(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7092953号
(P7092953)

(45)発行日　令和4年6月28日(2022.6.28)

(24)登録日　令和4年6月20日(2022.6.20)

(51)国際特許分類

Ｇ１０Ｌ　15/32　(2013.01)	Ｇ１０Ｌ	15/32	２１０Ｅ
Ｇ１０Ｌ　15/16　(2006.01)	Ｇ１０Ｌ	15/16	
Ｇ１０Ｌ　15/08　(2006.01)	Ｇ１０Ｌ	15/08	２００Ｚ

ＦＩ

請求項の数　20　（全20頁）

(21)出願番号　特願2021-564950(P2021-564950)
(86)(22)出願日　令和2年4月28日(2020.4.28)
(65)公表番号　特表2022-523883(P2022-523883A)
(43)公表日　令和4年4月26日(2022.4.26)
(86)国際出願番号　PCT/US2020/030321
(87)国際公開番号　WO2020/226948
(87)国際公開日　令和2年11月12日(2020.11.12)
審査請求日　令和4年1月24日(2022.1.24)
(31)優先権主張番号　62/842,571
(32)優先日　令和1年5月3日(2019.5.3)
(33)優先権主張国・地域又は機関
米国(US)
特許法第３０条第２項適用　令和１年６月２１日にウェブサイトのアドレス　ｈｔｔｐｓ：／／ａｒｘｉｖ．ｏｒ
最終頁に続く

(73)特許権者　502208397
グーグル エルエルシー
Ｇｏｏｇｌｅ ＬＬＣ
アメリカ合衆国 カリフォルニア州 ９４０４３ マウンテン ビュー アンフィシアター パークウェイ １６００
１６００ Ａｍｐｈｉｔｈｅａｔｒｅ Ｐａｒｋｗａｙ ９４０４３ Ｍｏｕｎｔａｉｎ Ｖｉｅｗ， ＣＡ Ｕ．Ｓ．Ａ．
(74)代理人　100142907
弁理士　本田 淳
(72)発明者　フー、キー
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】　エンドツーエンドモデルによる多言語音声認識のための音素に基づく文脈解析

(57)【特許請求の範囲】
【請求項１】
データ処理ハードウェア（６１０）において、第１言語のネイティブスピーカ（１１０）によって話される発話（１０６）を符号化する音声データを受け取る工程と、
前記データ処理ハードウェア（６１０）において、前記第１言語とは異なる第２言語の１つまたは複数の用語を備えているバイアス用語リスト（１０５）を受け取る工程と、
前記データ処理ハードウェア（６１０）において、音声認識モデル（２００）を用いて、前記音声データから得られる音響特徴（１０４）を処理して、前記第１言語の語句と、対応する音素シーケンスとの両方に対する音声認識スコアを生成する工程と、
前記データ処理ハードウェア（６１０）によって、前記バイアス用語リスト（１０５）内の前記１つまたは複数の用語に基づき、前記音素シーケンスに対する前記音声認識スコアを再スコアリングする工程と、
前記データ処理ハードウェア（６１０）によって、前記語句に対する前記音声認識スコアと、前記音素シーケンスに対する再スコアリングされた音声認識スコアとを用いて、復号グラフ（４００）を実行して、前記発話（１０６）に対する転写（１１６）を生成する工程と、
を備えている方法（５００）。
【請求項２】
前記音素シーケンスに対する前記音声認識スコアを再スコアリングする工程は、バイアス有限状態変換器（ＦＳＴ）を使用して、前記音素シーケンスに対する前記音声認識スコア

を再スコアリングする工程を備えている、
請求項１に記載の方法（５００）。
【請求項３】
前記方法はさらに、
前記データ処理ハードウェア（６１０）によって、前記バイアス用語リスト（１０５）の各用語を、前記第２言語の対応する音素シーケンスにトークン化する工程と、
前記データ処理ハードウェア（６１０）によって、前記第２言語における各対応する音素シーケンスを、前記第１言語における対応する音素シーケンスに写像する工程と、
前記データ処理ハードウェア（６１０）によって、前記第１言語における各対応する音素シーケンスに基づき、前記バイアス有限状態変換器（３００）を生成する工程と、
を備えている、請求項２に記載の方法（５００）。
【請求項４】
前記音声認識モデル（２００）は、エンドツーエンド語句－音素モデル（２００）を備えている、
請求項１～３のいずれか一項に記載の方法（５００）。
【請求項５】
前記エンドツーエンド語句－音素モデル（２００）は、リカレントニューラルネットワーク－変換器（ＲＮＮ－Ｔ）を備えている、
請求項４に記載の方法（５００）。
【請求項６】
前記復号グラフ（４００）の実行中に、前記復号グラフ（４００）は、前記バイアス用語リスト（１０５）内の前記１つまたは複数の用語のいずれかを有利にするように、前記転写（１１６）をバイアスする、
請求項１～５のいずれか一項に記載の方法（５００）。
【請求項７】
前記音声認識モデル（２００）は、前記第１言語のみの学習発話で学習される、
請求項１～６のいずれか一項に記載の方法（５００）。
【請求項８】
前記バイアス用語リスト（１０５）内の用語のいずれも、前記音声認識モデル（２００）を学習するために使用されなかった、
請求項１～７のいずれか一項に記載の方法（５００）。
【請求項９】
前記データ処理ハードウェア（６１０）および前記音声認識モデル（２００）は、ユーザ装置（１０２）上に存在する、
請求項１～８のいずれか一項に記載の方法（５００）。
【請求項１０】
前記データ処理ハードウェア（６１０）および前記音声認識モデル（２００）は、リモート計算装置（２０１）上に存在し、
前記発話（１０６）を符号化する前記音声データを受け取る工程は、前記リモート計算装置（２０１）に通信しているユーザ装置（１０２）から、前記発話（１０６）を符号化する前記音声データを受け取る工程を備えている、
請求項１～９のいずれか一項に記載の方法（５００）。
【請求項１１】
データ処理ハードウェア（６１０）と、
前記データ処理ハードウェア（６１０）に通信するメモリハードウェア（６２０）であって、前記メモリハードウェア（６２０）は、前記データ処理ハードウェア（６１０）上で実行されると前記データ処理ハードウェア（６１０）に、以下を備えている動作を実行させる命令を格納する、前記メモリハードウェア（６２０）と、
を備えているシステム（１００）であって、前記動作は、
第１言語のネイティブスピーカ（１１０）によって話される発話（１０６）を符号化する

音声データを受け取る工程と、
前記第１言語とは異なる第２言語による１つまたは複数の用語を備えているバイアス用語リスト（１０５）を受け取る工程と、
音声認識モデル（２００）を用いて、前記音声データから得られる音響特徴（１０４）を処理して、前記第１言語の語句と、対応する音素シーケンスとの両方に対する音声認識スコアを生成する工程と、
前記バイアス用語リスト（１０５）の前記１つまたは複数の用語に基づき、前記音素シーケンスに対する前記音声認識スコアを再スコアリングする工程と、
前記語句に対する前記音声認識スコアと、前記音素シーケンスに対する再スコアリングされた音声認識スコアとを用いて、復号グラフ（４００）を実行して、前記発話（１０６）に対する転写（１１６）を生成する工程と、
を備えている、システム（１００）。
【請求項１２】
前記音素シーケンスに対する前記音声認識スコアを再スコアリングする工程は、バイアス有限状態変換器（ＦＳＴ）を使用して、前記音素シーケンスに対する前記音声認識スコアを再スコアリングする工程を備えている、
請求項１１に記載のシステム（１００）。
【請求項１３】
前記動作はさらに、
前記バイアス用語リスト（１０５）の各用語を、前記第２言語の対応する音素シーケンスにトークン化する工程と、
前記第２言語における各対応する音素シーケンスを、前記第１言語における対応する音素シーケンスに写像する工程と、
前記第１言語における各対応する音素シーケンスに基づき、前記バイアス有限状態変換器（３００）を生成する工程と、
を備えている、請求項１２に記載のシステム（１００）。
【請求項１４】
前記音声認識モデル（２００）は、エンドツーエンド語句－音素モデル（２００）を備えている、
請求項１１～１３のいずれか一項に記載のシステム（１００）。
【請求項１５】
前記エンドツーエンド語句－音素モデル（２００）は、リカレントニューラルネットワーク－変換器（ＲＮＮ－Ｔ）を備えている、
請求項１４に記載のシステム（１００）。
【請求項１６】
前記復号グラフ（４００）の実行中に、前記復号グラフ（４００）は、前記バイアス用語リスト（１０５）内の前記１つまたは複数の用語のいずれかを有利にするように、前記転写（１１６）をバイアスする、
請求項１１～１５のいずれか一項に記載のシステム（１００）。
【請求項１７】
前記音声認識モデル（２００）は、前記第１言語のみの学習発話で学習される、
請求項１１～１６のいずれか一項に記載のシステム（１００）。
【請求項１８】
前記バイアス用語リスト（１０５）内の用語のいずれも、前記音声認識モデル（２００）を学習するために使用されなかった、
請求項１１～１７のいずれか一項に記載のシステム（１００）。
【請求項１９】
前記データ処理ハードウェア（６１０）および前記音声認識モデル（２００）は、ユーザ装置（１０２）上に存在する、
請求項１１～１８のいずれか一項に記載のシステム（１００）。

【請求項２０】
前記データ処理ハードウェア（６１０）および前記音声認識モデル（２００）は、リモート計算装置（２０１）上に存在し、
前記発話（１０６）を符号化する前記音声データを受け取る工程は、前記リモート計算装置（２０１）に通信しているユーザ装置（１０２）から、前記発話（１０６）を符号化する前記音声データを受け取る工程を備えている、
請求項１１～１９のいずれか一項に記載のシステム（１００）。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、エンドツーエンドモデルにおける多言語（クロスリンガル）音声認識のための音素（ｐｈｏｎｅｍｅ）ベースのコンテキスト化（文脈解析）に関する。
【背景技術】
【０００２】
音声の文脈（コンテキスト）を認識することは、自動音声認識（ＡＳＲ）システムの目標である。しかし、人が話す言葉は多種多様であり、アクセントや発音にも違いがあるので、音声の文脈を認識することは困難である。多くの場合、人が話す単語やフレーズの種類（タイプ）は、その人が置かれている文脈に応じて変化する。
【０００３】
文脈的（コンテクスチュアル）自動音声認識ＡＳＲは音声認識を、ユーザ自身のプレイリスト、連絡先、地理的な地名など、与えられた文脈（コンテキスト）に偏らせる（バイアスする）。文脈情報には、通常、認識すべき関連フレーズのリストが含まれており、このリストには、珍しいフレーズや、学習（トレーニング）ではあまり見られない外国語が含まれていることが多い。文脈バイアスを行うべく、従来の自動音声認識ＡＳＲシステムでは、文脈情報をｎ－ｇｒａｍ重み付き有限状態変換器（ＷＦＳＴ：ｗｅｉｇｈｔｅｄ Ｆｉｎｉｔｅ Ｓｔａｔｅ Ｔｒａｎｓｄｕｃｅｒ）を用いて、独立した文脈言語モデル（ＬＭ：Ｌａｎｇｕａｇｅ Ｍｏｄｅｌ）でモデル化し、その独立した文脈言語モデルＬＭをベースライン言語モデルＬＭと合成して、オンザフライ（ＯＴＦ）再スコアリングを行うことがある。
【先行技術文献】
【特許文献】
【０００４】
【文献】米国特許出願公開第２０１６／１０４４８２号明細書
【発明の概要】
【発明が解決しようとする課題】
【０００５】
近年、エンドツーエンド（Ｅ２Ｅ）モデルが自動音声認識ＡＳＲに大きな期待を寄せており、従来のオンデバイスモデルと比較して、ワードエラーレート（ＷＥＲ）やレイテンシの指標（メトリックス）が改善されている。これらのＥ２Ｅモデルは、音響モデル（ＡＭ）、発音モデル（ＰＭ：Ｐｒｏｎｕｎｃｉａｔｉｏｎ Ｍｏｄｅｌ）、および言語モデルＬＭを単一のネットワークに折り畳んで、音声とテキストの写像（スピーチツーテキストマッピング）を直接学習するものであり、音響モデルＡＭ、発音モデルＰＭ、および言語モデルＬＭを個別に有する従来のＡＳＲシステムと比較して、競争力のある結果を示している。代表的なＥ２Ｅモデルには、単語ベースのＣＴＣ（Ｃｏｎｎｅｃｔｉｏｎｉｓｔ ｔｅｍｐｏｒａｌ Ｃｌａｓｓｉｆｉｃａｔｉｏｎ）モデルと、ＲＮＮ－Ｔ（リカレントニューラルネットワークトランスデューサ）モデルと、ＬＡＳ（Ｌｉｓｔｅｎ，Ａｔｔｅｎｄ，およびＳｐｅｌｌ）などの注意ベースモデル（アテンションベースモデル）とがある。Ｅ２Ｅモデルは、ビーム検索復号時（ビームサーチデコーディング時）に限られた数の認識候補を保持しているので、文脈的自動音声認識ＡＳＲはＥ２Ｅモデルにとって困難である。
【課題を解決するための手段】

【０００６】
本開示の一態様は、バイアス用語（ｂｉａｓｉｎｇ ｔｅｒｍ）リストに存在する用語に音声認識結果をバイアスする（偏らせる）方法を提供する。この方法は、データ処理ハードウェアにおいて、第１言語のネイティブスピーカによって話される発話を符号化（エンコーディング）する音声データ（オーディオデータ）を受け取る工程と、データ処理ハードウェアにおいて、第１言語とは異なる第２言語の１つまたは複数の用語（ｔｅｒｍｓ）を備えているバイアス用語リストを受け取る工程とを備えている。本方法は、データ処理ハードウェアによって、音声認識モデルを用いて、音声データから得られた音響特徴を処理して、第１言語における語句（ワードピース（Ｗｏｒｄｐｉｅｃｅ））と、対応する音素シーケンス（音素列）との両方に対する音声認識スコアを生成する工程も備えている。また、本方法は、データ処理ハードウェアによって、バイアス用語リスト内の１つまたは複数の用語に基づき、音素シーケンスに対する音声認識スコアを再スコアリングする工程を備えている。本方法はまた、データ処理ハードウェアによって、語句に対する音声認識スコアと、音素シーケンスに対する再スコアリングされた音声認識スコアとを用いて、発話に対する転写（トランスクリプション）を生成するための復号グラフ（デコーディンググラフ）を実行する工程を備えている。
【０００７】
本開示の実装は、以下のオプション機能のうちの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、音素シーケンスに対する音声認識スコアを再スコアリングする工程は、バイアスのかかった有限状態変換器（ｂａｉａｓｉｎｇ Ｆｉｎｉｔｅ Ｓｔａｔｅ Ｔｒａｎｓｄｕｃｅｒ。バイアスＦＳＴ）を使用して、音素シーケンスに対する音声認識スコアを再スコアリングする工程を備えている。これらの実装では、本方法は、データ処理ハードウェアによって、バイアス用語リスト内の各用語を、第２言語における対応する音素シーケンスにトークン化する工程と、データ処理ハードウェアによって、第２言語における各対応する音素シーケンスを、第１言語における対応する音素シーケンスに写像する工程と、データ処理ハードウェアによって、第１言語における各対応する音素シーケンスに基づき、バイアス有限状態変換器ＦＳＴを生成する工程と、を備えていることもできる。
【０００８】
いくつかの例では、音声認識モデルは、エンドツーエンドの語句－音素モデルを備えている。特定の例では、エンドツーエンドの語句－音素モデルは、リカレントニューラルネットワーク－変換器（ＲＮＮ－Ｔ）を備えている。
【０００９】
いくつかの実装では、復号グラフの実行中に、復号グラフは、バイアス用語リスト内の１つまたは複数の用語のいずれかを有利にするように転写（トランスクリプション）をバイアスする。音声認識モデルは、第１言語のみの学習発話で学習されてもよい。さらに、バイアス用語リスト内のいずれの用語も、音声認識モデルの学習に使用されなくてもよい。
【００１０】
データ処理ハードウェアおよび音声認識モデルは、ユーザ装置上、またはユーザ装置に通信するリモート計算装置上に存在してもよい。データ処理ハードウェアおよび音声認識モデルがリモート計算装置上に存在する場合、発話を符号化する音声データを受け取る工程は、ユーザ装置から、発話を符号化する音声データを受け取る工程を備えてもよい。
【００１１】
本開示の別の態様は、バイアス用語リストに存在する用語に音声認識結果を偏らせるシステムを提供する。このシステムは、データ処理ハードウェアと、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、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を実行させる命令を格納するメモリハードウェアとを備えている。この動作は、第１言語のネイティブスピーカによって話された発話を符号化する音声データを受け取る工程と、第１言語とは異なる第２言語の１つまたは複数の用語を備えているバイアス用語リストを受け取る工程と、音声認識モデルを使用して、音声データから

得られた音響特徴を処理して、第１言語の語句と、対応する音素シーケンスとの両方に対する音声認識スコアを生成する工程とを備えている。また、動作は、バイアス用語リスト内の１つまたは複数の用語に基づき、音素シーケンスに対する音声認識スコアを再スコアリングする工程と、語句に対する音声認識スコアと、音素シーケンスに対する再スコアリングされた音声認識スコアとを使用して、復号グラフを実行して、発話に対する転写を生成する工程とを備えている。

【００１２】

この態様は、以下のオプション機能の１つまたは複数を備えていることができる。いくつかの実装において、音素シーケンスに対する音声認識スコアを再スコアリングする工程は、バイアスのかかった有限状態変換器（ＦＳＴ）を使用して、音素シーケンスに対する音声認識スコアを再スコアリングする工程を備えている。これらの実装では、動作は、バイアス用語リスト内の各用語を第２言語の対応する音素シーケンスにトークン化する工程と、第２言語の各対応する音素シーケンスを第１言語の対応する音素シーケンスに写像する工程と、第１言語の各対応する音素シーケンスに基づきバイアス有限状態変換器ＦＳＴを生成する工程と、を備えていることもできる。

【００１３】

いくつかの例では、音声認識モデルは、エンドツーエンドの語句－音素モデルを備えている。特定の例では、エンドツーエンドの語句－音素モデルは、リカレントニューラルネットワーク－変換器（ＲＮＮ－Ｔ）を備えている。

【００１４】

いくつかの実装では、復号グラフの実行中に、復号グラフは、バイアス用語リスト内の１つまたは複数の用語のいずれかに有利になるように転写（トランスクリプション）をバイアスする。音声認識モデルは、第１言語のみの学習発話で学習されてもよい。さらに、バイアス用語リスト内のいずれの用語も、音声認識モデルの学習に使用されなくてもよい。

【００１５】

データ処理ハードウェアおよび音声認識モデルは、ユーザ装置上、またはユーザ装置に通信するリモート計算装置上に存在してもよい。データ処理ハードウェアおよび音声認識モデルがリモート計算装置上に存在する場合、発話を符号化する音声データを受け取る工程は、ユーザ装置から、発話を符号化する音声データを受け取る工程を備えてもよい。

【００１６】

本開示の１つまたは複数の実装の詳細は、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】

【００１７】

【図１】バイアス用語リストに存在する用語に向けて音声認識結果をバイアスする音声認識モデルを備えている、自動音声認識システムの例を示す概略図。

【図２】図１の音声認識モデルのアーキテクチャの一例を示す概略図。

【図３】例示的なバイアス有限状態変換器の模式図。

【図４】語句と、対応する音素シーケンスとに基づく、復号グラフの例の概略図。

【図５】バイアス用語リストに存在する用語に向けて音声認識結果をバイアスする方法のための動作の、例示的な配置のフローチャート。

【図６】本明細書に記載されたシステムおよび方法を実施するべく使用することができる、例示的な計算装置の概略図。

【発明を実施するための形態】

【００１８】

様々な図面における同様の参照記号は、同様の要素を示す。

本明細書の実装は、他の動作の中でも、外国語音素セットを自動音声認識ＡＳＲモデルの言語（例えば、アメリカ英語）の音素セットに写像（マッピング）して、音素レベルでバイアスをかける有限状態変換器（ＦＳＴ）において外国語のモデリングを可能にすること

で、外国語を認識する文脈的（コンテクチュアル）自動音声認識（ＡＳＲ）モデルを強化することに向けられている。さらなる実装は、自動音声認識ＡＳＲモデルが、モデリング空間における自動音声認識ＡＳＲモデルの言語（例えば、アメリカ英語）のための語句（ワードピース）および音素を備えている語句（ワードピース）－音素モデルを組み込むことに向けられている。例として、文脈的自動音声認識ＡＳＲモデルは、語句－音素モデルおよび文脈的バイアス有限状態変換器ＦＳＴを使用して音声発話（スポークン発話）を復号（デコード）し、発話の転写を文脈的に１つまたは複数の外国語に偏らせるように構成される。たとえば、アメリカ英語を話す人が、クレテイユ（Ｃｒｅｔｅｉｌ。Ｃｒｅのｅの上にアクソンテグュが付されている）という単語がフランス語である、「クレテイユまでの道順」（Ｄｉｒｅｃｔｉｏｎｓ ｔｏ Ｃｒｅｔｅｉｌ）という発話をすると、文脈的自動音声認識ＡＳＲモデルは、アメリカ英語以外の言語の単語で学習されていないにもかかわらず、語句－音素モデルと文脈的バイアス有限状態変換器ＦＳＴを利用して、外国語であるクレテイユ（Ｃｒｅｔｅｉｌ）を認識するように転写を偏らせることができる。この例では、外国語のクレテイユ（Ｃｒｅｔｅｉｌ）は、現在の文脈に基づきバイアスをかけた単語リストに含まれる複数のフランス語のうちの１つである可能性がある。例えば、ユーザが現在フランスにいて車を運転している場合、現在の文脈（コンテキスト）は、フランスの都市名／地域名が関連していることを示している可能性があり、したがって、文脈的（コンテキストに基づく）自動音声認識ＡＳＲモデルは、これらのフランスの都市名／地域名に偏っている（バイアスしている）可能性がある。

【００１９】

図１を参照すると、いくつかの実装では、強化（エンハンスト）された自動音声認識ＡＳＲシステム１００は、外国語の単語（ワード）を認識するように強化されている。示された例では、自動音声認識ＡＳＲシステム１００は、ユーザ１１０のユーザ装置１０２上、および／または、ユーザ装置に通信するリモート計算装置（リモート計算装置）２０１（例えば、クラウド計算環境で実行される分散システムの１つまたは複数のサーバ）上に存在する。ユーザ装置１０２は、モバイル計算装置（例えば、スマートフォン）として描かれているが、ユーザ装置１０２は、限定されないが、タブレットデバイス、ラップトップ／デスクトップコンピュータ、ウェアラブルデバイス、デジタルアシスタントデバイス、スマートスピーカ／ディスプレイ、スマートアプライアンス、自動車インフォテイメントシステム、またはＩｏＴ（インターネットオブシングス）デバイスなどの任意のタイプの計算装置に対応してもよい。

【００２０】

ユーザ装置１０２は音声サブシステム１０３を備えており、音声サブシステム１０３は、ユーザ１１０によって話された発話１０６を受け取り（例えば、ユーザ装置１０２は、話された発話１０６を記録するための１つまたは複数のマイクロフォンを備えてもよい）、発話１０６を、自動音声認識ＡＳＲシステム１００によって処理可能なパラメータ化された入力音響フレーム１０４に関連する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。示されている例では、ユーザは「クレテイユまでの道順」（Ｄｉｒｅｃｔｉｏｎｓ ｔｏ Ｃｒｅｔｅｉｌ）というフレーズに対するそれぞれの発話１０６を話し、音声サブシステム１０３は発話１０６を、自動音声認識ＡＳＲシステム１００に入力するための対応する音響フレーム１０４に変換する。例えば、音響フレーム１０４は、短い、例えば２５ｍｓのウィンドウで計算され、数ミリ秒、例えば１０ミリ秒ごとにシフトされた、それぞれが８０次元のｌｏｇ－Ｍｅｌ特徴を備えている一連（シリーズ）のパラメータ化された入力音響フレームであってもよい。

【００２１】

その後、自動音声認識ＡＳＲシステム１００は、入力として、発話１０６に対応する音響フレーム１０４を受け取り、出力として、発話１０６に対応する転写（トランスクリプション。例えば、認識結果／認識仮説）１１６を生成／予測する。図示の例では、ユーザ装置１０２および／またはリモート計算装置２０１は、ユーザ装置１０２のユーザインタフェース１３６において、発話１０６の転写１１６の表現（レプレゼンテーション）をユー

ザ１１０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システム１０７も実行する。いくつかの例では、ユーザインタフェース１３６は、ユーザ装置１０２に通信しているスクリーン上に表示されてもよい。

【００２２】

いくつかの構成では、自動音声認識ＡＳＲシステム１００から出力された転写１１６は、例えば、ユーザ装置１０２またはリモート計算装置２０１上で実行される自然言語理解（ＮＬＵ）モジュールによって、ユーザコマンドを実行するべく処理される。さらに、または代替として、音声合成（テキストツースピーチ）システム（例えば、ユーザ装置１０２またはリモート計算装置２０１の任意の組み合わせ上で実行される）は、別のデバイスによる可聴出力のために、転写を合成音声に変換してもよい。例えば、元の発話１０６は、ユーザ１１０が友人に送信しているメッセージに対応していてもよく、その場合、転写１１６は、元の発話１０６で伝えられたメッセージを聞くべく、友人への可聴出力のために合成音声に変換される。

【００２３】

強化された自動音声認識ＡＳＲシステム１００は、バイアス構成要素１１５と、語句－音素モデル２００およびバイアス有限状態変換器ＦＳＴ３００を有する音声認識装置１５０と、学習構成要素１１４とを備えている。バイアス構成要素１１５は、バイアス有限状態変換器ＦＳＴ３００を生成するように構成され、学習構成要素１１４は、音素レベルで外国語を再スコアリングすることで文脈的バイアスを実行するように、語句－音素モデル２００およびバイアス有限状態変換器ＦＳＴ３００を学習するように構成される。明らかになるように、音声認識装置１５０は、学習された語句－音素モデル２００およびバイアス有限状態変換器ＦＳＴ３００を使用して、外国語の単語に向かってバイアスをかけることで、文脈的な音声認識を実行する。

【００２４】

学習構成要素１１４は、単一の言語、例えば、アメリカ英語のテキストのコーパスを有する辞書（レキシコン、語彙集）１１７と、頻度チェッカ１１８と、モデル学習器１２０とを備えている。頻度（フリーケンシ）チェッカ１１８は、コーパスのテキストの中での単一言語の用語の相対的な頻度を決定するように構成され、モデル学習器１２０は、テキストコーパスの用語の語句と音素の両方に基づき語句－音素モデル２００を学習し、モデリング空間に語句と音素の両方を含めるように構成される。いくつかの例では、語句－音素モデル２００は、単一の言語のみ、例えば、アメリカ英語のみからの語句－音素セットを含む一方で、他の言語からの語句－音素セットを除外した学習データを用いて、モデル学習器１２０によってエンドツーエンドで学習される。モデル学習器１２０は、単語頻度ベースのサンプリング戦略を採用して、辞書１１７を用いて、稀な単語をターゲットシーケンスの音素にランダムにトークン化してもよい。段階（ステージ）Ａにおいて、学習構成要素１１４は、辞書１１７からのテキストを使用して、語句－音素モデル２００を学習する。

【００２５】

いくつかの例では、辞書１１７は約５０万個の単語を含み、その頻度は音素シーケンスを使用するタイミングを決定するべく使用される。辞書１１７は、学習データからの単語とその頻度を含み、同音異義語（ホモフォン。例えば、「ｆｌｏｗｅｒ」（花）と「ｆｌｏｕｒ」（小麦粉））、同形異義語（ホモグラフ。例えば、動詞または形容詞としての「ｌｉｖｅ」（生きる、生の））、および発音変種（プロナンシエイションバリエント。例えば、「ｅｉｔｈｅｒ」（イーザーまたはアイザー））を除去してトリミングされる。このように、辞書１１７には、綴りから発音へまたはその逆の場合に、曖昧さがない項目のみが含まれている。

【００２６】

いくつかの実装では、モデル学習器（トレーナ）１２０は、学習入力発話を２５ｍｓのフレームに分割し、１０ｍｓのレート（速度）で窓を開けシフトする。各フレームで８０次元のｌｏｇ－Ｍｅｌ特徴が抽出され、現在のフレームと左隣の２つのフレームが連結され

て２４０次元のｌｏｇ－Ｍｅｌ特徴が生成される。これらの特徴は、その後、３０ｍｓのレートでダウンサンプリングされる。

【００２７】

いくつかの実装では、語句－音素モデル２００は、シーケンスツーシーケンスモデルを備えている。いくつかの例では、語句－音素モデル２００は、ＲＮＮ－Ｔ（リカレントニューラルネットワーク－トランスデューサ）シーケンスツーシーケンスモデルアーキテクチャを備えている。他の例では、語句－音素モデル２００は、リッスン、アテンド、スペルのシーケンスツーシーケンスモデルアーキテクチャを備えている。

【００２８】

語句－音素モデル２００は、学習において少数の語句を選択的に音素に分解することができる点で、語句のみのモデルとは異なる。このモデルの出力は、記号セット（シンボルセット）が語句記号と音素記号との組合わせである、単一のソフトマックスである。単語の音素シーケンスを得るためには、発音辞書（レキシコン）が用いられる。音素は、希少な単語の認識に強みを発揮するので、これらの単語はより頻繁に音素として提示される。ターゲット文では、ｉ番目の単語が確率ｐ（ｉ）＝p_0・ｍｉｎ（Ｔ／（ｃ（ｉ）），１．０）でランダムに音素として提示される。ここでp_0とＴは定数であり、ｃ(ｉ)は、学習コーパス全体での単語の出現回数を表す整数である。出現回数がＴ回以下の単語は、確率p_0で音素として提示される。Ｔ回よりも多く出現する単語については、頻度が高いほど音素として提示されないことになる。いくつかの例では、Ｔは１０に等しく、p_0は０．５に等しくなっているが、他の例では異なる値を選択することができる。なお、単語と音素のどちらを使用するかの決定は、勾配のイテレーションごとにランダムに行われるので、ある文は、異なるエポックで異なるターゲットシーケンスを持つ可能性がある。いくつかの実装では、音素は文脈に依存しない音素である。

【００２９】

図２を参照すると、語句－音素モデル２００は、インタラクティブアプリケーションに関連付けられたレイテンシ制約に準拠したエンドツーエンド（Ｅ２Ｅ）のＲＮＮ－Ｔモデル２００を備えていることができる。ＲＮＮ－Ｔモデル２００は、小さな計算フットプリントを提供し、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、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（例えば、リモートサーバとの通信は必要とされない）。ＲＮＮ－Ｔモデル２００は、符号化器ネットワーク（エンコーダネットワーク）２１０と、予測ネットワーク２２０と、結合ネットワーク（ジョイントネットワーク）２３０とを備えている。符号化器ネットワーク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）にほぼ類似しており、積層されたＬＳＴＭ（Ｌｏｎｇ Ｓｈｏｒｔ－Ｔｅｒｍ Ｍｅｍｏｒｙ）層のリカレントネットワークを備えている。例えば符号化器は、$x_t \in R^d$（Ｒは白抜き文字）であるｄ次元特徴ベクトル（例えば、音響フレーム１０４（図１））のシーケンスｘ＝（x_1，x_2，・・・，x_T）を読み込み、各時間ステップで高次の特徴表現を生成する。この高次の特徴表現は、h_1^{enc}，・・・，h_T^{enc}のように示される。

【００３０】

同様に、予測ネットワーク２２０もＬＳＴＭネットワークであり、言語モデル（ＬＭ）のように、これまでに最終ソフトマックス層２４０が出力した非空白記号のシーケンスy_0，・・・，y_{ui-1}を処理して、高密度の表現P_{ui}にする。最後に、ＲＮＮ－Ｔモデルのアーキテクチャでは、符号化器ネットワーク２１０および予測ネットワーク２２０によって生成された表現同士は、結合ネットワーク２３０によって結合される。結合ネットワーク２３０は、次の出力記号に対する分布である予測Ｐ（y_i｜x_1，・・・，x_{ti}，y_0，・・・，y_{ui-1}）を行う。別の言い方をすると、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、「可能性のある音声認識仮説」（ポシブルスピーチレコグニションヒポセシス）は、指定された自然言語の記号／文字（キャラクタ）をそれぞれ

が表す出力ラベルの第１セットと、指定された自然言語の音素をそれぞれが表す出力ラベルの第２セットとに対応する。したがって、結合ネットワーク２３０は、所定の出力ラベルのセットのそれぞれの発生の可能性（ライクリフッドオブオカレンス）を示す一連の値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、出力ラベルのセットに対する確率分布を示すことができる。いくつかのケースでは、出力ラベルは、第１セットでは書記素（ｇｒａｐｈｅｍｅｓ。例えば、個々の文字、および潜在的には句読点および他の記号）であり、第２セットでは音素であるが、出力ラベルのセットはそのように限定されない。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それぞれに対する事後確率値（ポステリアプロバビリティバリュー）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す１００個の異なる出力ラベルがある場合、結合ネットワーク２３０の出力ｙ｟は、各出力ラベルに対して１つずつになるように、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、転写１１６を決定するためのビーム探索プロセス（例えば、ソフトマックス層２４０による）において、正書法の候補要素（ｃａｎｄｉｄａｔｅ ｏｒｔｈｇｒａｐｈｉｃ ｅｌｅｍｅｎｔ）（例えば、書記素、語句、単語、音素）を選択し、スコアを割り当てるべく使用することができる。

【００３１】

ソフトマックス層２４０は、対応する出力ステップでモデル２００によって予測される次の出力記号として、分布内で最も高い確率を持つ出力ラベル／記号を選択するべく、任意の技術を採用することができる。このようにして、ＲＮＮ－Ｔモデル２００は条件付き独立性仮定を行わず、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ている。ＲＮＮ－Ｔモデル２００は、出力記号が将来の音響フレーム１０４から独立していると仮定しており、これによって、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。

【００３２】

いくつかの例では、ＲＮＮ－Ｔモデル２００の符号化器ネットワーク２１０は、８個の２０４８次元ＬＳＴＭ層で構成され、それぞれの後に６４０次元の投影（プロジェクション）層が続く。モデルのレイテンシを低減するべく、符号化器の第２ＬＳＴＭ層の後に、低減（リダクション）係数が２の時間低減層を挿入してもよい。また、予測ネットワーク２２０は、２個の２０４８次元ＬＳＴＭ層を有していてもよく、それぞれの後に６４０次元の投影層が続いている。最後に、結合ネットワーク２３０は、６４０個の隠れユニットと、それの後に続く４０９６個のソフトマックス出力も有していてもよい。具体的には、出力ユニットは、４１個の文脈非依存音素を含み、残りは語句（ワードピース）である。

【００３３】

図１に戻って、自動音声認識ＡＳＲシステム１００のバイアス構成要素１１５は、バイアスされるべき外国語のバイアス用語リスト１０５からの用語を外国語音素にトークン化するように構成されたトークン化器１２１と、トークン化された用語の外国語音素を単一言語、例えば、アメリカ英語に関連する類似の音素に写像（マッピング）するように構成された音素写像器（マッパー）１２３とを備えている。音素写像器１２３は、人間が生成したソース言語からターゲット言語への音素ペアを備えている辞書によって表されてもよく、Ｘ－ＳＡＭＰＡ音素セットはすべての言語に使用される。注目すべきは、音素写像器１２３は、語句－音素モデル２００が、単一の言語、例えば、アメリカ英語に関連する音素のみを備えている場合に有用である。

【００３４】

例えば、ナビゲーションクエリ「クレテイユまでの道順」（ｄｉｒｅｃｔｉｏｎｓ ｔｏ Ｃｒｅｔｅｉｌ）の発話１０６と、フランス語の単語「クレテイユ」（Ｃｒｅｔｅｉｌ）がバイアス用語リスト１０５内にあるという仮定が与えられた場合、「クレテイユ」（Ｃｒｅｔｅｉｌ）は、まずトークン化器１２１によって「ｋ Ｒ ｅ ｔ Ｅ ｊ」としてフランス語の音素にトークン化され、次に音素写像器１２３によって「ｋ ｒ￥ Ｅ ｔ Ｅ ｊ」として英語の音素に写像されて、音素レベルのバイアス有限状態変換器ＦＳＴ３００の生成

に使用される。語句－音素モデル２００が単一の言語、例えば、アメリカ英語からの音素のみをモデリングユニットとして備えているので、音素写像は使用される。

【００３５】

本開示は、どのような用語がバイアス用語リスト１０５に含まれるか、または用語がバイアス用語リスト１０５に含まれるようにどのように選択されるかに限定されない。バイアス用語リスト１０５は、関連する文脈（コンテキスト）に基づき、動的に更新されてもよい。例えば、文脈情報は、ユーザ装置１０２上でどのようなアプリケーションが開いていて使用中であるか、ユーザの連絡先リストからの連絡先名、ユーザ１１０のメディアライブラリ内のアーティスト名／アルバム名、ユーザ１１０の位置などを示してもよい。例えば、ユーザ１１０はアメリカ英語を話すことができ、ナビゲーション／地図アプリケーションがユーザ装置１０２上で開かれていることと、ユーザ１１０の場所がフランスであることとを示す文脈情報に基づき、バイアス用語リスト１０５は、フランスの都市名および／または地域名に関連する用語を備えていることができる。

【００３６】

また、バイアス構成要素１１５は、音素レベルのバイアス有限状態変換器ＦＳＴ生成器１２５を備えており、音素レベルのバイアス有限状態変換器ＦＳＴ生成器１２５は、バイアス用語リスト１０５内の外国語（例えば、フランス語）用語のそれぞれを表す母語（例えば、アメリカ英語）の音素シーケンスに基づき、バイアス有限状態変換器ＦＳＴ３００を生成するように構成されている。いくつかの例では、バイアス有限状態変換器ＦＳＴ生成器１２５は、音素レベルで重みを割り当てるべく押す重み（ｗｅｉｇｈｔ　ｐｕｓｈｉｎｇ）を使用し、過剰バイアスを避けるべく失敗アーク（ｆａｉｌｕｒｅ　ａｒｃｓ）を追加する。いくつかの実装では、復号化において、すべてのバイアス語を使用して、各アーク（ａｒｃ）が同じ重みを持つ文脈的有限状態変換器ＦＳＴを構築する。これらの重みは、異なるモデルに対して独立して、調整することができる。

【００３７】

音声認識装置１５０は、バイアス構成要素１１５によって生成されたバイアス有限状態変換器ＦＳＴ３００を使用して、語句－音素モデル２００によって出力された音素を再スコアリングし、一方、復号グラフ４００は、バイアス有限状態変換器ＦＳＴ３００からの再スコアリングされた音素と、語句－音素モデル２００によって出力された語句とを消費して、転写１１６に含めるための語句を生成する。復号グラフ４００は、発話１０６に対する１つまたは複数の転写候補を決定するビーム探索復号処理に対応してもよい。

【００３８】

いくつかの例では、語句－音素モデル２００による復号中に、バイアス有限状態変換器ＦＳＴ３００は、語句－音素モデル２００によって出力された英語音素記号を消費し、外国語辞書および音素写像を使用して語句を生成してもよく、すなわち「ｋｒ￥ＥｔＥｊ」→クレテイユ（Ｃｒｅｔｅｉｌ）である。復号グラフ４００によって出力された語句は、連結器（コンカチネータ）１３４によって、ユーザ装置１０２の他の構成要素に出力される転写１１６の単語（ワード）に連結され、ここでユーザ装置１０２の他の構成要素は、例えば、ユーザインタフェース生成システム１０７や、他の自然言語処理構成要素である。

【００３９】

図３は、音素レベルでの単語「クレテイユ」（Ｃｒｅｔｅｉｌ）に対する、例示的なバイアス有限状態変換器ＦＳＴ３００を示す。そして、このバイアス有限状態変換器ＦＳＴは、以下の式（１）を用いて、語句－音素モデルの音素出力をオンザフライで再スコアリングするべく使用される。

【００４０】

【数１】

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} \log P(\boldsymbol{y}|\boldsymbol{x}) + \lambda \log P_C(\boldsymbol{y}) \qquad (1)$$

【００４１】
式（１）において、ｘは音響観測値であり、ｙはサブ単語（サブワード）ユニットシーケンスであり、ＰはＥ２Ｅモデルからの確率推定であり、P_cはバイアス再スコアリング確率である。λは、再スコアリングにおける文脈言語モデルＬＭの重みを制御する。
【００４２】
図１に戻ると、語句－音素モデル２００は、モデリングユニットとして語句だけでなく音素も組み込み、バイアス用語リスト１０５内の外国語用語に向けた文脈上のバイアスのためにバイアス有限状態変換器ＦＳＴ３００を使用する。すべての音素モデルとは対照的に、音素と語句の両方をモデル化する語句－音素モデル２００は、通常の単語（レギュラーワード）を認識する際の回帰を緩和する。
【００４３】
語句－音素モデル２００が段階Ａで学習（トレーニング）された後、段階Ｂで、ユーザ１１０は、発話１０６「クレテイユへの道順」（ｄｉｒｅｃｔｉｎｏｓ ｔｏ Ｃｒｅｔｅｉｌ）をユーザ装置１０２に話す。段階Ｃにおいて、音声サブシステム１０３は、例えばマイクロフォンを使用して、発話を受け取り、受け取った発話を、一連のパラメータ化された入力音響フレーム１０４に変換する。例えば、パラメータ化された入力音響フレーム１０４はそれぞれ、８０次元のｌｏｇ－Ｍｅｌ特徴を備えてもよく、ここで８０次元のｌｏｇ－Ｍｅｌ特徴は、短い、例えば２５ｍｓのウィンドウで計算されるとともに、数ミリ秒ごと、例えば１０ミリ秒ごとにシフトされる。
【００４４】
段階Ｄにおいて、自動音声認識ＡＳＲシステム１００は、上述したようにパラメータ化された入力音響フレームを処理し、文脈的にバイアスされた転写１１６、すなわちテキスト「クレテイユ（Ｃｒｅｔｅｉｌ）への道順」を出力する。段階Ｅにおいて、ユーザインタフェース生成システム１０７は、転写の表現を備えているグラフィカルユーザインタフェース１３６のためのコンピュータコードを生成し、段階Ｆにおいて、ユーザインタフェース１３６に表示するべく、そのコンピュータコードをモバイル装置（１０２）に送信する。
【００４５】
自動音声認識ＡＳＲシステム１００によって実行される追加の詳細は、段階Ｄの期間内に発生する可能性がある。例えば、段階Ｄ′の期間内に、バイアス構成要素１１５は、用語「クレテイユ」（Ｃｒｅｔｅｉｌ）を備えているバイアス用語リスト１０５の受け取りに基づき、バイアス有限状態変換器ＦＳＴ３００を生成する。段階Ｄ′′において、音声認識装置１５０の学習された語句－音素モデル２００は、ユーザ１１０の発話１０６に基づき、語句と、対応する音素シーケンスとの両方に対する音声認識スコアを生成し、音素に対する音声認識スコアは、バイアス有限状態変換器ＦＳＴ３００によって再スコアリングおよび再写像され、語句と、再スコアリング／再写像された音素とは、段階Ｄ′′において、転写１１６で出力するための語句を生成するべく、復号グラフ４００によって消費される。復号グラフ４００および連結器１３４は、文脈的にバイアスされた転写１１６を生成し、出力用の転写を、例えば、ユーザ装置１０２のＧＵＩ１３６に表示するべくユーザインタフェース生成システム１０７に提供する。注目すべきは、バイアス用語リスト１０５内の用語のいずれかに対応する音素シーケンスをバイアス有限状態変換器ＦＳＴ３００が再スコアリングした後に、復号グラフ４００は実行されることである。このように、バイアス用語リスト１０５内の外国語に対応する低い音声認識スコアを有する語句は、早々には剪定（ｐｒｕｎｅ）されない。
【００４６】
テスト中、語句－音素モデル２００およびバイアス有限状態変換器ＦＳＴ３００を採用して認識結果をバイアス用語リスト１０５内の用語に向けて文脈的にバイアスする音声認識装置１５０は、書記素のみのバイアスモデルと語句のみのバイアスモデルとの両方よりも顕著に優れたＷＥＲ率で、外国語単語の認識に成功することが示された。また、語句－音素モデル２００は、モデルのスケーラビリティの問題なく、他の外国語に直接適用してバイアスをかけることができるという利点がある。

【００４７】
図４は、音声認識装置１５０が音声認識結果を文脈的にバイアスするべく実行する、例示的な復号グラフ４００を示す。具体的には、例示的な復号グラフ４００は、英語のクロスリンガル発音「ｋ ｒ￥ Ｅ Ｓ」を有する単語「クレイシュ」（ｃｒｅｃｈｅ。ｃｒｅのｅの上にアクソングラーヴが付されている。英語では「デイケア」（ｄａｙｃａｒｅ））と、発音「ｋ ｒ￥ Ｅ ｔ Ｅ ｊ」を有する単語「クレテイユ」（Ｃｒｅｔｅｉｌ。フランスの都市）とに対する復号を描いている。なお、わかりやすくするべく、「０」という状態の語句はほとんど省略している。
【００４８】
復号グラフ４００は、語句－音素モデル２００から出力された音素と語句の両方を入力として受け取るように構成されている。音声復号化処理は、復号グラフ（デコーディンググラフ）４００を検索して、出力として単語を生成する。図示の例では、復号有限状態変換器ＦＳＴは、状態０を中心とした語句ループを有するが、発音有限状態変換器ＦＳＴ、すなわち状態１～１４を有し、それら状態は音素を入力とし、対応する語句を出力とする接頭辞（ｐｒｅｆｉｘ）ツリーを備えている。発音有限状態変換器ＦＳＴは、すべてのバイアス用語について、バイアス時に使用された発音と同じ発音を用いて構築される。常に語句である最終出力記号は、（例えば、図１の連結器１３４によって）単語（ワード）に連結される。
【００４９】
図４の復号グラフ４００は、全体的な復号戦略に２つの追加の改善をもたらす。第１に、復号グラフ４００の性質を考慮すると、同じコストで同じ入力を消費するが、同じ出力を持たないいくつかの仮説が存在する可能性がある。例えば、状態７で終了する仮説は、状態９で終了する仮説と同じコストを持つことになる。このため、すべてが等価な多くの仮説によって、ビームが埋め尽くされてしまうという問題が生じる。本明細書に記載されている強化された自動音声認識ＡＳＲ技術は、このように、状態９で終わる仮説のみを保持することで、ビームを刈り取る（ｐｒｕｎｅする）。
【００５０】
第２改善点は、結合（マージ）された経路（ｐａｔｈ）に関する。学習と復号との性質を考慮すると、与えられた単語は、直接語句で出力されるか、または、音素から語句に変換される。同等の仮説が追跡され、それらの確率を加算することで再結合され、最も可能性の高い仮説に合計確率を割り当て、他のものをビームから削除する。
【００５１】
語句－音素モデル２００のバイアスの結果を、語句のみのモデルと、書記素のみのモデルとに対して比較するテストが行われた。後者の２つのモデルは語句－音素モデル２００と同じ構造を有しており、違いは、書記素モデルが出力として７６個の書記素を有している一方で、語句モデルが４０９６個の語句（ワードピース）を有することである。この違いによって、書記素モデルと語句モデルのパラメータは、それぞれ約１１７Ｍ個と１２０Ｍ個になる。なお、この２つのモデルの出力記号は英語であり、全英語データを用いて学習されている。これらの２つのモデルでは、フランス語のバイアス単語（ワード）の英語音訳版を使用して、書記素レベルまたは語句レベルのみでバイアスが行われる。
【００５２】
一般的に、テストでは、３つのモデルはバイアスをかけなくても同じように動作することが示された。これは、地名がフランス語であり、それらが学習では見られたことがないためであり、すなわち、ほぼ１００％の単語ＯＯＶ率である。さらに、すべてのモデルは、バイアスをかけることで大幅に性能が向上する。バイアスをかけない場合と比較して、ＷＥＲの減少が顕著である。
【００５３】
異なるバイアス戦略を比較すると、語句－音素モデル２００が最も優れた性能を示し、書記素モデルおよび語句モデルの両方よりも有意に良好に動作した。語句－音素モデルの優れた性能は、ＯＯＶの単語に対する音素のロバスト性に起因する。語句－音素モデル２０

０は、モデリングユニットとして語句と音素の両方を備えているので、音素有限状態変換器ＦＳＴに加えて語句有限状態変換器ＦＳＴを構築することで、音素ベースのバイアスに加えて語句バイアスを実行することができる。この語句単位の有限状態変換器ＦＳＴを追加することで、ＷＥＲがさらに減少することが実証されており、語句単位のバイアスと音素単位のバイアスとは相互に補完し合う関係にあることがわかる。音素と語句のバイアスに使用する重みは、同じでもよいし、異なっていてもよい。観察によると、長いユニットをマッチングする際のスパース性の問題から、語句単位の方が、書記素（ｇｒａｐｈｅｍｅ）単位よりも性能が高い場合がある。

【００５４】

テストの結果、バイアスは外国の地名を認識するのに役立つことがわかった。例えば、バイアスをかけると、正しいフランス語の単語（ｗｏｒｄ）が生成され、逆にバイアスをかけないと、音韻的には似ているが間違った英語の単語が生成される。誤りは、フランス語の音韻的に類似した単語が原因であることが多い。

【００５５】

バイアスなしのシナリオでの回帰がないことをより確実にするべく、通常の英語の発話の復号（デコーディング）で３つのモデルを比較した。復号では、バイアスフレーズの空リストを使用することで、バイアスメカニズムをオフにした。テストの結果、語句モデルは、書記素（ｇｒａｐｈｅｍｅ）モデルよりも優れた性能を示すことがわかった。語句－音素モデルは、書記素モデルよりもやや良好な結果となったが、これは学習時に語句の頻度が高かったことに起因していると考えられる。語句モデルと比較して、語句－音素モデルは非常にわずかに劣化している。これは、モデリングに電話を導入したことによる。回帰性を向上させるための潜在的なアプローチとしては、語句ベースの再スコアリングと同様に、再スコアリングに音素の英語外部言語モデルを組み込むことが考えられる。しかし、全音素（ａｌｌ－ｐｈｏｎｅｍｅ）モデルに比べて、回帰が著しく小さくなる。

【００５６】

図５は、バイアス用語リスト内の外国語用語に向かって転写を文脈的にバイアスする方法の動作の例示的な配置のフローチャートである。動作５０２において、方法５００は、第１言語のネイティブスピーカ（１１０）によって話される発話１０６を符号化（エンコーディング）する音声データを受け取る工程を備えている。発話１０６は、第１言語とは異なる第２言語の１つまたは複数の外国語を備えてもよい。動作５０４において、方法５００は、第２言語の１つまたは複数の用語を備えているバイアス用語リスト１０５を受け取る工程を備えている。

【００５７】

動作５０６において、方法５００は、音声認識モデル２００を使用して、音声データから導出された音響特徴（１０４）を処理して、第１言語における語句と、対応する音素シーケンスとの両方に対する音声認識スコアを生成する工程も備えている。動作５０８において、方法５００は、バイアス用語リスト内の１つまたは複数の用語に基づき、音素シーケンスに対する音声認識スコアを再スコアリングする工程も備えている。動作５０６において、方法５００は、語句に対する音声認識スコアと、音素シーケンスに対する再スコアリングされた音声認識スコアとを用いて、復号グラフ（デコーディンググラフ）４００を実行して、発話１０６に対する転写（トランスクリプション）１１６を生成する工程を備えている。

【００５８】

ソフトウェアアプリケーション（すなわち、ソフトウェアリソース）は、計算装置にタスクを実行させるコンピュータソフトウェアを指すことがある。いくつかの例では、ソフトウェアアプリケーションは、「アプリケーション」、「アプリ」、または「プログラム」と呼ばれることがある。アプリケーションの例としては、システム診断アプリケーション、システム管理アプリケーション、システムメンテナンスアプリケーション、ワープロアプリケーション、表計算アプリケーション、メッセージングアプリケーション、メディアストリーミングアプリケーション、ソーシャルネットワーキングアプリケーション、ゲー

ムアプリケーションなどがあるが、これらに限定されない。
【００５９】
非一過性メモリは、計算装置が使用するためのプログラム（例えば、命令のシーケンス）またはデータ（例えば、プログラムの状態情報）を一時的または永久的に保存するべく使用される物理デバイスであってもよい。非一時的メモリは、揮発性および／または不揮発性のアドレス可能な半導体メモリであってもよい。不揮発性メモリの例としては、フラッシュメモリ、リードオンリーメモリ（ＲＯＭ）／プログラマブルリードオンリーメモリ（ＰＲＯＭ）／消去可能プログラマブルリードオンリーメモリ（ＥＰＲＯＭ）／電子的消去可能プログラマブルリードオンリーメモリ（ＥＥＰＲＯＭ）（例えば、ブートプログラムなどのファームウェアに典型的に使用される）などがあるが、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）、フェイズチェンジメモリ（ＰＣＭ）のほか、ディスクやテープなどが挙げられるが、これらに限定されるものではない。
【００６０】
図６は、本書で説明したシステムおよび方法を実施するべく使用することができる例示的な計算装置６００の概略図である。計算装置６００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、様々な形態のデジタルコンピュータを表すことを意図している。ここに示されている構成要素、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を意図しており、本書に記載および／または請求されている発明の実施を制限することを意図していない。
【００６１】
計算装置６００は、プロセッサ６１０と、メモリ６２０と、記憶装置（ストレージデバイス）６３０と、メモリ６２０および高速拡張ポート６５０に接続する高速インタフェース／コントローラ６４０と、および低速バス６７０および記憶装置（ストレージデバイス）６３０に接続する低速インタフェース／コントローラ６６０とを備えている。構成要素６１０、６２０、６３０、６４０、６５０、６６０のそれぞれは、様々なバスを用いて相互に接続されており、共通のマザーボードに搭載されていてもよいし、適宜他の態様で搭載されていてもよい。プロセッサ６１０は、高速インタフェース６４０に結合されたディスプレイ６８０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６２０または記憶装置６３０に格納された命令を備えている、計算装置６００内で実行するための命令を処理することができる。他の実装では、複数のプロセッサおよび／または複数のバスが、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、複数の計算装置６００が接続され、各デバイスが必要な動作の一部を提供してもよい（例えば、サーババンク、ブレードサーバ群、またはマルチプロセッサシステムとして）。
【００６２】
メモリ６２０は、計算装置６００内の情報を非一時的に格納する。メモリ６２０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６２０は、計算装置６００による使用のために、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）を一時的または永久的に格納するべく使用される物理デバイスであってもよい。不揮発性メモリの例には、フラッシュメモリおよびリードオンリーメモリ（ＲＯＭ）／プログラマブルリードオンリーメモリ（ＰＲＯＭ）／消去可能プログラマブルリードオンリーメモリ（ＥＰＲＯＭ）／電子的消去可能プログラマブルリードオンリーメモリ（ＥＥＰＲＯＭ）（例えば、ブートプログラムなどのファームウェアに典型的に使用される）が含まれるが、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）、フェイズチェンジメモリ（ＰＣＭ）のほか、ディスクやテープな

どが挙げられるが、これらに限定されるものではない。

【００６３】

記憶装置６３０は、計算装置６００に大容量記憶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６３０は、コンピュータ可読媒体である。様々な異なる実装において、記憶装置（ストレージデバイス）６３０は、フロッピー（登録商標）ディスクデバイス、ハードディスクデバイス、光ディスクデバイス、またはテープデバイス、フラッシュメモリまたは他の類似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。追加の実装では、コンピュータプログラム製品が、情報キャリアに有形的に具現化される。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（情報担体）は、メモリ６２０、記憶装置６３０、またはプロセッサ６１０上のメモリなどの、コンピュータまたは機械可読媒体である。

【００６４】

高速コントローラ６４０は、計算装置６００のための帯域幅集中型の動作を管理し、低速コントローラ６６０は、より低い帯域幅集中型の動作を管理する。このような職務の割り当ては、例示的なものに過ぎない。いくつかの実装では、高速コントローラ６４０は、メモリ６２０と、ディスプレイ６８０（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）と、および、様々な拡張カード（図示せず）を受け入れ得る高速拡張ポート６５０とに結合される。いくつかの実装では、低速コントローラ６６０は、記憶装置６３０および低速拡張ポート６９０に結合される。様々な通信ポート（例えば、ＵＳＢ、Ｂｌｕｅｔｏｏｔｈ（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））を備えてもよい低速拡張ポート６９０は、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、またはスイッチやルータなどのネットワークデバイスに、例えばネットワークアダプタを介して結合されてもよい。

【００６５】

計算装置６００は、図に示すように、いくつかの異なる形態で実装されてもよい。例えば、計算装置は、標準的なサーバ６００ａまたはそのようなサーバ６００ａのグループにおける複数倍として、ラップトップコンピュータ６００ｂとして、またはラックサーバシステム６００ｃの一部として、実装されてもよい。

【００６６】

本明細書に記載されたシステムおよび技術の様々な実装は、デジタル電子および／または光学回路、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、プログラム可能な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での実装を備えていることができ、ここでプログラム可能なシステムは、データおよび命令を記憶装置から受け取り、データおよび命令を記憶装置に送信するように記憶装置に結合された、特殊目的または汎用の少なくとも１つのプログラム可能なプロセッサと、少なくとも１つの入力装置と、および少なくとも１つの出力装置とを備えている。

【００６７】

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、またはコードとも呼ばれる）は、プログラマブルプロセッサのための機械命令を含み、高レベルの手続き型および／またはオブジェクト指向の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。本明細書において、「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、機械可読信号として機械命令を受け取る機械可読媒体を含んでいる、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、非一時的なコンピュータ可読媒体、装置（アパレイタス）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））を意味する。「

機械可読信号」とは、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。

【００６８】

本明細書に記載されてい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して、入力データを操作して出力を生成することで機能を実行することができる。また、ＦＰＧＡ（Ｆｉｅｌｄ Ｐｒｏｇｒａｍｍａｂｌｅ Ｇａｔｅ Ａｒｒａｙ）やＡＳＩＣ（Ａｐｐｌｉｃａｔｉｏｎ Ｓｐｅｃｉｆｉｃ Ｉｎｔｅｇｒａｔｅｄ Ｃｉｒｃｕｉｔ）などの特殊な論理回路によっても処理や論理フローを実行することができる。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、汎用および特殊目的のマイクロプロセッサ、および任意の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デバイスである。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、または光ディスクを備えているか、またはデータを受け取るか、またはデータを転送するか、もしくは両方であるように動作可能に結合される。しかし、コンピュータはそのようなデバイスを持っている必要はない。コンピュータプログラムの命令やデータを格納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、例として、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ＣＤ－ＲＯＭおよびＤＶＤ－ＲＯＭディスクが挙げられる。プロセッサとメモリは、特別な目的の論理回路によって補完されるか、またはそれに組み込まれることができる。

【００６９】

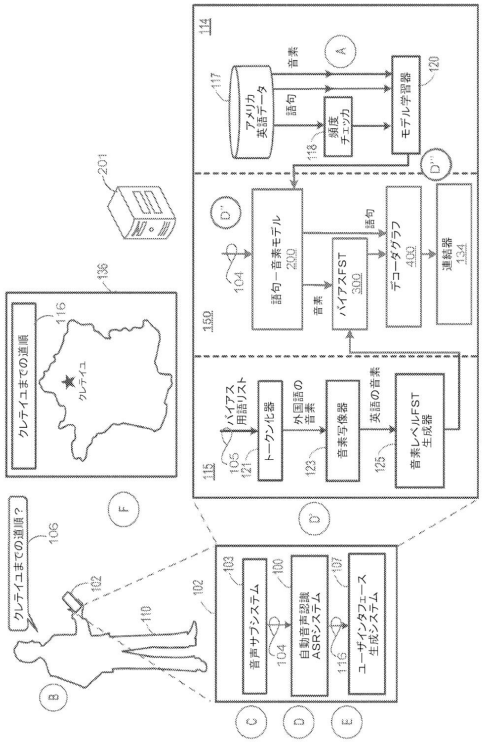
ユーザとの対話（相互作用）を提供するべく、本開示の１つまたは複数の態様は、ユーザに情報を表示するためのディスプレイデバイス、例えばＣＲＴ（ｃａｔｈｏｄｅ ｒａｙ ｔｕｂｅ）、ＬＣＤ（ｌｉｑｕｉｄ ｃｒｙｓｔａｌ ｄｉｓｐｌａｙ）モニタ、またはタッチスクリーンと、任意でキーボードおよびポインティングデバイス、例えばマウスまたはトラックボールを有し、それによってユーザがコンピュータに入力を提供することができるコンピュータ上で実装することができる。同様にユーザに相互作用を提供できる多の種類の装置が使用でき、例えば、ユーザに提供されるフィードバックは、視覚的なフィードバック、聴覚的なフィードバック、触覚的なフィードバックなど、あらゆる形態の感覚的なフィードバックであり、ユーザからの入力は、音響的な入力、音声的な入力、触覚的な入力など、あらゆる形態で受け取ることができる。さらに、コンピュータは、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け取ったりすることで、ユーザと対話することができる。例えば、ウェブブラウザから受け取った要求に応答して、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で、ユーザと対話することができる。

【００７０】

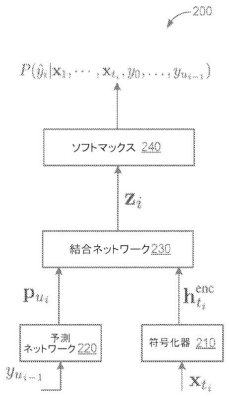
多数の実施例を説明してきた。それにもかかわらず、本開示の精神および範囲から逸脱することなく、様々な変更を行うことができることが理解されるであろう。したがって、他の実施態様は、以下の請求項の範囲内にある。

【図面】

【図１】



【図２】



【図３】

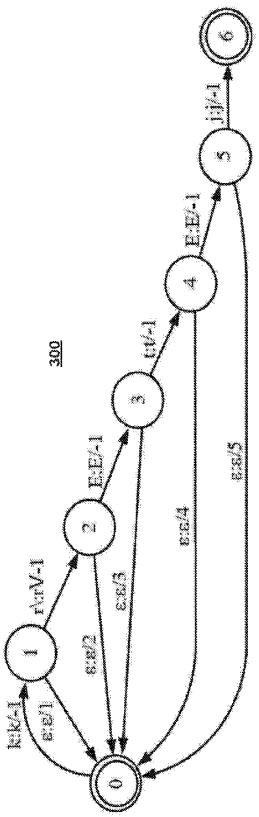


FIG. 3

【図４】

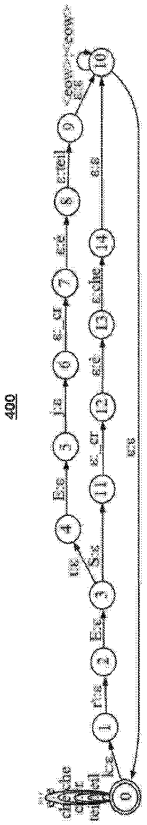
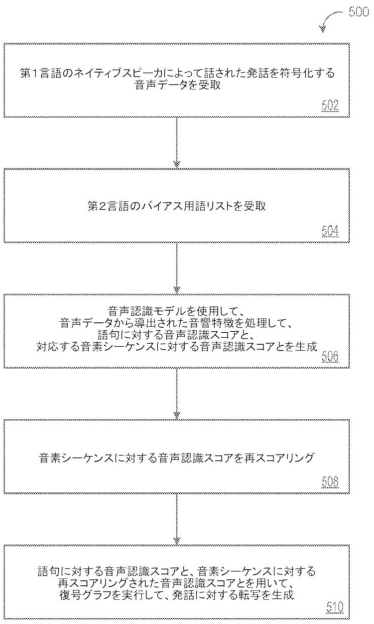


FIG. 4

【図５】


500
第１言語のネイティブスピーカによって話された発話を符号化する音声データを受取
502
第2言語のバイアス用語リストを受取
504
音声認識モデルを使用して、音声データから導出された音響特徴を処理して、話句に対する音声認識スコアと、対応する音素シーケンスに対する音声認識スコアとを生成
506
音素シーケンスに対する音声認識スコアを再スコアリング
508
語句に対する音声認識スコアと、音素シーケンスに対する再スコアリングされた音声認識スコアとを用いて、復号グラフを実行して、発話に対する転写を生成
510

【図６】

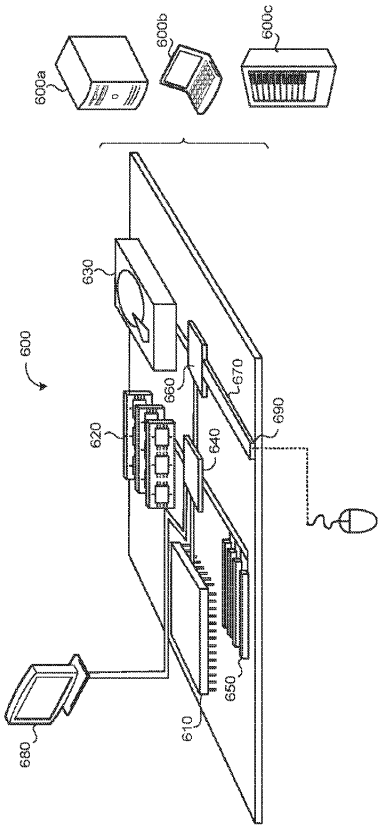

600a
600b
600c
600
610
620
630
640
650
660
670
680
690

FIG. 6

フロントページの続き

ｇ／ａｂｓ／１９０６．０９２９２にて発表
早期審査対象出願
(72)発明者　ブルギエ、アントワーヌ ジャン
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　サイナス、ターラ エヌ．
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　プラバーバルカル、ロヒット プラカーシュ
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　プンダック、ゴラン
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
審査官　菊池 智紀
(56)参考文献　特表２０１９－５０７３６２（ＪＰ，Ａ）
PATEL, Ami et al.，"CROSS-LINGUAL PHONEME MAPPING FOR LANGUAGE ROBUST CONTEXTUAL SPEECH RECOGNITION"，Proc. of the 2018 IEEE ICASSP，2018年04月15日，pp. 5924-5928
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１５／３４
ＩＥＥＥ Ｘｐｌｏｒｅ