

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7635194号
(P7635194)

(45)発行日 令和7年2月25日(2025.2.25)

(24)登録日 令和7年2月14日(2025.2.14)

(51)国際特許分類

G 10 L	15/16 (2006.01)	G 10 L	15/16
G 10 L	15/187 (2013.01)	G 10 L	15/187
G 10 L	15/19 (2013.01)	G 10 L	15/19
G 10 L	15/065 (2013.01)	G 10 L	15/065

F I

Z

請求項の数 20 外国語出願 (全27頁)

(21)出願番号 特願2022-204682(P2022-204682)
 (22)出願日 令和4年12月21日(2022.12.21)
 (62)分割の表示 特願2021-575440(P2021-575440)
)の分割
 原出願日 令和2年4月30日(2020.4.30)
 (65)公開番号 特開2023-29416(P2023-29416A)
 (43)公開日 令和5年3月3日(2023.3.3)
 審査請求日 令和5年1月11日(2023.1.11)
 (31)優先権主張番号 62/863,308
 (32)優先日 令和1年6月19日(2019.6.19)
 (33)優先権主張国・地域又は機関
 米国(US)

前置審査

(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 94
 043 マウンテン ピュー アンフィシ
 アター パークウェイ 1600
 1600 Amphitheatre P
 arkway 94043 Mounta
 in View, CA U.S.A.
 (74)代理人 100108453
 弁理士 村山 靖彦
 (74)代理人 100188558
 弁理士 飯田 雅人
 (74)代理人 100110364
 弁理士 実広 信哉

最終頁に続く

(54)【発明の名称】 音声認識のためのコンテキストバイアス

(57)【特許請求の範囲】

【請求項1】

データ処理ハードウェアにおいて実行されたときに、前記データ処理ハードウェアに複数の動作を実行させるコンピュータ実装方法であって、前記複数の動作が、

発話の1つまたは複数の単語をエンコードするオーディオデータを受信する動作と、

前記発話の前記1つまたは複数の単語のうちのバイアスフレーズの書記素データおよび音素データを取得する動作であって、前記バイアスフレーズが、前記発話のコンテキストに対応し、固有名詞を含む、動作と、

前記オーディオデータの音響特徴を受信するように構成されたオーディオエンコーダを使用して、オーディオエンコーディングを生成する動作と、

前記書記素データを受信するように構成された書記素エンコーダを使用して、書記素エンコーディングを生成する動作と、

前記音素データを受信するように構成された音素エンコーダを使用して、音素エンコーディングを生成する動作と、

前記オーディオエンコーディングを受信するように構成された第1の注意モジュールを使用して、第1の注意出力を生成する動作であって、前記第1の注意出力が、前記オーディオデータのベクトル要約を含む加重オーディオエンコーディングベクトルを含む、動作と、

前記書記素エンコーダから出力された前記書記素エンコーディングと前記音素エンコーダから出力された前記音素エンコーディングとの表現を受信するように構成された第2の注意モジュールを使用して、第2の注意出力を生成する動作と、

10

20

前記オーディオデータに対応する出力書記素のセットの可能性を決定するために、前記第1および第2の注意出力を、デコーダを使用して処理する動作とを含む、コンピュータ実装方法。

【請求項 2】

前記複数の動作が、前記出力書記素のセットの可能性に基づいて、前記発話のトランスクリプションを生成する動作をさらに含む、請求項1に記載のコンピュータ実装方法。

【請求項 3】

前記書記素エンコーダおよび前記音素エンコーダの各々が、ニューラルネットワークを備える、請求項1に記載のコンピュータ実装方法。

【請求項 4】

前記書記素エンコーダが、特定の単語に対応する書記素エンコーディングを生成するように構成され、

前記音素エンコーダが、前記特定の単語に対応する音素エンコーディングを生成するように構成され、

前記第2の注意モジュールが、前記特定の単語に対応する前記書記素および音素エンコーディングに基づいて、前記特定の単語に対応するコンテキストバイアスペクトルを備える、前記第2の注意出力をエンコードするように構成される、請求項1に記載のコンピュータ実装方法。

【請求項 5】

前記書記素エンコーダから出力された前記書記素エンコーディングと前記音素エンコーダから出力された前記音素エンコーディングとの表現が、前記書記素エンコーディングと前記音素エンコーディングとの間の連結を含む、請求項1に記載のコンピュータ実装方法。

【請求項 6】

前記書記素エンコーディングと前記音素エンコーディングとの間の前記連結が、前記第2の注意モジュールへ入力される射影ベクトルによって表される、請求項5に記載のコンピュータ実装方法。

【請求項 7】

前記書記素エンコーダ、前記音素エンコーダ、前記第1および第2の注意モジュール、および前記デコーダが、オーディオデータに対応する出力書記素のセットの可能性を決定するように共同でトレーニングされる、請求項1に記載のコンピュータ実装方法。

【請求項 8】

前記出力書記素のセットが書記素を備える、請求項1に記載のコンピュータ実装方法。

【請求項 9】

前記出力書記素のセットが、単語または単語のピースを含む、請求項1に記載のコンピュータ実装方法。

【請求項 10】

前記複数の動作が、

前記発話を話したユーザの位置、

前記発話を話したユーザに関連付けられるユーザデバイスで開いている1つまたは複数のアプリケーション、ならびに

発話の現在の日付および/または時刻

のうちの少なくとも1つに基づいて、前記バイアスフレーズを決定する動作をさらに含む、請求項1に記載のコンピュータ実装方法。

【請求項 11】

システムであって、

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と通信するメモリハードウェアとを備え、前記メモリハードウェアは、前記データ処理ハードウェアにおいて実行されたときに、前記データ処理ハードウェアに複数の動作を実行させる命令を記録し、前記複数の動作が、

10

20

30

40

50

発話の1つまたは複数の単語をエンコードするオーディオデータを受信する動作と、前記発話の前記1つまたは複数の単語のうちのバイアスフレーズの書記素データおよび音素データを取得する動作であって、前記バイアスフレーズが、前記発話のコンテキストに対応し、固有名詞を含む、動作と、

前記オーディオデータの音響特徴を受信するように構成されたオーディオエンコーダを使用して、オーディオエンコーディングを生成する動作と、

前記書記素データを受信するように構成された書記素エンコーダを使用して、書記素エンコーディングを生成する動作と、

前記音素データを受信するように構成された音素エンコーダを使用して、音素エンコーディングを生成する動作と、

前記オーディオエンコーディングを受信するように構成された第1の注意モジュールを使用して、第1の注意出力を生成する動作であって、前記第1の注意出力が、前記オーディオデータのベクトル要約を含む加重オーディオエンコーディングベクトルを含む、動作と、

前記書記素エンコーダから出力された前記書記素エンコーディングと前記音素エンコーダから出力された前記音素エンコーディングとの表現を受信するように構成された第2の注意モジュールを使用して、第2の注意出力を生成する動作と、

前記オーディオデータに対応する出力書記素のセットの可能性を決定するために、前記第1および第2の注意出力を、デコーダを使用して処理する動作とを含む、システム。

【請求項 1 2】

前記複数の動作が、前記出力書記素のセットの可能性に基づいて、前記発話のトランスクリプションを生成する動作をさらに含む、請求項11に記載のシステム。

【請求項 1 3】

前記書記素エンコーダおよび前記音素エンコーダの各々が、ニューラルネットワークを備える、請求項11に記載のシステム。

【請求項 1 4】

前記書記素エンコーダが、特定の単語に対応する書記素エンコーディングを生成するように構成され、

前記音素エンコーダが、前記特定の単語に対応する音素エンコーディングを生成するように構成され、

前記第2の注意モジュールが、前記特定の単語に対応する前記書記素および音素エンコーディングに基づいて、前記特定の単語に対応するコンテキストバイアスベクトルを備える、前記第2の注意出力をエンコードするように構成される、請求項11に記載のシステム。

【請求項 1 5】

前記書記素エンコーダから出力された前記書記素エンコーディングと前記音素エンコーダから出力された前記音素エンコーディングとの表現が、前記書記素エンコーディングと前記音素エンコーディングとの間の連結を含む、請求項11に記載のシステム。

【請求項 1 6】

前記書記素エンコーディングと前記音素エンコーディングとの間の前記連結が、前記第2の注意モジュールへ入力される射影ベクトルによって表される、請求項15に記載のシステム。

【請求項 1 7】

前記書記素エンコーダ、前記音素エンコーダ、前記第1および第2の注意モジュール、および前記デコーダが、オーディオデータに対応する出力書記素のセットの可能性を決定するように共同でトレーニングされる、請求項11に記載のシステム。

【請求項 1 8】

前記出力書記素のセットが書記素を備える、請求項11に記載のシステム。

【請求項 1 9】

前記出力書記素のセットが、単語または単語のピースを含む、請求項11に記載のシステム。

10

20

30

40

50

ム。

【請求項 20】

前記複数の動作が、

前記発話を話したユーザの位置、

前記発話を話したユーザに関連付けられるユーザデバイスで開いている1つまたは複数のアプリケーション、ならびに

発話の現在の日付および/または時刻

のうちの少なくとも1つに基づいて、前記バイアスフレーズを決定する動作をさらに含む、請求項11に記載のシステム。

【発明の詳細な説明】

10

【技術分野】

【0001】

本開示は、音声認識のためのコンテキストバイアスに関する。

【背景技術】

【0002】

音声の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。しかしながら、人々が話す可能性のある単語の多様性とアクセントや発音の多くのバリエーションを考えると、スピーチのコンテキストを認識する能力は困難である。多くの場合、人が話す単語やフレーズの種類は、その人が置かれているコンテキストによって異なる。

【0003】

20

コンテキスト自動音声認識(ASR)は、ユーザ自身のプレイリスト、連絡先、または地理的な地名に向けてなど、所与のコンテキストに向けて音声認識にバイアスをかけることを含む。コンテキスト情報は通常、認識される関連フレーズのリストを含み、これは、トレーニングにおいてめったに見られないまれなフレーズまたは外国語さえを含むことがよくある。コンテキストバイアスを実行するために、従来のASRシステムは、Nグラム加重有限状態トランスデューサ(WFST)を使用して、独立したコンテキスト言語モデル(LM)においてコンテキスト情報をモデル化し、オンザフライ(OTF)再スコアリング用のベースラインLMを使用して独立したコンテキストLMを構成することがある。

【0004】

最近、エンドツーエンド(E2E)モデルは、従来のオンデバイスモデルと比較して、改善された単語誤り率(WER)および遅延メトリックを示し、ASRに大きな期待を寄せている。音声からテキストへのマッピングを直接学習するために、音響モデル(AM)、発音モデル(PM)、およびLMを単一のネットワークに折りたたむこれらのE2Eモデルは、個別のAM、PM、およびLMを備えた従来の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モデルは、単語ベースのコネクショニスト時分類(CTC)モデル、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、およびリップン、アテンド、およびスペル(LAS)などの注意ベースのモデルを含む。E2Eモデルは、ビーム検索のデコード中に限られた数の認識候補を維持するため、コンテキストASRはE2Eモデルにとって困難な場合がある。

30

【発明の概要】

40

【課題を解決するための手段】

【0005】

本明細書の実装形態は、音声認識モデルを現在のコンテキストに関連するフレーズのセットに向けてバイアスをかけることを目的としている。フレーズのセットは、コンテキストの変化に応じてフレーズが変化する可能性があるという点で動的である。音声認識モデルは、フレーズの書かれた表現に加えて、バイアスに使用されるフレーズの発音をさらに考慮に入れることができる。たとえば、モデルは、認識されている音声に最も適したフレーズを選択するために、バイアスフレーズを表すテキスト情報(たとえば、書記素シーケンス)および発音情報(たとえば、音素シーケンス)を使用することができる。これにより、音声モデルのコンテキストバイアスの態様は、個々のバイアスフレーズの発音を認識できる

50

ようになり、それによって、まれな単語および異常な発音の単語を認識する際に、モデルがはるかに高い精度を実現できるようになる。本明細書で説明するコンテキストバイアス技法は、個別の音響モデル、発音モデル、および言語モデルなしで文字起こしデータを生成できるエンドツーエンドモデルなどを含む、様々なタイプの音声認識モデルで使用することができる。

【 0 0 0 6 】

高い音声認識精度を提供するために、汎用自動音声認識(ASR)システムは、音声が発生するコンテキストを示す情報を使用する場合がある。コンテキストは、たとえば、ユーザの連絡先、カレンダの予定、開いているアプリ、および場所に関連している場合がある。ASRシステムの1つの例は、個別の音響、発音、および言語モデルを含む。他のASRシステムは、音響、発音、および言語モデルを単一のニューラルネットワークとして結合する。単一のニューラルネットワークモデルは、単純さと品質を向上させ、単語誤り率(WER)を最適化する。

【 0 0 0 7 】

従来のニューラルネットワークベースのASRモデルの場合、言語にめったに出現しない単語を認識することは困難である。言語における単語の分布は、通常ジップの分布に従い、ジップの分布では少数の単語が非常に頻繁に使用され、膨大な数の単語は使用されることはない。まれな単語の例を効果的に提供するために、オーディオおよび対応するテキストを含む十分な量のトレーニングデータを取得するのは難しい場合がある。たとえ十分なトレーニングデータが利用できる場合でも、トレーニングの例をどんどん追加すると、多くの場合、改善の度合いがますます低くなる。さらに、これらのモデルが、スペルに比べて異常な発音を有する単語を認識することは困難である。

【 0 0 0 8 】

ニューラルネットワークベースのASRモデルにコンテキストバイアスを組み込むと、まれな単語および異常な発音の単語の認識を向上させることができる。有用なアプリケーションの1つは、固有の名前(すなわち、人の名前、歌の名前、都市の名前などの固有名詞)をよりよく認識することであり、これは、トレーニングデータにおいて比較的まれである場合もあり、および/または一般的な発音規則に従わない場合もある。たとえば、スマートフォンまたは他のユーザデバイスは、多くの場合、ユーザの連絡先を記憶する。ユーザがメッセージングまたは電話通話アプリケーションを使用している場合、ASRシステムがユーザによって話された名前を認識することを支援するために、この情報をコンテキストとして使用することができる。同様に、たとえば、ユーザがメディアプレーヤーアプリケーションを使用しているときに、音声認識にバイアスをかけるために、ユーザの音楽ライブラリ内の曲のリストを使用することができる。本明細書の実装形態は、バイアスフレーズの記述形式と発音との両方に関連付けられる情報をASRモデルに注入することによって、コンテキストバイアスをASRモデルに適用することを目的としている。たとえASRモデルのトレーニング中にその単語が観察されなかった場合でも、ASRモデルにコンテキストバイアスをかけると、まれな単語の発音に関する知識が組み込まれる。

【 0 0 0 9 】

本開示の一態様は、データ処理ハードウェアにおいて、発話をエンコードするオーディオデータを受信するステップと、データ処理ハードウェアによって、発話のコンテキストに対応する1つまたは複数のバイアスフレーズのセットを取得するステップであって、1つまたは複数のバイアスフレーズのセット内の各バイアスフレーズが、1つまたは複数の単語を含む、ステップとを含む、音声認識にバイアスをかけるための方法を提供する。本方法はまた、音声認識モデルの出力を生成するために、データ処理ハードウェアによって、音声認識モデルを使用して、オーディオデータから得られる音響特徴、ならびに1つまたは複数のバイアスフレーズのセットから得られる書記素および音素データを処理するステップを含む。本方法はまた、データ処理ハードウェアによって、音声認識モデルの出力に基づいて発話のトランスクリプションを決定するステップを含む。

【 0 0 1 0 】

10

20

30

40

50

本開示の実装形態は、以下の任意の特徴のうちの1つまたは複数を含み得る。いくつかの実装形態では、音声認識モデルは、第1のエンコーダ、第1の注意モジュール、書記素エンコーダ、音素エンコーダ、第2の注意モジュール、およびデコーダを含む。第1のエンコーダは、音響特徴を入力として受信し、音響特徴からオーディオエンコーディングを出力として生成するように構成されている。第1の注意モジュールは、第1のエンコーダから出力されたオーディオエンコーディングを入力として受信し、第1の注意出力を出力として生成するように構成されている。書記素エンコーダは、1つまたは複数のバイアスフレーズのセット内の各単語の書記素を示す書記素データを入力として受信し、書記素エンコーディングを出力として生成するように構成されている。音素エンコーダは、1つまたは複数のバイアスフレーズのセット内の各単語の音素を示す音素データを入力として受信し、音素エンコーディングを出力として生成するように構成されている。第2の注意モジュールは、書記素エンコーダから出力された書記素エンコーディングおよび音素エンコーダから出力された音素エンコーディングの表現を入力として受信し、第2の注意出力を出力として生成するように構成されている。デコーダは、第1の注意出力および第2の注意出力に基づいて、音声要素のシーケンスの可能性(likelihood)を決定するように構成されている。

【0011】

いくつかの例では、1つまたは複数のバイアスフレーズのセット内の各バイアスフレーズの特定の単語ごとに、書記素エンコーダは、特定の単語に対応する書記素エンコーディングを生成するように構成されており、音素エンコーダは、特定の単語に対応する音素エンコーディングを生成するように構成されており、第2の注意モジュールは、特定の単語に対応する書記素および音素エンコーディングに基づいて、特定の単語の対応するコンテキストバイアスベクトルを備える、対応する第2の注意出力をエンコードするように構成されている。さらに、例として、書記素エンコーダから出力される書記素エンコーディングおよび音素エンコーダから出力される音素エンコーディングの表現は、書記素エンコーディングと音素エンコーディングとの間の連結を表す射影ベクトルを含む。さらに、第1のエンコーダ、第1の注意モジュール、書記素エンコーダ、音素エンコーダ、第2の注意モジュール、およびデコーダは、音響特徴フレームのシーケンスから書記素のシーケンスを予測するように共同でトレーニングされ得る。

【0012】

いくつかの実装形態では、第1の注意モジュールは、デコーダの以前の非表示状態(hidden state)と、第1のエンコーダによって出力されたオーディオエンコーディングの完全なシーケンスの関数として注意を計算するように構成されており、第2の注意モジュールは、デコーダの以前の非表示状態と、1つまたは複数のバイアスフレーズのセット内のバイアスフレーズから得られる書記素と音素データとの両方の特性を表す射影ベクトルの完全なシーケンスの関数として注意を計算するように構成されている。追加の実装形態では、第2の注意モジュールは、1つまたは複数のバイアスフレーズのセット内のバイアスフレーズのいずれにも対応しないコンテキストバイアスベクトルを受信するように構成されており、コンテキストバイアスベクトルは、音声認識モデルの出力にバイアスをかけないオプションを表す。いくつかの例では、音声要素は書記素を含む。他の例では、音声要素は単語または単語ピース(wordpieces)を含む。任意で、音声要素は音素を含み得る。

【0013】

1つまたは複数のバイアスフレーズのセットは、特定のユーザのためにパーソナライズされた1つまたは複数の連絡先名を含み得る。追加的または代替的に、1つまたは複数のバイアスフレーズのセットは、特定のユーザのためにパーソナライズされた1つまたは複数のカレンダイベントを含み得る。いくつかの例では、本方法はまた、データ処理ハードウェアによって、発話を話したユーザの位置、発話を話したユーザに関連付けられるユーザデバイスで開いている1つまたは複数のアプリケーション、あるいは発話を現在の日付および/または時刻のうちの少なくとも1つに基づいて発話をコンテキストを決定するステップを含む。発話をコンテキストは、追加的または代替的に、ユーザによって発行された1

つまたは複数の以前のコマンドに基づき得る。

【0014】

いくつかの実装形態では、音声認識モデルは、埋め込みベクトルと、デコーダの以前の非表示状態と、第1のベクトルと、第2のベクトルとに基づいて、非表示状態および音声認識モデルの出力を決定するように構成されたデコーダを含む。ここで、埋め込みベクトルは、音声認識モデルによって出力された以前の書記素に対するものであり、第1のベクトルは第1の注意モジュールによって出力され、第2のベクトルは第2の注意モジュールによって出力される。

【0015】

本開示の別の態様は、発話のコンテキストに基づいて音声認識にバイアスをかけるためのシステムを提供する。本システムは、データ処理ハードウェアと、データ処理ハードウェアと通信し、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記憶するメモリハードウェアとを含む。本動作は、発話をエンコードするオーディオデータを受信することと、発話のコンテキストに対応する1つまたは複数のバイアスフレーズのセットを取得することとを含み、1つまたは複数のバイアスフレーズのセット内の各バイアスフレーズが、1つまたは複数の単語を含む。本動作はまた、音声認識モデルの出力を生成するために、音声認識モデルを使用して、オーディオデータから得られる音響特徴、ならびに1つまたは複数のバイアスフレーズのセットから得られる書記素および音素データを処理することを含む。本動作はまた、音声認識モデルの出力に基づいて発話のトランスクリプションを決定することを含む。

10

【0016】

この態様は、以下の任意の機能のうちの1つまたは複数を含み得る。いくつかの実装形態では、音声認識モデルは、第1のエンコーダ、第1の注意モジュール、書記素エンコーダ、音素エンコーダ、第2の注意モジュール、およびデコーダを含む。第1のエンコーダは、音響特徴を入力として受信し、音響特徴からオーディオエンコーディングを出力として生成するように構成されている。第1の注意モジュールは、第1のエンコーダから出力されたオーディオエンコーディングを入力として受信し、第1の注意出力を出力として生成するように構成されている。書記素エンコーダは、1つまたは複数のバイアスフレーズのセット内の各単語の書記素を示す書記素データを入力として受信し、書記素エンコーディングを出力として生成するように構成されている。音素エンコーダは、1つまたは複数のバイアスフレーズのセット内の各単語の音素を示す音素データを入力として受信し、音素エンコーディングを出力として生成するように構成されている。第2の注意モジュールは、書記素エンコーダから出力された書記素エンコーディングおよび音素エンコーダから出力された音素エンコーディングの表現を入力として受信し、第2の注意出力を出力として生成するように構成されている。デコーダは、第1の注意出力および第2の注意出力に基づいて、音声要素のシーケンスの可能性を決定するように構成されている。

20

【0017】

いくつかの例では、1つまたは複数のバイアスフレーズのセット内の各バイアスフレーズの特定の単語ごとに、書記素エンコーダは、特定の単語に対応する書記素エンコーディングを生成するように構成されており、音素エンコーダは、特定の単語に対応する音素エンコーディングを生成するように構成されており、第2の注意モジュールは、特定の単語に対応する書記素および音素エンコーディングに基づいて、特定の単語の対応するコンテキストバイアスベクトルを備える、対応する第2の注意出力をエンコードするように構成されている。さらに、例として、書記素エンコーダから出力される書記素エンコーディングおよび音素エンコーダから出力される音素エンコーディングの表現は、書記素エンコーディングと音素エンコーディングとの間の連結を表す射影ベクトルを含む。さらに、第1のエンコーダ、第1の注意モジュール、書記素エンコーダ、音素エンコーダ、第2の注意モジュール、およびデコーダは、音響特徴フレームのシーケンスから書記素のシーケンスを予測するように共同でトレーニングされ得る。

30

【0018】

40

50

いくつかの実装形態では、第1の注意モジュールは、デコーダの以前の非表示状態と、第1のエンコーダによって出力されたオーディオエンコーディングの完全なシーケンスとの関数として注意を計算するように構成されており、第2の注意モジュールは、デコーダの以前の非表示状態と、1つまたは複数のバイアスフレーズのセット内のバイアスフレーズから得られる書記素と音素データの両方の特性を表す射影ベクトルの完全なシーケンスの関数として注意を計算するように構成されている。追加の実装形態では、第2の注意モジュールは、1つまたは複数のバイアスフレーズのセット内のバイアスフレーズのいずれにも対応しないコンテキストバイアスベクトルを受信するように構成されており、コンテキストバイアスベクトルは、音声認識モデルの出力にバイアスをかけないオプションを表す。いくつかの例では、音声要素は書記素を含む。他の例では、音声要素は単語または単語ピースを含む。任意で、音声要素は音素を含み得る。

【 0 0 1 9 】

1つまたは複数のバイアスフレーズのセットは、特定のユーザのためにパーソナライズされた1つまたは複数の連絡先名を含み得る。追加的または代替的に、1つまたは複数のバイアスフレーズのセットは、特定のユーザのためにパーソナライズされた1つまたは複数のカレンダイベントを含み得る。いくつかの例では、本動作はまた、発話を話したユーザの位置、発話を話したユーザに関連付けられるユーザデバイスで開いている1つまたは複数のアプリケーション、あるいは発話の現在の日付および/または時刻のうちの少なくとも1つに基づいて発話のコンテキストを決定することを含む。発話のコンテキストは、追加的または代替的に、ユーザによって発行された1つまたは複数の以前のコマンドに基づき得る。

【 0 0 2 0 】

いくつかの実装形態では、音声認識モデルは、埋め込みベクトルと、デコーダの以前の非表示状態と、第1のベクトルと、第2のベクトルとにに基づいて、非表示状態および音声認識モデルの出力を決定するように構成されたデコーダを含む。ここで、埋め込みベクトルは、音声認識モデルによって出力された以前の書記素に対するものであり、第1のベクトルは第1の注意モジュールによって出力され、第2のベクトルは第2の注意モジュールによって出力される。

【 0 0 2 1 】

本開示の1つまたは複数の実装形態の詳細は、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】

【 0 0 2 2 】

【図1】発話を関連するコンテキストに基づいて音声認識結果にバイアスをかける音声認識モデルを含む、例示的な自動音声認識システムの概略図である。

【図2】バイアスフレーズを選択するための例示的な選択プロセスの概略図である。

【図3】図1の音声認識モデルの例示的なアーキテクチャの概略図である。

【図4】図1の音声認識モデルをトレーニングするための例示的なトレーニングプロセスの概略図である。

【図5】音声認識モデルをコンテキスト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である。

【図6】本明細書で説明されるシステムおよび方法を実装するために使用され得る例示的なコンピューティングデバイスの概略図である。

【発明を実施するための形態】

【 0 0 2 3 】

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。

【 0 0 2 4 】

従来の自動音声認識(ASR)システムは、音響モデル、発音モデル、および言語モデルの3つのモデルに依存している。発音モデルの設計上の決定は、音響モデルと言語モデルとの

10

20

30

40

50

両方に影響を与える。発音モデルの場合、考えられるすべての異なる音の単位を表すために音素のセットが選択され、音響モデルが予測する必要のあるラベルのセットを決定する。言語モデルにおけるすべての単語には発音がある。

【 0 0 2 5 】

正確な発音モデルは広い範囲をカバーしている。言語の所与のバリエントでは、音素セットは地域ごとに異なる。たとえば、アメリカ英語では、ネイティブスピーカは「ペン(pen)」および「ピン(pin)」などの単語が同じように発音されるかどうかについては同意しない。さらに、特定の単語は省略された発音を有する場合があり、この場合、話者は「トウェンティ(twenty)」などの単語における文字「t」を発音することもあり発音しないこともある。「ザ(the)」のような非常に一般的な単語の場合、発音をハードコード化するの 10 は不自然である。

【 0 0 2 6 】

エンドツーエンド(E2E)音声認識モデルは、音響、発音、および言語モデルを単一のニューラルネットワークに結合する。単一のニューラルネットワークモデルは、単純さと品質を向上させ、単語誤り率(WER)を最適化する。しかしながら、E2E音声認識モデルにおける課題は、言語でめったに出現しない単語、および/またはスペルに比べて異常な発音を有する単語を認識する際のパフォーマンスを最適化することである。トレーニングデータは、人間がトランスクライブしたボイスデータとテキストのみのデータとの両方を含むことができるが、これらのE2E音声認識モデルをトレーニングするために大規模なトレーニングデータセットを使用することは非効率的である。言語における単語の分布は、通常ジップの分布に従い、ジップの分布では少数の単語が非常に頻繁に使用され、膨大な数の単語は使用されることはないので、トレーニング例の数を増やすと、典型的には改善の度合いがますます低くなる。 20

【 0 0 2 7 】

ニューラルネットワークASRモデルにコンテキストバイアスを組み込むと、まれな単語および異常な発音の単語の認識を向上させることができる。たとえば、ユーザの連絡先はスマートフォンに記憶されることが多いため、ASRシステムがユーザによって話された連絡先の名前を認識することを支援するためのコンテキストとして使用することができる。バイアスコンテキストと発音との両方をモデルに注入することによって、コンテキストバイアスをASRモデルに適用することができる。コンテキスト的にバイアスされたモデルは、まれな単語の発音の単純で統一されたトレーニングおよび暗黙の学習を含む、ニューラルネットワークモデルの利点を保持している。コンテキスト的にバイアスされたモデルは、たとえトレーニング中にその単語が存在したことがない場合でも、まれな単語の発音の知識を組み込んでいる。 30

【 0 0 2 8 】

図1を参照すると、例示的なASRシステム100は、音声認識精度を改善するためにコンテキストバイアスを組み込んだ音声認識モデル300を含む。音声認識モデル300は、オーディオエンコーダ110、書記素エンコーダ122、音素エンコーダ126、オーディオ注意モジュール132、コンテキストバイアス注意モジュール134、および、オーディオ注意モジュール132およびコンテキストバイアス注意モジュール134との両方からの出力を入力として受信するデコーダ142を含む。いくつかの例では、デコーダ142の出力は、ソフトマックス層144で正規化され、それによって、ソフトマックス層144は、出力書記素のシーケンスにわたる確率分布を出力する。ソフトマックス層144は、書記素のセットなどの言語単位(たとえば、音声要素)のセットにわたる確率分布を提供し得る。たとえば、ソフトマックス層144の出力は、言語における書記素、ならびに潜在的に他の記号(たとえば、句読点、スペースなど)の各々の確率値を有するベクトルであり得る。ソフトマックス層144からのベクトルのシーケンスは、ユーザデバイス106によって記録された音声発話104のトランスクリプション150を生成するために使用される。いくつかの例では、音声認識モデル300は、ユーザ102に関連付けられるユーザデバイス106上に存在する。他の例では、音声認識モデル300は、ユーザデバイス106と通信するリモートサーバ上に存在するか 40

10

20

30

40

50

、または音声認識モデル300の機能が、リモートサーバとユーザデバイス106との間で分割される。

【 0 0 2 9 】

図示される例において、ユーザ102は、ユーザデバイス106の1つまたは複数のマイクロフォンによってキャプチャされた発話104を話す。ユーザデバイス106は、スマートフォン、タブレット、スマートヘッドホン、スマートウォッチなどのモバイルデバイスを含み得る。ユーザ102によって話される発話104は、「ベア裁判所にナビゲートしてください(Navigate to Bexar Courthouse)」というコマンドであり得、ここで、「ベア(Bexar)」は、「熊(bear)」および「むきだしの(bare)」という単語に類似した、/ber/と発音される郡の名前である。ユーザデバイス106の1つまたは複数のマイクロフォンは、話されたコマンドからオーディオ信号108を生成する。オーディオ信号(交換可能に「オーディオデータ」と呼ばれる)は、発話104をエンコードし、音声認識モデル300のニューラルネットワーク要素の入力を決定するために様々な方法のいずれかで処理することができる。たとえば、特徴抽出モジュール(図示せず)は、オーディオ信号108の異なるウィンドウまたはフレームのメル周波数ケプストラム成分などの、オーディオ信号108から導出された音響特徴109を生成することができる。

【 0 0 3 0 】

ユーザデバイス106はまた、音声認識モデル300にコンテキスト情報114を提供する。コンテキスト情報114は、たとえば、ユーザの位置、デバイス106上で開かれているアプリ、ユーザの連絡先、および今後のカレンダの予定を含むことができる。この場合、コンテキスト情報114は、ベア郡内のユーザの現在のGPS位置、ユーザデバイス106上で開かれるナビゲーションアプリケーション107、および裁判所におけるユーザの今後のカレンダの予定を含み得る。カレンダの予定は、ユーザデバイス106上で実行可能なカレンダアプリケーション107にアクセスすることによって取得することができる。ユーザ102は、各タイプのコンテキスト情報114を個別に共有することに同意することに明示的にオプトインし得、1つまたは複数のタイプのコンテキスト情報114の共有に以前に同意したことを取り消すためにいつでもオプトアウトし得る。

【 0 0 3 1 】

音声認識モデル300は、特定のバイアスフレーズ116を選択するために、ユーザデバイス106から受信したコンテキスト情報114を使用し、その各々は、書記素データ(grapheme data)118と音素データ(phoneme data)120との両方を含み得る。書記素データ118は書記素エンコーダ122に入力され、音素データ120は、音素エンコーダ126に入力される。バイアスフレーズ116の選択に関する追加の詳細は、図2に提供される。

【 0 0 3 2 】

いくつかの実装形態では、オーディオエンコーダ110、書記素エンコーダ122、および音素エンコーダ126は、長短期記憶(LSTM)要素または他のリカレントニューラルネットワーク(RNN)要素を有するシーケンスモデルを含む。オーディオエンコーダ110は、オーディオ信号108に基づいてエンコードされたオーディオベクトル(たとえば、オーディオエンコーディング)112を生成するように構成された第1のエンコーダを含み得る。たとえば、オーディオエンコーダ110は、発話104をエンコードするオーディオ信号108から導出された音響特徴109を受信し、音響特徴109からエンコードされたオーディオベクトル112を生成し得る。エンコードされたベクトル112は、オーディオ注意モジュール132に入力される。書記素エンコーダ122は、選択されたバイアスフレーズ116に関連付けられる書記素データ118からエンコードされた書記素ベクトル124を生成するように構成されるが、音素エンコーダ126は、選択されたバイアスフレーズ116に関連付けられる音素データ120からエンコードされた音素ベクトル128を生成するように構成される。図示されている例では、音声認識モデル300は、エンコードされた書記素ベクトル124および、エンコードされた音素ベクトル128を、コンテキストバイアス注意モジュール134に入力される射影ベクトル130に連結する。したがって、射影ベクトル130は、エンコードされた書記素ベクトル124および音素ベクトル128の表現に対応し得る。

【 0 0 3 3 】

オーディオ注意モジュール132は、注意の重みを計算し、モデル300が現在アテンドしているオーディオフレームのベクトル要約を含む加重オーディオエンコーディングベクトル136を生成するように構成される。たとえば、オーディオエンコーダ110は、追加の音響特徴109が受信されると、オーディオ信号108のウィンドウまたはフレームごとにオーディオベクトル112を生成することができ、したがって、より多くのオーディオが受信されると、追加のオーディオベクトル112を提供し続ける。いくつかの例では、オーディオ注意モジュール132は、以前の時間ステップからのデコーダオーディオ状態141およびエンコードされたオーディオベクトル112を使用して、オーディオベクトル112のシーケンスに対する注意の重みを計算する。デコーダオーディオ状態141は、最新のデコーダ出力を出力するときのデコーダ142の状態の一部、たとえば、直前のウィンドウまたはオーディオのフレームを表す以前の時間ステップのデコーダ状態の一部を表し得る。デコーダ状態141は、現在の時間ステップにおいてオーディオ注意モジュール132に入力されるオーディオベクトル112の相対的な重要性をオーディオ注意モジュール132に通知するために役立つ。デコーダオーディオ状態141は、デコーダ142の出力の指定されたサブセット、典型的にはデコーダコンテキスト状態143において使用されるものとは異なる出力のセットなどの、デコーダ142の出力の一部であり得る。いくつかの実装形態では、状態ベクトル141、143は、デコーダ142の出力の重複しない部分を表す。他の実装形態では、状態ベクトル141、143は、1つまたは複数の重複するセクションを含むか、または両方の状態ベクトル141、143は、それぞれデコーダ142の出力全体を表す。

10

【 0 0 3 4 】

同時に、コンテキストバイアス注意モジュール134は、射影ベクトル130に含まれるエンコードされた書記素ベクトル124と音素ベクトル128との両方を使用して注意の重みを計算するように構成される。コンテキストバイアス注意モジュール134は、コンテキストバイアスベクトル(contextual biasing vector)138として出力される書記素埋め込みの加重和のみを計算するために、これらの注意の重みを使用する。より具体的には、コンテキストバイアス注意モジュール134は、コンテキストバイアスベクトル138を生成するために、以前の時間ステップからの射影ベクトル130およびデコーダコンテキスト状態143を処理する。デコーダコンテキスト状態143は、以前のコンテキストバイアスベクトル138に関して以前の時間ステップにおけるデコーダ142の状態を含む。デコーダコンテキスト状態143は、たとえば、デコーダ142の出力の指定されたサブセットなどのデコーダ142の出力の一部を表し得る。デコーダコンテキスト状態143は、現在の時間ステップにおいてコンテキストバイアス注意モジュール134に入力された射影ベクトル130の相対的な重要性をコンテキストバイアス注意モジュール134に通知する。したがって、デコーダコンテキスト状態143が変化すると、コンテキストバイアス注意モジュール134は、射影ベクトル130の異なる要約または集約を計算し、したがって、デコーダ142が異なるバイアスフレーズ116の要素に対して示す確率を変更する。

20

【 0 0 3 5 】

コンテキストバイアス注意モジュール134に書記素埋め込みの加重和のみを計算させることにより、モデル300はスペルと発音との間の事前の関係を学習せず、それによって、モデルが任意の音素のシーケンスを任意の単語に割り当てる 것을可能にする。しかしながら、コンテキストバイアス注意モジュール134は、注意の重みを計算するために、エンコードされた書記素ベクトル124に加えて、エンコードされた音素ベクトル128を使用するので、注意の重みから計算された書記素埋め込みの加重和は、書記素情報に加えて音素情報を含む。

30

【 0 0 3 6 】

特に、コンテキストバイアス注意モジュール134から出力されるコンテキストバイアスベクトル138は、音声認識モデル300のコンテキストバイアスを表し、書記素情報と音素情報との両方を含む。したがって、コンテキストバイアスベクトル138を加重オーディオエンコーディングベクトル136と連結して加重ベクトルにすると、コンテキストバイアス

40

50

が音声認識モデル300に「注入」される。加重ベクトル140は、オーディオ、書記素、および音素情報を集合的に表す。加重ベクトル140は、デコーダ142に入力される。

【 0 0 3 7 】

いくつかの構成では、デコーダ142は、LSTMネットワークなどのシーケンスモデルを含み、加重ベクトル140から出力シーケンスを抽出するように構成される。デコーダ142の出力はまた、図示の例において書記素である出力ターゲットのセットにわたる確率分布を生成するために、ソフトマックス層144で正規化される。他の例では、確率分布の出力ターゲットのセットは、単語、単語ピース、または音素を含み得る。ビーム探索プロセスは、発話104のトランスクリプション(transcription)146を生成するために、デコーダ142からの様々な出力ベクトルを使用し得る。現在の時間ステップにおいてどの書記素がオーディオを表すかについて行われた決定は、音声要素出力148としてデコーダ142にフィードバックされ、デコーダ142からの次の出力を計算するために使用される。図示される例において、音声要素出力148は書記素出力を含む。他の例では、音声要素出力148は、単語出力または単語ピース出力を含む。さらに他の例では、音声要素出力148は音素出力を含む。

【 0 0 3 8 】

図1の例において、ユーザ102の位置、開いているアプリ、および今後のカレンダの予定のコンテキスト情報114は、音声認識モデル300をバイアスフレーズ「ベア」および発音/ber/にバイアスをかける。これにより、音声認識モデル300は、ユーザ102のコマンドを「ベア裁判所にナビゲートしてください」として正しくトランスクリプトすることができる。次いで、ユーザデバイス106は、ナビゲーションアプリ内の裁判所へのルートのマッピングなど、このトランスクリプションに基づいてアクションを実行することができる。

【 0 0 3 9 】

図2は、発話104のコンテキストに関する1つまたは複数のバイアスフレーズ116のセットを生成するための例示的なプロセス200の概略図を示している。プロセス200は、音声発話104のコンテキスト202を識別することと、書記素データ118および音素データ120の両方を含むバイアスフレーズ116を生成するためにデータソース210からコンテキスト情報114を選択することとを含む。1つまたは複数のバイアスフレーズのセット内の各バイアスフレーズ116は、1つまたは複数の単語を含み、その結果、書記素データ118は、バイアスフレーズ116の各単語に関連付けられる書記素を含み、音素データ118は、バイアスフレーズ116の各単語に関連付けられる音素を含む。

【 0 0 4 0 】

より詳細には、プロセス200は、ユーザが発話104を話すときに開始する。ASRシステムが、話された発話104からオーディオ入力を受信すると、発話104の1つまたは複数のコンテキスト202が決定される。コンテキスト202は、コンテキストバイアスにおいて使用するために選択されるバイアスフレーズ116を決定する。

【 0 0 4 1 】

コンテキスト202は、たとえば、ユーザの位置であってよい。ユーザの位置は、GPS座標によって決定することができる。たとえば、ユーザが自宅、職場、またはユーザの自宅もしくは職場に近い地理的位置にいるというコンテキストが考えられる。ユーザの位置は、レストラン、コンサートホール、またはスタジアムなどの会場として識別される場合もある。ユーザが旅行している場合、ユーザがおそらく見知らぬ都市または外国の都市に休暇または仕事のために旅行しているというコンテキストが考えられる。

【 0 0 4 2 】

コンテキスト202の別の例は、ユーザのモバイルデバイスで開いているアプリである。ユーザが電話アプリを開くと、連絡先に電話をかけようとしている可能性がある。ユーザがナビゲーションアプリを開くと、ユーザが目的地にナビゲートしようとしている可能性がある。

【 0 0 4 3 】

10

20

30

40

50

コンテキスト202の例はまた、話された発話の日付および時刻であり得る。たとえば、月曜日の早朝の場合、ユーザは車で通勤する準備をしている可能性がある。したがって、ユーザは、天候または交通状況を尋ねたり、オフィスにナビゲートしたりするなど、彼または彼女が就業日の準備のために役立つ発話104を話す場合がある。

【0044】

選択モジュール203は、様々なデータソース210から情報114を選択するために、コンテキスト202を参照する。データソース210は、たとえば、連絡先204、カレンダイベント206、および以前に話されたコマンド208を含むことができる。一例では、コンテキスト202がオープン電話アプリを含む場合、選択モジュール203は、記憶された連絡先204から情報114を選択し得る。コンテキスト202がオープンナビゲーションアプリを含む場合、選択モジュール203は、今後のカレンダイベントの場所に関する情報114を選択し得る。選択モジュール203はまた、以前のコマンド208を参照し得る。たとえば、ユーザは、平日の毎朝に仕事にナビゲートするように要求し得る。したがって、月曜日の朝に話される発話104は、ユーザの職場に関連してユーザによって最近話された単語が含まれる可能性が高い。コンテキスト情報114はまた、話された発話104に関連する1つまたは複数のコンテキスト202を指定し得る。

【0045】

選択モジュール203によって取得/選択された情報114は、バイアスフレーズ116を作成するために使用される。バイアスフレーズ116は、書記素データ118と音素データ120との両方を含む。音素データ120は、発音辞書216を使用して生成され得る。

【0046】

例として、ユーザはほとんどの平日の朝に家を出て、メリーランド州ボウイに車で行き、デイケアで子供を降ろし、次いで、メリーランド州ボルチモアに仕事に行く場合がある。ユーザが「メリーランド州ボウイにナビゲートしてください(Navigate to Bowie, Maryland)」という発話104を話すとき、発話104のコンテキスト202は、メリーランド州の自宅にいるユーザの位置を含むことができる。コンテキスト104はまた、開いているナビゲーションアプリ、および月曜日の朝であり得る日付および時刻を含むことができる。選択モジュール203は、データソース210にアクセスする。カレンダイベント206は、ユーザがその日遅くにメリーランド州ボルチモアで会議があることを示している。以前のコマンド208は、ほとんどの平日の朝に、ユーザがボウイ、次にボルチモアにナビゲートするように要求することを示している。選択モジュール203は、バイアスフレーズ212を作成するために、このコンテキスト情報114を使用する。

【0047】

この例では、ユーザがナビゲーションアプリを開いているので、書記素バイアスデータ118は、「ボルチモア(Baltimore)」、「メリーランド(Maryland)」、および「ボウイ(Bowie)」という単語を含み得、これらの3つの単語は、ユーザの可能性のある目的地に関連している。発音辞書216は、音素バイアスデータ120を作成するために使用される。ユーザデバイス106は、発音辞書216にアクセスし、単語の一般的な発音を表す音素のシーケンスなどの発音を検索する。発音辞書は、単語の発音のバリエーションを含み得る。たとえば、ボルチモア出身の人は、別の州から訪問している人とは「ボルチモア」の発音が異なる場合がある。同様に、話されている言語が母国語ではない国を旅行者が訪問している場合、地名の旅行者の発音は、ネイティブスピーカの発音とは異なる可能性がある。

【0048】

このプロセス200は、書記素バイアスデータ118および音素バイアスデータ120を生成する。次いで、バイアスフレーズ116は、図1に示されるように、ASRモデル300に注入するために書記素エンコーダ122および音素エンコーダ126に供給される。

【0049】

図3は、書記素および音素を使用するコンテキストバイアスのための図1のASRシステム100の音声認識モデル300の例示的なアーキテクチャを示している。図3に示されるモデル300の例示的なアーキテクチャは、リッスン、アテンド、およびスペル(LAS)モデルを

10

20

30

40

50

含む。LASアーキテクチャは、オーディオエンコーダ(たとえば、第1のエンコーダ)110と、オーディオ注意モジュール(たとえば、第1の注意モジュール)132と、デコーダ142との3つの主要なコンポーネントを含む。オーディオエンコーダ110は、入力音声オーディオ信号108の時間周波数表現(すなわち、音響特徴)109を受信するように構成され、入力109をより高いレベルの特徴表現、すなわち、エンコードされたオーディオベクトル(たとえば、オーディオエンコーディング)112にマッピングするために、ニューラルネットワーク層のセットを使用する。オーディオエンコーダ110のエンコードされたオーディオベクトル112出力は、オーディオ注意モジュール132に渡され、オーディオ注意モジュール132は、入力音響特徴109と予測された書記素148との間のアラインメントを学習するために、エンコードされたオーディオベクトル112を使用する。注意オーディオモジュール132の出力136は、デコーダ142に渡され、デコーダ142は、仮定された書記素のセットにわたる確率分布を生成する。LASモデルのコンポーネントは、単一のエンドツーエンドニューラルネットワークとして共同でトレーニングされる。

【0050】

例示的なLASアーキテクチャは、モデル300が所与のコンテキスト202に関連するバイアスフレーズ116に向けて音声認識結果146、148にバイアスをかけることを可能にする追加のコンポーネントを含むように拡張される。これらの追加のコンポーネントは、書記素エンコーダ(たとえば、第2のエンコーダ)122、音素エンコーダ(たとえば、第3のエンコーダ)126、およびバイアス注意モジュール(たとえば、第2の注意モジュール)134を含む。

【0051】

音声認識モデル300の拡張アーキテクチャは、オーディオ、書記素、および音素入力 x_u 、 g_n 、および p_n を受け入れ、下付き文字はU個のオーディオフレーム109およびN個のバイアスフレーズ116に対応する。エンコーダ110、122、126は、それぞれの入力をエンコーディング h^x 、 h^g 、および h^p にエンコードし、次いで、これらは、対応する注意モジュール132、134に供給される。次に、注意モジュール132、134による注意は、デコーダ142に供給される対応するコンテキスト注意ベクトル ct^x および ct^g を出力する。デコーダ142は、次の注意の出力を計算するための次のデコーディングステップに使用される dt^x および dt^g を出力する。例示的なLASモデル300の出力は、1からTまでのすべてのデコーディングステップtに対するベクトル y_t である。

【0052】

いくつかの実装形態では、オーディオエンコーダ110は、オーディオフレーム $x_1 \dots x_U$ (たとえば、対数メル特徴などの音響特徴109)を入力として受信するように構成された双方向LSTMを含む。一例では、オーディオエンコーダ110は、10個のLSTM層を含み、それが256ノードを有し、非表示サイズは1400である。さらに、書記素エンコーダ122および音素エンコーダ126はそれぞれ、512寸法の単層LSTMを含み得る。書記素エンコーダ122および音素エンコーダ126はそれぞれ、N個のランダムにシャッフルされたバイアスフレーズ116に関連付けられるそれぞれの書記素データ118および音素データ120を入力として受信し得る。したがって、発音情報を伝達する音素入力(たとえば、音素データ)120の数は、書記素入力(たとえば、書記素データ)118の数と同一であり得、エンコーダ122、126から出力される対応する埋め込み/エンコーディング124、128は互いに独立して計算される。

【0053】

いくつかの例では、オーディオ注意モジュール132は、512の座標を有する4つのヘッドを含む。図示される例では、オーディオ注意モジュール132は、オーディオフレーム109のエンコーディングを含むオーディオエンコーディング112、 h^x と、以前のタイムステップにおけるデコーダ142の出力141、 dt_{t-1} との2つの入力を受け取る。上付き文字 x は、出力141が、最新のデコーダ出力、たとえば、直前のウィンドウまたはオーディオのフレームを表す以前の時間ステップのデコーダ状態の一部を出力するときのデコーダ142の状態の一部を表すデコーダオーディオ状態に対応することをオーディオ注意モジュール132に通知する。したがって、デコーダオーディオ状態に関連付けられる出力141は、現在

10

20

30

40

50

の時間ステップにおいてオーディオ注意モジュール132に入力されるオーディオベクトル h^x 、 h^y の相対的な重要性をオーディオ注意モジュール132に通知するために役立つ。デコーダオーディオ状態に対応する出力141は、デコーダ142の出力の指定されたサブセットなどの、デコーダ142の出力の一部であり得、典型的には、デコーダのコンテキスト状態に対応する出力143において使用されるものとは異なる出力のセットである。

【0054】

他方、バイアス注意モジュール134は、書記素入力118および音素入力120に基づいて、書記素エンコーダ122から出力された書記素エンコーディング(grapheme encoding)124、 h^g と、音素エンコーダ126から出力された音素エンコーディング128、 h^p との間の連結に対応するエンコードされた射影130を入力として受信するように構成される。射影層132は、エンコーディング124、128を連結し、対応するエンコードされた射影130を出力し得る。より具体的には、バイアス注意モジュール134は、すべてのデコーディングステップ t において、エンコードされた射影130の加重和にアテンドする。ここで、バイアス注意モジュール134は、注意の重みを計算するために、書記素エンコーディング124と音素エンコーディング128との両方を使用し得る。注意の重みが計算されると、それらは、書記素埋め込みの加重和 c_t^g を計算するために使用される。したがって、バイアス注意モジュール134の出力138は、書記素エンコーディング124、 h^g の加重和 c_t^g のみを含む。

【0055】

より詳細には、バイアス注意モジュール134の出力138は、3つのステップで計算される。第1のステップにおいて、バイアス注意モジュール134は、書記素エンコーディング h_i^g 、音素エンコーディング h_i^p 、および以前のデコーディングステップ、 d^g_{t-1} からの出力143を入力として受信する。上付き文字 g は、出力143が、最新のデコーダ出力を出力するときのデコーダ142の状態の一部を表すデコーダ書記素状態に対応することをバイアス注意モジュール134に通知する。いくつかの例では、バイアス注意モジュール134は、すべてのバイアス単語およびすべてのデコーディングステップ u^g_{it} の重みのセットを次のように計算する。

【0056】

【数1】

$$u_{it}^g = v^{gT} \cdot \tanh(W_h^g h_i^g + W_h^p h_i^p + W_d^g d_t + b^g) \quad (1)$$

10

20

30

【0057】

第2のステップにおいて、 c_t^g を取得するために、次のように出力がソフトマックス144で正規化される。

【0058】

【数2】

$$\alpha_t^g = \text{softmax}(u_t^g)$$

【0059】

第3のステップにおいて、出力138は、埋め込み h_i^g の加重和として計算される。

40

【0060】

【数3】

$$c_t^g = \sum_{i=1}^N \alpha_{ti}^g h_i^g$$

【0061】

バイアス注意出力(bias attention output)138、 c_t^g は、書記素の埋め込みを表している。したがって、いくつかの例では、デコーダ142は、書記素のみで動作するように構成

50

されており、したがって、デコーダ入力 c_t^g は、書記素のみに関するものである。これにより、スペルから発音を予測するのが難しい単語の注入が可能になる。この手段では、スペルと発音との間の事前の関係はなく、任意の音素のシーケンスを任意の単語に割り当てることができる。

【 0 0 6 2 】

しかしながら、デコーダ142が、単語ピースなどの書記素以外の出力ターゲットのセットにわたって確率分布を生成するように構成されている構成においては、バイアス注意出力138は、デコーダ142が生成するように構成されている出力ターゲットのタイプに関してであり得る。任意で、バイアス注意モジュールの出力は、音素の埋め込み、たとえば、音素の加重和を表し得、デコーダ142は、代わりに、音素に関連付けられる出力ターゲットのセットを生成するためにのみ音素で動作し得る。

10

【 0 0 6 3 】

いくつかの例では、デコーダ142は、4つのLSTM層を含み、それぞれが256のサイズを有し、1024の非表示寸法を有する。図示されている例では、デコーダ142は3つの入力を受信する。第1の入力は、以前の書記素出力138、 y_{t-1} である。予測を行うためにモデルが使用される場合、 y_{t-1} はソフトマックス144の出力である。モデルがトレーニング中の場合、 y_{t-1} は、真の書記素のワンホットエンコーディング、または出力のサンプリングと真の書記素のワンホットエンコーディングのいずれかであってよい。デコーダ142への別の入力136、 c_t^x は、モデル300が現在注目しているオーディオフレーム109の要約を含む。デコーダ142への第3の入力138は、バイアス注意出力 c_t^g を含む。

20

【 0 0 6 4 】

図3を引き続き参照すると、例示的なLASモデル300の出力は、1からTまでのすべてのデコーディングステップ t のためのベクトル y_t を含む。ASRにおいて、ベクトル y_t は、出力シンボル(書記素、スペース、およびシーケンスの終わり)のワンホットエンコーディングに対応する。モデル300は、T番目のステップにおいてシーケンスの終わりのシンボルを予測するまで、T-1のシンボルを出力する。

【 0 0 6 5 】

いくつかの実装形態では、トレーニングプロセスは、トレーニング例の複数のバッチで音声認識モデル300をトレーニングする。ここでは、バッチごとに、参照トランスクリプトが確率 P_{keep} でランダムに保持される。これにより、一致するバイアスがない例が可能になる。残りの参照トランスクリプトについては、Nグラムがランダムに選択され、ここで、nは[1:N_{order}]から均一にサンプリングされる。選択されたNグラムの数はNであり、ここで、Nは[1:N_{phrases}]から均一にサンプリングされる。値の例は、 $P_{keep}=0.5$ 、 $N_{phrases}=1$ 、および $N_{order}=4$ である。したがって、各バッチにおけるバイアスフレーズ116の量は可変であり、モデル300のトレーニング時に指定されない。さらに、トレーニング中に適用される1つまたは複数のバイアスフレーズ116は、他のバイアスフレーズ116とは異なる長さを有し得る。モデル300は、10ミリ秒ごとの対数メル音響特徴を使用してトレーニングすることができ、25ミリ秒のウィンドウに対して計算され、3つのグループでスタックされる。例示的な音響特徴ベクトルは80の座標を有する。

30

【 0 0 6 6 】

バイアスのある書記素と音素は空の場合、バイアスのないオプションを追加することができる。バイアスのないオプションは、書記素を含まないコンテキストバイアスベクトル138と、任意のバイアスフレーズ116に関連付けられるバイアスデータ118、120を適用することによって適用することができる。このサンプリング手法により、一部の例に一致する単語がないことを確認するが、他の例には一致する単語がある。モデルの収束を支援するために、特別なトークンが導入されている。一致するバイアストークンごとに、特殊文字 /bias が追加される。バイアスフレーズを使用することによってのみ修正できるバイアス誤りが発生する。

40

【 0 0 6 7 】

音素のトレーニングのために、辞書と書記素から音素(G2P)の組合せを使用することが

50

できる。単語が辞書にある場合、対応する発音を使用することができる。単語が辞書にならない場合は、発音を予測するために、G2Pを使用することができる。

【0068】

発音の知識を含むバイアスコンテキストをニューラルネットワークモデルに注入すると、バイアスコンテキストが注入されていない場合よりも14%の相対的な改善が得られる。コンテキスト的にバイアスされたモデルは、まれな単語の発音の単純で統一されたトレーニングおよび暗黙の学習を含む、ニューラルネットワークモデルの利点を保持する。同時に、コンテキストバイアスモデルは、たとえトレーニング中に存在したことがない場合でも、まれな単語の発音の知識を組み込む。しかしながら、コンテキストが同じように聞こえるフレーズを含む場合(たとえば、「ジョアン(Joan)」と「ジョン(John)」との両方がユーザの連絡先リストに含まれている場合)、正しいフレーズの曖昧性解消は依然として困難である。曖昧性解消は、難しい否定的な例を使用してASRコンテキストメカニズムをトレーニングすることによって改善することができる。

【0069】

上記のトレーニング例では、トレーニングバイアスフレーズは、参照トランスク립トおよびトレーニングデータ内の他の発話からNグラムをランダムにサンプリングすることによって作成される。テスト時には、バイアスフレーズは事前に修正されており、同じセマンティックカテゴリ(たとえば、連絡先名)からのものである。この場合、ジョアンやジョンなどのファーストネームを区別することは、通常、ランダムな無関係のNグラムを区別するよりも難しいため、テストタスクはトレーニングタスクよりも困難である。

【0070】

図4は、ランダムにサンプリングされたNグラムを用いたトレーニングから生じるトレインテストの不一致を閉じるために使用することができる例示的な方法400を示している。方法400は、ランダムなNグラムをサンプリングする代わりに、参照トランスク립トから固有名詞を検出し、トレーニング用のバイアスフレーズのセットを作成するために、それらを音声的に類似した、または「ファジー(fuzzy)」な代替物で補強する。

【0071】

方法400は、参照トランスク립トにおける固有名詞(たとえば、人および場所の名前などの一意のエンティティ)に焦点を当て、発音的に類似したフレーズを否定的な例として使用し、ニューラルネットワークモデルがより識別力のある表現を学習するように促す。この手法は、WERを最大53.1%向上させるために、正しいコンテキスト項目をトランスクライプして選択することを共同で学習するニューラルネットワークコンテキストASRモデルに適用することができる。

【0072】

ASRにおいて、2つのフレーズは音声的に(すなわち、発音の仕方が)互いに非常に似ている場合があるが、間違いなく異なる(たとえば、「ジョアンに電話する(call Joan)」と「ジョンに電話する(call John)」)。ニューラルネットワークASRモデルの場合、これらの名前の学習された表現は非常に類似している可能性があり、モデルが間違った表現を予測することになる。この問題は、まれでスペルが難しい単語が表示されるASRモデルにとって特に困難であり、これは、モデルがトレーニング中にこれらの単語をまったく観察しない可能性があり、したかってテスト時にこれらの単語を正しくスペルできないためである。

【0073】

否定的な例を使用したトレーニングは、これらのまれでスペルが難しいフレーズを音声的に類似したフレーズと区別するようにモデルに教える方法である。この手法を使用すると、モデルは、テスト時のパフォーマンスを向上させるために役立つ、より堅牢な表現を学習する。否定的な例を使用してトレーニングするための方法は(i)焦点のターゲットとして入力発話におけるトランスクライプが困難でまれな単語を検出するステップと、(ii)モデルに難しい否定的な例を提供することによって、これらの単語をより厳しくトレーニングするステップとを含む。

【0074】

10

20

30

40

50

((i))に焦点を当てるフレーズを検出する場合、固有名詞(「NNP」ともタグ付けされる)は、比較的簡単に検出できる一方で、まれで通常はトランスクライプトが難しいフレーズの一般的なカテゴリである。これらのフレーズ((ii))をより厳しくトレーニングするために、音声的に類似した代替フレーズが抽出され、否定的な例としてモデルに供給される。

【0075】

この手法は、音声のデータ拡張とを考えることができる。データ拡張は通常、突然変異した肯定的な例を生成するための機械学習において使用されるが、この手法は、突然変異した否定的な例を生成するために使用される。この手法は、WERを改善するために、コンテキスト的にバイアスされたLASモデルに適用される。

【0076】

図4の例では、発話「ジョアンのモバイルに電話する(Call Joan's mobile)」がASRモデルに提供される。例示的な発話410が提示されると、方法400は、参照トランスクライプト405から固有名詞を検出する(415)。

【0077】

この例では、「ジョアン(Joan)」420という単語がバイアスフレーズとして選択されている。「ジョアン(Joan)」420は人の名前であるため、固有名詞として簡単に識別される。固有名詞を識別するために、参照トランスクライプトは品詞タグ付け(part-of-speech tagger)で分析される。

【0078】

次いで、方法400は、代替物として、音声的に類似した、またはファジーなフレーズを追加する(425)。この場合、単語「ジョン(John)」および「ジーン(Jean)」430が、代替バイアスフレーズとして選択される。単語Nグラムは、両方が音声的に類似しており、トレーニングデータにおける同じ発話の異なるデコーディングビームにおいて頻繁に共起する場合、第2の単語Nグラムのファジーな代替として識別される。ファジーな代替物を識別するプロセスを以下に説明する。

【0079】

第1に、ファジーNグラムの代替物のセットであるNグラムごとに記憶するファジーインベントリが作成される。ファジーインベントリは、発話の大規模なコーパスをデコードするために外部の従来のモデルを使用することによって、教師なしの方法で構築される。デコードされた各発話は、仮説のセットになる。残りの仮説が同一である異なる仮説に現れる単語Nグラムのペアのすべての共起がカウントされ、各Nグラムのペアがそのカウントに従ってスコアリングされる。たとえば、「ジョンレモン(John Lemon)」は、同じ発話の異なる仮説において頻繁に見つかる場合、ターゲットNグラムの「ジョンレノン(John Lennon)」に対して高いスコアを取得する。

【0080】

次に、ファジーインベントリは、コンテキスト的にバイアスされたLASトレーニング中に使用される。具体的には、単語Nグラム(この場合はバイアスフレーズを表す)が与えられると、ファジーな代替がファジーインベントリから選択され、共起スコアによってソートされる。上位の候補は、ターゲットNグラムの音声に類似しているものだけを保持することによってフィルタリングされ、ここで、音声の類似性はHixonメトリックを使用して測定される。ファジーバイアスの代替物の選択は、データ準備フェーズの一部として行われる。

【0081】

抽出された固有名詞のセットとそれらのファジーな代替物は、バイアスフレーズ435として書記素エンコーダおよび音素エンコーダに供給される。このように、バイアスフレーズ435は、参照トランスクライプト405における例でASRモデルをトレーニングするために使用される。この手法は、モデルが、音声的に類似した名前を区別するという難しいタスクでトレーニングすることを可能にする。

【0082】

各トレーニング例から、最大3つの固有名詞がランダムに選択され、固有名詞ごとに3つ

10

20

30

40

50

のファジーな代替物が追加される。他のランダムな固有名詞が残りのトレーニングデータから追加され、例ごとに最大64個のバイアスフレーズが生成される。

【 0 0 8 3 】

テスト時に音声的に類似したフレーズが提示された場合、ファジーモデルは、非ファジーモデルよりも微妙な音声の違いをより適切にキャプチャする。これは、より正確な予測の両方によって表される。また、結果として、ファジーな代替物ではなく、参照トランスクリプトに実際に表示されるバイアスフレーズにより多くの注意が向けられる。したがって、ファジーなディストラクタを使用したトレーニングにより、モデルは音声的に類似したフレーズをより適切に識別できるようになる。さらに、ファジーモデルは、その注意に誤ったフレーズを含む、非ファジーモデルに比べてより鋭くアテンドし、その分布は、はるかにクリーンである。

10

【 0 0 8 4 】

難しい否定的な例を用いてコンテキスト化されたニューラル音声認識モデルをトレーニングすると、結果としてWERが改善される。中心的なアイデアは、参照トランスクリプト内の固有名詞(「NNP」)を検出して焦点を合わせ、それらの否定的な例として音声的に類似した(「ファジーな」)フレーズをモデルに提示することである。音声バイアスタスクに適用すると、この手法はWERを最大53.1%向上させる。

【 0 0 8 5 】

図5は、音声認識モデル300を話された発話104のコンテキスト202に関連する1つまたは複数のバイアスフレーズ116に向けてバイアスをかける方法500のための動作の例示的な配置のフローチャートである。方法500は、発話104を話したユーザ102に関連付けられるユーザデバイス106上にあるデータ処理ハードウェア610(図6)上で実行され得る。任意で、データ処理ハードウェア610は、たとえば、ネットワークを介して、ユーザデバイス106と通信するリモートデバイス(たとえば、クラウドベースのコンピューティング環境のサーバ)上に常駐し得る。動作502において、方法500は、データ処理ハードウェア610において、発話104をエンコードするオーディオデータ106を受信するステップを含み、動作504において、方法500は、データ処理ハードウェア610によって、発話104のコンテキスト202に対応する1つまたは複数のバイアスフレーズ116のセットを取得するステップを含む。ここで、1つまたは複数のバイアスフレーズ116のセット内の各バイアスフレーズ116は、1つまたは複数の単語を含む。本明細書で使用される場合、各バイアスフレーズ116の1つまたは複数の「単語」は、限定ではないが、ユーザの連絡先リスト204内の人の名前、都市名、音楽アーティスト名、アルバム/映画(または、他のメディアコンテンツタイプ)のタイトルなどの固有名詞を含み得る。コンテキスト202は、発話の現在の日付および/または時刻、ユーザデバイス106上で開かれている1つまたは複数のアプリケーション、ユーザ102の位置のうちの少なくとも1つに基づいて決定され得る。

20

【 0 0 8 6 】

動作506において、方法500は、音声認識モデル300の出力148を生成するために、データ処理ハードウェア610によって、音声認識モデル300を使用して、オーディオデータから得られる音響特徴109、ならびに1つまたは複数のバイアスフレーズ116のセットから得られる書記素および音素データ118、120を処理するステップを含む。出力148は、音響特徴109の各フレームの書記素出力を含み得る。動作508において、方法500は、データ処理ハードウェア610によって、音声認識モデル300の出力148に基づいて、発話104のトランスクリプション146を決定するステップを含む。トランスクリプション146は、データ処理ハードウェア610と通信するユーザデバイス106および/あるいは1つまたは複数の他のデバイスのグラフィカルユーザインターフェース106上に表示され得る。追加の例では、トランスクリプション146は、動作を実行するための検索エンジンまたはソフトウェアアプリケーションへのコマンドとして提供される。

30

【 0 0 8 7 】

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、コンピューティングデバイス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場合がある。いくつ

40

50

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。例示的なアプリケーションは、これらに限定されないが、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、ワードプロセッシングアプリケーション、スプレッドシートアプリケーション、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、およびゲームアプリケーションを含む。

【0088】

非一時的メモリは、コンピューティングデバイスによって使用するために一時的または永続的にプログラム(たとえば、命令のシーケンス)またはデータ(たとえば、プログラム状態情報)を記憶するために使用される物理デバイスであり得る。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り得る。不揮発性メモリの例は、これらに限定されないが、フラッシュメモリおよび読み取り専用メモリ(ROM)/プログラム可能読み取り専用メモリ(PROM)/消去可能プログラム可能読み取り専用メモリ(EPROM)/電子的に消去可能プログラム可能読み取り専用メモリ(EEPROM)(たとえば、通常、ブートプログラムなどのファームウェアに使用される)を含む。揮発性メモリの例は、これらに限定されないが、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)、ならびにディスクまたはテープを含む。

【0089】

図6は、本明細書で説明されるシステムおよび方法を実装するために使用され得る例示的なコンピューティングデバイス600の概略図である。コンピューティングデバイス600は、ラップトップ、デスクトップ、ワークステーション、携帯情報端末、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すことを意図している。本明細書に示されているコンポーネント、それらの接続と関係、およびそれらの機能は、単なる例示を意味するものであり、本明細書で説明および/または主張されている発明の実装形態を制限することを意味するものではない。

【0090】

コンピューティングデバイス600は、プロセッサ610、メモリ620、ストレージデバイス630、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ーフェース/コントローラ640、ならびに低速バス670およびストレージデバイス630に接続する低速インターフェース/コントローラ660を含む。コンポーネント610、620、630、640、650、および660の各々は、様々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に、または必要に応じて他の方法で取り付けられ得る。プロセッサ610は、高速インターフェース640に結合されたディスプレイ680などの外部入力/出力デバイス上のグラフィカルユーザインターフェース(GUI)のグラフィカル情報を表示するために、メモリ620またはストレージデバイス630に記憶された命令を含む、コンピューティングデバイス600内で実行するための命令を処理することができる。他の実装形態では、複数のメモリおよびメモリのタイプとともに、必要に応じて、複数のプロセッサおよび/または複数のバスが使用され得る。また、複数のコンピューティングデバイス600が接続され得、各デバイスは、必要な動作の一部を提供する(たとえば、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)。

【0091】

メモリ620は、情報を非一時的にコンピューティングデバイス600内に記憶する。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、または不揮発性メモリユニットであり得る。非一時的メモリ620は、コンピューティングデバイス600によって使用するために一時的または永続的にプログラム(たとえば、命令のシーケンス)またはデータ(たとえば、プログラム状態情報)を記憶するために使用される物理デバイスであり得る。不揮発性メモリの例は、これらに限定されないが、フラッシュメモリおよび読み取り専用メモリ(

10

20

30

40

50

ROM) / プログラム可能読み取り専用メモリ(PROM) / 消去可能プログラム可能読み取り専用メモリ(EPROM) / 電子的に消去可能プログラム可能読み取り専用メモリ(EEPROM)(たとえば、通常、ブートプログラムなどのファームウェアに使用される)を含む。揮発性メモリの例は、これらに限定されないが、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)、静态ディクラランダムアクセスメモリ(SRAM)、相変化メモリ(PCM)、ならびにディスクまたはテープを含む。

【 0 0 9 2 】

ストレージデバイス630は、コンピューティングデバイス600に大容量ストレージを提供することができる。いくつかの実装形態では、ストレージデバイス630は、コンピュータ可読媒体である。様々な異なる実装形態では、ストレージデバイス630は、フロッピーディスクデバイス、ハードディスクデバイス、光ディスクデバイス、あるいはテープデバイス、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、あるいはストレージエリアネットワークまたはその他の構成におけるデバイスを含む、デバイスのアレイであり得る。追加の実装では、コンピュータプログラム製品は、情報キャリアに具体的に具現化される。コンピュータプログラム製品は、実行されると、上記のような1つまたは複数の方法を実行する命令を含む。情報キャリアは、メモリ620、ストレージデバイス630、またはプロセッサ610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。

【 0 0 9 3 】

高速コントローラ640は、コンピューティングデバイス600の帯域幅を大量に消費する動作を管理し、一方、低速コントローラ660は、より低い帯域幅を大量に消費する動作を管理する。そのような役目の割当ては例示的なものにすぎない。いくつかの実装形態では、高速コントローラ640は、メモリ620、ディスプレイ680(たとえば、グラフィックプロセッサまたは加速器を通じて)、および様々な拡張カード(図示せず)を受け入れることができる高速拡張ポート650に結合されている。いくつかの実装形態では、低速コントローラ660は、ストレージデバイス630および低速拡張ポート690に結合されている。様々な通信ポート(たとえば、USB、ブルートゥース(登録商標)、イーサネット、ワイヤレスイーサネット)を含み得る低速拡張ポート690は、キー・ボード、ポインティングデバイス、スキャナ、あるいはネットワークアダプタを通じたスイッチまたはルータなどのネットワークデバイスなどの、1つまたは複数の入力/出力デバイスに結合され得る。

【 0 0 9 4 】

コンピューティングデバイス600は、図面に示されるように、いくつかの異なる形態で実装され得る。たとえば、それは、標準サーバ600aとして、またはそのようなサーバ600aのグループ内で複数回、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得る。

【 0 0 9 5 】

本明細書に記載のシステムおよび技法の様々な実装形態は、デジタル電子および/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/またはそれらの組合せにおいて実現することができる。これらの様々な実装形態は、ストレージシステム、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と命令を受信することと、それらにデータおよび命令を送信することとを行うために結合された、専用または汎用であり得る、少なくとも1つのプログラム可能なプロセッサを含む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における実装形態を含むことができる。

【 0 0 9 6 】

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、プログラム可能なプロセッサ用の機械命令を含み、高レベルの手続き型および/またはオブジェクト指向プログラミング言語において、および/またはアセンブリ/機械語において実装することができる。本明細書で使用される

10

20

30

40

50

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む、機械命令および/またはデータをプログラム可能なプロセッサに提供するために使用される、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、装置および/またはデバイス(たとえば、磁気ディスク、光ディスク、メモリ、プログラム可能な論理デバイス(PLD))を指す。「機械可読信号」という用語は、プログラム可能なプロセッサに機械命令および/またはデータを提供するために使用される任意の信号を指す。

【0097】

本明細書に記載のプロセス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ム可能なプロセッサによって実行することができ、入力データを操作し、出力を生成することによって機能を実行するために、1つまたは複数のコンピュータプログラムを実行する。プロセスおよび論理フローは、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)などの専用論理回路によって実行することもできる。コンピュータプログラムの実行に適したプロセッサは、例として、汎用マイクロプロセッサと専用マイクロプロセッサとの両方、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を含む。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およびデータを受信する。コンピュータの重要な要素は、命令を実行するためのプロセッサと、命令およびデータを記憶するための1つまたは複数のメモリデバイスである。一般に、コンピュータはまた、データを受信するか、またはデータを転送するか、あるいはその両のために、データを記憶するための1つまたは複数の大容量ストレージデバイス、たとえば、磁気ディスク、光磁気ディスク、または光ディスクを含むか、動作可能に結合される。しかしながら、コンピュータはそのようなデバイスが必要なわけではない。コンピュータプログラム命令およびデータを記憶するために適したコンピュータ可読媒体は、たとえば、半導体メモリデバイス、たとえば、EPROM、EEPROM、およびフラッシュメモリデバイス、磁気ディスク、たとえば、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、ならびにCD-ROMおよびDVD-ROMディスクなどの、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を含む。プロセッサおよびメモリは、専用論理回路によって補完することもでき、専用論理回路に組み込むこともできる。

【0098】

ユーザとの対話を提供するために、本開示の1つまたは複数の態様を、ディスプレイデバイス、たとえば、CRT(陰極線管)、LCD(液晶ディスプレイ)モニタ、またはユーザに情報を表示するためのタッチスクリーン、ならびに任意でユーザがコンピュータに入力を提供できるマウスまたはトラックボールなどのキーボードおよびポインティングデバイスを有するコンピュータに実装することができる。ユーザとの対話を提供するために他の種類のデバイスを使用することもでき、たとえば、ユーザに提供されるフィードバックは、視覚的フィードバック、聴覚的フィードバック、または触覚的フィードバックなど、任意の形態の感覚的フィードバックであってよく、ユーザからの入力は、音響、音声、触覚入力などを含む、あらゆる形式で受信することができる。さらに、コンピュータは、ユーザによって使用されるデバイスとの間でドキュメントを送受信することによって、たとえば、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、ユーザと対話することができる。

【0099】

いくつかの実装形態が説明してきた。それにもかかわらず、本開示の趣旨および範囲から逸脱することなしに、様々な変更が行われてよいことが理解されるであろう。したがって、他の実装形態は、以下の特許請求の範囲内にある。

【符号の説明】

【0100】

100 ASRシステム

102 ユーザ

10

20

30

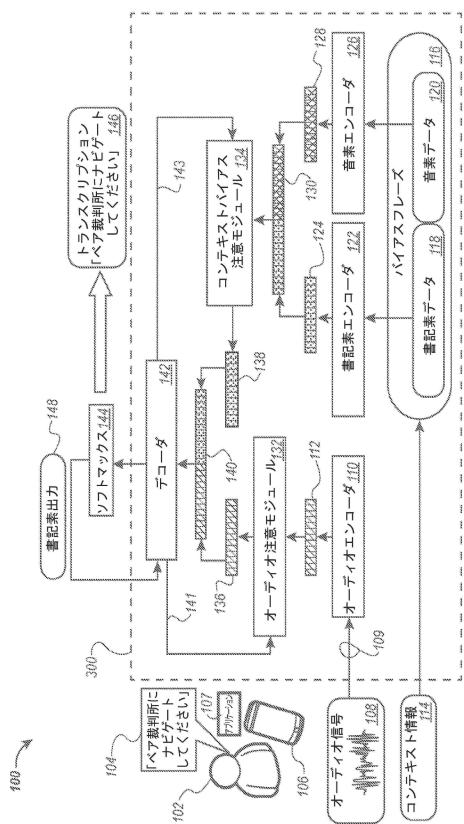
40

50

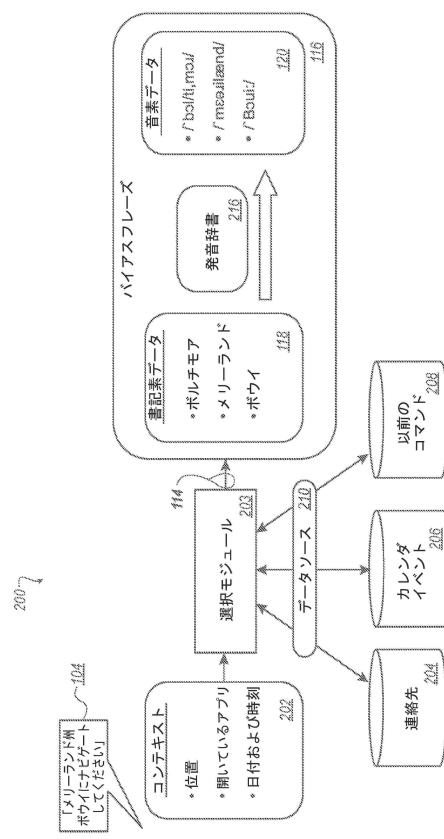
104	音声発話	
106	ユーザデバイス	
107	ナビゲーションアプリケーション	
107	カレンダーアプリケーション	
108	オーディオ信号	
109	音響特徴	
109	オーディオフレーム	10
110	オーディオエンコーダ	
112	オーディオベクトル	
114	コンテキスト情報	
116	バイアスフレーズ	
118	書記素データ	
118	書記素入力	
118	書記素バイアスデータ	
120	音素データ	
120	音素バイアスデータ	
120	音素入力	
122	書記素エンコーダ	
124	書記素ベクトル	
124	書記素エンコーディング	20
126	音素エンコーダ	
128	音素ベクトル	
128	音素エンコーディング	
130	射影ベクトル	
130	射影	
132	オーディオ注意モジュール	
132	射影層	
134	コンテキストバイアス注意モジュール	
136	加重オーディオエンコーディングベクトル	
136	出力	30
138	コンテキストバイアスベクトル	
138	バイアス注意出力	
138	出力	
140	加重ベクトル	
141	状態ベクトル	
141	デコーダオーディオ状態	
141	出力	
142	デコーダ	
143	デコーダコンテキスト状態	
143	状態ベクトル	40
143	出力	
144	ソフトマックス層	
146	トランスクリプション	
146	音声認識結果	
148	音声要素出力	
148	書記素	
148	音声認識結果	
150	トランスクリプション	
200	プロセス	
202	コンテキスト	50

203	選択モジュール	
204	連絡先	
204	連絡先リスト	
206	カレンダーアイベント	
208	コマンド	
210	データソース	
212	バイアスフレーズ	
216	発音辞書	
300	音声認識モデル	
300	LASモデル	10
400	方法	
405	参照トランскriプト	
410	発話	
420	「ジョアン」	
430	単語「ジョン」および「ジーン」	
435	バイアスフレーズ	
500	方法	
502	動作	
506	動作	
600	コンピューティングデバイス	20
600a	標準サーバ	
600b	ラップトップコンピュータ	
600c	ラックサーバシステム	
610	データ処理ハードウェア	
610	プロセッサ	
620	メモリ	
630	ストレージデバイス	
640	高速インターフェース/コントローラ	
650	高速拡張ポート	
660	低速インターフェース/コントローラ	30
670	低速バス	
680	ディスプレイ	
690	低速拡張ポート	

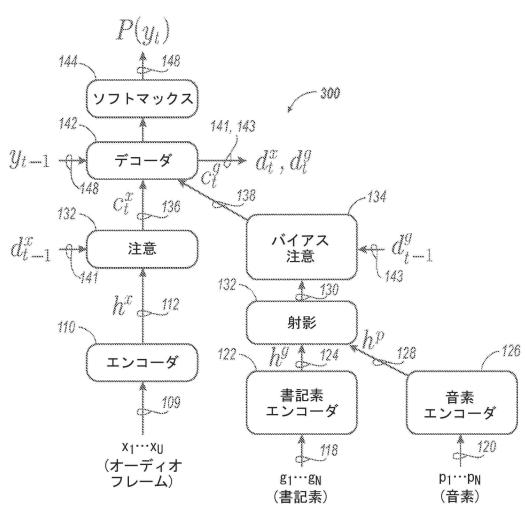
【図面】
【図 1】



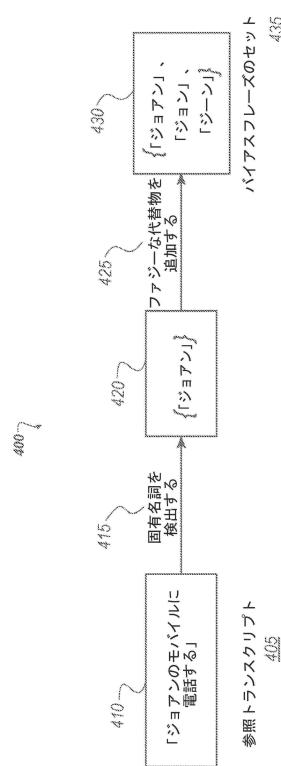
【 四 2 】



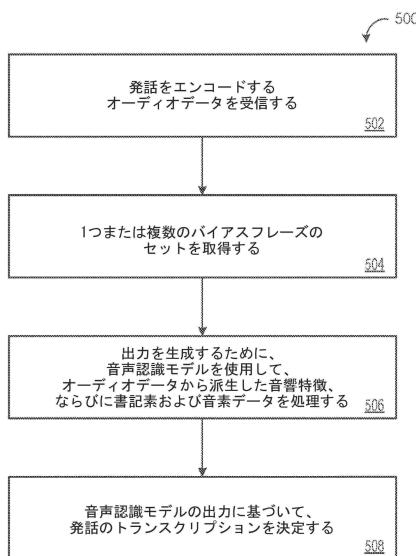
【 図 3 】



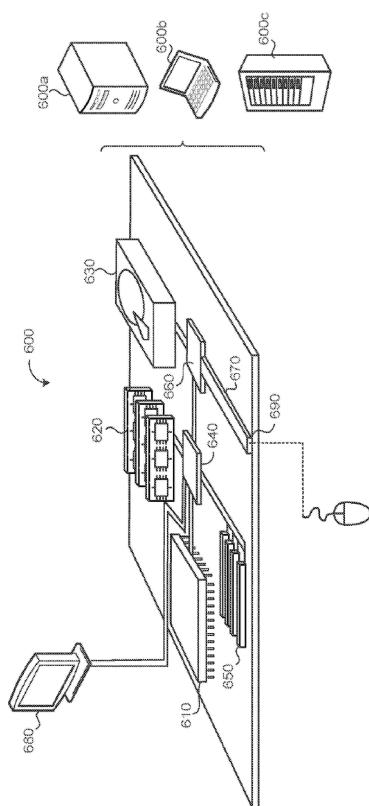
【 义 4 】



【図 5】



【図 6】



10

20

30

40

50

フロントページの続き

(72)発明者 ロヒット・プラカシュ・プラバヴァルカル
アメリカ合衆国・カリフォルニア・94043・マウンテン・ビュー・アンフィシアター・パーク
ウェイ・1600

(72)発明者 ゴラン・ブンダク
アメリカ合衆国・カリフォルニア・94043・マウンテン・ビュー・アンフィシアター・パーク
ウェイ・1600

(72)発明者 タラ・エヌ・サイナス
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 アントワーヌ・ジーン・ブルガイアー
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 菊池 智紀

(56)参考文献 特開2018-109760 (JP, A)
特開2019-120841 (JP, A)
特開2013-250379 (JP, A)
特開平01-081999 (JP, A)
特表2018-513399 (JP, A)
米国特許出願公開第2018/0330725 (US, A1)
BRUGUIER, Antoine et al. , "Phoebe:Pronunciation-aware Contextualization for End-to-end Speech Recognition" , Proc. of the 2019 IEEE ICASSP , 2019年05月12日 , pp.6171-6175
PUNDAK, Golan et al. , "Deep Context: End-to-end Contextual Speech Recognition" , Proc. of the 2018 IEEE SLT , 2018年12月18日 , pp.418-425
HU, Ke et al. , "Phoneme-Based Contextualization for Cross-Lingual Speech Recognition in End-to-End Models" , [online] , 2019年06月21日 , [retrieved on 2022.07.29], Retrieved from the Internet: URL:<https://arxiv.org/abs/1906.09292v1>
田中智大 他 , 電子情報通信学会技術研究報告 , 2018年08月20日 , Vol.118 , No.198 , pp. 31-36

(58)調査した分野 (Int.Cl. , D B名)
G 10 L 15 / 00 - 15 / 34
I E E E X p l o r e