブシステム
914 プロセッサ
916 ネットワークインターフェース
920 ユーザインターフェース出力デバイス
922 ユーザインターフェース入力デバイス
924 ストレージサブシステム
925 メモリサブシステム
926 ファイルストレージサブシステム

【図面】

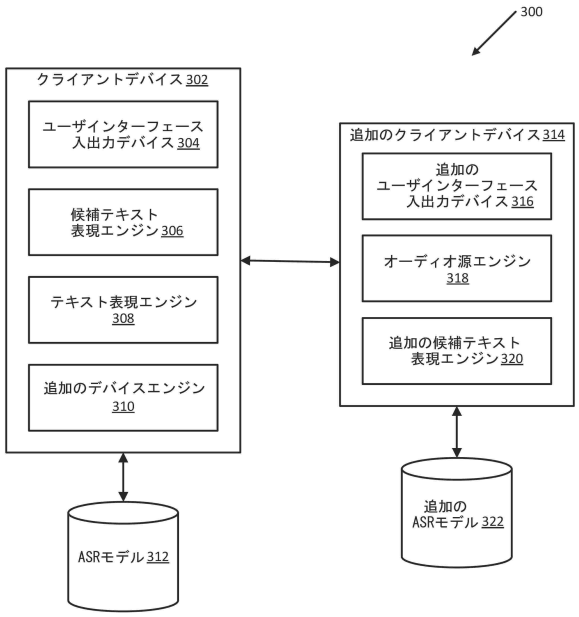
【図１】



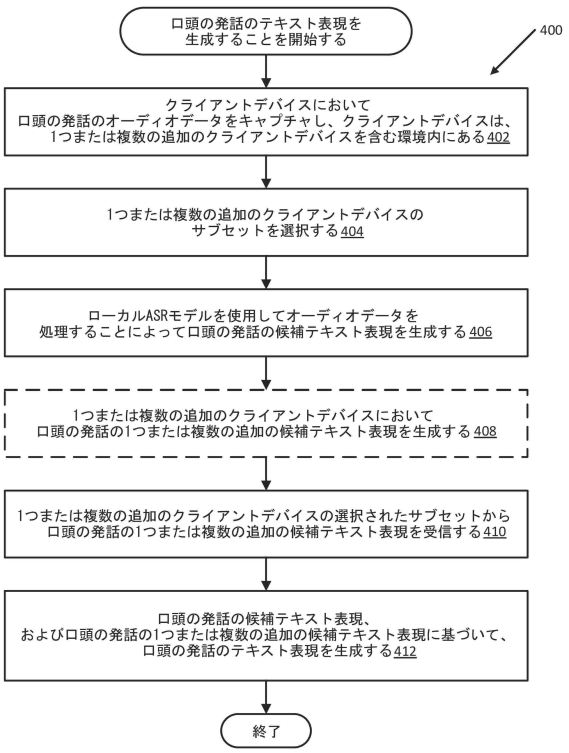
【図２】



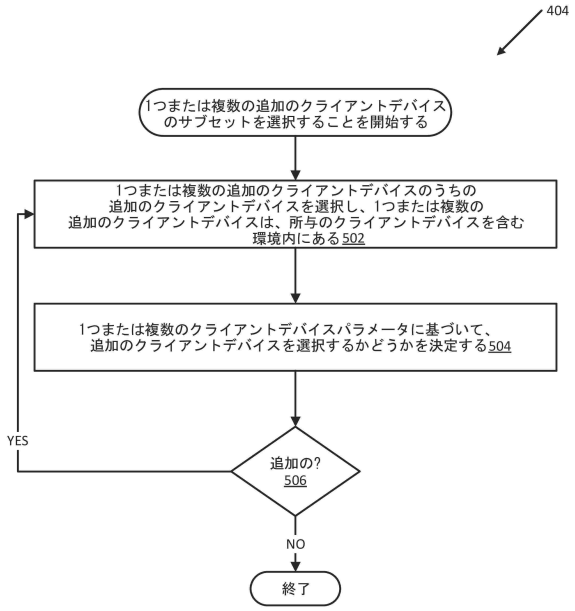
【図３】


300
クライアントデバイス 302
ユーザインターフェース 入出力デバイス 304
候補テキスト 表現エンジン 306
テキスト表現エンジン 308
追加のデバイスエンジン 310
ASRモデル 312
追加のクライアントデバイス 314
追加の ユーザインターフェース 入出力デバイス 316
オーディオ源エンジン 318
追加の候補テキスト 表現エンジン 320
追加の ASRモデル 322

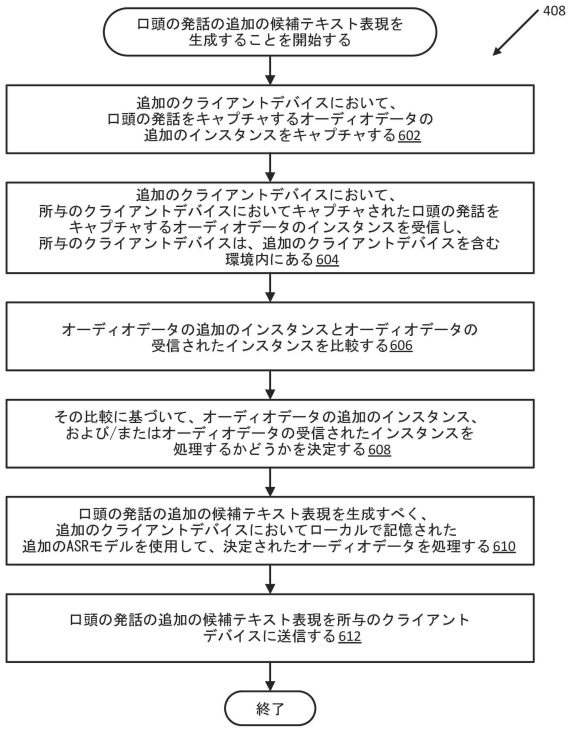
【図４】


400
口頭の発話のテキスト表現を 生成することを開始する
クライアントデバイスにおいて 口頭の発話のオーディオデータをキャプチャし、クライアントデバイスは、 1つまたは複数の追加のクライアントデバイスを含む環境内にある 402
1つまたは複数の追加のクライアントデバイスの サブセットを選択する 404
ローカルASRモデルを使用してオーディオデータを 処理することによって口頭の発話の候補テキスト表現を生成する 406
1つまたは複数の追加のクライアントデバイスにおいて 口頭の発話の1つまたは複数の追加の候補テキスト表現を生成する 408
1つまたは複数の追加のクライアントデバイスの選択されたサブセットから 口頭の発話の1つまたは複数の追加の候補テキスト表現を受信する 410
口頭の発話の候補テキスト表現、 および口頭の発話の1つまたは複数の追加の候補テキスト表現に基づいて、 口頭の発話のテキスト表現を生成する 412
終了

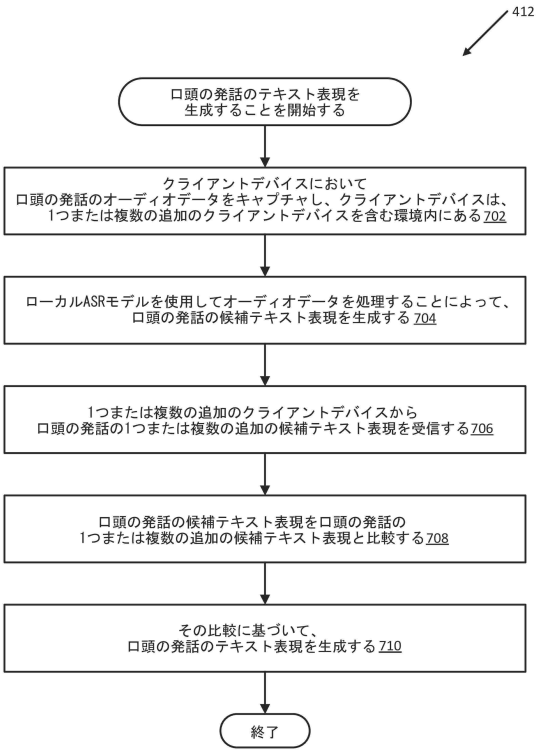
【図５】


404
1つまたは複数の追加のクライアントデバイス のサブセットを選択することを開始する
1つまたは複数の追加のクライアントデバイスのうちの 追加のクライアントデバイスを選択し、1つまたは複数の 追加のクライアントデバイスは、所与のクライアントデバイスを含む 環境内にある 502
1つまたは複数のクライアントデバイスパラメータに基づいて、 追加のクライアントデバイスを選択するかどうかを決定する 504
追加の? 506
YES
NO
終了

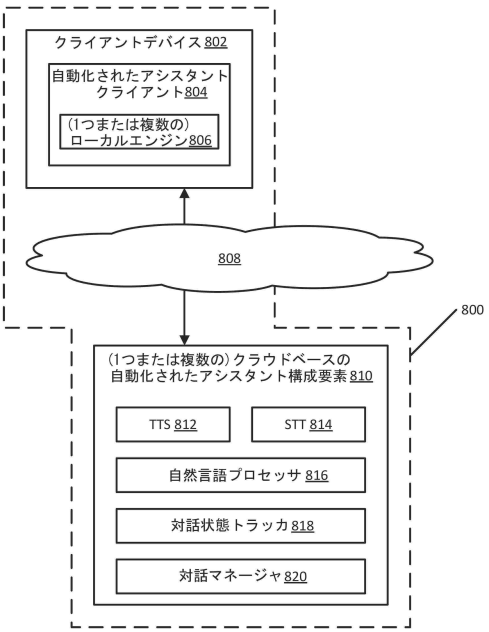
【図６】


408
口頭の発話の追加の候補テキスト表現を 生成することを開始する
追加のクライアントデバイスにおいて、 口頭の発話をキャプチャするオーディオデータの 追加のインスタンスをキャプチャする 602
追加のクライアントデバイスにおいて、 所与のクライアントデバイスにおいてキャプチャされた口頭の発話を キャプチャするオーディオデータのインスタンスを受信し、 所与のクライアントデバイスは、追加のクライアントデバイスを含む 環境内にある 604
オーディオデータの追加のインスタンスとオーディオデータの 受信されたインスタンスを比較する 606
その比較に基づいて、オーディオデータの追加のインスタンス、 および/またはオーディオデータの受信されたインスタンスを 処理するかどうかを決定する 608
口頭の発話の追加の候補テキスト表現を生成すべく、 追加のクライアントデバイスにおいてローカルで記憶された 追加のASRモデルを使用して、決定されたオーディオデータを処理する 610
口頭の発話の追加の候補テキスト表現を所与のクライアント デバイスに送信する 612
終了

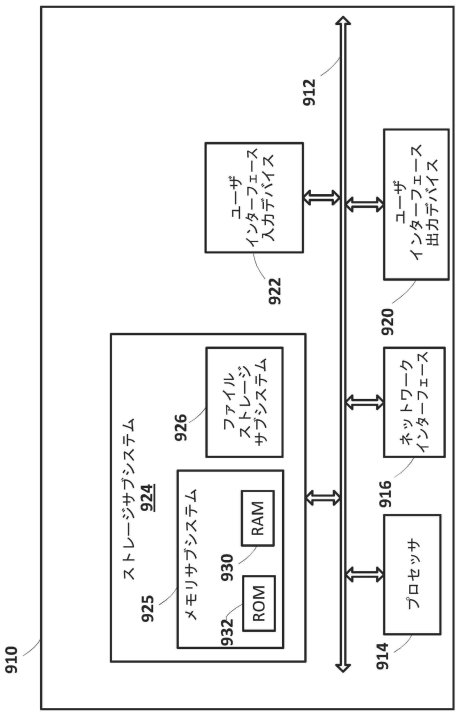
【図７】


412
口頭の発話のテキスト表現を
生成することを開始する
クライアントデバイスにおいて
口頭の発話のオーディオデータをキャプチャし、クライアントデバイスは、
1つまたは複数の追加のクライアントデバイスを含む環境内にある 702
ローカルASRモデルを使用してオーディオデータを処理することによって、
口頭の発話の候補テキスト表現を生成する 704
1つまたは複数の追加のクライアントデバイスから
口頭の発話の1つまたは複数の追加の候補テキスト表現を受信する 706
口頭の発話の候補テキスト表現を口頭の発話の
1つまたは複数の追加の候補テキスト表現と比較する 708
その比較に基づいて、
口頭の発話のテキスト表現を生成する 710
終了

【図８】


クライアントデバイス 802
自動化されたアシスタント
クライアント 804
(1つまたは複数の)
ローカルエンジン 806
808
800
(1つまたは複数の)クラウドベースの
自動化されたアシスタント構成要素 810
TTS 812
STT 814
自然言語プロセッサ 816
対話状態トラッカ 818
対話マネージャ 820

【図９】


910
912
914
916
920
922
924
925
926
930
932
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
ファイル
ストレージ
サブシステム
ネットワーク
インターフェース
ストレージサブシステム
メモリサブシステム
RAM
ROM
プロセッサ

フロントページの続き

(72)発明者　マシュー・シャリフィ
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　ヴィクター・カルブネ
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　山下 剛史
(56)参考文献　特開２０２０－１２９１３０（ＪＰ，Ａ）
米国特許出願公開第２０１９／０３１８７４２（ＵＳ，Ａ１）
特開２０１１－９０１００（ＪＰ，Ａ）
特開２００５－２６６１９２（ＪＰ，Ａ）
国際公開第２０２０／０４０７７５（WO，Ａ１）
特表２０１８－５３２１５１（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１７／２６