(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)公開番号
特開2024-50983
(P2024-50983A)
(43)公開日　令和6年4月10日(2024.4.10)

(51)国際特許分類　FI
G10L 15/32 (2013.01)　G10L 15/32　200Z
G10L 15/16 (2006.01)　G10L 15/16

審査請求　有　請求項の数　20　OL　（全22頁）

(21)出願番号　特願2024-24188(P2024-24188)
(22)出願日　令和6年2月21日(2024.2.21)
(62)分割の表示　特願2023-558803(P2023-558803)の分割
原出願日　令和4年3月22日(2022.3.22)
(31)優先権主張番号　63/166,916
(32)優先日　令和3年3月26日(2021.3.26)
(33)優先権主張国・地域又は機関
米国(US)

(71)出願人　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　100142907
弁理士　本田 淳
(72)発明者　ガウル、ニーラジ
アメリカ合衆国 ９４０４３ カリフォルニア州 マウンテン ビュー アンフィシアター パークウェイ １６００
最終頁に続く

(54)【発明の名称】　自動音声認識のための多言語再スコアリングモデル

(57)【要約】
【課題】Ｎ個の候補仮説の中から、最も高い各々の総合スコアを有している候補仮説を、発話の最終トランスクリプション（１２０）として選択する。
【解決手段】方法（４００）は、発話（１０６）に対応する音声データから抽出済みの音響フレーム（１１０）のシーケンスを受信する。第１パス（３０１）中、音響フレームのシーケンスを処理して、発話に対するＮ個の候補仮説（２０４）を生成する。第２パス（３０２）中、および各候補仮説に対して、本方法は：各々の非正規化尤度スコア（３２５）を生成する工程と、各々の外部言語モデルスコア（３１５）を生成する工程と、対応する候補仮説の事前統計をモデル化する単体スコア（２０５）を生成する工程と、非正規化尤度スコア、外部言語モデルスコア、および単体スコアに基づき、候補仮説に対する各々の総合スコア（３５５）を生成する工程と、を備える。
【選択図】図１

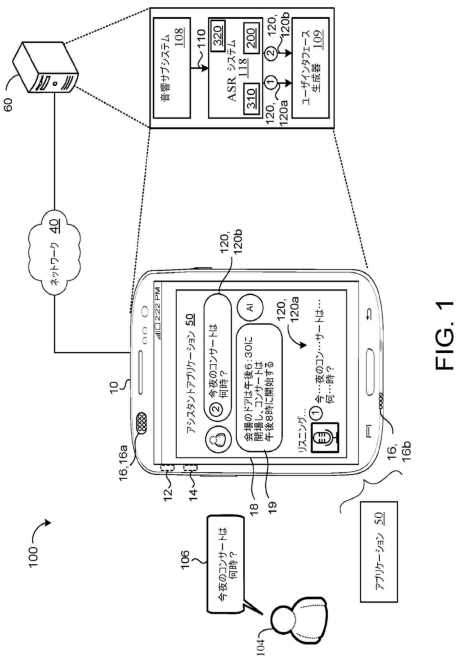

FIG. 1

【特許請求の範囲】
【請求項１】
データ処理ハードウェア上で実行されたときに前記データ処理ハードウェアに動作を実行させるコンピュータ実装方法としての方法であって、前記動作は、
発話のグラウンドトゥルーストランスクリプションとでペアにされた前記発話に対応するトレーニング音声データを備えている、書き起こされた音声トレーニングデータを受信する工程と、
第１パス中に、音声認識モデルを使用することで前記トレーニング音声データを処理することによって、前記発話のためのＮ個の候補仮説を生成する工程であって、Ｎ個の前記候補仮説のうちの対応する前記各候補仮説は各々の第１パススコアを有している、Ｎ個の前記候補仮説を生成する工程と、
第２パス中に、Ｎ個の前記候補仮説のうちの対応する前記候補仮説ごとに
ニューラルネットワーク再スコアリングモデルを使用することで、対応する前記候補仮説についての各々の前記第１パススコアに基づき、各々の第２パススコアを生成するとともに、
対応する前記候補仮説と前記グラウンドトゥルーストランスクリプションとの間の各々の負の編集距離に、ソフトマックス関数を適用する工程と、
前記グラウンドトゥルーストランスクリプションとＮ個の前記候補仮説のうちの対応する前記各候補仮説との間の各々の前記負の編集距離に適用される前記ソフトマックス関数に基づき、前記ニューラルネットワーク再スコアリングモデルのモデルパラメータを最適化する工程と、
を備えている、方法。
【請求項２】
前記音声認識モデルは、リカレントニューラルネットワーク－トランスデューサ（ＲＮＮ－Ｔ）アーキテクチャを備えている、
請求項１に記載の方法。
【請求項３】
Ｎ個の前記候補仮説のうちの対応する前記各候補仮説は、単語ラベルの各々のシーケンスを備えており、
前記各単語ラベルは、各々の埋込ベクトルによって表わされる、
請求項１に記載の方法。
【請求項４】
Ｎ個の前記候補仮説のうちの対応する前記各候補仮説は、サブワードラベルの各々のシーケンスを備えており、
前記各サブワードラベルは、各々の埋込ベクトルによって表わされる、
請求項１に記載の方法。
【請求項５】
前記第１パス中に前記音声認識モデルを使用することで生成済みのＮ個の前記候補仮説は、候補仮説のＮベストリストを備えている、
請求項１に記載の方法。
【請求項６】
前記音声認識モデルは、複数のコンフォーマ層を有しているコンフォーマエンコーダを備えているエンコーダデコーダアーキテクチャを備えている、
請求項１に記載の方法。
【請求項７】
前記音声認識モデルは、複数のトランスフォーマ層を有しているトランスフォーマエンコーダを備えているエンコーダデコーダアーキテクチャを備えている、
請求項１に記載の方法。
【請求項８】
前記動作はさらに、Ｎ個の前記候補仮説について生成済みの前記各第２パススコアに基

づき、前記発話の最終トランスクリプションとしてＮ個の前記候補仮説のうちの１つを選択する工程を備えている、
請求項１に記載の方法。
【請求項９】
前記ニューラルネットワーク再スコアリングモデルは、言語固有のニューラルネットワーク再スコアリングモデルを備えている、
請求項１に記載の方法。
【請求項１０】
前記ニューラルネットワーク再スコアリングモデルは、多言語ニューラルネットワーク再スコアリングモデルを備えている、
請求項１に記載の方法。
【請求項１１】
システムであって、前記システムは、
データ処理ハードウェアと、
前記データ処理ハードウェアに通信するメモリハードウェアであって、前記データ処理ハードウェア上で実行されると前記データ処理ハードウェアに動作を実行させる命令を記憶している前記メモリハードウェアと、
を備えており、前記動作は、
発話のグラウンドトゥルーストランスクリプションとでペアにされた前記発話に対応するトレーニング音声データを備えている、書き起こされた音声トレーニングデータを受信する工程と、
第１パス中に、音声認識モデルを使用することで前記トレーニング音声データを処理することによって、前記発話のためのＮ個の候補仮説を生成する工程であって、Ｎ個の前記候補仮説のうちの対応する前記各候補仮説は各々の第１パススコアを有している、Ｎ個の前記候補仮説を生成する工程と、
第２パス中に、Ｎ個の前記候補仮説のうちの対応する前記候補仮説ごとに、
ニューラルネットワーク再スコアリングモデルを使用することで、対応する前記候補仮説についての各々の前記第１パススコアに基づき、各々の第２パススコアを生成するとともに、
対応する前記候補仮説と前記グラウンドトゥルーストランスクリプションとの間の各々の負の編集距離に、ソフトマックス関数を適用する工程と、
前記グラウンドトゥルーストランスクリプションとＮ個の前記候補仮説のうちの対応する前記各候補仮説との間の各々の前記負の編集距離に適用される前記ソフトマックス関数に基づき、前記ニューラルネットワーク再スコアリングモデルのモデルパラメータを最適化する工程と、
を備えている、システム。
【請求項１２】
前記音声認識モデルは、リカレントニューラルネットワーク－トランスデューサ（ＲＮＮ－Ｔ）アーキテクチャを備えている、
請求項１１に記載のシステム。
【請求項１３】
Ｎ個の前記候補仮説のうちの対応する前記各候補仮説は、単語ラベルの各々のシーケンスを備えており、
前記各単語ラベルは、各々の埋込ベクトルによって表わされる、
請求項１１に記載のシステム。
【請求項１４】
Ｎ個の前記候補仮説のうちの対応する前記各候補仮説は、サブワードラベルの各々のシーケンスを備えており、
前記各サブワードラベルは、各々の埋込ベクトルによって表わされる、
請求項１１に記載のシステム。

【請求項１５】
前記第１パス中に前記音声認識モデルを使用することで生成済みのＮ個の前記候補仮説は、前記候補仮説のＮベストリストを備えている、
請求項１１に記載のシステム。
【請求項１６】
前記音声認識モデルは、複数のコンフォーマ層を有しているコンフォーマエンコーダを備えているエンコーダデコーダアーキテクチャを備えている、
請求項１１に記載のシステム。
【請求項１７】
前記音声認識モデルは、複数のトランスフォーマ層を有しているトランスフォーマエンコーダを備えているエンコーダデコーダアーキテクチャを備えている、
請求項１１に記載のシステム。
【請求項１８】
前記動作はさらに、Ｎ個の前記候補仮説について生成済みの前記各第２パススコアに基づき、前記発話の最終トランスクリプションとしてＮ個の前記候補仮説のうちの１つを選択する工程を備えている、
請求項１１に記載のシステム。
【請求項１９】
前記ニューラルネットワーク再スコアリングモデルは、言語固有のニューラルネットワーク再スコアリングモデルを備えている、
請求項１１に記載のシステム。
【請求項２０】
前記ニューラルネットワーク再スコアリングモデルは、多言語ニューラルネットワーク再スコアリングモデルを備えている、
請求項１１に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、自動音声認識のための多言語再スコアリングモデルに関する。
【背景技術】
【０００２】
自動音声認識（ＡＳＲ）システムは、一般的にモバイル機器（移動装置）やその他の機器で使用される技術を提供する。一般に、自動音声認識ＡＳＲシステムは、ユーザがモバイル機器に話した内容の正確なトランスクリプション（転写、書き起こし、採録）を提供しようとする。より具体的には、自動音声認識ＡＳＲシステムは複数のトランスクリプション候補を生成するとともに、音声入力に一致する可能性が最も高いトランスクリプション候補を出力する。場合によっては、自動音声認識ＡＳＲシステムは、ユーザが実際に話した内容には一致しない不正確なトランスクリプションを出力する。このような場合、自動音声認識ＡＳＲシステムは複数のトランスクリプション候補を再スコアリングするとともに、音声入力に一致する正確なトランスクリプションを出力する。
【先行技術文献】
【非特許文献】
【０００３】
【非特許文献１】OGAWA ATSUNORI ET AL, "Rescoring N-Best Speech Recognition List Based on One-on-One Hypothesis Comparison Using Encoder-Classifier Model", 2018 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), IEEE,15 April 2018 (2018-04-15), page 6099-6103,XP033403971,DOI: 10.1109/ICASSP.2018.8461405
【発明の概要】

【発明が解決しようとする課題】
【０００４】
しかし、再スコアリングの１つの課題は、自動音声認識ＡＳＲシステムが複数のトランスクリプション候補を正確に再スコアリングするべく、音声入力の言語情報に依存することである。そのため、自動音声認識ＡＳＲシステムが多言語の音声環境で再スコアリングを行なうのは、しばしば面倒な作業となる。
【課題を解決するための手段】
【０００５】
本開示の一態様は、データ処理ハードウェア上で実行されるとデータ処理ハードウェアに、自動音声認識のための多言語再スコアリングモデルを使用する動作（操作、オペレーション）を実行させる、コンピュータ実装方法を提供する。動作は、発話に対応する音声データから抽出済みの音響フレームのシーケンスを受信する工程を備えている。第１パス中に、動作は、発話に対するＮ個の候補仮説（ヒポセシス）を生成するべく、多言語音声認識モデルを使用することで、音響フレームのシーケンスを処理する工程を備えている。第２パス中に、Ｎ個の候補仮説の各候補仮説について、本方法は以下を備えている：すなわち、ニューラルオラクルサーチ（ＮＯＳ）モデルを使用することで、音響フレームのシーケンスと、対応する候補仮説と、に基づき各々の非正規化（正規化されていない）尤度スコアを生成する工程と、言語モデルを使用することで、各々の外部言語モデルスコアを生成する工程と、第１パス中に生成済みの対応する候補仮説の事前統計（プライアスタティスティクス）をモデル化する単体（スタンドアロン、単独）スコアを生成する工程と、非正規化スコア、外部言語モデルスコア、および単体スコア、に基づき候補仮説の各々の総合（オーバーオール、全体）スコアを生成する工程と、を備えている。動作はまた、Ｎ個の候補仮説の中から、最も高い各々の総合（全体）スコアを有している候補仮説を、発話の最終トランスクリプションとして選択する工程も備えている。
【０００６】
本開示の実施形態は、以下の任意の特徴の１つまたは複数を備えていることができる。いくつかの実装では、Ｎ個の候補仮説の各候補仮説は、単語またはサブワードラベル（ワードラベルまたはサブワードラベル）の各々のシーケンスを備えている。ここで、各単語またはサブワードラベルは、各々の埋込ベクトルによって表わされる。外部言語モデルは、テキストのみのデータでトレーニング（訓練、学習）されてもよい。いくつかの例では、ニューラルオラクルサーチＮＯＳモデルは言語固有（言語特異的）のニューラルオラクルサーチＮＯＳモデルを備えている。これらの例では、動作はさらに、発話の言語を示す言語識別子を受信する工程と、異なる各々の言語について各々が訓練（トレーニング）された複数の言語固有ニューラルオラクルサーチＮＯＳモデルの中から、言語固有ニューラルオラクルサーチＮＯＳモデルを選択する工程と、を備えている。
【０００７】
オプションとして、ニューラルオラクルサーチＮＯＳモデルは、多言語ニューラルオラクルサーチＮＯＳモデルを備えていることができる。いくつかの実装では、外部言語モデルは、言語固有の外部言語モデルを備えている。これらの実装では、動作はさらに、発話の言語を示す言語識別子を受信する工程と、異なる各々の言語で各々が訓練された複数の言語固有の外部言語モデルの中から言語固有の外部言語モデルを選択する工程と、を備えている。ニューラルオラクルサーチＮＯＳモデルは、２つの単方向長短期記憶（一方向ＬＳＴＭ）層を備えていることができる。いくつかの例では、音声認識モデルは、複数のコンフォーマ層を有しているコンフォーマエンコーダと、２つのＬＳＴＭ層を有しているＬＳＴＭデコーダと、を備えているエンコーダデコーダアーキテクチャを備えている。
【０００８】
本開示の別の態様は、データ処理ハードウェアと、データ処理ハードウェア上で実行されるとデータ処理ハードウェアに動作（オペレーション、操作）を実行させる命令を記憶するメモリハードウェアと、を備えているシステムを提供する。動作は、発話に対応する音声データから抽出済みの音響フレームのシーケンスを受信する工程を備えている。第１

パス中に、動作は、発話に対するＮ個の候補仮説を生成するべく、多言語音声認識モデルを使用することで、音響フレームのシーケンスを処理する工程を備えている。第２パス中に、Ｎ個の候補仮説の各候補仮説について、本方法は以下を備えている：すなわち、ニューラルオラクルサーチ（ＮＯＳ）モデルを使用することで、音響フレームのシーケンスと、対応する候補仮説と、に基づき各々の非正規化尤度スコアを生成する工程と、言語モデルを使用することで、各々の外部言語モデルスコアを生成する工程と、第１パス中に生成済みの対応する候補仮説の事前統計をモデル化する単体スコアを生成する工程と、非正規化スコア、外部言語モデルスコア、および単体スコア、に基づき候補仮説の各々の総合スコアを生成する工程と、を備えている。動作はまた、Ｎ個の候補仮説の中から最も高い各々の総合スコアを有している候補仮説を、発話の最終トランスクリプションとして選択する工程も備えている。

【０００９】

本開示の実施形態は、以下の任意の特徴の１つまたは複数を備えていることができる。いくつかの実装では、Ｎ個の候補仮説の各候補仮説は、単語またはサブワードラベルの各々のシーケンスを備えている。ここで、各単語またはサブワードラベルは、各々の埋込ベクトルによって表わされる。外部言語モデルは、テキストのみのデータでトレーニングされてもよい。いくつかの例では、ニューラルオラクルサーチＮＯＳモデルは言語固有のニューラルオラクルサーチＮＯＳモデルを備えている。これらの例では、動作はさらに、発話の言語を示す言語識別子を受信する工程と、異なる各々の言語について各々が訓練された複数の言語固有ニューラルオラクルサーチＮＯＳモデルの中から、言語固有ニューラルオラクルサーチＮＯＳモデルを選択する工程と、を備えている。

【００１０】

オプションとして、ニューラルオラクルサーチＮＯＳモデルは多言語ニューラルオラクルサーチＮＯＳモデルを備えていることができる。いくつかの実装では、外部言語モデルは、言語固有の外部言語モデルを備えている。これらの実装では、動作はさらに、発話の言語を示す言語識別子を受信する工程と、異なる各々の言語で各々が訓練された複数の言語固有の外部言語モデルの中から言語固有の外部言語モデルを選択する工程と、を備えている。ニューラルオラクルサーチＮＯＳモデルは、２つの単方向長短期記憶（一方向ＬＳＴＭ）層を備えていることができる。いくつかの例では、音声認識モデルは、複数のコンフォーマ層を有しているコンフォーマエンコーダと、２つのＬＳＴＭ層を有しているＬＳＴＭデコーダと、を備えているエンコーダデコーダアーキテクチャを備えている。

【００１１】

本開示の１つまたは複数の実施態様の詳細は、添付の図面および以下の説明に記載されている。他の態様、特徴、および利点、は説明および図面、ならびに特許請求の範囲から明らかになるであろう。

【図面の簡単な説明】

【００１２】

【図１】一例の音声認識モデルを実行する、音声環境の概略図。

【図２】図１の例示的な音声認識モデルの概略図。

【図３Ａ】複数の言語固有のニューラルオラクル探索（ＮＯＳ）モデルを使用する、例示的な再スコアリング処理の概略図。

【図３Ｂ】多言語ニューラルオラクルサーチＮＯＳモデルを用いた、再スコアリング処理の一例を示す概略図。

【図４】自動音声認識のために多言語の再スコアリングモデルを使用する、方法の動作の配置例のフローチャート。

【図５】本明細書に記載のシステムおよび方法を実施するべく使用され得る、例示的なコンピューティング装置の概略図。

【発明を実施するための形態】

【００１３】

様々な図面における同様の参照符号は、同様の要素を示す。

自動音声認識（ＡＳＲ）システムは、ユーザが話した内容のより正確なトランスクリプション（転写、書き起こし、音声記録）を提供するべく、ユーザ装置（ユーザデバイス、ユーザ機器）にますます普及している。しかし、自動音声認識ＡＳＲシステムは、ユーザが実際に話した内容を誤認識した不正確なトランスクリプションを生成する場合もありうる。いくつかの構成では、自動音声認識ＡＳＲシステムは、発声済みの発話に対してＮ個の最良候補仮説を生成するとともに、最良の候補仮説を最終トランスクリプションとして出力する。しかし、Ｎ個の最良候補仮説の構成は、１つの最良仮説の構成とで比較して、単語誤り率（ＷＥＲ）がほぼ５０％低くなる。したがって、いくつかの実装では、自動音声認識ＡＳＲシステムは、単語誤り率ＷＥＲを高めるべく追加情報を統合することによって、Ｎ個の最良候補仮説を再スコアリングする。このような再スコアリングの実装は、多言語音声環境における言語情報（すなわち、ユーザが話した言語識別子）に依存しており、わずかな単語誤り率ＷＥＲの改善しか提供しない。上述した課題によって、Ｎ個の最良候補仮説構成を使用する自動音声認識ＡＳＲシステムと、１個の最良候補仮説構成を使用する自動音声認識ＡＳＲシステムと、の間に単語誤り率ＷＥＲ性能のギャップがあることが明らかになった。

【００１４】

したがって、本明細書の実装は、対応する発話に対してＮ個の候補仮説を生成する再スコアリング処理を実行するとともに、最も可能性の高い候補仮説を選択して最終トランスクリプションとして出力するような、方法およびシステムに向けられている。特に、第１パス中、再スコアリング処理は、多言語音声認識モデルを使用することで、Ｎ個の候補仮説を生成する。その後、第２パス中、再スコアリング処理は、各候補仮説について、ニューラルオラクルサーチ（ＮＯＳ）モデルを使用することで各々の非正規化（正規化されていない）尤度スコアを生成したり、外部言語モデルのスコアを生成したり、候補仮説の事前統計をモデル化した単体（スタンドアロン、単独）スコアを生成したりする。以下で明らかになるように、ニューラルオラクルサーチＮＯＳモデルは、言語固有のニューラルオラクルサーチＮＯＳモデルまたは多言語ニューラルオラクルサーチＮＯＳモデルであってもよい。さらに、第２パス中、再スコアリング処理は、非正規化尤度スコア、外部言語モデルスコア、および単体スコア、に基づき各候補仮説の総合スコア（オーバーオールスコア）を生成する。再スコアリング処理は、総合スコアが最も高い候補仮説を、発話に対する最終トランスクリプションとして選択する。

【００１５】

図１は、音声（スピーチ、発話）環境１００の一例である。音声環境１００において、ユーザ１０４がユーザ装置１０などのコンピューティング装置とで対話する方法は、音声入力であってもよい。ユーザ装置１０は、音声環境１００内の１人または複数のユーザ１０４からの音（例えば、ストリーミング音響データ）を取り込む（キャプチャする、捕捉する）ように構成される。ここで、ストリーミング音響（オーディオ）データは、可聴クエリ、ユーザ装置１０に対するコマンド、またはユーザ装置１０によって捕捉された可聴コミュニケーション、として機能するユーザ１０４による発話（話し言葉、音声発話）１０６を指す場合がある。ユーザ装置１０の音声対応システムは、クエリに応答することによって、および／またはコマンドを１つまたは複数の下流アプリケーションによって実行／履行させることによって、クエリまたはコマンドをフィールド化することができる。

【００１６】

ユーザ装置１０は、ユーザ１０４に関連付けられているとともに音響データを受信することができる任意のコンピューティング装置に対応することができる。ユーザ装置１０のいくつかの例には、モバイル機器（例えば、携帯電話、タブレット、ラップトップ、など）、コンピュータ、ウェアラブル機器（例えば、スマートウォッチ）、スマート家電、モノのインターネット（ＩｏＴ）機器、車両インフォテインメントシステム、スマートディスプレイ、スマートスピーカ、などが含まれるが、これらに限定されない。ユーザ装置１０は、データ処理ハードウェア１２と、データ処理ハードウェア１２に通信するメモリハードウェア１４と、を備えている。メモリハードウェア１４は、データ処理ハードウェア

１２によって実行されるとデータ処理ハードウェア１２に１つまたは複数の動作を実行させる命令を記憶する。ユーザ装置１０はさらに音声システム１６を備えている。音声システム１６は、音声環境１００内の発話１０６を捕捉およびカバーして電気信号に変換するための音声キャプチャ装置（例えば、マイクロフォン）１６、１６ａと、可聴音声信号を（例えば、ユーザ装置１０からの出力音声データとして）伝達するための音声出力装置（例えば、スピーカ）１６、１６ｂと、を有している。ユーザ装置１０は、図示の例では単一の音声（音響）キャプチャ装置１６ａを実装しているが、ユーザ装置１０は、本開示の範囲から逸脱することなく音声キャプチャ装置１６ａのアレイを実装してもよい。それによって、アレイ内の１つまたは複数の音声キャプチャ装置１６ａは、ユーザ装置１０上に物理的に存在しなくてもよいが、音声（音響）システム１６に通信していてもよい。

【００１７】

音声環境１００において、音声認識モデル（すなわち、自動音声認識ＡＳＲモデル）２００を実装する自動音声認識（ＡＳＲ）システム１１８は、ユーザ１０４のユーザ装置１０上に、および／または、ネットワーク４０を介してユーザ装置１０に通信するリモートコンピューティング装置６０（例えば、クラウドコンピューティング環境で実行される分散システムの１つまたは複数のリモートサーバ）上に、存在（常駐）する。自動音声認識ＡＳＲシステム１１８はまた、１つまたは複数の外部言語モデル３１０およびニューラルオラクルサーチ（ＮＯＳ）モデル３２０を実装することができる。ユーザ装置１０および／またはリモートコンピューティング装置（すなわち、リモートサーバ）６０は、音声サブシステム１０８も備えている。音声サブシステム１０８は、ユーザ１０４によって発声済みの発話１０６であって、音声キャプチャ装置１６ａによってキャプチャ済みの発話１０６を、受信するとともに、当該発話１０６を、自動音声認識ＡＳＲシステム１１８によって処理可能な入力音響フレーム（音響フレーム１１０）に関連付けられた対応するデジタルフォーマットに変換するように構成されている。図示の例では、ユーザが各々の発話１０６を話している一方で、音声サブシステム１０８は当該発話１０６を、自動音声認識ＡＳＲシステム１１８に入力するための対応する音声データ（例えば、音響フレーム１１０）変換する。その後、音声認識モデル２００は、入力として、発話１０６に対応する音声データ（１１０）を受信するとともに、出力として、発話１０６の対応するトランスクリプション１２０（例えば、音声認識結果／仮説）を生成／予測する。以下にさらに詳細に説明するように、音声認識モデル２００は、発話１０６によって指定済みのクエリが待ち時間に対してどの程度敏感であるか、および／またはユーザ１０４が待ち時間に対してどの程度寛容であるか、に応じて音声認識を実行する際に、音声認識モデル２００が推論中に、先読み音声コンテキストの異なる継続時間を設定できるようにするべく、可変の先読み音声コンテキストでトレーニング済みのエンドツーエンドの音声認識モデル２００を備えていることができる。例えば、ユーザ装置１０上で実行されるデジタルアシスタントアプリケーション５０は、発話１０６によって指定済みのクエリが待ち時間（レイテンシ）に対してどの程度敏感であるか、および／またはユーザ１０４が待ち時間に対してどの程度の許容範囲を持っているか、に応じて音声認識を要求することができる。

【００１８】

いくつかの実装では、音声認識モデル２００は、Ｎ個の候補仮説２０４（図３Ａ、図３Ｂ）を生成するための第１パス中、音声データ１１０に対してストリーミング音声認識を実行している。ニューラルオラクルサーチＮＯＳモデル３２０および言語モデル３１０は、最終トランスクリプション（転写、書き起こし）１２０を生成するための第２パス中、Ｎ個の候補仮説２０４を再スコアリングする。例えば、図示の例では、音声認識モデル２００は、（Ｎ個の候補仮説２０４に基づき）部分音声認識結果（すなわち、部分トランスクリプション）１２０、１２０ａを生成するべく、音声データ１１０に対してストリーミング音声認識を実行している。言語モデル３１０およびニューラルオラクルサーチＮＯＳモデル３２０は、最終音声認識結果（すなわち、最終トランスクリプション）１２０、１２０ｂを生成するべく、Ｎ個の候補仮説２０４を再スコアリングする。特に、音声認識モデル２００は、部分音声認識結果１２０ａを生成するべく、ゼロ（または約２４０ミリ秒

）に設定され得る先読み音声コンテキストを使用してもよい。したがって、入力発話（発話１０６）に対する最終音声認識結果１２０ｂは、入力発話に対する部分音声認識結果１２０ａから遅れてもよい。

【００１９】

ユーザ装置１０および／またはリモートコンピューティング装置６０はまた、発話１０６のトランスクリプション１２０の表現を、ユーザ装置１０のユーザ１０４に提示するように構成されたユーザインタフェース生成部１０９を実行する。以下にさらに詳細に説明するように、ユーザインタフェース生成部１０９は、第１時間１中に部分音声認識結果１２０ａをストリーミング方式で表示することができる。その後、ユーザインタフェース生成部１０９は、第２時間２中に最終音声認識結果１２０ｂを表示することができる。いくつかの構成では、自動音声認識ＡＳＲシステム１１８から出力されたトランスクリプション１２０は、例えば、ユーザ装置１０またはリモートコンピューティング装置６０上で実行される自然言語理解（ＮＬＵ）モジュールによって処理されることで、発話１０６によって指定済みのユーザコマンド／クエリを実行する。追加的または代替的に、テキスト音声合成システム（図示せず）（例えば、ユーザ装置１０またはリモートコンピューティング装置６０の任意の組み合わせ上で実行される）は、ユーザ装置１０および／または別の機器による可聴出力用に、トランスクリプションを合成音声に変換してもよい。

【００２０】

図示の例では、ユーザ１０４がデジタルアシスタントアプリケーション５０に通信している。デジタルアシスタントアプリケーション５０は、ユーザ１０４とデジタルアシスタントアプリケーション５０との間の会話を描写するべく、ユーザ装置１０の画面上にデジタルアシスタントインタフェース１８を表示する。この例では、ユーザ１０４はデジタルアシスタントアプリケーション５０に、「今夜のコンサートは何時？」と質問する。ユーザ１０４からのこの質問は、音声キャプチャ装置１６ａによってキャプチャされるとともに、ユーザ装置１０の音声システム１６によって処理される発話１０６である。この例では、音声システム１６は発話１０６を受信するとともに、当該発話１０６を、自動音声認識ＡＳＲシステム１１８に入力するための音響フレーム１１０に変換する。

【００２１】

この例を続けると、音声認識モデル２００は、ユーザ１０４が話す発話１０６に対応する音響フレーム（すなわち、音声データ）１１０を受信しながら、音響フレーム１１０を符号化（エンコード）するだけでなく、さらに符号化（エンコード）済みの音響フレーム１１０を部分音声認識結果１２０ａに復号化（デコード）する。第１時間１中、ユーザインタフェース生成部１０９は、デジタルアシスタントインタフェース１８を介して、発話１０６の部分音声認識結果１２０ａの表現を、単語、単語片、および／または個々の文字、が発声されるとすぐに画面上に現れるストリーミング方式で、ユーザ装置１０のユーザ１０４に提示する。

【００２２】

第２パス中、および発話１０６に対応する全ての音響フレーム１１０が受信された後、自動音声認識ＡＳＲシステム１１８は、言語モデル３１０およびニューラルオラクルサーチＮＯＳモデル３２０を使用することで、Ｎ個の候補仮説２０４のうちの各候補仮説２０４を再スコアリングするとともに、Ｎ個の候補仮説２０４の中から、発話１０６の正確なトランスクリプション（転写）１２０である可能性（尤度）が最も高い候補仮説２０４を選択する。第２時間２中、ユーザインタフェース生成部１０９は、デジタルアシスタントインタフェース１８を介して、発話１０６の最終音声認識結果１２０ｂの表現を、ユーザ装置１０のユーザ１０５に提示する。いくつかの実装では、ユーザインタフェース生成部１０９は、部分音声認識結果１２０ａの表現を、最終音声認識結果１２０ｂの表現によって置き換える。例えば、最終音声認識結果１２０ｂは、先読み音声コンテキストを活用せずに生成済みの部分音声認識結果１２０ａよりも、正確であると推定される。よって、最終的にトランスクリプション１２０として表示される最終音声認識結果１２０ｂは、部分音声認識結果１２０ａにおいて誤認識されたかもしれない用語を、修正することができる

。この例では、ストリーミングの部分音声認識結果１２０ａは、音声認識モデル２００によって出力されているとともに、第１時間１にユーザ装置１０の画面上に表示されており、低レイテンシ（低い待ち時間）に関連付けられている。よって、自分のクエリが処理されているという応答性をユーザ１０４に提供する。一方、第２時間２に画面上に表示される最終音声認識結果１２０ｂは、精度の点で音声認識品質を向上させるが、待ち時間（レイテンシ）が増大する。しかし、部分音声認識結果１２０ａはユーザが発話１０６を話すときに表示されるので、最終認識結果を生成するとともに最終的に表示することに関連する高い待ち時間は、ユーザ１０４には気づかれない。

【００２３】

図１に示す例では、デジタルアシスタントアプリケーション５０は、自然言語処理を使用することで、ユーザ１０４によって提起された質問に応答することができる。自然言語処理は、一般に、書かれた言語（例えば、部分音声認識結果１２０ａおよび／または最終音声認識結果１２０ｂ）を解釈するとともに、書かれた言語が何らかのアクションを促すかどうかを決定する、といった処理を指す。この例では、デジタルアシスタントアプリケーション５０は自然言語処理を使用することで、ユーザ１０４からの質問が、ユーザのスケジュールに関するものであることを、より詳細にはユーザのスケジュール上のコンサートに関するものであることを、認識する。自然言語処理でこれらの詳細を認識することによって、自動アシスタントは、ユーザの問い合わせ（クエリ）に対して、応答１９を返している。応答１９は、「会場のドアは午後６時３０分に開き、コンサートは午後８時に始まります」と述べる。いくつかの構成では、自然言語処理は、ユーザ装置１０のデータ処理ハードウェア１２に通信しているリモートサーバ６０上で、行なわれる。

【００２４】

図２を参照すると、例示的なフレームアライメントベースのトランスデューサモデル２００ａは、対話型アプリケーションに関連する待ち時間制約を遵守するリカレントニューラルネットワーク－トランスデューサ（ＲＮＮ－Ｔ）モデルアーキテクチャを備えている。リカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデルアーキテクチャの使用は例示的なものであり、フレームアライメントベースのトランスデューサモデル２００は、特に、トランスフォーマ－トランスデューサおよびコンフォーマ－トランスデューサモデルアーキテクチャなどの、他のアーキテクチャを含み得る。リカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデル２００は、小さな計算フットプリントを提供しているとともに、従来の自動音声認識ＡＳＲアーキテクチャよりも少ないメモリ要件を利用するので、リカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデルアーキテクチャは、ユーザ装置１０２上で完全に音声認識を実行するのに適している（たとえば、リモートサーバとの通信は必要無い）。リカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデル２００は、エンコーダネットワーク２１０、予測ネットワーク２２０、および結合（ジョイント）ネットワーク２３０、を備えている。エンコーダネットワーク２１０は、従来の自動音声認識ＡＳＲシステムにおける音響モデル（ＡＭ：アコースティックモデル）にほぼ類似しているので、スタックされたロングショートターム（ＬＳＴＭ）層のリカレントネットワークを備えていることができる。例えば、エンコーダは、ｄ次元特徴ベクトルのシーケンス（例えば、音響フレーム１１０（図１）$x=(X_1, X_2, \cdots, X_T)$、ここで$X_t \in R^d$、Ｒは白抜き文字）を読み取ることによって、各出力ステップで高次特徴表現を生成する。この高次特徴表現は、$h_1^{enc}, \cdots, h_T^{enc}$と表記される。

【００２５】

同様に、予測ネットワーク２２０もＬＳＴＭネットワーク（すなわち、ＬＳＴＭデコーダ）であり、言語モデル（ＬＭ）のように、これまでの最終ソフトマックス層２４０によって出力された非ブランク記号シーケンス（すなわち、ラベル履歴）２４５（$y_0, \cdots, y_{ui-1}$）を、密な表現p_{ui}に変換する。最後に、リカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデルアーキテクチャでは、エンコーダおよび予測／デコーダネットワーク２１０、２２０によって生成済みの表現同士は、結合ネットワーク２

３０によって結合される。予測ネットワーク２２０は、密な表現同士を処理する代わりに、ルックアップ（先読み）されたスパース（疎）な埋め込みを出力することによってレイテンシ（待ち時間）を改善するべく、埋込ルックアップテーブルによって置き換えられてもよい。次に、結合ネットワークは、次の出力記号に対する分布である、Ｐ（$y_i | X_{t_i}, y_0, \cdots, y_{u_i-1}$）を予測する。別の言い方をすれば、結合ネットワーク２３０は、各出力ステップ（例えば、時間ステップ）において、可能性のある音声認識仮説に対する確率分布を生成する。ここで、「可能性のある（ポシブル）音声認識仮説」は、指定済みの自然言語における記号（シンボル）／文字を各々表わす、出力ラベルのセットに対応する。例えば、自然言語が英語である場合、出力ラベルのセットは、２７個のシンボル（記号）を備えていることができ、例えば、英語のアルファベットにおける２６個の文字の各々に対する１個のラベルと、スペースを指定する１個のラベルと、を備えていることができる。従って、結合ネットワーク２３０は、所定の出力ラベルの集合（セット）の各々の発生の尤度を示す値の集合を出力することができる。この値のセットはベクトルとすることができるとともに、出力ラベルのセットにわたる確率分布を示すことができる。場合によっては、出力ラベルは書記素（例えば、個々の文字であり、潜在的には句読点や他の記号）であるが、出力ラベルのセットはそれほど限定されない。例えば、出力ラベルのセットは、書記素（グラフェムズ）に加えて、または書記素の代わりに、単語片および／または単語全体を備えていることができる。結合ネットワーク２３０の出力分布は、異なる出力ラベル同士の各々に対する事後確率値を備えていることができる。したがって、異なる書記素または他の記号を表わす１００個の異なる出力ラベルが存在する場合、結合ネットワーク２３０の出力ｙｉは、出力ラベルごとに１つずつの確率値であるように、１００個の異なる確率値を備えていることができる。次に、確率分布は、トランスクリプション１２０を決定するためのビーム探索処理（例えば、ソフトマックス層２４０による）において、候補となる正書法（オルソグラフィック）要素（例えば、書記素（グラフェムズ）、単語片（ワードピース）、および／または単語（ワード））を選択するとともにスコアを割り当てるべく、使用することができる。

【００２６】

ソフトマックス層２４０は、対応する出力ステップにおいてリカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデル２００によって予測される次の出力記号（アウトプットシンボル）として、分布において最も高い確率を有している出力ラベル／記号を選択する任意の技術を採用することができる。このように、リカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデル２００は、条件付き独立性の仮定を行なわず、むしろ、各記号の予測は、音響だけでなく、これまでに出力されたラベルのシーケンスにも条件付けられる。リカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデル２００は、出力記号が将来の音響フレーム１１０から独立していると仮定している。これによって、リカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデルをストリーミング方式で採用することができる。

【００２７】

いくつかの例では、リカレントニューラルネットワーク－トランスデューサＲＮＮ－Ｔモデル２００のエンコーダネットワーク（例えば、音響エンコーダ）２１０は、コンフォーマ層のスタックを備えているコンフォーマベースのエンコーダを有しているエンコーダデコーダアーキテクチャである。ここで、各コンフォーマ層は、一連（シリーズ）の多頭自己アテンション層、深度ワイズ畳み込み層、およびフィードフォワード層、を備えている。いくつかの例では、コンフォーマベースのエンコーダは、１７個のコンフォーマ層のスタックを備えていることができる。エンコーダネットワーク２１０は、多頭自己アテンション機構を有している、他のタイプのエンコーダを含んでもよい。たとえば、エンコーダネットワーク２１０は、トランスフォーマベースのエンコーダ、または軽量畳み込み（ＬＣｏｎｖ）ベースのエンコーダであってもよい。また、エンコーダネットワーク２１０は、ＬＳＴＭ層のシーケンスを備えているＲＮＮベースであってもよい。予測ネットワーク２２０は、２つの２０４８次元ＬＳＴＭ層を有しているＬＳＴＭデコーダであってもよ

く、各ＬＳＴＭ層には６４０次元の投影層も続く。あるいは、予測ネットワーク２２０は、ＬＳＴＭ層の代わりに、変換器（トランスフォーマ）またはコンフォーマブロックのスタック、または埋込ルックアップテーブルを含んでもよい。最後に、結合ネットワーク２３０も６４０個の隠れユニットを持つことがある。ソフトマックス層２４０は、複数のトレーニングデータセットに含まれる全ての固有の単語片（ワードピース）または書記素を用いて生成済みの、統一された単語片または書記素セットで構成されてもよい。

【００２８】

ここで図３Ａおよび図３Ｂを参照すると、いくつかの実装では、リモートサーバ６０（図１）は、第１パス３０１中に自動音声認識ＡＳＲモデル２００によって生成済みのＮ個の候補仮説２０４を再スコアリングするための例示的な再スコアリング処理３００を実行する。あるいは、ユーザ装置１０（図１）は、リモートサーバ６０（図１）に加えてまたはリモートサーバ６０（図１）の代わりに、例示的な再スコアリング処理３００を実行してもよい。再スコアリング処理３００は、発話１０６に対応する音響フレーム１１０のシーケンス（X_1，X_2，…，X_T）に対して、Ｎ個の候補仮説２０４、２０４ａ～２０４ｎ（H_1，H_2，…，H_N）を生成する第１パス３０１を備えている。さらに、再スコアリング処理３００は、Ｎ個の候補仮説２０４のうちの各候補仮説２０４を、以下でさらに詳細に説明する追加情報源（インフォメーションリソース）を統合することによって再スコアリングする第２パス３０２を備えている。このように、第２パス３０２は、Ｎ個の候補仮説２０４の中から、発話１０６の正確なトランスクリプションである可能性（尤度）が最も高い候補仮説２０４を選択するように構成されたシーケンス分類オブジェクトを備えている。

【００２９】

特に、自動音声認識ＡＳＲモデル２００は、発話１０６に対応する音声データから抽出済みの音響フレーム１１０のシーケンスを受信する。第１パス３０１中、自動音声認識ＡＳＲモデル２００は、音響フレーム１１０のシーケンスを処理することで、発話１０６に対するＮ個の候補仮説２０４を生成する。ここで、各候補仮説２０４は、発話１０６の候補トランスクリプション１２０に対応しており、各々の埋込ベクトルによって表わされる単語、サブワード、および／または書記素ラベルの各々のシーケンスによって表わされる。さらに各候補仮説２０４は、対応する候補仮説２０４の事前統計をモデル化する単体スコア２０５を備えている。すなわち、単体スコア２０５は、対応する候補仮説２０４が発話１０６の正確なトランスクリプションであるという信頼度（確信度、コンフィデンス）を示すことができる。単体スコア２０５の信頼度は、以前に実現された発話１０６の頻度（例えば、候補仮説２０４が以前に発声された回数）を示すこともある。

【００３０】

自動音声認識ＡＳＲモデル２００は、任意の数の候補仮説２０４を生成してもよい（例えば、Ｎは任意の整数値であってもよい）。いくつかの例では、自動音声認識ＡＳＲモデル２００は、予め定義されたパラメータに基づき、指定済みの数の候補仮説２０４を出力する。例えば、自動音声認識ＡＳＲモデル２００は、全ての発話１０６に対して５つの候補仮説２０４（すなわち、Ｎ＝５）を出力する。例えば、Ｎ個の候補仮説２０４は、最も高い単体スコア２０５を有しているＮ個の候補仮説に関連する候補仮説のＮベストリストに対応することができる。他の例では、自動音声認識ＡＳＲモデル２００は、閾値を満たす単体スコア２０５を有している全ての候補仮説２０４を出力する。

【００３１】

図示の例では、自動音声認識ＡＳＲモデル２００は、ユーザ１０４によって発声済みの発話１０６「プレイ＿ネクスト＿ソング」（次の歌を再生して）に対応する音響フレーム１１０のシーケンスを処理する。そして自動音声認識ＡＳＲモデル２００は、３つの候補仮説２０４（すなわち、Ｎ＝３）を生成する。すなわち、複数の候補仮説２０４は、０．６の単体スコア２０５を有している「プレイ＿ネクスト＿ソング」（次の歌を再生して）と、０．３の単体スコア２０５を有している「ヘイ＿ネクスト＿ロング」（やあ次の長い）と、および０．８の単体スコア２０５を有している「プレイ＿ネクスト＿ポン」（次の

ポン（ｐｏｎｇ）を再生して）と、を備えている。ここで、再スコアリング処理３００は、候補仮説２０４「プレイ＿ネクスト＿ポン」（次のポン（ｐｏｎｇ）を再生して）が最も高い単体スコア２０５を有しているので、当該候補仮説２０４「プレイ＿ネクスト＿ポン」を部分（部分的な、パーシャルな）トランスクリプション１２０ａ（図１）として出力することができる。あるいは、再スコアリング処理３００は、再スコアリング処理が最終トランスクリプションを生成するまで、部分トランスクリプションの出力を控えてもよい。特にこの例では、最も高い単体スコア２０５を有している候補仮説２０４は、ユーザ１０４によって発声済みの発話１０６の、不正確なトランスクリプションである。

【００３２】

自動音声認識ＡＳＲモデル２００は、多言語（複数の言語）で話された発話１０６を認識するように構成された、多言語自動音声認識ＡＳＲモデルであってもよい。すなわち、単一の自動音声認識ＡＳＲモデル２００は、第１言語で発話１０６を受信することによって第１言語でＮ個の候補仮説２０４を生成したり、異なる第２言語で別の発話１０６を受信することによって第２言語でＮ個の候補仮説２０４を生成したり、することができる。さらに、単一の自動音声認識ＡＳＲモデルは、第１言語および第２言語の両方の用語を備えているコード混合音声を備えている発話１０６を受信することができる。このように、再スコアリング処理３００は、単一の多言語自動音声認識ＡＳＲモデル２００を、多言語音声環境で実装することができる。

【００３３】

いくつかの実装では、第２パス３０２は、第１パス３０１からＮ個の候補仮説２０４を受け取るとともに、各候補仮説２０４の追加情報を統合することによって、対応する総合スコア３５５を生成することができる。総合スコア３５５は、各候補仮説２０４が正確なトランスクリプションであるかどうかについての、第１パス３０１からの単体スコア２０５よりも一層正確な信頼度を示すことがある。その後、第２パス３０２は、最も高い総合スコア３５５を有している候補仮説２０４を、トランスクリプション１２０（すなわち、最終トランスクリプション１２０ｂ（図１））として選択することができる。

【００３４】

より具体的には、第２パス３０２中に、外部言語モデル（ＬＭ）３１０は、Ｎ個の候補仮説２０４を受信することで、各候補仮説２０４について各々の外部言語モデルスコア３１５を生成する。いくつかの実装では、外部言語モデルＬＭ３１０は、リカレントニューラルネットワーク言語モデルＲＮＮ＿ＬＭを備えている。ここで、外部言語モデルＬＭ３１０は、各々が特定の言語のテキストのみのデータ（すなわち、ペア（対）になっていないデータ）上でトレーニング済みの、複数の言語固有の外部言語モデルＬＭ３１０、３１０ａ～３１０ｎを備えていることができる。このように、「外部言語モデルＬＭ３１０」および「言語固有の外部言語モデルＬＭ３１０」は、本明細書において互換的に使用される場合がある。したがって、各言語固有の外部言語モデルＬＭ３１０は、各々の言語の発話１０６の外部言語モデルスコア（すなわち、言語モデルスコア）３１５を生成するように構成される。例えば、英語のテキストのみのデータでトレーニング済みの第１言語固有の外部言語モデルＬＭ３１０、３１０ａは、英語で発声済みの発話１０６の言語モデルスコア３１５を生成する。スペイン語のテキストのみのデータでトレーニング済みの第２言語固有の外部言語モデルＬＭ３１０、３１０ｂは、スペイン語で発声済みの発話１０６の言語モデルスコア３１５を生成する。複数の外部言語モデルＬＭ３１０は、任意の数の言語でトレーニングすることができる。ここで各外部言語モデルＬＭ３１０は、異なる各々の言語のテキストのみのデータでトレーニングされる。

【００３５】

したがって、外部言語モデルＬＭ３１０は、発話１０６の言語を示す言語識別子１０７を受信することで、複数の言語固有の外部言語モデルＬＭ３１０の中から、発話１０６の言語に対応する言語固有の外部言語モデルＬＭ３１０を選択することができる。別の言い方をすれば、再スコアリング処理３００は、言語識別子１０７に基づき、言語固有の外部言語モデルＬＭ３１０を選択することができる。いくつかの例では、自動音声認識ＡＳＲ

モデル２００は、発話１０６の音響フレーム１１０のシーケンスの処理に基づき、言語識別子１０７を決定する。他の例では、自動音声認識ＡＳＲモデル２００は、外部ソースから言語識別子１０７を取得する。例えば、ユーザは、特定の言語用に自動音声認識ＡＳＲモデルを設定（構成）することができる。他の例では、自動音声認識ＡＳＲモデル２００は、発話１０６を発声したユーザ１０４のアイデンティティを決定するとともに、識別済みのユーザ１０４に関連付けられた言語に基づき言語識別子１０７を識別することができる。

【００３６】

したがって、第２パス３０２中、再スコアリング処理３００は、言語識別子１０７に基づき発話１０６の言語に対応する外部言語モデルＬＭ３１０を選択するとともに、各候補仮説（仮説候補）２０４の言語モデルスコア３１５を生成する。言語モデルスコア３１５は、候補仮説２０４のうちの一連（シーケンス）の仮説用語が、ユーザ１０４によって発声される尤度（可能性）を示す。例えば、外部言語モデルＬＭ３１０は、候補仮説２０４「今日の天気は何？」（ワット＿イズ＿ザ＿ウェザー＿トゥデイ）に対して、候補仮説２０４「天気は何？ホーレイ」（ワット＿イズ＿ザ＿ウェザー＿ホーレイ（ｈｏｏｒａｙ、万歳））とは対照的に、より高い言語モデルスコア３１５を生成する。特に、外部言語モデルＬＭ３１０は、「今日の天気は何？」（ワット＿イズ＿ザ＿ウェザー＿トゥデイ）に対して、より高い言語モデルスコア３１５を生成する。なぜなら、この仮説用語のシーケンスは、「天気は何？ホーレイ」（ワット＿イズ＿ザ＿ウェザー＿ホーレイ（ｈｏｏｒａｙ、万歳））よりも頻繁に、テキストのみのトレーニング（学習）データに含まれている可能性があるからである。

【００３７】

例示的な再スコアリング処理３００はまた、ニューラルオラクルサーチ（ＮＯＳ）モデル３２０を備えている。ニューラルオラクルサーチＮＯＳモデル３２０は、Ｎ個の候補仮説２０４、音響フレーム１１０のシーケンス、およびラベル履歴２４５（例えば、以前に出力された単語、単語片、および／または書記素）、を受信する。ラベル履歴２４５（ｙ０：ｉ－１）は、自動音声認識ＡＳＲモデル２００、（例えば、再スコアラー３５０を介した）再スコアリング処理３００の第２パス３０２、またはそれらの組み合わせ、によって出力され得る。いくつかの例では、ラベル履歴２４５は、ユーザ１０４によって発声済みの以前の発話１０６のトランスクリプション（文字起こし）を備えている。例えば、ユーザ１０４は、「明日はどうする？」（ワット＿アバウト＿トゥデイ？）という現在の発話１０６に対するラベル履歴２４５を表わす、「今日は私に何か会議がある？」（ドゥー＿アイ＿ハブ＿エニー＿ミーティング＿トゥデイ？）という以前の発話１０６を以前に発声したことがあるかもしれない。他の例では、ラベル履歴２４５は、発話の現在のラベルに先行する、全ての用語を備えている。例えば、発話１０６の「プレイ＿マイ＿プレイリスト」（私の再生リストを再生して）について、ラベル履歴２４５は、発話１０６の現在の用語（例えば、次の仮説用語）が「プレイリスト」（再生リスト）である、用語「プレイ＿マイ」（私の再生して）に対応することがある。任意選択で、ニューラルオラクルサーチＮＯＳモデル３２０は、ユーザ１０４によって発声済みの発話１０６の言語を示す言語識別子１０７を受信することができる。

【００３８】

図３Ａは、複数の言語固有ニューラルオラクルサーチＮＯＳモデル３２０Ｓ、３２０Ｓａ～３２０Ｓｎを備えている再スコアリング処理３００、３００ａの一例を示す。ここで、各言語固有ニューラルオラクルサーチＮＯＳモデル３２０Ｓは、特定の言語のペアワイズデータ（すなわち、書き起こされた音響トレーニングデータ）上でトレーニングされる。したがって、第２パス３０２中、再スコアリング処理３００は、言語識別子１０７に基づき、複数の言語固有ニューラルオラクルサーチＮＯＳモデル３２０Ｓの中から、発話１０６の言語に対応する言語固有ニューラルオラクルサーチＮＯＳモデル３２０Ｓを選択する。このように、例示的な再スコアリング処理３００ａは、正しい言語固有ニューラルオラクルサーチＮＯＳモデル３２０Ｓを選択するべく、言語識別子１０７が利用可能にされ

ていると仮定する。
【００３９】
代替的に、図３Ｂは、多言語ニューラルオラクルサーチＮＯＳモデル３２０、３２０Ｍを備えている例示的な再スコアリング処理３００、３００ｂを示す。この例では、多言語ニューラルオラクルサーチＮＯＳモデル３２０Ｍは、任意の数の言語のペアワイズデータ（すなわち、書き起こされた音声トレーニングデータ）上でトレーニングされる。したがって、例示的な再スコアリング処理３００ｂは、多言語音声環境において、単一の多言語ニューラルオラクルサーチＮＯＳモデル３２０Ｍを実装することができる。注目すべきことに、発話１０６の言語に関連付けられた言語固有のニューラルオラクルサーチＮＯＳモデル３２０Ｓ（図３Ａを参照して説明したような）の選択が要求されないので、例示的な再スコアリング処理３００ｂは、任意の言語識別子１０７の使用を必要にしない。したがって、発話１０６は、２つ以上の言語にまたがる音声のコードミキシングを備えている多言語発話を備えていることができる。本明細書で使用されるように、ニューラルオラクルサーチＮＯＳモデル３２０は、再スコアリング処理３００ａが言語識別子１０７に基づき選択する言語固有ニューラルオラクルサーチＮＯＳモデル３２０Ｓ（図３Ａ）か、または多言語ニューラルオラクルサーチＮＯＳモデル（図３Ｂ）か、のいずれかを備えていることができる。
【００４０】
図３Ａおよび図３Ｂを引き続き参照すると、ニューラルオラクルサーチＮＯＳモデル３２０は、ラベル履歴２４５を与えられた次のラベルＹ｡を予測する、事前モデルを備えている。すなわち、事前モデルは、以前に認識された単語、単語片、および／または書記素、に基づき次のラベルの事前スコアを予測する。ニューラルオラクルサーチＮＯＳモデル３２０の事前モデルは、１層あたり５１２ユニットである、２層の一方向ＬＳＴＭを備えていることができる。事前モデルは、ラベル付けされた音声トレーニングデータとクロスエントロピー損失とを使用することで、トレーニング（学習）する。さらに、ニューラルオラクルサーチＮＯＳモデル３２０は、ラベル履歴２４５と、第１パス３０１からの音響フレーム１１０のシーケンスと、をラベル同期方式で組み合わせることによって、事後スコアを予測する事後モデルを備えている。ニューラルオラクルサーチＮＯＳモデル３２０の事後モデルは、１層あたり５１２ユニットである２層の一方向ＬＳＴＭと、１層あたり１２８ユニットである２層のラベル同期アテンション（注意）メカニズムと、を備えていることができる。事後モデルは、ラベル履歴２４５と、音響フレーム１１０のシーケンスと、が与えられると、次のラベルＹ｡を予測するべく、ラベル付き音響トレーニングデータとクロスエントロピー損失とでトレーニング（学習）する。ニューラルオラクルサーチＮＯＳモデル３２０は、トークンレベルの事前スコアと、トークンレベルの事後スコアと、を合計することで非正規化尤度スコア３２５を生成する。このように、非正規化尤度スコア３２５は、以下のような和によって表わされるシーケンスレベルのスコアである。
【００４１】
【数１】

$$S_{\theta_1}(X|Y = y_{0:U}) = \sum_{i=0}^{U} \phi\,(X, y_{0:i-1}|Y_i = y)$$

$$\propto \sum_{i=0}^{U} \log P(Y_i = y|X, y_{0:i-1})$$

$$- \sum_{i=0}^{U} \log P(Y_i = y|y_{0:i-1})$$

$$= \log P(Y|X) - \log P(Y)$$

$$\propto \log P(X|Y) \qquad (1)$$

【００４２】

式１において、$S_{\theta 1}$は非正規化尤度スコア３２５を表わす。

再スコアラー３５０は、Ｎ個の候補仮説２０４のうちの各候補仮説２０４について、単体（スタンドアロン、単独）スコア２０５、言語モデルスコア３１５、および非正規化尤度スコア３２５、を受け取るとともに、各々の総合スコア３５５を生成する。特に、再スコアラー３５０は、単体スコア２０５、言語モデルスコア３１５、および非正規化尤度スコア３２５、の任意の組み合わせに基づき、各候補仮説２０４の総合スコア３５５を生成する。いくつかの例では、再スコアラー３５０は、単体（スタンドアロン、単独）スコア２０５、言語モデルスコア３１５、および非正規化尤度スコア３２５、を線形に合計することで、以下の式で表わされるシーケンスレベルの総合スコア３５５を決定する。

【００４３】

【数２】

$$\tilde{P}(\text{Oracle} = i|X, H_{1:N}) = \frac{\exp\left(Score(H_i,\ X)\right)}{\sum_j \exp\left(Score\left(H_j, X\right)\right)} \qquad (2)$$

$$Score(H_i, X) = \lambda_1 S_{\theta_1}(X|H_i) + \lambda_2 S_{\theta_2}(H_i) + S_{\theta_3}(i) \qquad (3)$$

【００４４】

式（３）において、$S_{\theta 1}$は、非正規化尤度スコア３２５を表している。$S_{\theta 2}$は外部言語モデルスコア３１５を表している。$S_{\theta 3}$は単体スコア２０５を表わす。トレーニング中に再スコアラー３５０のモデルパラメータを最適化するべく、再スコアリング処理３００は、事後スコアとシーケンスレベルのグラウンドトゥルース分布との間の、クロスエントロピーオブジェクトを使用する。いくつかの例では、トレーニング処理は、全（トータル）グラウンドトゥルース分布をグラウンドトゥルーストランスクリプションに割り当てる一方で、他の全ての候補仮説をゼロに割り当てる。他の例では、トレーニング処理は、最良の候補仮説（すなわち、グラウンドトゥルーストランスクリプション）未満の単語誤り率（ＷＥＲ）を有している全ての候補仮説にわたって一様に、全グラウンドトゥルース分布を割り当てる。さらに他の例では、トレーニング処理は、各候補仮説とグラウンド

トゥルーストランスクリプションとの間の負の編集距離に、ソフトマックス関数を適用する。

【００４５】

その後、再スコアラー３５０は、Ｎ個の候補仮説２０４の中から、最も高い総合スコア３５５を有している候補仮説２０４を、発話１０６の最終トランスクリプション１２０として選択する。図示の例では、候補仮説２０４には、０．９の総合スコア３５５を有している「プレイ＿ネクスト＿ソング」（次の歌を再生して）、０．３の総合スコア３５５を有している「ヘイ＿ネクスト＿ロング」（やあ次の長い）、および０．５の総合スコア３５５を有している「プレイ＿ネクスト＿ポン」（次のポン（ｐｏｎｇ）を再生して）、が含まれる。この例を続けると、再スコアラー３５０は、０．９という最も高い総合スコア３５５を有している「プレイ＿ネクスト＿ソング」（次の歌を再生して）（実線のボックスで示される）の候補仮説２０４を、トランスクリプション１２０（例えば、最終トランスクリプション１２０ｂ（図１））として選択する。注目すべきは、最も高い単体スコア２０５（すなわち、正しいトランスクリプションである可能性）を持つ候補仮説２０４は、正しい候補仮説２０４ではないということ。そして、最も高い総合スコア３５５を持つ候補仮説が、第２パス３０２からの正しいトランスクリプションであるということである。

【００４６】

図４は、自動音声認識のために多言語の再採点（再スコアリング）モデルを使用するコンピュータ実装方法４００の動作（操作、オペレーション）の、例示的な配置（アレンジメント、構成）のフローチャートである。動作４０２において、方法４００は、発話１０６に対応する音声データから抽出済みの音響フレーム１１０のシーケンスを受信する工程を備えている。動作４０４において、第１パス３０１中、方法４００は、音響フレーム１１０のシーケンスを処理して、多言語音声認識モデル（すなわち、自動音声認識ＡＳＲモデル）２００を使用することで、発話１０６に対するＮ個の候補仮説２０４、２０４ａ～２０４ｎを生成する工程を備えている。第２パス３０２中、Ｎ個の候補仮説２０４のうちの各候補仮説２０４について、方法４００は動作４０６～４１２を実行する。動作４０６において、方法４００は、ニューラルオラクルサーチＮＯＳモデル３２０を用いて、各々の非正規化尤度スコア３２５を生成する工程を備えている。ここで、ニューラルオラクルサーチＮＯＳモデル３２０は、音響フレーム１１０のシーケンスと、対応する候補仮説２０４と、に基づき非正規化尤度スコア３２５を生成する。動作４０８において、方法４００は、言語モデル３１０を用いて各々の外部言語モデルスコア３１５を生成する工程を備えている。動作４１０において、方法４００は、第１パス３０１中に生成済みの対応する候補仮説２０４の事前統計をモデル化する、単体スコア２０５を生成する工程を備えている。動作４１２において、方法４００は、非正規化尤度スコア３２５、外部言語モデルスコア３１５、および単体スコア２０５、に基づき候補仮説２５５の各々の総合スコア３５５を生成する工程を備えている。動作４１４において、方法４００は、Ｎ個の候補仮説２０４の中から最も高い（最高の）各々の総合スコア３５５を有している候補仮説２０４を、発話１０６の最終トランスクリプション１２０として選択する工程を備えている。

【００４７】

図５は、本書に記載されるシステムおよび方法を実施するべく使用され得る例示的なコンピューティング装置５００の概略図である。コンピューティング装置５００は、ラップトップ、デスクトップ、ワークステーション、パーソナルデジタルアシスタント、サーバ、ブレードサーバ、メインフレーム、および他の適切なコンピュータ、などの様々な形態のデジタルコンピュータを表わすことを意図している。ここに示された構成要素、それらの接続および関係、ならびにそれらの機能は、例示的なものであることのみを意図しているのであり、本書で説明および／または特許請求される発明の実施を制限することを意図していない。

【００４８】

コンピューティング装置５００は、プロセッサ５１０と、メモリ５２０と、ストレージ

デバイス５３０と、メモリ５２０および高速拡張ポート５５０に接続する高速インタフェース／コントローラ５４０と、および低速バス５７０およびストレージデバイス５３０に接続する低速インタフェース／コントローラ５６０と、を備えている。各構成要素５１０、５２０、５３０、５４０、５５０、および５６０、はさまざまなバスを使用して相互接続されており、共通のマザーボード上に、または適切な他の方法で、実装することができる。プロセッサ５１０は、高速インタフェース５４０に結合されたディスプレイ５８０などの外部入出力デバイスにグラフィカルユーザインタフェース（ＧＵＩ）のためのグラフィカル情報を表示するべく、メモリ５２０または記憶デバイス５３０に記憶された命令を備えている、コンピューティング装置５００内で実行するための命令を処理することができる。他の実施態様では、複数のプロセッサおよび／または複数のバスは、複数のメモリおよびメモリの種類とともに、適宜、使用されてもよい。また、複数のコンピューティング装置５００が接続されることで、各装置（デバイス）が必要な動作（操作）の一部を提供してもよい（例えば、サーババンク、ブレードサーバのグループ、またはマルチプロセッサシステムとして）。

【００４９】

メモリ５２０は、コンピューティング装置５００内で情報を非遷移的（非一時的、非一過性）に記憶する。メモリ５２０は、コンピュータ可読媒体、揮発性メモリユニット（複数可）、または不揮発性メモリユニット（複数可）、であってもよい。不揮発性メモリ５２０は、コンピューティング装置５００によって使用されるプログラム（例えば、命令のシーケンス）またはデータ（例えば、プログラム状態情報）を一時的または永続的に記憶するべく使用される物理的デバイスであってよい。不揮発性メモリの例としては、フラッシュメモリ、読み出し専用メモリ（ＲＯＭ）／プログラマブル読み出し専用メモリ（ＰＲＯＭ）／消去可能プログラマブル読み出し専用メモリ（ＥＰＲＯＭ）／電子消去可能プログラマブル読み出し専用メモリ（ＥＥＰＲＯＭ）（例えば、ブートプログラムなどのファームウェアに通常使用される）が挙げられるが、これらに限定されない。揮発性メモリの例としては、ランダムアクセスメモリ（ＲＡＭ）、ダイナミックランダムアクセスメモリ（ＤＲＡＭ）、スタティックランダムアクセスメモリ（ＳＲＡＭ）、相変化メモリ（ＰＣＭ）、およびディスクやテープ、などがあるが、これらに限定されるものではない。

【００５０】

記憶装置５３０は、コンピューティング装置５００に大容量記憶装置を提供することができる。いくつかの実施態様において、記憶装置５３０は、コンピュータ読み取り可能な媒体である。様々な異なる実装において、記憶装置５３０は、フロッピー（登録商標）ディスク装置、ハードディスク装置、光ディスク装置、またはテープ装置、フラッシュメモリまたは他の同様のソリッドステートメモリデバイス、またはストレージエリアネットワークまたは他の構成のデバイスを備えているデバイスのアレイ、であってもよい。追加の実施態様において、コンピュータプログラム製品は、情報キャリアに具体化される。コンピュータプログラム製品は、実行されると上述したような１つまたは複数の方法を実行する命令を備えている。情報キャリアは、メモリ５２０、記憶装置５３０、またはプロセッサ５１０上のメモリ、などのコンピュータ可読媒体または機械可読媒体である。

【００５１】

高速コントローラ５４０は、コンピューティング装置５００の帯域幅集約的な動作を管理しており、低速コントローラ５６０は、帯域幅集約的ではない（低帯域幅集約的な）動作を管理する。このような任務の割り当ては、例示的なものに過ぎない。一部の実装では、高速コントローラ５４０は、メモリ５２０、ディスプレイ５８０（例えば、グラフィックプロセッサまたはアクセラレータを介して）、および高速拡張ポート５５０に結合されているとともに、様々な拡張カード（図示せず）を受け入れることができる。いくつかの実装では、低速コントローラ５６０は、ストレージデバイス５３０および低速拡張ポート５９０に結合される。低速拡張ポート５９０は、様々な通信ポート（例えば、ＵＳＢ、ブルートゥース（登録商標）、イーサネット（登録商標）、ワイヤレスイーサネット（登録商標））を備えているとともに、キーボード、ポインティングデバイス、スキャナ、など

の１つまたは複数の入出力デバイスに、またはネットワークアダプタを介してスイッチやルータなどのネットワークデバイスに、結合される。

【００５２】

コンピューティング装置５００は、図示のように、多数の異なる形態で実装されてもよい。例えば、標準サーバ５００ａとして、またはそのようなサーバ５００ａのグループ内の複数倍（回）、ラップトップコンピュータ５００ｂとして、またはラックサーバシステム５００ｃの一部として、実装することができる。

【００５３】

本明細書で説明するシステムおよび技術の様々な実装は、デジタル電子回路および／または光回路、集積回路、特別に設計されたＡＳＩＣ（特定用途向け集積回路）、コンピュータハードウェア、ファームウェア、ソフトウェア、および／またはそれらの組み合わせ、で実現することができる。これらの様々な実装は、特殊目的であっても汎用目的であってもよく、記憶システム、少なくとも１つの入力デバイス、および少なくとも１つの出力デバイス、からデータおよび命令を受信したり、記憶システム、少なくとも１つの入力デバイス、および少なくとも１つの出力デバイス、にデータおよび命令を送信したり、するように結合された少なくとも１つのプログラマブルプロセッサを備えているプログラマブルシステム上で実行可能および／または解釈可能な１つまたは複数のコンピュータプログラムにおける実装を備えていることができる。

【００５４】

これらのコンピュータプログラム（プログラム、ソフトウェア、ソフトウェアアプリケーションまたはコードとしても知られる）は、プログラマブルプロセッサ用の機械命令を備えており、高レベルの手続き型および／またはオブジェクト指向プログラミング言語、および／またはアセンブリ／機械言語で実装することができる。本明細書で使用される場合、「機械可読媒体」および「コンピュータ可読媒体」という用語は、機械命令を機械可読信号として受信する機械可読媒体を含めて、機械命令および／またはデータをプログラマブルプロセッサに提供するべく使用される、任意のコンピュータプログラム製品、非一過性コンピュータ可読媒体、装置および／またはデバイス（例えば、磁気ディスク、光ディスク、メモリ、プログラマブルロジックデバイス（ＰＬＤ））を指す。「機械可読信号」という用語は、機械命令および／またはデータをプログラマブルプロセッサに提供するべく使用される、あらゆる信号を指す。

【００５５】

本明細書で説明する処理および論理フローは、データ処理ハードウェアとも呼ばれる１つまたは複数のプログラマブルプロセッサが、１つまたは複数のコンピュータプログラムを実行することで、入力データに対して動作するとともに出力を生成することによって機能を実行することで実行することができる。処理および論理フローは、特殊用途の論理回路、例えばＦＰＧＡ（フィールドプログラマブルゲートアレイ）またはＡＳＩＣ（特定用途向け集積回路）によっても実行できる。コンピュータプログラムの実行に適したプロセッサには、一例として、汎用および特殊用途のマイクロプロセッサが、およびあらゆる種類のデジタルコンピュータの任意の１つまたは複数のプロセッサが、含まれる。一般に、プロセッサは、読み取り専用メモリまたはランダムアクセスメモリ、あるいはその両方から命令とデータを受け取る。コンピュータの本質的な要素は、命令を実行するためのプロセッサと、命令やデータを格納するための１つまたは複数のメモリ装置と、である。一般に、コンピュータは、データを格納するための１つまたは複数の大容量記憶装置、例えば磁気ディスク、光磁気ディスク、光ディスク、などからもデータを受け取るか、それら記憶装置にデータを転送するか、あるいはその両方、を行なうようにそれら記憶装置に動作可能に結合されている。しかし、コンピュータがそのような装置を備えている必要はない。コンピュータプログラム命令およびデータを記憶するのに適したコンピュータ可読媒体には、あらゆる形態の不揮発性メモリ、媒体およびメモリデバイスが含まれ、例えば、半導体メモリデバイス、例えば、ＥＰＲＯＭ、ＥＥＰＲＯＭおよびフラッシュメモリデバイス；磁気ディスク、例えば、内蔵ハードディスクまたはリムーバブルディスク；光磁気デ

ィスク；およびＣＤＲＯＭおよびＤＶＤ－ＲＯＭディスクが含まれる。プロセッサとメモリは、特殊用途の論理回路によって補足されるか、または特殊用途の論理回路に組み込まれる。

【００５６】

ユーザとの相互作用を提供するべく、本開示の１つまたは複数の態様は、ユーザに情報を表示するためのディスプレイデバイス、例えばＣＲＴ（陰極線管）、ＬＣＤ（液晶ディスプレイ）モニタ、またはタッチスクリーンと、任意選択で、ユーザがコンピュータに入力を提供することができるキーボードおよびポインティングデバイス、例えばマウスまたはトラックボールと、を有しているコンピュータ上で実施することができる。他の種類のデバイスも同様に、ユーザとの対話を提供するべく使用することができる。例えば、ユーザに提供されるフィードバックは、視覚フィードバック、聴覚フィードバック、または触覚フィードバック、などの任意の形式の感覚フィードバックとすることができる。ユーザからの入力は、音響入力、音声入力、または触覚入力、を備えている任意の形式で受信することができる。さらに、コンピュータは、例えば、ウェブブラウザから受信した要求に応じて、ユーザのクライアントデバイス上のウェブブラウザにウェブページを送信することによって、ユーザが使用するデバイスにドキュメントを送信したり、デバイスからドキュメントを受信したり、することによってユーザとで対話することができる。

【００５７】

多くの実施態様を説明してきた。それにもかかわらず、本開示の精神および範囲から逸脱することなく、様々な変更がなされ得ることが理解されるであろう。従って、他の実施態様も以下の特許請求の範囲に含まれる。

【図面】

【図１】

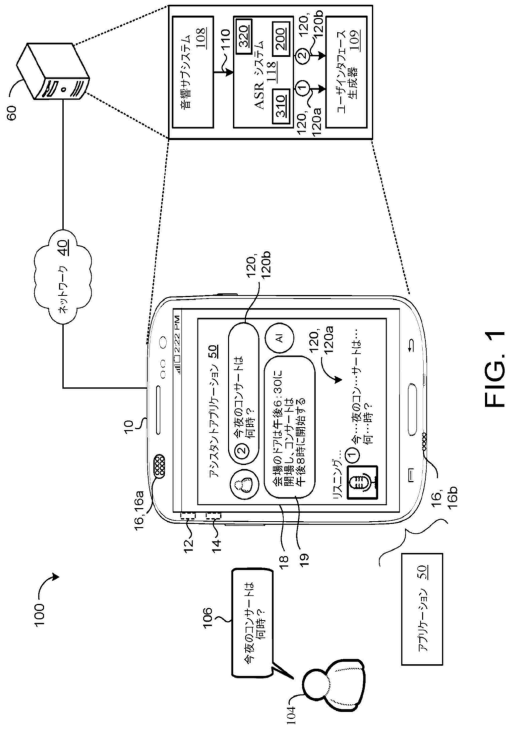

FIG. 1

【図２】

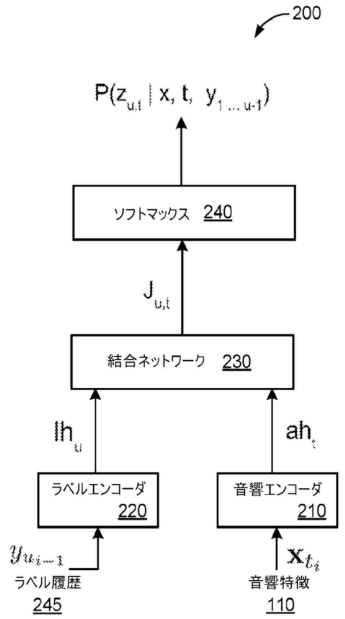

FIG. 2

【図３Ａ】

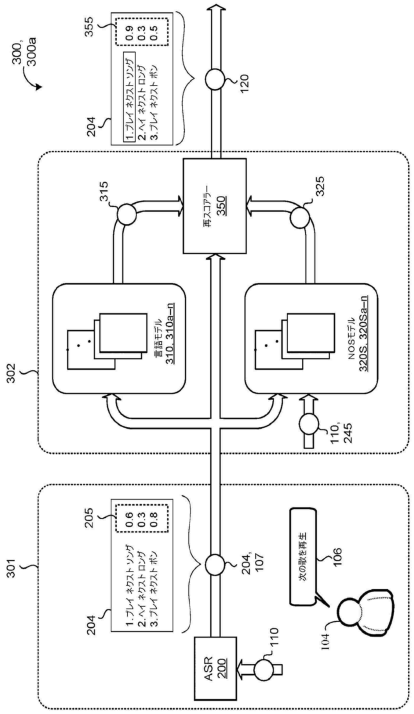

FIG. 3A

【図３Ｂ】

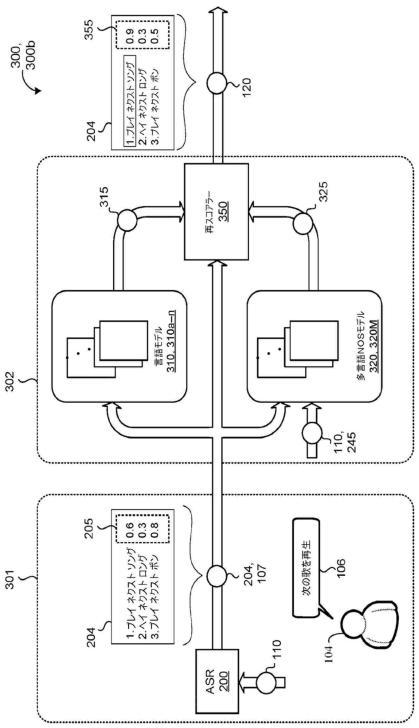

FIG. 3A

【図４】

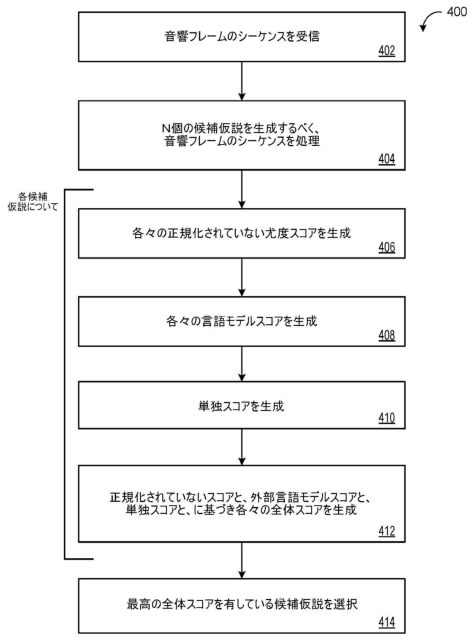

FIG. 4

【図５】

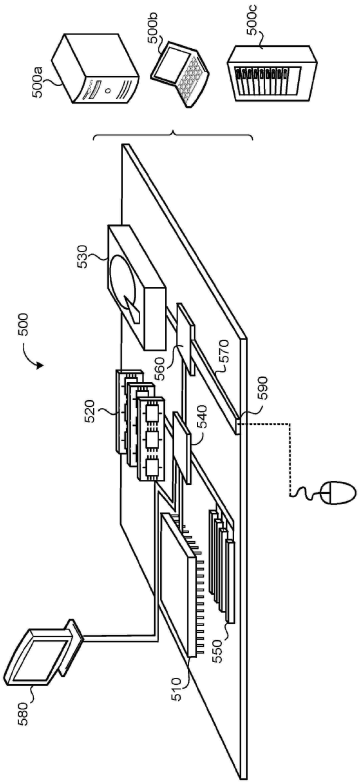

FIG. 5

フロントページの続き

(72)発明者　チェン、トンジョウ
アメリカ合衆国 ９４０４３ カリフォルニア州 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　ヴァリアニ、エフサン
アメリカ合衆国 ９４０４３ カリフォルニア州 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　ラマバドラン、ブバナ
アメリカ合衆国 １０５４９ ニューヨーク州 マウント キスコ ビクトリア ドライブ ４００２
(72)発明者　ハガニ、パリサ
アメリカ合衆国 ９４０４３ カリフォルニア州 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　モレノ メンヒバル、ペドロ ジェイ.
アメリカ合衆国 ９４０４３ カリフォルニア州 マウンテン ビュー アンフィシアター パークウェイ １６００