(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7544989号
(P7544989)

(45)発行日　令和6年9月3日(2024.9.3)

(24)登録日　令和6年8月26日(2024.8.26)

(51)国際特許分類

Ｇ１０Ｌ	*15/197(2013.01)*	Ｇ１０Ｌ	15/197
Ｇ１０Ｌ	*15/16 (2006.01)*	Ｇ１０Ｌ	15/16

ＦＩ

請求項の数　22　（全19頁）

(21)出願番号　特願2023-558607(P2023-558607)
(86)(22)出願日　令和4年2月10日(2022.2.10)
(65)公表番号　特表2024-512579(P2024-512579A)
(43)公表日　令和6年3月19日(2024.3.19)
(86)国際出願番号　PCT/US2022/015956
(87)国際公開番号　WO2022/203773
(87)国際公開日　令和4年9月29日(2022.9.29)
審査請求日　令和5年11月21日(2023.11.21)
(31)優先権主張番号　63/165,725
(32)優先日　令和3年3月24日(2021.3.24)
(33)優先権主張国・地域又は機関
米国(US)
早期審査対象出願

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

最終頁に続く

(54)【発明の名称】　ルックアップテーブルリカレント言語モデル

(57)【特許請求の範囲】

【請求項１】

データ処理ハードウェア(103)上で実行されるときに前記データ処理ハードウェア(103)に動作を実行させるコンピュータによって実施される方法(400)であって、前記動作が、

ユーザ(10)によって話され、ユーザデバイス(102)によってキャプチャされた発話(119)に対応するオーディオデータ(120)を受け取ることと、

音声認識器(200)を使用して、前記話された発話(119)の候補文字起こし(132)を決定するために前記オーディオデータ(120)を処理する動作であって、前記候補文字起こし(132)が、トークン(133)のシーケンスを含む、ことと、

トークン(133)の前記シーケンスの各トークン(133)に関して、

第1の埋め込みテーブル(310)を使用して、対応するトークン(133)のトークン埋め込み(312)を決定することと、

第2の埋め込みテーブル(320)を使用して、現在の出力ステップにおける前のn-gramトークンのシーケンスのn-gramトークン埋め込み(322)を決定することと、

前記対応するトークン(133)の連結された出力(335)を生成するために、前記トークン埋め込み(312)と前記n-gramトークン埋め込み(322)とを連結することと、

外部言語モデル(300)を使用して、トークン(133)の前記シーケンスのそれぞれの対応するトークン(133)に関して生成された前記連結された出力(335)を処理することによって、前記話された発話(119)の前記候補文字起こし(132)に再びスコアを付けることとを含む、方法(400)。

【請求項２】
前記外部言語モデル(300)が、リカレントニューラルネットワーク(340)言語モデルを含む、請求項1に記載の方法(400)。
【請求項３】
前記外部言語モデル(300)が、前記音声認識器(200)をハイブリッド自己回帰トランスデューサ(HAT)分解することによって、前記音声認識器(200)と統合される、請求項1または2に記載の方法(400)。
【請求項４】
前記音声認識器(200)が、コンフォーマオーディオエンコーダ(201)およびリカレントニューラルネットワーク-トランスデューサデコーダ(220)を含む、請求項1から3のいずれか一項に記載の方法(400)。
【請求項５】
前記音声認識器(200)が、トランスフォーマオーディオエンコーダ(201)およびリカレントニューラルネットワーク-トランスデューサデコーダ(220)を含む、請求項1から3のいずれか一項に記載の方法(400)。
【請求項６】
前記候補文字起こし(132)のトークン(133)の前記シーケンスの各トークン(133)が、前記候補文字起こし(132)内の単語を表す、請求項1から5のいずれか一項に記載の方法(400)。
【請求項７】
前記候補文字起こし(132)のトークン(133)の前記シーケンスの各トークン(133)が、前記候補文字起こし(132)内のワードピースを表す、請求項1から6のいずれか一項に記載の方法(400)。
【請求項８】
前記候補文字起こし(132)のトークン(133)の前記シーケンスの各トークン(133)が、前記候補文字起こし(132)内のn-gram、音素、または書記素を表す、請求項1から7のいずれか一項に記載の方法(400)。
【請求項９】
前記第1の埋め込みテーブル(310)および前記第2の埋め込みテーブル(320)が、前記データ処理ハードウェア(103)と通信するメモリハードウェア(105)上に記憶される、請求項1から8のいずれか一項に記載の方法(400)。
【請求項１０】
前記対応するトークン(133)の前記トークン埋め込み(312)を決定することが、いかなるグラフィックス処理ユニットおよび/またはテンソル処理ユニットへのアクセスも必要とすることなく、検索によって前記第1の埋め込みテーブル(310)から前記トークン埋め込み(312)を取り出すことを含む、請求項1から9のいずれか一項に記載の方法(400)。
【請求項１１】
前記データ処理ハードウェア(103)が、前記ユーザデバイス(102)上に存在する、請求項1から10のいずれか一項に記載の方法(400)。
【請求項１２】
データ処理ハードウェア(103)と、
前記データ処理ハードウェア(103)と通信するメモリハードウェア(105)であって、前記データ処理ハードウェア(103)によって実行されるときに前記データ処理ハードウェア(103)に、
ユーザ(10)によって話され、ユーザデバイス(102)によってキャプチャされた発話(119)に対応するオーディオデータ(120)を受け取ること、
音声認識器(200)を使用して、前記話された発話(119)の候補文字起こし(132)を決定するために前記オーディオデータ(120)を処理することであって、前記候補文字起こし(132)が、トークン(133)のシーケンスを含む、こと、
トークン(133)の前記シーケンスの各トークン(133)に関して、

第1の埋め込みテーブル(310)を使用して、前記対応するトークン(133)のトークン埋め込み(312)を決定すること、

第2の埋め込みテーブル(320)を使用して、現在の出力ステップにおける前のn-gramトークンのシーケンスのn-gramトークン埋め込み(322)を決定すること、および

前記対応するトークン(133)の連結された出力(335)を生成するために、前記トークン埋め込み(312)と前記n-gramトークン埋め込み(322)とを連結すること、ならびに

外部言語モデル(300)を使用して、トークン(133)の前記シーケンスのそれぞれの対応するトークン(133)に関して生成された前記連結された出力(335)を処理することによって、前記話された発話(119)の前記候補文字起こし(132)に再びスコアを付けることを含む動作を実行させる命令を記憶する、メモリハードウェア(105)とを含む、システム(100)。

【請求項１３】

前記外部言語モデル(300)が、リカレントニューラルネットワーク(340)言語モデルを含む、請求項12に記載のシステム(100)。

【請求項１４】

前記外部言語モデル(300)が、前記音声認識器(200)をハイブリッド自己回帰トランスデューサ(HAT)分解することによって、前記音声認識器(200)と統合される、請求項12または13に記載のシステム(100)。

【請求項１５】

前記音声認識器(200)が、コンフォーマオーディオエンコーダ(201)およびリカレントニューラルネットワーク-トランスデューサデコーダ(220)を含む、請求項12から14のいずれか一項に記載のシステム(100)。

【請求項１６】

前記音声認識器(200)が、トランスフォーマオーディオエンコーダ(201)およびリカレントニューラルネットワーク-トランスデューサデコーダ(220)を含む、請求項12から14のいずれか一項に記載のシステム(100)。

【請求項１７】

前記候補文字起こし(132)のトークン(133)の前記シーケンスの各トークン(133)が、前記候補文字起こし(132)内の単語を表す、請求項12から16のいずれか一項に記載のシステム(100)。

【請求項１８】

前記候補文字起こし(132)のトークン(133)の前記シーケンスの各トークン(133)が、前記候補文字起こし(132)内のワードピースを表す、請求項12から17のいずれか一項に記載のシステム(100)。

【請求項１９】

前記候補文字起こし(132)のトークン(133)の前記シーケンスの各トークン(133)が、前記候補文字起こし(132)内のn-gram、音素、または書記素を表す、請求項12から18のいずれか一項に記載のシステム(100)。

【請求項２０】

前記第1の埋め込みテーブル(310)および前記第2の埋め込みテーブル(320)が、前記データ処理ハードウェア(103)と通信するメモリハードウェア(105)上に記憶される、請求項12から19のいずれか一項に記載のシステム(100)。

【請求項２１】

前記対応するトークン(133)の前記トークン埋め込み(312)を決定することが、いかなるグラフィックス処理ユニットおよび/またはテンソル処理ユニットへのアクセスも必要とすることなく、検索によって前記第1の埋め込みテーブル(310)から前記トークン埋め込み(312)を取り出すことを含む、請求項12から20のいずれか一項に記載のシステム(100)。

【請求項２２】

前記データ処理ハードウェア(103)が、前記ユーザデバイス(102)上に存在する、請求項12から21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、ルックアップテーブルリカレント言語モデルに関する。
【背景技術】
【０００２】
自動音声認識(ASR)システムは、近年、アシスタントが使用可能なデバイスのために人気が高まってきた。話される頻度の低い単語の認識を向上させることは、ASRシステムにとって継続中の問題である。話される頻度の低い単語は、音響的訓練データにめったに含まれず、したがって、ASRシステムがスピーチ内で正確に認識することが困難である。場合によっては、ASRシステムは、話される頻度の低い単語の認識を向上させるために、テキストのみのデータで訓練する言語モデルを含む。しかし、これらの言語モデルは、ASRシステムの効率の低下につながる大規模なメモリおよび計算の要件を含むことが多い。
【発明の概要】
【課題を解決するための手段】
【０００３】
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きにデータ処理ハードウェアに、ルックアップテーブルリカレント言語モデルを使用して音声認識を実行するための動作を実行させるコンピュータによって実施される方法を提供する。動作は、ユーザによって話され、ユーザデバイスによってキャプチャされた発話に対応するオーディオデータを受け取ることを含む。動作は、音声認識器を使用して、話された発話のトークンのシーケンスを含む候補文字起こし(transcription)を決定するためにオーディオデータを処理することも含む。トークンのシーケンスの各トークンに関して、動作は、第1の埋め込みテーブルを使用して、対応するトークンのトークン埋め込みを決定することと、第2の埋め込みテーブルを使用して、n-gramトークンの前のシーケンスのn-gramトークン埋め込みを決定することと、対応するトークンの連結された出力を生成するために、トークン埋め込みとn-gramトークン埋め込みとを連結することとを含む。動作は、外部言語モデルを使用して、トークンのシーケンスのそれぞれの対応するトークンに関して生成された連結された出力を処理することによって話された発話の候補文字起こしに再びスコアを付けることも含む。
【０００４】
本開示の実装は、以下の任意の特徴のうちの1つまたは複数を含む場合がある。一部の実装において、外部言語モデルは、リカレントニューラルネットワーク言語モデルを含む。外部言語モデルは、音声認識器をハイブリッド自己回帰トランスデューサ(HAT: Hybrid Autoregressive Transducer)分解することによって、音声認識器と統合されてよい。一部の例において、音声認識器は、コンフォーマ(conformer)オーディオエンコーダおよびリカレントニューラルネットワーク-トランスデューサデコーダを含む。その他の例において、音声認識器は、トランスフォーマオーディオエンコーダおよびリカレントニューラルネットワーク-トランスデューサデコーダを含む。任意で、候補文字起こしのトークンのシーケンスの各トークンは、候補文字起こし内の単語を表す場合がある。候補文字起こしのトークンのシーケンスの各トークンは、候補文字起こし内のワードピース(wordpiece)を表す場合がある。
【０００５】
一部の実装において、候補文字起こしのトークンのシーケンスの各トークンは、候補文字起こし内のn-gram、音素、または書記素を表す。一部の例において、第1の埋め込みテーブルおよび第2の埋め込みテーブルは、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上に疎らに記憶される。対応するトークンのトークン埋め込みを決定することは、いかなるグラフィックス処理ユニットおよび/またはテンソル処理ユニットへのアクセスも必要とすることなく、検索によって第1の埋め込みテーブルからトークン埋め込みを取り出すことを含んでよい。任意で、データ処理ハードウェアは、ユーザデバイス上に存在

してよい。

【０００６】

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きに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システムを提供する。動作は、ユーザによって話され、ユーザデバイスによってキャプチャされた発話に対応するオーディオデータを受け取ることを含む。動作は、音声認識器を使用して、話された発話のトークンのシーケンスを含む候補文字起こしを決定するためにオーディオデータを処理することも含む。トークンのシーケンスの各トークンに関して、動作は、第1の埋め込みテーブルを使用して、対応するトークンのトークン埋め込みを決定することと、第2の埋め込みテーブルを使用して、n-gramトークンの前のシーケンスのn-gramトークン埋め込みを決定することと、対応するトークンの連結された出力を生成するために、トークン埋め込みとn-gramトークン埋め込みとを連結することとを含む。動作は、外部言語モデルを使用して、トークンのシーケンスのそれぞれの対応するトークンに関して生成された連結された出力を処理することによって話された発話の候補文字起こしに再びスコアを付けることも含む。

【０００７】

本開示の実装は、以下の任意の特徴のうちの1つまたは複数を含む場合がある。一部の実装において、外部言語モデルは、リカレントニューラルネットワーク言語モデルを含む。外部言語モデルは、音声認識器をハイブリッド自己回帰トランスデューサ(HAT)分解することによって、音声認識器と統合されてよい。一部の例において、音声認識器は、コンフォーマオーディオエンコーダおよびリカレントニューラルネットワーク-トランスデューサデコーダを含む。その他の例において、音声認識器は、トランスフォーマオーディオエンコーダおよびリカレントニューラルネットワーク-トランスデューサデコーダを含む。任意で、候補文字起こしのトークンのシーケンスの各トークンは、候補文字起こし内の単語を表す場合がある。候補文字起こしのトークンのシーケンスの各トークンは、候補文字起こし内のワードピースを表す場合がある。

【０００８】

一部の実装において、候補文字起こしのトークンのシーケンスの各トークンは、候補文字起こし内のn-gram、音素、または書記素を表す。一部の例において、第1の埋め込みテーブルおよび第2の埋め込みテーブルは、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上に疎らに記憶される。対応するトークンのトークン埋め込みを決定することは、いかなるグラフィックス処理ユニットおよび/またはテンソル処理ユニットへのアクセスも必要とすることなく、検索によって第1の埋め込みテーブルからトークン埋め込みを取り出すことを含んでよい。任意で、データ処理ハードウェアは、ユーザデバイス上に存在してよい。

【０００９】

本開示の1つまたは複数の実装の詳細が、添付の図面および以下の説明に記載されている。その他の態様、特徴、および利点は、説明および図面から、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】

【００１０】

【図１】音声認識モデルをn-gram埋め込みルックアップテーブルを有する言語モデルと統合するための例示的なシステムの図である。

【図２】例示的な音声認識モデルの図である。

【図３】図1の例示的な言語モデルの概略図である。

【図４】ルックアップテーブルリカレント言語モデルを使用して音声認識を実行する方法の動作の例示的な配列の図である。

【図５】本明細書に記載のシステムおよび方法を実装するために使用されてよい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】

【００１１】
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。
【００１２】
稀な単語またはシーケンスの認識を向上させることは、音響データにゼロまたは低い頻度で出現する多くの入力テキストの発話を誤認識する音声認識システムにおける継続中の問題である。特に、通りの名前、都市などの固有名詞は、めったに話されず(すなわち、ロングテールコンテンツ(long tail content))、音響的訓練データに含まれないことが多く、音声認識システムにとってロングテールコンテンツを認識することを難しくする。一部の実装において、音声認識システムは、めったに話されないロングテールコンテンツを含むテキストのみのデータで訓練する言語モデルを統合する。つまり、言語モデルは、音響データにはないロングテールコンテンツを含むテキストのみのデータのコーパスで訓練される場合があり、ロングテールコンテンツを正しく復号するように音声認識システムにバイアスをかけることができる。
【００１３】
膨大な量のロングテールコンテンツを正確にモデリングするために、言語モデルは、埋め込み語彙(embedding vocabulary)のサイズを大きくする必要がある。埋め込み語彙は、言語モデルのトークン語彙(token vocabulary)(すなわち、単語、ワードピース、n-gramなど)の各トークンに関連する埋め込み識別情報を表す。ほとんどの場合、埋め込み語彙を増やすことは、トークン語彙を増やすことを含み、トークン語彙を増やすことは、大量のメモリおよび計算リソースを必要とし、それは、音声認識システムに負担をかける。さらに、音声認識システムの下流のタスクも、増加したトークン語彙によって影響を受ける。
【００１４】
本明細書の実装は、トークン語彙のサイズを一定に保ちながら、埋め込み語彙を増やすリカレント言語モデルを対象とする。つまり、埋め込み語彙が、トークン語彙と無関係に増加し、音声認識システムの計算リソースに負担をかけることなく、より正確なモデリングを可能にする。特に、リカレント言語モデルは、トークンシーケンスの現在のトークン(たとえば、単語、ワードピース、n-gramなど)のトークン埋め込みを生成する第1の埋め込みテーブルと、トークンシーケンス(たとえば、n-gramシーケンス)の前のトークンのシーケンス埋め込みを生成する第2の埋め込みテーブルとを含む。ここで、トークン埋め込みは、音声認識システムに特定の単語を文字起こしする尤度(likelihood)を提供し、シーケンス埋め込みは、以前に文字起こしされた単語の特定のシーケンスに基づいて特定の単語を文字起こしする尤度を提供する。したがって、第2の埋め込みテーブルは、トークン語彙が一定のままでありながら、言語モデルの埋め込み語彙を増やす。
【００１５】
第2の埋め込みテーブルを用いて言語モデルの埋め込み語彙をスケールアップすることは、埋め込み語彙のサイズが各出力ステップにおける埋め込み検索または動作の数に影響を与えないので、音声認識システムの追加の動作(たとえば、計算リソース)を必要としない。したがって、埋め込み語彙のサイズに関する唯一の実際的な制約は、メモリ容量である。一部の例において、埋め込みテーブルは、疎らにアクセスされ、グラフィックス処理ユニット(GPU)および/またはテンソル処理ユニット(TPU)に記憶されなくてよい。むしろ、埋め込みテーブルは、GPUおよび/またはTPUのメモリよりもはるかに大きな容量を含む、コンピュータ処理ユニット(CPU)のメモリ、ディスク、またはその他のストレージに記憶されてよい。
【００１６】
ここで図1を参照すると、一部の実装において、例示的な音声認識システム100が、それぞれのユーザ10に関連するユーザデバイス102を含む。ユーザデバイス102は、ネットワーク104を介してリモートシステム110と通信してよい。ユーザデバイス102は、モバイル電話、コンピュータ、ウェアラブルデバイス、スマート家電、オーディオインフォテインメントシステム、スマートスピーカなどのコンピューティングデバイスに対応する場

合があり、データ処理ハードウェア103およびメモリハードウェア105を備える。リモートシステム110は、スケーラブルな/融通の利くコンピューティングリソース112(たとえば、データ処理ハードウェア)および/またはストレージリソース114(たとえば、メモリハードウェア)を有する単一のコンピュータ、複数のコンピュータ、または分散システム(たとえば、クラウド環境)であってよい。

【００１７】

ユーザデバイス102は、ユーザ10によって話された発話119に対応する、ユーザデバイス102の1つまたは複数のマイクロフォン106によってキャプチャされたストリーミングオーディオ118を受け取り、ストリーミングオーディオ118から音響的特徴を抽出して、発話119に対応するオーディオデータ120を生成してよい。音響的特徴は、発話119に対応するオーディオデータ120のウィンドウ(window)にわたって計算されたメル周波数ケプストラム係数(MFCC)またはフィルタバンクエネルギーを含んでよい。ユーザデバイス102は、発話119に対応するオーディオデータ120を音声認識器200(本明細書においては自動音声認識器(ASR)モデル200とも呼ばれる)に伝達する。ASRモデル200は、ユーザデバイス102に存在し、実行される場合がある。その他の実装において、ASRモデル200は、リモートシステム110に存在し、実行される。

【００１８】

図2を参照すると、ASRモデル200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統合することによってエンドツーエンド(E2E)の音声認識を提供してよく、辞書(lexicon)または別のテキスト正規化構成要素を必要としない。様々な構造および最適化メカニズムが、向上した精度および削減されたモデル訓練時間を提供することができる。ASRモデル200は、インタラクティブアプリケーションに関連するレイテンシの制約を遵守するコンフォーマ-トランスデューサモデルアーキテクチャを含む場合がある。ASRモデル200は、小さな計算フットプリントを提供し、通常のASRアーキテクチャよりも低いメモリ要件を利用し、ASRモデルアーキテクチャを、音声認識をすべてユーザデバイス102上で実行するのに適するようにする(たとえば、リモートシステム110との通信が必要とされない)。ASRモデル200は、オーディオエンコーダ201、ラベルエンコーダ220、およびジョイントネットワーク(joint network)230を含む。従来のASRシステムにおける音響モデル(AM)に概ね似ているオーディオエンコーダ201は、複数のコンフォーマ層を有するニューラルネットワークを含む。たとえば、オーディオエンコーダ201は、d次元特徴ベクトル(たとえば、ストリーミングオーディオ118(図1)の音響フレーム)のシーケンス$x = (x_1, x_2, \cdots, x_T)$を読み取り、ただし、$x_t \in$

【００１９】

【数１】

$$\mathbb{R}_d$$

【００２０】

であり、オーディオエンコーダ201は、各時間ステップにおいて高次特徴表現を生成する。この高次特徴表現は、$ah_1, \ldots, ah_T$と表記される。任意で、オーディオエンコーダ201は、コンフォーマ層の代わりにトランスフォーマ層を含む場合がある。同様に、ラベルエンコーダ220も、トランスフォーマ層のニューラルネットワークまたはルックアップテーブル埋め込みモデル(look-up table embedding model)を含む場合があり、これは、言語モデル(LM)のように、これまでに最終的なソフトマックス層240によって出力されたブランクでないシンボルのシーケンス$y_0, \ldots, y_{ui-1}$を処理して、予測されたラベルの履歴を符号化する密な表現lh_uにする。

【００２１】

最後に、オーディオエンコーダ201およびラベルエンコーダ220によって生成された表現は、密な層$J_{u,t}$を使用してジョイントネットワーク230によって組み合わされる。そして、ジョイントネットワーク230は、次の出力シンボルの分布である$P(Z_{u,t} \mid x, t, y_{\cdots}, \ldots$

, y_{u-1})を予測する。別の言い方をすれば、ジョイントネットワーク230は、各出力ステップ(たとえば、時間ステップ)において、可能な音声認識仮説の確率分布を生成する。ここで、「可能な音声認識仮説」は、指定された自然言語における書記素(たとえば、記号/文字)またはワードピースもしくは単語をそれぞれ表す出力ラベル(「スピーチ単位(speech unit)」とも呼ばれる)の集合に対応する。たとえば、自然言語が英語であるとき、出力ラベルの集合は、27個の記号、たとえば、英語のアルファベットの26文字の各々の1つのラベルと、スペースを指定する1つのラベルとを含む場合がある。したがって、ジョイントネットワーク230は、出力ラベルの所定の集合の各々の発生の尤度を示す値の集合を出力してよい。値のこの集合は、ベクトルであることが可能であり、出力ラベルの集合の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(たとえば、個々の文字、ならびに潜在的に句読点およびその他の記号)であるが、出力ラベルの集合は、そのように限定されない。たとえば、出力ラベルの集合は、書記素に加えてまたは書記素の代わりに、ワードピースおよび/または単語全体を含むことが可能である。ジョイントネットワーク230の出力分布は、異なる出力ラベルの各々に関する事後確率値を含み得る。したがって、異なる書記素またはその他の記号を表す100個の異なる出力ラベルがある場合、ジョイントネットワーク230の出力$z_{u,t}$は、各出力ラベルに関して1つずつ、100個の異なる確率値を含み得る。それから、確率分布は、文字起こしを決定するための(たとえば、ソフトマックス層240による)ビームサーチプロセスにおいて、候補の正書法要素(orthographic element)(たとえば、書記素、ワードピース、および/または単語)を選択し、スコアを割り振るために使用され得る。

【００２２】

ソフトマックス層240は、対応する出力ステップにおいてASRモデル200によって予測された次の出力シンボルとして、分布において最も高い確率を持つ出力ラベル/シンボルを選択するために、任意の技術を採用してよい。このようにして、ASRモデル200は、条件付き独立性の仮定をせず、むしろ、各シンボルの予測は、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。

【００２３】

再び図1を参照すると、ASRモデル200は、話された発話119の候補文字起こし132を決定するためにオーディオデータ120を処理するように構成される。ここで、候補文字起こしは、トークン133、133a～nのシーケンスを含み、各トークン133は、発話119の候補文字起こし132の一部を表す。すなわち、トークン133のシーケンスの各トークン133は、候補文字起こし132内の潜在的な単語、ワードピース、n-gram、音素、および/または書記素を表す場合がある。たとえば、ASRモデル200は、「driving directions to Beaubien」という話された発話119に関して「driving directions to bourbon」という候補文字起こし132を生成する。この例において、候補文字起こし132のトークン133のシーケンスの各トークン133は、1つの単語を表す場合がある。したがって、トークン133のシーケンスは、候補文字起こし132(たとえば、「driving directions to bourbon」)内の単一の単語をそれぞれ表す4つのトークン133を含む。特に、語「bourbon」を表す第4のトークン133は、正しい語「Beaubien」に関してASRモデル200によって誤認識される。

【００２４】

さらに、ASRモデル200は、トークン133のシーケンスの各トークン133を、可能な候補トークンの対応する確率分布として生成する場合がある。たとえば、トークン133のシーケンスの第4のトークン133に関して、ASRモデル200は、ASRモデル200がそれぞれの第4のトークン133に関して可能なトークンを認識したという確信を示す対応する確率または尤度をそれぞれ有する対応する可能なトークンとして、トークン「bourbon」および「Beaubien」を生成する場合がある。ここで、第1の候補トークンおよび第2の候補トークンは音声学的に類似しているが、「Beaubien」がASRモデル200の音響的訓練データに含まれていない可能性が高い固有名詞(たとえば、希な単語またはロングテールコンテンツ)であるので、ASRモデルは、語「Beaubien」に関連する第2の候補トークンよりも、

語「bourbon」に関連する第1の候補トークンを優先する場合がある。つまり、「Beaubien」が音響的訓練データに含まれていなかったか、または音響的訓練データのいくつかの事例にしか含まれていなかったので、ASRモデル200は、発話119内のこの特定の語を正しく認識しない場合がある。別の言い方をすれば、ASRモデル200は、第2の候補トークン(たとえば、Beaubien)よりも第1の候補トークン(たとえば、bourbon)に関してより高い確率/尤度スコアを出力するように構成される場合がある。したがって、ASRモデルは、より高い確率/尤度スコアにより、トークン133のシーケンスの第4のトークン133を「bourbon」として出力する。例は、簡単にするために、可能な候補トークンの確率分布の中の2つの候補トークンのみを説明するが、確率分布の中の可能な候補トークンの数は、3以上の任意の数であることが可能である。一部の例において、トークン133のシーケンスの各トークン133は、最も高い確率/尤度スコアを有するn個の可能な候補トークン133のランク付けされたリストに関連する可能な候補トークン133のnベストリスト(n-best list)によって表される。各候補トークン133は、音声認識仮説と呼ばれる場合がある。追加の例において、トークン133のシーケンスの各トークン133は、可能な候補トークン133の確率分布において最も高い確率/尤度スコアを有する可能な候補トークン133によって表される。

【００２５】

ASRモデル200は、トークン133のシーケンスを含む候補文字起こし132を言語モデル(すなわち、リカレント言語モデル)300に伝達する。言語モデル300は、ユーザデバイス102のメモリハードウェア105、もしくは任意でリモートシステム110のストレージリソース114、またはそれらの何らかの組合せに存在する場合がある。言語モデル300は、候補トークン133のシーケンスの各トークン133を出力する尤度を決定するように構成される。すなわち、言語モデル300は、トークン133のシーケンスの各トークン133の可能な候補トークン133のうち、どの候補トークン133が、話された発話119の候補文字起こし132内の対応するトークン133を正しく表す可能性が最も高いかを決定することによって、ASRモデル200によって出力された候補文字起こし132に再びスコアを付ける。言語モデル300は、ASRモデル200を訓練するために使用される訓練オーディオデータにめったに含まれないかまたは含まれない、固有名詞またはロングテールコンテンツなどの希な単語に向けて、ASRモデル200によって出力された音声認識仮説にバイアスをかけるのを支援してよい。上の例において、言語モデル300は、トークン133のシーケンスの第4のトークン133の可能な候補トークン133の確率分布において単語「Beaubien」に関連する候補トークン133の確率/尤度スコアを引き上げることによって、音声認識にバイアスをかけてよい。ここで、単語「Beaubien」に関連する候補トークン133の引き上げられた確率/尤度スコアは、今や候補トークン「bourbon」の確率/尤度スコアよりも高くなり、それによって、話された発話119から単語「Beaubien」を今や正しく認識するための再びスコアを付けられた文字起こし345をもたらす場合がある。

【００２６】

言語モデル300は、リカレントニューラルネットワーク(RNN)言語モデルを含んでよい。より詳細には、言語モデル300は、埋め込みテーブルの行数を増やすことによってRNN言語モデルのサイズをスケールアップするように構成されたルックアップテーブル言語モデルを含む場合があるが、浮動小数点演算の増加は最小限に過ぎない。つまり、ルックアップテーブルは、埋め込みがメモリハードウェア(たとえば、CPUのメモリまたはディスク)上に疎らに記憶され、テーブルのサイズが各フォワードパス(forward pass)に追加の演算を追加しないような検索によってテーブルから取り出されることを可能にし、制限された/制約されたGPU/TPUのメモリ上の記憶の必要性を減らす。

【００２７】

言語モデル300は、第1の埋め込みテーブル310、第2の埋め込みテーブル320、連結器330、およびリカレントニューラルネットワーク(RNN)340を含む。一部の例において、言語モデル300は、ロングテールコンテンツを正しく文字起こしするようにASRモデル200の候補トークンにバイアスをかけるために、めったに話されない単語(すなわち、ロング

テールコンテンツ)を含むテキストのみのデータで訓練される。

【００２８】

第1の埋め込みテーブル310は、トークン133のシーケンスの現在のトークン133、t_iのトークン埋め込み312を生成するように構成される。ここで、t_iは、トークン133のシーケンスの現在のトークン133を示す。第1の埋め込みテーブル310は、トークン133のシーケンスのトークン133の残りとは独立して、各トークン133のそれぞれのトークン埋め込み312を決定する。一方、第2の埋め込みテーブル320は、現在の出力ステップ(たとえば、時間ステップ)における前のn-gramトークンシーケンス133、$t_0, ..., t_{n-1}$を受け取り、それぞれのn-gramトークン埋め込み322を生成するn-gram埋め込みテーブルを含む。前のn-gramトークンシーケンス(たとえば、$t_0, ..., t_{n-1}$)は、各時間ステップにおいて以前に生成されたトークン133についてのコンテキスト(context)情報を提供する。前のn-gramトークンシーケンス$t_0, ..., t_{n-1}$は、各時間ステップにおいてnによって指数関数的に大きくなる。したがって、各出力ステップにおけるn-gramトークン埋め込み322は、稀な単語をスペルアウトするなど、短期的な依存関係を支援し、それによって、サブワードモデルにおけるロングテールトークン(long tail token)(たとえば、単語)のモデリングを改善する。

【００２９】

各時間ステップにおいて、連結器330は、トークン埋め込み312とn-gramトークン埋め込み322とを連結して連結された出力335にする。RNN 340は、連結された出力335を受け取り、最終的に、再びスコアを付けられた文字起こし(たとえば、最終的な文字起こし)345を生成するために、連結された出力335を使用して、ASRモデル200によって出力された候補文字起こし132に再びスコアを付ける。さらに、言語モデル300は、再びスコアを付けられた文字起こし345をネットワーク104を介してユーザデバイス102に提供してよい。ユーザデバイス102は、再びスコアを付けられた文字起こし345をユーザデバイス102のディスプレイ上に視覚的に表示するか、または再びスコアを付けられた文字起こし345をユーザデバイス102の1つもしくは複数のスピーカによって聴覚的に提示してよい。その他の例において、再びスコアを付けられた文字起こし345は、ユーザデバイス102にアクションを実行するよう命じるクエリである場合がある。

【００３０】

示された例の単一の出力ステップにおいて、第2の埋め込みテーブル320は、第1の埋め込みテーブル310によって検索されたそれぞれのトークン埋め込み312によって表される現在のトークン133(たとえば、「bourbon」)から、3つの前のn-gramトークンのシーケンス全体(たとえば、「driving directions to」)を表すn-gramトークン埋め込み322を生成してよい。その後、連結器330は、連結された出力335(たとえば、「driving directions to bourbon」)を生成するために、n-gramトークン埋め込み322とそれぞれのトークン埋め込み312とを連結してよい。RNN 340は、連結された出力335を処理することによって、候補文字起こし132に再びスコアを付ける。たとえば、RNN 340は、今や「bourbon」よりも高い第4のトークン133に関する尤度/確率スコアを有するように「Beaubien」の尤度/確率スコアを引き上げるように、候補文字起こし132(たとえば、driving directions to bourbon)に再びスコアを付ける場合がある。RNN 340は、現在のコンテキスト情報(たとえば、前のn-gramトークンシーケンス$t_0, ..., t_{n-1}$)によって、「bourbon」が正しいトークン133である可能性が高くないという判定に部分的に基づいて、「Beaubien」の尤度/確率スコアを引き上げる。したがって、RNN 340は、再びスコアを付けられた文字起こし345(たとえば、driving directions to Beaubien)を生成する。特に、第4のトークン133は、たとえ候補文字起こし132において誤認識されたとしても、再びスコアを付けられた文字起こし345においては今や正しく認識される。

【００３１】

図3は、図1のリカレント言語モデル300の概略図を示す。第1の埋め込みテーブル310は、複数のトークン埋め込み312、312a～nを含み、第2の埋め込みテーブル320は、複数のn-gramトークン埋め込み322、322a～nを含む。第1の埋め込みテーブル310は、U

×E行列によって表され、Eは、埋め込み次元(すなわち、埋め込み長(embedding length))を表し、Uは、第1の埋め込みテーブル310内のトークン埋め込み322の数を表す。概して、埋め込みの数は、一意のトークンの数V(たとえば、単語、ワードピース、n-gramなどの数)に等しい。すなわち、第1の埋め込みテーブル310は、それぞれの可能なトークン133(たとえば、単語、ワードピースなど)のそれぞれのトークン埋め込み312を記憶する。たとえば、トークン133がワードピースを含むとき、リカレント言語モデル300は、512幅の2つのLSTM層を有する4,096ワードピースモデルを含む場合があり、トークン埋め込み312の次元Eは、96に等しい場合がある。n-gramトークン埋め込み322は、nが4に設定されるとき、それぞれ、2,048の次元を含む場合がある。第2の埋め込みテーブル(たとえば、n-gram埋め込みテーブル)320は、それぞれの前のn-gramトークンシーケンス(たとえば、$t_0, \ldots, t_{n-1}$)に、以下のように、モジュラーハッシュ(modular hash)によって埋め込みn-gramトークン埋め込み(たとえば、埋め込み識別子)322を割り振る場合がある。

【００３２】

【数２】

$$\text{ngram2id}\ (t_0,\ldots,t_{n-1}) = \left(\sum_{i=0}^{n-1} t_i V^i\right)\ \text{mod}\ U\ \forall\ t_i\ \in [0,V) \qquad (1)$$

【００３３】

特に、モジュラーハッシュは、任意の異なるn-gram埋め込みが同じn-gramトークン埋め込み322にハッシュされるような衝突をともなう。しかし、一意のトークンの数Vを増やすことによって、衝突が減少し、性能が向上する。

【００３４】

示された例において、第1の埋め込みテーブル310および第2の埋め込みテーブル320は、話された発話119「Hello Pam」に対応する候補文字起こし132のトークン133のシーケンスを受け取る。特に、候補文字起こし132は、正しい語「Pam」を語「pan」と誤認識する。ここで、第1の埋め込みテーブル310は、トークン133のシーケンスの第3のトークン133cに対応する候補文字起こし132内の現在のトークン133(たとえば、t_2)を受け取る。第1の埋め込みテーブル310は、第3のトークン133cに関して、トークン133のシーケンスの残りとは独立に、複数のトークン埋め込み312からトークン埋め込み312を決定する。この例において、「pan」は0.8の尤度/確率スコアを有する場合があり、一方、「Pam」は0.2の尤度/確率スコアを有するので、第1の埋め込みテーブル310は、「pan」を第3のトークン133cとして決定する(黒い四角によって示される)。

【００３５】

第2の埋め込みテーブル320は、候補文字起こし132の前のn-gramトークンシーケンス133を受け取る。ここで、前のn-gramトークンシーケンス133は、「〈s〉 Hello」(たとえば、t_0、t_1)を含み、〈s〉は、シーケンスの始まりのトークン133を示す。第2の埋め込みテーブル320は、前のn-gramトークンシーケンス133に基づいて、「〈s〉 Hello」のn-gramトークン埋め込み322を決定する(黒い四角によって示される)。

【００３６】

連結器330は、現在のトークン「pan」の、第1の埋め込みテーブル310から出力されたトークン埋め込み312と、「〈s〉 Hello」のn-gramトークン埋め込み322とを連結し、連結された出力335(たとえば、「〈s〉Hello pan」)をRNN 340のRNNセルに提供する。特に、連結された出力335は、連結された出力335を処理するためにRNN 340に密なパラメータ(dense parameter)を増やさせるが、RNN出力の次元が固定されたままであるので、スケーリングは2次ではなく線形である。n-gramトークン埋め込み322からのコンテキストの情報は、RNN 340への入力として有用であるだけでなく、中間層/セルがその層に固有の隠れ状態を介してコンテキスト情報を受け取るように、中間層においても有用で

ある。したがって、RNN 340は、連結された出力335を、RNN内のあらゆる層の入力活性化(input activation)に注入し、その層に固有の埋め込みテーブルからそれぞれを引き出してよい。

【００３７】

上記の例について続けると、RNN 340は、連結された出力335を処理することによって、候補文字起こし132に再びスコアを付ける。すなわち、RNN 340は、連結された出力335に含まれるコンテキスト情報(たとえば、前のn-gramトークンシーケンス133)に基づいて、語「Pam」の確率を引き上げる場合がある。したがって、RNN 340は、「pan」の尤度/確率スコアを0.4に調整し、「Pam」の尤度/確率スコアを0.6に調整する場合がある。よって、RNNは、第3のトークン133cとして「Pam」を含むように再びスコアを付けられた文字起こしを生成し、話された発話119を正しく認識する。

【００３８】

n-gram埋め込みテーブル320は、ほぼ10億個のパラメータまで効果的にスケールアップする場合がある。リカレント言語モデル300は、30億のテキストのみの文コーパスで訓練される場合がある。言語モデル300は、100万のよく使用される単語のホワイトリストに対するスペルミスの除去、および末尾の表現を確実にするための文の頻度のlog(n)スケーリングでさらに前処理される場合がある。

【００３９】

リカレント言語モデル300は、以下のように、最初に、ハイブリッド自己回帰トランスデューサ(HAT)分解によってASRモデル200の内部言語モデルスコアから対数事後確率(log-posterior)を分離することによって有効尤度(effective likelihood)を取得することによって、エンドツーエンドのASRモデル200と統合されてよい。

$$\log p(y \mid x) \quad \log p(y \mid x) - \lambda_2 \log p_{ILM}(y) \qquad (2)$$

その後、言語モデルの対数事後確率スコアが、以下のように加えられる。

【００４０】

【数３】

$$y^* = \underset{y}{\operatorname{argmax}} [\log p(x|y) + \lambda_1 p_{LM}(y)]$$
$$= \underset{y}{\operatorname{argmax}} [\log p(y|x) - \lambda_2 \log p_{ILM}(y) + \lambda_1 p_{LM}(y) + \lambda_1 p_{LM}(y)] \qquad (3)$$

【００４１】

RNN-Tデコーダ(たとえば、予測ネットワーク220およびジョイントネットワーク230(図2))は、訓練中にHAT分解されてよい。

【００４２】

図4は、ルックアップテーブルリカレント言語モデルを使用して音声認識を実行する方法400の動作の例示的な配列の流れ図である。動作402において、方法400は、ユーザ10によって話され、ユーザデバイス102によってキャプチャされた発話119に対応するオーディオデータ120を受け取ることを含む。動作404において、方法400は、音声認識器200を使用して発話119の候補文字起こし132を決定するためにオーディオデータ120を処理することを含む。ここで、候補文字起こし132は、トークン133、133a～nのシーケンスを含む。トークン133のシーケンスの各トークン133に関して、方法400は、動作406～410を実行する。動作406において、方法400は、第1の埋め込みテーブル310を使用して、対応するトークン133のトークン埋め込み312を決定することを含む。動作408において、方法400は、第2の埋め込みテーブル320を使用して、前のn-gramトークンシーケンス133、$t_0, \ldots, t_{n-1}$のn-gramトークン埋め込み322を決定することを含む。動作410において、方法400は、対応するトークン133の連結された出力335を生成するために、トークン埋め込み312とn-gramトークン埋め込み322とを連結することを含む。動作412において、方法400は、外部言語モデル(たとえば、RNN)340を使用して、トークン133のシーケンスのそれぞれの対応するトークン133に関して生成された連結された出力3

35を処理することによって、話された発話119の候補文字起こし132に再びスコアを付けることを含む。

【００４３】

図5は、本明細書に記載のシステムおよび方法を実装するために使用されてよい例示的なコンピューティングデバイス500の概略図である。コンピューティングデバイス500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、およびその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ように意図される。本明細書に示される構成要素、それらの構成要素の接続および関係、ならびにそれらの構成要素の機能は、単に例示的であるように意図されており、本明細書において説明および/または特許請求される本発明の実装を限定するように意図されていない。

【００４４】

コンピューティングデバイス500は、プロセッサ510、メモリ520、ストレージデバイス530、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ーフェース/コントローラ540、ならびに低速バス570およびストレージデバイス530に接続する低速インターフェース/コントローラ560を含む。構成要素510、520、530、540、550、および560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搭載されるか、または適宜その他の方法で搭載される場合がある。プロセッサ510は、メモリ520内またはストレージデバイス530上に記憶された命令を含む、コンピューティングデバイス500内で実行するための命令を処理して、高速インターフェース540に結合されたディスプレイ580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)のためのグラフィカルな情報を表示することができる。その他の実装においては、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複数の種類のメモリと一緒に適宜使用される場合がある。また、複数のコンピューティングデバイス500が、各デバイスが必要な動作の一部を提供するようにして(たとえば、サーババンク、一群のブレードサーバ、またはマルチプロセッサシステムとして)接続される場合がある。

【００４５】

メモリ520は、コンピューティングデバイス500内で情報を非一時的に記憶する。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、または不揮発性メモリユニットであってよい。非一時的メモリ520は、コンピューティングデバイス500による使用のために、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、プログラムの状態情報)を一時的または永続的に記憶するために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は、フラッシュメモリおよび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)/電子的消去可能プログラマブル読み出し専用メモリ(EEPROM)(たとえば、典型的にはブートプログラムなどのファームウェアのために使用される)を含むがこれらに限定されない。揮発性メモリの例は、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)、およびディスクまたはテープを含むがこれらに限定されない。

【００４６】

ストレージデバイス530は、コンピューティングデバイス500に大容量ストレージを提供することができる。一部の実装において、ストレージデバイス530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、ストレージデバイス530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、またはテープデバイス、フラッシュメモリもしくはその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もしくはその他の構成内のデバイスを含むデバイスのアレイであってよい。追加的な実装においては、コンピュータプログラム製品が、情報担体内に有形で具現化される。コンピュータプログラム製品は、実行されるときに上述の方法などの1つまたは複数の方法を実行する命令を含む。情報担体は、メモリ520、ストレージデバイス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。

【００４７】

高速コントローラ540は、コンピューティングデバイス500に関する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560は、帯域幅をそれほど消費しない動作を管理する。役割のそのような割り当ては、例示的であるに過ぎない。一部の実装において、高速コントローラ540は、メモリ520に、(たとえば、グラフィックスプロセッサまたはアクセラレータを通じて)ディスプレイ580に、および様々な拡張カード(図示せず)を受け入れてよい高速拡張ポート550に結合される。一部の実装において、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。様々な通信ポート(たとえば、USB、Bluetooth、イーサネット、ワイヤレスイーサネット)を含んでよい低速拡張ポート590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つもしくは複数の入力/出力デバイスに結合される場合があり、またはたとえばネットワークアダプタを介してスイッチもしくはルータなどのネットワークデバイスに結合される場合がある。

【００４８】

コンピューティングデバイス500は、図に示されるように、多くの異なる形態で実装されてよい。たとえば、コンピューティングデバイス500は、標準的なサーバ500aとして、もしくはそのようなサーバ500aのグループ内で複数回、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の一部として実装されてよい。

【００４９】

本明細書に記載の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/もしくは光回路、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、ならびに/またはこれらの組合せで実現され得る。これらの様々な実装は、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および命令を受信し、それらにデータおよび命令を送信するために結合された、専用または汎用であってよい少なくとも1つのプログラミング可能なプロセッサを含むプログラミング可能なシステム上の、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への実装を含み得る。

【００５０】

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。一部の例において、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」と呼ばれる場合がある。例示的なアプリケーションは、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、文書処理アプリケーション、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、およびゲーミングアプリケーションを含むがこれらに限定されない。

【００５１】

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、プログラミング可能なプロセッサ用の機械命令を含み、高級手続き型プログラミング言語および/もしくはオブジェクト指向プログラミング言語、ならびに/またはアセンブリ/機械言語で実装され得る。本明細書において使用されるとき、用語「機械可読媒体」および「コンピュータ可読媒体」は、機械命令を機械可読信号として受け取る機械可読媒体を含む、プログラミング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コンピュータプログラム製品、非一時的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、プログラマブルロジックデバイス(PLD))を指す。用語「機械可読信号」は、プログラミング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。

【００５２】

本明細書に記載のプロセスおよび論理フローは、入力データに対して演算を行い、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ミング可能なプロセッサが実行することによって実行され得る。また、プロセスおよび論理フローは、専用の論理回路、たとえば、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)によって実行され得る。コンピュータプログラムの実行に好適なプロセッサは、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。概して、プロセッサは、読み出し専用メモリ、またはランダムアクセスメモリ、またはこれらの両方から命令およびデータを受け取る。コンピュータの必須の要素は、命令を実行するためのプロセッサ、ならびに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。また、概して、コンピュータは、データを記憶するための1つもしく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、もしくは光ディスクを含むか、またはそれらの大容量ストレージデバイスからデータを受信するか、もしくはそれらの大容量ストレージデバイスにデータを転送するか、もしくはその両方を行うために動作可能なように結合される。しかし、コンピュータは、そのようなデバイスを有していなくてもよい。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好適なコンピュータ可読媒体は、例として、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ディスクおよびDVD-ROMディスクを含む、すべての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、専用の論理回路によって補完され得るか、または専用の論理回路に組み込まれ得る。

【００５３】

ユーザとのインタラクションを提供するために、本開示の1つまたは複数の態様は、ユーザに対して情報を表示するためのディスプレイデバイス、たとえば、CRT(ブラウン管)、LCD(液晶ディスプレイ)モニタ、またはタッチスクリーンと、任意で、ユーザがコンピュータに入力を与え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、たとえば、マウスまたはトラックボールとを有するコンピュータ上に実装され得る。その他の種類のデバイスが、ユーザとのインタラクションを提供するためにやはり使用されることが可能であり、たとえば、ユーザに提供されるフィードバックは、任意の形態の感覚フィードバック、たとえば、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であることが可能であり、ユーザからの入力は、音響、スピーチ、または触覚による入力を含む任意の形態で受け取られることが可能である。加えて、コンピュータは、ユーザによって使用されるデバイスに文書を送信し、そのデバイスから文書を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された要求に応答してユーザのクライアントデバイス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インタラクションすることができる。

【００５４】

いくつかの実装が、説明された。しかしながら、本開示の精神および範囲を逸脱することなく様々な修正がなされてよいことは、理解されるであろう。したがって、その他の実装は、添付の請求項の範囲内にある。

【符号の説明】

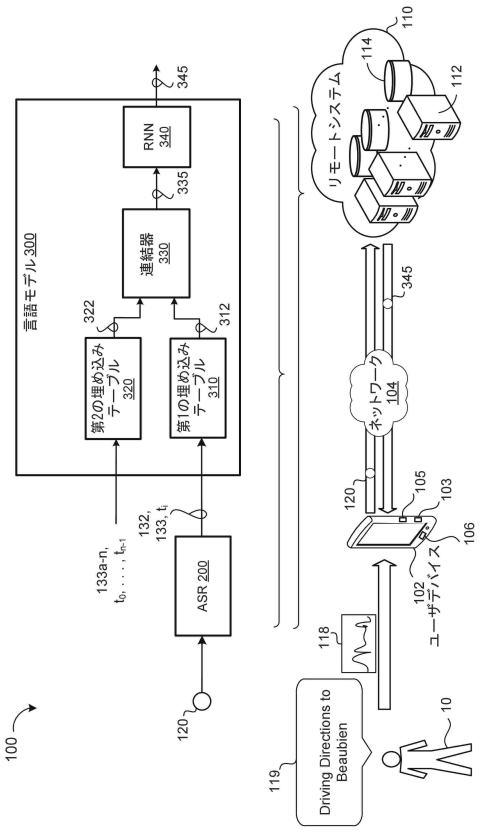
【００５５】

10 ユーザ
100 音声認識システム
102 ユーザデバイス
104 ネットワーク
103 データ処理ハードウェア
105 メモリハードウェア

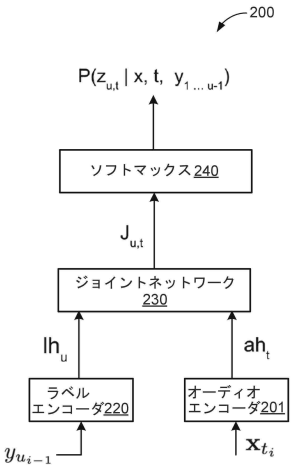
106 マイクロフォン
110 リモートシステム
112 コンピューティングリソース
114 ストレージリソース
118 ストリーミングオーディオ
119 発話
120 オーディオデータ
132 候補文字起こし
133 トークン、トークンシーケンス
133a～n トークン
200 音声認識器、自動音声認識器(ASR)モデル
201 オーディオエンコーダ
220 ラベルエンコーダ、予測ネットワーク
230 ジョイントネットワーク
240 ソフトマックス層
300 言語モデル
310 第1の埋め込みテーブル
312、312a～n トークン埋め込み
320 第2の埋め込みテーブル
322、322a～n n-gramトークン埋め込み
330 連結器
335 連結された出力
340 RNN、外部言語モデル
345 再びスコアを付けられた文字起こし
400 方法
500 コンピューティングデバイス
500a 標準的なサーバ
500b ラップトップコンピュータ
500c ラックサーバシステム
510 プロセッサ
520 メモリ
530 ストレージデバイス
540 高速インターフェース/コントローラ
550 高速拡張ポート
560 低速インターフェース/コントローラ
570 低速バス
580 ディスプレイ
590 低速拡張ポート

【図面】

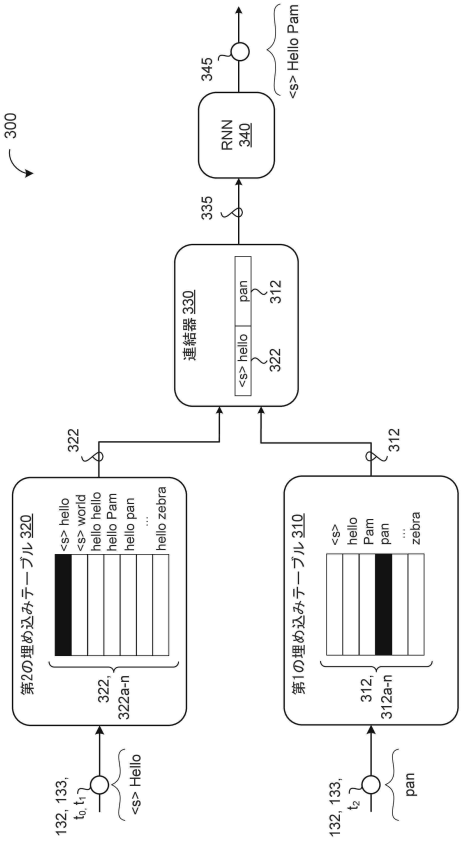
【図１】


100
言語モデル300
RNN 340
連結器 330
第2の埋め込みテーブル 320
第1の埋め込みテーブル 310
ASR 200
リモートシステム
ネットワーク 104
ユーザデバイス
Driving Directions to Beaubien

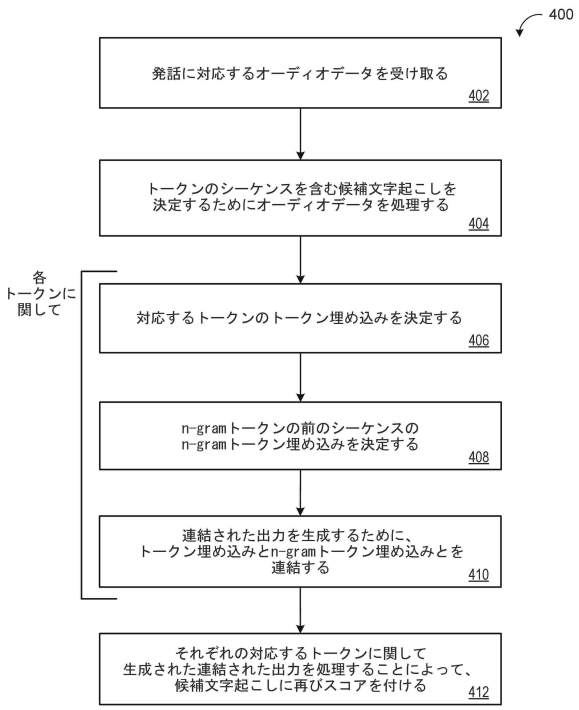
【図２】


200
$P(z_{u,t} \mid x, t, y_{1 \ldots u-1})$
ソフトマックス240
$J_{u,t}$
ジョイントネットワーク 230
lh_u
ah_t
ラベルエンコーダ220
オーディオエンコーダ201
$y_{u_{i-1}}$
$\mathbf{x}_{t_i}$

【図３】


300
RNN 340
連結器 330
<s> hello
pan
<s> Hello Pam
第2の埋め込みテーブル 320
第1の埋め込みテーブル 310
<s> hello
hello Pam
hello pan
hello zebra
<s> world
hello hello
<s>
Pam
zebra
<s> Hello
pan

【図４】


400
発話に対応するオーディオデータを受け取る 402
トークンのシーケンスを含む候補文字起こしを決定するためにオーディオデータを処理する 404
各トークンに関して
対応するトークンのトークン埋め込みを決定する 406
n-gramトークンの前のシーケンスのn-gramトークン埋め込みを決定する 408
連結された出力を生成するために、トークン埋め込みとn-gramトークン埋め込みとを連結する 410
それぞれの対応するトークンに関して生成された連結された出力を処理することによって、候補文字起こしに再びスコアを付ける 412

【図５】

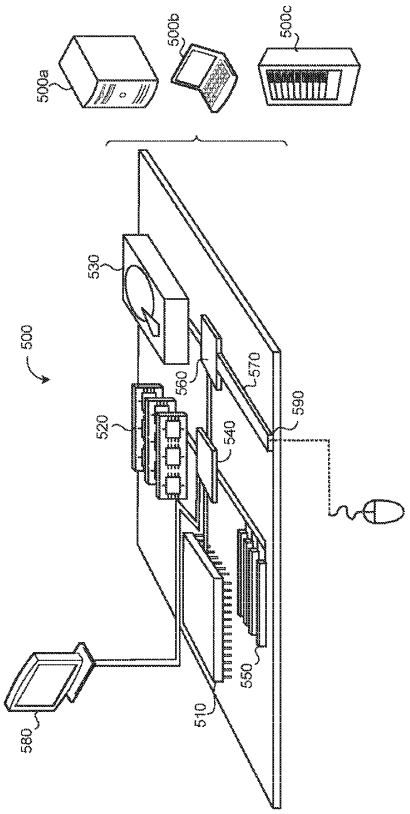


FIG. 5

フロントページの続き

(72)発明者　ロニー・ファン
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００
(72)発明者　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　トレヴァー・ストローマン
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００
(72)発明者　シャンカール・クマール
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００

審査官　中村 天真

(56)参考文献　特開２０１４－１４９４９０（ＪＰ，Ａ）
国際公開第２０２１／０２４４９１（WO，Ａ１）
W. Ronny Huang et al.，Lookup-Table Recurrent Language Models for Long Tail Speech Recognition，[online]，2021年04月09日，[2024.03.26検索], インターネット 〈URL: https://arxiv.org/pdf/2104.04552v1.pdf〉

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１５／３４