(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7351018号
(P7351018)

(45)発行日　令和5年9月26日(2023.9.26)

(24)登録日　令和5年9月15日(2023.9.15)

(51)国際特許分類

Ｇ１０Ｌ　15/06　(2013.01)	Ｇ１０Ｌ	15/06	３００Ｚ
Ｇ１０Ｌ　15/16　(2006.01)	Ｇ１０Ｌ	15/16	

ＦＩ

請求項の数　16　（全22頁）

(21)出願番号　特願2022-545879(P2022-545879)
(86)(22)出願日　令和3年1月15日(2021.1.15)
(65)公表番号　特表2023-503717(P2023-503717A)
(43)公表日　令和5年1月31日(2023.1.31)
(86)国際出願番号　PCT/US2021/013759
(87)国際公開番号　WO2021/154520
(87)国際公開日　令和3年8月5日(2021.8.5)
審査請求日　令和4年10月26日(2022.10.26)
(31)優先権主張番号　62/966,823
(32)優先日　令和2年1月28日(2020.1.28)
(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２年４月９日にウェブサイトのアドレス　ｈｔｔｐｓ：／／ｉｅｅｅｘｐｌｏｒ
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　ペイザー、チャールズ ケイレブ
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・エンド音声認識における固有名詞認識

(57)【特許請求の範囲】

【請求項１】

データ処理ハードウェア（５１０）上での実行時に、前記データ処理ハードウェア（５１０）に動作を実行させるコンピュータが実施する方法（４００）であって、前記動作は、

最小単語誤り率損失関数を使用して音声認識モデル（２００）をトレーニングすることであって、

固有名詞を含むトレーニング例（３０２）を受信すること、

前記トレーニング例（３０２）に対応する複数の仮説（２２２）を生成すること、前記複数の仮説（２２２）の各仮説（２２２）は、前記固有名詞を表し、かつ仮説（２２２）が前記固有名詞を表す可能性を示す対応する確率を含んでおり、

前記複数の仮説（２２２）のうちの１つに関連付けられた対応する確率がペナルティ基準を満たすことを決定すること、前記ペナルティ基準は、

前記対応する確率が確率しきい値を満たしていること、および

前記対応する確率を含む前記仮説（２２２）が固有名詞を誤って表していることを示しており、

前記最小単語誤り率損失関数にペナルティ（３３２）を適用すること

によって、前記音声認識モデル（２００）をトレーニングすることを含んでおり、

前記音声認識モデル（２００）が、

リカレントニューラルネットワークトランスデューサ（以下、ＲＮＮ－Ｔとする）デコーダ（２２０）を含む第１のパスのネットワーク（２０６）と、

リッスン・アテンド・スペル（以下、ＬＡＳとする）デコーダ（２３０）を含む第２のパスのネットワーク（２０８）とを備えており、
前記最小単語誤り率損失関数を使用したトレーニングは、前記ＬＡＳデコーダ（２３０）において行われる、コンピュータが実施する方法（４００）。
【請求項２】
前記対応する確率が他の仮説（２２２）に関連付けられた対応する確率よりも大きい場合、前記対応する確率が前記確率しきい値を満たす、請求項１に記載のコンピュータが実施する方法（４００）。
【請求項３】
前記音声認識モデルは、共有エンコーダをさらに備え、前記共有エンコーダは、前記第１のパスのネットワーク（２０６）および前記第２のパスのネットワーク（２０８）の各々に対して音響フレーム（２１２）をエンコードする、請求項１に記載の方法（４００）。
【請求項４】
前記動作は、
前記ＲＮＮ－Ｔデコーダ（２２０）をトレーニングすること、
前記最小単語誤り率損失関数を使用して前記ＬＡＳデコーダ（２３０）をトレーニングする前に、トレーニングされた前記ＲＮＮ－Ｔデコーダ（２２０）のパラメータが固定された状態で、前記ＬＡＳデコーダ（２３０）をトレーニングすること、をさらに含む、請求項１または３に記載のコンピュータが実施する方法（４００）。
【請求項５】
前記動作は、確率を前記複数の仮説（２２２）の各仮説（２２２）に割り当てることをさらに含む、請求項１乃至４のいずれか一項に記載のコンピュータが実施する方法（４００）。
【請求項６】
前記動作は、
誤った仮説（２２２）を受信すること、
個別の確率を前記誤った仮説（２２２）に割り当てること、をさらに含み、
前記ペナルティ基準は、仮説（２２２）が生成された誤った仮説（２２２）を含むという表示をさらに含む、請求項１乃至５のいずれか一項に記載のコンピュータが実施する方法（４００）。
【請求項７】
前記誤った仮説（２２２）は、固有名詞に対する音声学的類似性を含む、請求項６に記載のコンピュータが実施する方法（４００）。
【請求項８】
前記動作が、前記複数の仮説（２２２）のうちの生成された仮説に対して前記誤った仮説を置換することをさらに含む、請求項６または７に記載のコンピュータが実施する方法（４００）。
【請求項９】
システム（５００）であって、
データ処理ハードウェア（５１０）と、
前記データ処理ハードウェア（５１０）と通信するメモリハードウェア（５２０）と、を備え、前記メモリハードウェア（５２０）は、命令を格納しており、前記命令は、前記データ処理ハードウェア（５１０）上での実行時に、前記データ処理ハードウェア（５１０）に動作を実行させ、前記動作は、
最小単語誤り率損失関数を使用して音声認識モデル（２００）をトレーニングすることであって、
固有名詞を含むトレーニング例（３０２）を受信すること、
前記トレーニング例（３０２）に対応する複数の仮説（２２２）を生成すること、
前記複数の仮説（２２２）の各仮説（２２２）は、前記固有名詞を表し、かつ仮説（２２２）が前記固有名詞を表す可能性を示す対応する確率を含んでおり、

前記複数の仮説（２２２）のうちの１つに関連付けられた対応する確率がペナルティ基準を満たすことを決定すること、前記ペナルティ基準は、
前記対応する確率が確率しきい値を満たしていること、および
前記対応する確率を含む前記仮説（２２２）が固有名詞を誤って表していることを示しており、
前記最小単語誤り率損失関数にペナルティ（３３２）を適用すること
によって、前記音声認識モデル（２００）をトレーニングすることを含んでおり、
前記システムは、
リカレントニューラルネットワークトランスデューサ（以下、ＲＮＮ－Ｔとする）デコーダ（２２２）を含む第１のパスのネットワーク（２０６）と、
リッスン・アテンド・スペル（以下、ＬＡＳとする）デコーダ（２３０）を含む第２のパスのネットワーク（２０８）と、をさらに備えており、
前記音声認識モデル（２００）は、前記第１のパスのネットワーク（２０６）および前記第２のパスのネットワーク（２０８）を含んでおり、
前記最小単語誤り率損失関数を使用したトレーニングは、前記ＬＡＳデコーダ（２３０）において行われる、システム（５００）。

【請求項１０】
前記対応する確率が他の仮説（２２２）に関連付けられた対応する確率よりも大きい場合、前記対応する確率が前記確率しきい値を満たす、請求項９に記載のシステム（５００）。

【請求項１１】
前記第１のパスのネットワーク（２０６）および前記第２のパスのネットワーク（２０８）の各々に対して音響フレーム（２１２）をエンコードするように構成された共有エンコーダ（２１０）をさらに備える、請求項９に記載のシステム（５００）。

【請求項１２】
前記動作は、
前記ＲＮＮ－Ｔデコーダ（２２０）をトレーニングすること、
前記最小単語誤り率損失関数を使用してＬＡＳデコーダ（２３０）をトレーニングする前に、トレーニングされた前記ＲＮＮ－Ｔデコーダ（２２０）のパラメータが固定された状態で、前記ＬＡＳデコーダ（２３０）をトレーニングすること、をさらに含む、請求項９または１１に記載のシステム（５００）。

【請求項１３】
前記動作は、確率を前記複数の仮説（２２２）の各仮説（２２２）に割り当てることをさらに含む、請求項９乃至１２のいずれか一項に記載のシステム（５００）。

【請求項１４】
前記動作は、
誤った仮説（２２２）を受信すること、
個別の確率を前記誤った仮説（２２２）に割り当てること、をさらに含み、
前記ペナルティ基準は、仮説（２２２）が生成された誤った仮説（２２２）を含むという表示をさらに含む、請求項９乃至１３のいずれか一項に記載のシステム（５００）。

【請求項１５】
前記誤った仮説は、前記固有名詞に対する音声学的類似性を含む、請求項１４に記載のシステム（５００）。

【請求項１６】
前記動作は、前記複数の仮説（２２２）のうちの生成された仮説（２２２）に対して前記誤った仮説（２２２）を置換することをさらに含む、請求項１４または１５に記載のシステム（５００）。

【発明の詳細な説明】
【技術分野】
【０００１】

本開示は、エンド・ツー・エンド音声認識における固有名詞認識に関する。

【背景技術】

【０００２】

最新の自動音声認識（ＡＳＲ：ａｕｔｏｍａｔｅｄ　ｓｐｅｅｃｈ　ｒｅｃｏｇｎｉｔｉｏｎ）システムは、高品質（例えば、低い単語誤り率（ＷＥＲ：ｗｏｒｄ　ｅｒｒｏｒ　ｒａｔｅ））のみならず、低遅延（例えば、ユーザが話してから文字起こし（ｔｒａｎｓｃｒｉｐｔｉｏｎ）が表示されるまでの短い遅延）を提供することに重点を置いている。さらに、現在ＡＳＲシステムを使用する場合、ＡＳＲシステムは、リアルタイムに対応するか、またはリアルタイムよりもさらに高速に対応するストリーミング方式で発話をデコードすることが要求されている。例えば、ユーザとの直接対話を行う携帯電話にＡＳＲシステムが搭載されている場合、ＡＳＲシステムを使用する携帯電話上のアプリケーションは、単語が話されるとすぐに画面上に表示されるように音声認識がストリーミングされることを必要とする場合がある。ここで、携帯電話のユーザは、遅延に対する許容度が低い可能性もある。この低い許容度により、音声認識は、ユーザエクスペリエンスに悪影響を与える可能性のある遅延および不正確性による影響を最小限に抑えるようにモバイルデバイス上で動作することを目指している。

【発明の概要】

【０００３】

本開示の一態様は、データ処理ハードウェア上での実行時に、データ処理ハードウェアに、動作を実行させるコンピュータが実施する方法を提供し、動作は、最小単語誤り率損失関数を使用して音声認識モデルをトレーニングすることであって、固有名詞を含むトレーニング例を受信すること、トレーニング例に対応する複数の仮説を生成すること（複数の仮説の各仮説は、固有名詞を表し、かつ仮説が固有名詞を表す可能性を示す対応する確率を含んでいる）、複数の仮説のうちの１つに関連付けられた対応する確率がペナルティ基準を満たすと決定すること、最小単語誤り率損失関数にペナルティを適用することによって、音声認識モデルをトレーニングすることを含む。ペナルティ基準は、対応する確率が確率しきい値を満たしていること、および関連付けられた仮説が固有名詞を誤って表していることを示す。

【０００４】

本開示の実施形態は、以下の任意の特徴のうちの１つまたは複数を含み得る。いくつかの実施形態では、音声認識モデルは、リカレントニューラルネットワークトランスデューサ（ＲＮＮ－Ｔ）デコーダを含む第１のパスのネットワークと、リッスン・アテンド・スペル（ＬＡＳ）デコーダを含む第２のパスのネットワークとを含む２パスアーキテクチャを含む。これらの実施形態では、音声認識モデルは、第１のパスのネットワークおよび第２のパスのネットワークの各々に対して音響フレームをエンコードする共有エンコーダをさらに含み得る。これらの実施形態における最小単語誤り率損失関数を使用したトレーニングは、ＬＡＳエンコーダにおいて行われ得る。動作は、ＲＮＮ－Ｔデコーダをトレーニングすること、および最小単語誤り率損失関数を使用してＬＡＳデコーダをトレーニングする前に、トレーニングされたＲＮＮ－Ｔデコーダのパラメータが固定された状態で、ＬＡＳデコーダをトレーニングすることをさらに含み得る。

【０００５】

いくつかの例では、対応する確率が他の仮説に関連付けられた対応する確率よりも大きい場合、対応する確率が確率しきい値を満たす。動作は、確率を複数の仮説の各仮説に割り当てることをさらに含み得る。いくつかの実施形態では、動作は、誤った仮説を受信すること、および個別の確率を誤った仮説に割り当てることをさらに含み、ペナルティ基準は、仮説が生成された誤った仮説を含むという表示をさらに含む。これらの例では、誤った仮説は、固有名詞に対する音声学的（ｐｈｏｎｅｔｉｃａｌｌｙ）類似性を含み得、かつ／または動作は、複数の仮説のうちの生成された仮説に対して誤った仮説に置換することをさらに含み得る。

【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と通信するメモリハードウェアとを含むシステムを提供し、メモリハードウェアは、命令を格納し、命令は、データ処理ハードウェアによる実行時に、データ処理ハードウェアに動作を実行させ、動作は、最小単語誤り率損失関数を使用して音声認識モデルをトレーニングすることであって、固有名詞を含むトレーニング例を受信すること、トレーニング例に対応する複数の仮説を生成すること（複数の仮説の各仮説は固有名詞を表し、かつ仮説が固有名詞を表す可能性を示す対応する確率を含んでいる）、複数の仮説のうちの１つに関連付けられた対応する確率がペナルティ基準を満たすと決定すること、最小単語誤り率損失関数にペナルティを適用することによって、音声認識モデルをトレーニングすることを含む。ペナルティ基準は、対応する確率が確率しきい値を満たしていること、および関連付けられた仮説が固有名詞を誤って表していることを示す。

【０００７】

この態様は、以下の任意の特徴のうちの１つまたは複数を含み得る。いくつかの実施形態では、システムは、リカレントニューラルネットワークトランスデューサ（ＲＮＮ－Ｔ）デコーダを備える第１のパスのネットワークと、リッスン・アテンド・スペル（ＬＡＳ）デコーダを備える第２のパスのネットワークとをさらに含み、音声認識モデルは、第１のパスのネットワークと第２のパスのネットワークとを備える。これらの実施形態では、システムは、第１のパスのネットワークおよび第２のパスのネットワークの各々に対して音響フレームをエンコードするように構成された共有エンコーダをも含み得る。これらの実施形態における最小単語誤り率損失関数を使用したトレーニングは、ＬＡＳデコーダで行われ得る。動作は、ＲＮＮ－Ｔデコーダをトレーニングすること、および最小単語誤り率損失関数を使用してＬＡＳデコーダをトレーニングする前に、トレーニングされたＲＮＮ－Ｔデコーダのパラメータが固定された状態で、ＬＡＳデコーダをトレーニングすることをさらに含み得る。

【０００８】

いくつかの例では、対応する確率が他の仮説に関連付けられた対応する確率よりも大きい場合、対応する確率は確率しきい値を満たす。動作は、確率を複数の仮説の各仮説に割り当てることをさらに含み得る。いくつかの実施形態では、動作は、誤った仮説を受信すること、および個別の確率を誤った仮説に割り当てることをさらに含み、ペナルティ基準は、仮説が生成された誤った仮説を含むという表示をさらに含む。これらの例では、誤った仮説は、固有名詞に対する音声学的類似性を含み得、かつ／または動作は、複数の仮説のうちの生成された仮説に対して誤った仮説に置換することをさらに含み得る。

【０００９】

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

【図面の簡単な説明】

【００１０】

【図１Ａ】ジョイント音響モデルおよびテキストモデルを備えた２パス音声認識アーキテクチャを使用する例示的な音声環境の概略図である。

【図１Ｂ】ジョイント音響モデルおよびテキストモデルを備えた２パス音声認識アーキテクチャを使用する例示的な音声環境の概略図である。

【図２】音声認識のための例示的な２パス音声認識アーキテクチャの概略図である。

【図３Ａ】図２の２パス音声認識アーキテクチャをトレーニングするための例示的なトレーニング手順の概略図である。

【図３Ｂ】図２の２パス音声認識アーキテクチャをトレーニングするための例示的なトレーニング手順の概略図である。

【図３Ｃ】図２の２パス音声認識アーキテクチャをトレーニングするための例示的なトレーニング手順の概略図である。

【図４】図２の２パス音声認識アーキテクチャをトレーニングする方法のための動作の例

示的な構成のフローチャートである。
【図５】本明細書で説明されるシステムおよび方法を実施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】
【００１１】
様々な図面の同様の参照記号は、同様の構成要素を示す。
音声認識は、モバイル環境の非拘束性および機敏性の要求を満たすために進化し続けている。自動音声認識システム（ＡＳＲ）の品質を向上させるために、新たな音声認識アーキテクチャまたは既存のアーキテクチャの改良が引き続き開発されている。例えば、音声認識は、当初、各モデルが専用の目的を持つ複数のモデルを採用していた。例えば、ＡＳＲシステムは、音響モデル（ＡＭ）、発音モデル（ＰＭ）、および言語モデル（ＬＭ）を含んでいた。音響モデルは、音声のセグメント（即ち、音声のフレーム）を音素（ｐｈｏｎｅｍｅｓ）にマッピングした。発音モデルは、これらの音素をつなぎ合わせて単語を形成し、言語モデルは、所与のフレーズの可能性（即ち、単語のシーケンスの確率）を表現するために使用された。これらの個々のモデルは連携して機能したが、各モデルは個別にトレーニングされ、多くの場合、異なるデータセットで手動で設計された。
【００１２】
個別のモデルの手法により、特に所与のモデルに対するトレーニングコーパス（即ち、トレーニングデータの集合体）がモデルの有効性に対応している場合に、音声認識システムの精度をかなり向上させることが可能になったが、個別のモデルを個別にトレーニングする必要性により、それ自体が複雑になるため、統合モデルを備えたアーキテクチャが採用された。これらの統合モデルは、単一のニューラルネットワークを使用して、音声波形（即ち、入力シーケンス）を出力センテンス（即ち、出力シーケンス）に直接マッピングしようとするものである。これにより、任意の音声特徴のシーケンスが与えられると、単語（または書記素（ｇｒａｐｈｅｍｅｓ））のシーケンスが生成されるシーケンス・ツー・シーケンスの手法が実現された。シーケンス・ツー・シーケンスモデルの例には、「アテンションベース」モデルおよび「リッスン・アテンド・スペル」（ＬＡＳ）モデルが含まれる。ＬＡＳモデルは、リスナー（ｌｉｓｔｅｎｅｒ）コンポーネント、アテンダ（ａｔｔｅｎｄｅｒ）コンポーネント、およびスペラー（ｓｐｅｌｌｅｒ）コンポーネントを使用して、音声の発話を文字に変換する。ここで、リスナーは、音声入力（例えば、音声入力の時間周波数表現）を受信し、音声入力をより高レベルの特徴表現にマッピングするリカレントニューラルネットワーク（ＲＮＮ：ｒｅｃｕｒｒｅｎｔ ｎｅｕｒａｌ ｎｅｔｗｏｒｋ）エンコーダである。アテンダは、より高レベルの特徴をアテンションして、入力特徴と予測されるサブワード単位（例えば、書記素または単語ピース）との間のアラインメントを学習する。スペラーは、アテンションベースのＲＮＮデコーダであり、仮定単語のセットに対して確率分布を生成することによって、入力から文字シーケンスを生成する。統合化された構造により、モデルの全てのコンポーネントを単一のエンド・ツー・エンド（Ｅ２Ｅ：ｅｎｄ－ｔｏ－ｅｎｄ）ニューラルネットワークとして共同でトレーニングさせることができる。ここで、Ｅ２Ｅモデルとは、アーキテクチャが全てニューラルネットワークで構成されているモデルを指す。完全なニューラルネットワークは、外部コンポーネントおよび／または手動で設計したコンポーネント（例えば、有限状態トランスデューサ、辞書（ｌｅｘｉｃｏｎ）、またはテキスト正規化モジュール）なしで機能する。さらに、Ｅ２Ｅモデルをトレーニングする場合、これらのモデルは通常、決定木からのブートストラップ、または別のシステムからの時間調整を必要としない。
【００１３】
初期のＥ２Ｅモデルは正確であり、個別にトレーニングされたモデルよりもトレーニングが改善されたが、ＬＡＳモデルなどのこれらのＥ２Ｅモデルは、出力テキストを生成する前に入力シーケンス全体を確認することによって機能していたため、入力が受信されたときに出力をストリーミングすることはできなかった。ストリーミング機能がないと、ＬＡＳモデルは、リアルタイムの音声文字起こし（ｖｏｉｃｅ ｔｒａｎｓｃｒｉｐｔｉｏ

ｎ）を実行することができない。この欠陥のため、遅延に敏感な、かつ／またはリアルタイムの音声文字起こしを必要とする音声アプリケーションに対してＬＡＳモデルを搭載すると、問題が発生する可能性がある。このため、リアルタイムアプリケーション（例えば、リアルタイム通信アプリケーション）に依存することが多いモバイル技術（例えば、携帯電話）にとって、ＬＡＳモデルだけでは、理想的なモデルではない。

【００１４】

さらに、音響モデル、発音モデル、および言語モデル、またはそれらが共に構成されているモデルを有する音声認識システムは、これらのモデルに関連する比較的大規模のサーチグラフをサーチする必要があるデコーダに依存し得る。大規模のサーチグラフでは、この種の音声認識システムを完全オンデバイスでホストするのに有利ではない。ここで、音声認識システムが「オンデバイス（ｏｎ－ｄｅｖｉｃｅ）」でホストされている場合、音声入力を受信するデバイスは、そのプロセッサ（単数または複数）を使用して音声認識システムの機能を実行する。例えば、音声認識システムが完全にオンデバイスでホストされている場合、デバイスのプロセッサは、音声認識システムの機能を実行するために、デバイス外のコンピューティングリソースと連携する必要はない。完全にオンデバイスではない音声認識を実行するデバイスは、音声認識システムの少なくとも一部の機能を実行するために、リモートコンピューティング（例えば、リモートコンピューティングシステムまたはクラウドコンピューティング）、従ってオンライン接続に依存している。例えば、音声認識システムは、サーバベースのモデルとのネットワーク接続を使用して、大規模なサーチグラフによりデコーディングを実行する。

【００１５】

残念ながら、リモート接続に依存している状態では、音声認識システムは、遅延の問題および／または通信ネットワークに固有の信頼性の低さに対して脆弱になる。これらの問題を回避することによって音声認識の有用性を向上させるために、音声認識システムは、リカレントニューラルネットワークトランスデューサ（ＲＮＮ－Ｔ）として知られるシーケンス・ツー・シーケンス（ｓｅｑｕｅｎｃｅ－ｔｏ－ｓｅｑｕｅｎｃｅ）モデルの形態に再び進化した。ＲＮＮ－Ｔはアテンション機構を採用しておらず、かつ出力（例えば、センテンス）を生成するためにシーケンス全体（例えば、音声波形）を処理する必要がある他のシーケンス・ツー・シーケンスモデルとは異なり、ＲＮＮ－Ｔは、入力サンプルを連続的に処理して、出力シンボルをストリーミングするという、リアルタイム通信にとって特に魅力的な特徴を有している。例えば、ＲＮＮ－Ｔを使用した音声認識では、話した通りに文字が１つずつ出力され得る。ここで、ＲＮＮ－Ｔは、モデルによって予測されたシンボルを自身にフィードバックするフィードバックループを使用して、次のシンボルを予測する。ＲＮＮ－Ｔのデコーディングは、大規模なデコーダグラフではなく、単一のニューラルネットワークを介したビームサーチを含むため、ＲＮＮ－Ｔは、サーバベースの音声認識モデルの数分の１のサイズにスケーリングすることができる。サイズの縮小により、ＲＮＮ－Ｔは完全にオンデバイスで搭載され、オフラインで（即ち、ネットワーク接続なしで）動作させることができるため、通信ネットワークの信頼性の問題を回避することができる。

【００１６】

音声認識システムが低遅延で動作することに加えて、音声認識システムには、音声を正確に認識することが求められる。音声認識を実行するモデルの場合、モデルの精度を定義するメトリックとして、単語誤り率（ＷＥＲ）が用いられることが多い。ＷＥＲは、実際に話された単語の数と比較して、どれだけ単語が変更されたかを示す尺度である。一般に、これらの単語の変更は、置換（即ち、単語が置き換えられる場合）、挿入（即ち、単語が追加される場合）、および／または削除（即ち、単語が省略される場合）を指す。例えば、話者は「カー（ｃａｒ）」と言っているが、ＡＳＲシステムは、「カー（ｃａｒ）」という単語を「バー（ｂａｒ）」と文字起こしする。これは、音声の（ｐｈｏｎｅｔｉｃ）類似性による置換の例である。他のＡＳＲシステムと比較してＡＳＲシステムの能力を測定する場合、ＷＥＲは、別のシステムまたはあるベースラインと比較して、改善または

品質保証能力の尺度を示すことができる。

【００１７】

ＲＮＮ－Ｔモデルは、オンデバイスの音声認識に関する有力な候補モデルとして有望であることを示したが、ＲＮＮ－Ｔモデルのみでは、品質（例えば、音声認識精度）の観点で、大規模な最先端の従来モデル（例えば、別個のＡＭ、ＰＭ、およびＬＭを備えたサーバベースのモデル）に遅れをとっている。しかし、非ストリーミングＥ２Ｅ、ＬＡＳモデルは、大規模な最先端の従来モデルに匹敵する音声認識品質を備えている。非ストリーミングＥ２Ｅ　ＬＡＳモデルの品質を活用するために、ＲＮＮ－Ｔネットワークの第１のパスのコンポーネントと、それに続くＬＡＳネットワークの第２のパスのコンポーネントとを含む２パス音声認識システム（例えば、図２に示す）が開発された。この設計により、２パスモデルは、低遅延のＲＮＮ－Ｔモデルのストリーミング特性の恩恵を受け、ＬＡＳネットワークを組み込んだ第２のパスを通じてＲＮＮ－Ｔモデルの精度を向上させている。ＬＡＳネットワークは、ＲＮＮ－Ｔモデルのみと比較して遅延を増加させるが、遅延の増加は、適度にわずかであり、かつオンデバイス動作に関する遅延制約に適合している。精度に関しては、２パスモデルは、ＲＮＮ－Ｔ単独と比較した場合に１７～２２％のＷＥＲ低減を達成し、大規模な従来モデルと比較した場合に同程度のＷＥＲを有している。

【００１８】

ＲＮＮ－Ｔネットワークの第１のパスとＬＡＳネットワークの第２のパスを備えた２パスモデルでも、特に稀少な単語または一般的でない単語の場合に、トレードオフがある。これらのタイプの単語は、テイル発話（ｔａｉｌ　ｕｔｔｅｒａｎｃｅｓ）と呼ばれることがあり、曖昧さ、トレーニングでの希少性、または特殊な言語化によって、音声システムが文字起こしするのが本質的に困難である。テイル発話の例には、アクセントのある話し言葉、異言語間の話し言葉、数字、および固有名詞が含まれる。例えば、固有名詞は、２パスモデルを使用してＡＳＲをストリーミングする際の課題を提示する。これは、特定の名前がトレーニング中にまれにしか登場にないか、またはまったく登場しない場合があるものの、より一般的な単語に似た発音を有する可能性があるためである。これまで、従来のモデルは、固有名詞の発音に関する知識を注入することによって、発音モデル（ＰＭ：ｐｒｏｎｕｎｃｉａｔｉｏｎ　ｍｏｄｅｌ）を最適化して、テイル性能を改善することができる。残念ながら、２パスアーキテクチャには、固有名詞の発音で特別にトレーニングすることができる明示的な発音モデル（ＰＭ）と、固有名詞に多く触れる大規模なコーパスでトレーニングすることができる言語モデル（ＬＭ：ｌａｎｇｕａｇｅｍｏｄｅｌ）を欠いている。ストリーミング２パスシステムにおいて、適切な既知の知識を注入するための特定の場所としてのＰＭがなければ、固有名詞の発音などの特定の要件をモデル化することはより困難でなる。一部のモデルでは、追加のトレーニングデータまたはモデルを組み込むことによって、一般的でない単語／稀少な単語による問題を改善しようという試みがなされているが、これらの技法では、モデルのサイズ、トレーニング時間、および／または推論コストが増加する。

【００１９】

固有名詞および／またはその他のテイル発話に対する２パスモデルの有効性を高めるために、２パスアーキテクチャでは、カスタマイズされた最小単語誤り率（ＭＷＥＲ：ｍｉｎｉｍｕｍ　ｗｏｒｄ　ｅｒｒｏｒ　ｒａｔｅ）損失基準が使用される。この損失基準は、特に固有名詞の認識を強調することを目的としている。損失基準を使用して固有名詞の認識を向上させることによって、音声認識システムは、トレーニング時の新たなデータまたは推論時の外部モデルを必要としない。ここで、損失基準の２つの異なる方法を固有名詞認識に関して使用することができる。第１の方法は、グラウンドトゥルースの文字起こしにおいて固有名詞を識別して、トレーニング中に固有名詞を外している仮説の損失を増加させるエンティティタグ付けシステムを含む。第２の方法は、ＭＷＥＲビームに追加の仮説を注入することであり、追加の仮説は、音声学的に類似した代替語（ａｌｔｅｒｎａｔｉｖｅｓ）に置換された固有名詞に対応する。例えば、「ウォルマート（Ｗａｌｍａｒｔ）」に音声学的に類似した代替語として、「ホールマーク（Ｈａｌｌｍａｒｋ）」という

追加の仮説が追加される。第２の手法では、トレーニングのプロセスによって、可能性のある間違いと潜在的な代替語をモデルに認識させる。様々な固有名詞テストセットにおいて、これらのカスタム損失基準方法は、カスタム損失基準のない従来の２パスアーキテクチャと比較して、ＷＥＲを相対的に２～７％削減することができる。

【００２０】

図１Ａおよび図１Ｂは、発話環境１００の例である。発話環境１００において、ユーザデバイス１１０などのコンピューティングデバイスと対話するユーザの１０の方法は、音声入力を介するものであり得る。ユーザデバイス１１０（一般にデバイス１１０とも呼ばれる）は、発話対応環境１００内の１人または複数人のユーザ１０からの音（例えば、ストリーミング音声データ）をキャプチャするように構成されている。ここで、ストリーミング音声データ１２は、デバイス１１０によってキャプチャされる可聴の問い合わせ（クエリ）、デバイス１１０に対する命令（コマンド）、または可聴の会話（コミュニケーション）としての役割を持つ、ユーザ１０によって話された発話を指すことができる。デバイス１１０の発話対応システムは、問い合わせに応答することによって、かつ／またはコマンドを実行させることによって、問い合わせまたは命令を処理し得る。

【００２１】

ユーザデバイス１１０は、ユーザ１０に関連付けられ、かつ音声データ１２を受信することが可能な任意のコンピューティングデバイスに対応し得る。ユーザデバイス１１０のいくつかの例は、モバイルデバイス（例えば、携帯電話、タブレット、ラップトップなど）、コンピュータ、ウェアラブルデバイス（例えば、スマートウォッチ）、スマート家電、モノのインターネット（ＩｏＴ）デバイス、スマートスピーカなどを含むが、これらに限定されない。ユーザデバイス１１０は、データ処理ハードウェア１１２と、データ処理ハードウェア１１２と通信するメモリハードウェア１１４とを含み、メモリハードウェア１１４は、命令を格納し、命令は、データ処理ハードウェア１１２による実行時に、データ処理ハードウェア１１２に１つまたは複数の動作を実行させる。ユーザデバイス１１０は、発話対応システム１００内で話された発話１２をキャプチャして電気信号に変換するための音声キャプチャデバイス（例えば、マイクロフォン）１１６、１１６ａと、可聴音声信号を（例えば、デバイス１１０からの出力音声データとして）伝達するための発話出力デバイス（例えばスピーカ）１１６、１１６ｂとを有する音声サブシステム１１６をさらに含む。図示される例では、ユーザデバイス１１０は単一の音声キャプチャデバイス１１６ａを実装しているが、ユーザデバイス１１０は、本開示の範囲から逸脱することなく、音声キャプチャデバイス１１６ａのアレイを実装してもよく、それにより、アレイ内の１つまたは複数のキャプチャデバイス１１６ａは、ユーザデバイス１１０上に物理的に存在していないが、音声サブシステム１１６と通信状態になり得る。（例えば、ハードウェア１１２、１１４を使用する）ユーザデバイス１１０は、音声認識器２００を使用して、ストリーミング音声データ１２に対して音声認識処理を実行するようにさらに構成される。いくつかの例では、音声キャプチャデバイス１１６ａを含むユーザデバイス１１０の音声サブシステム１１６は、音声データ１２（例えば、話された発話）を受信し、音声データ１２を音声認識器２００と互換性のあるデジタル形式に変換するように構成される。デジタル形式は、メルフレーム（ｍｅｌ　ｆｒａｍｅｓ）などの音響フレーム（例えば、パラメータ化された音響フレーム）に対応し得る。例えば、パラメータ化された音響フレームは、ログメルフィルタバンク（ｌｏｇ－ｍｅｌ　ｆｉｔｅｒｂａｎｋ）エネルギーに対応する。

【００２２】

図１Ａなどのいくつかの例では、ユーザ１０は、音声認識器２００を使用するユーザデバイス１１０のプログラムまたはアプリケーション１１８と対話する。例えば、図１Ａは、ユーザ１０が自動アシスタントアプリケーションと通信している状態を示している。この例では、ユーザ１０が自動アシスタントに「今夜のコンサートは何時から？（Ｗｈａｔ　ｔｉｍｅ　ｉｓ　ｔｈｅ　ｃｏｎｃｅｒｔ　ｔｏｎｉｇｈｔ？）」と尋ねている。ユーザ１０からのこの質問は、音声キャプチャデバイス１１６ａによってキャプチャされ、ユーザ

デバイス１１０の音声サブシステム１１６によって処理される話された発話１２である。この例では、ユーザデバイス１１０の音声認識器２００は、「今夜のコンサートは何時から」という音声入力２０２を（例えば、音響フレームとして）受信し、音声入力２０２を文字起こし２０４（例えば、「今夜のコンサートは何時から？」というテキスト表現）に転写する。ここで、アプリケーション１１８の自動アシスタントは、自然言語処理を使用して、ユーザ１０によって投げかけられた質問に対して回答し得る。自然言語処理とは、一般に、書き言葉（例えば、文字起こし２０４）を解釈し、書き言葉が何らかのアクションを促しているかどうかを判断するプロセスを指す。この例では、自動アシスタントは、自然言語処理を使用して、ユーザ１０からの質問がユーザのスケジュール、より具体的にはユーザのスケジュールでのコンサートに関するものであることを認識する。自動アシスタントは、自然言語処理でこれらの詳細を認識することによって、ユーザの問い合わせに対して、「今夜のコンサートは午後８時３０分に開場します（Ｄｏｏｒｓ ｏｐｅｎ ａｔ ８：３０ ｐｍ ｆｏｒ ｔｈｅ ｃｏｎｃｅｒｔ ｔｏｎｉｇｈｔ）」という回答を返す。いくつかの構成では、自然言語処理は、ユーザデバイス１１０のデータ処理ハードウェア１１２と通信するリモートシステム上で行われ得る。

【００２３】

図１Ｂは、音声認識器２００を用いた音声認識の別の例である。この例では、ユーザデバイス１１０に関連付けられたユーザ１０は、通信アプリケーション１１８を用いてジェーン・ドゥという名前の友人と通信している。ここで、テッドという名前のユーザ１０は、音声認識器２００に自身の音声入力を文字起こしさせることによって、ジェーンと通信する。音声キャプチャデバイス１１６は、これらの音声入力をキャプチャし、それらを音声認識器２００にデジタル形式（例えば、音響フレーム）で伝達する。音声認識器２００は、これらの音響フレームを、通信アプリケーション１１８を介してジェーンに送信されるテキストに文字起こしする。この種類のアプリケーション１１８はテキストを介して通信するため、音声認識器２００からの文字起こし２０４は、さらなる処理（例えば、自然言語処理）なしでジェーンに送信され得る。

【００２４】

図２などのいくつかの例では、音声認識器２００は、２パス音声認識アーキテクチャ（または単に「２パスアーキテクチャ」）で構成される。一般的に、音声認識器２００の２パスアーキテクチャは、少なくとも１つのエンコーダ２１０、ＲＮＮ－Ｔデコーダ２２０、およびＬＡＳデコーダ２３０を含む。２パスデコーディングにおいて、第２のパス２０８（例えば、ＬＡＳデコーダ２３０として示される）は、第１のパス２０６（例えば、ＲＮＮ－Ｔデコーダ２２０として示される）からの初期出力を格子再スコアリング（ｌａｔｔｉｃｅ ｒｅｓｃｏｒｉｎｇ）またはｎベスト再ランク付け（ｎ－ｂｅｓｔ ｒｅ－ｒａｎｋｉｎｇ）などの技術を用いて改善し得る。言い換えれば、ＲＮＮ－Ｔデコーダ２２０はストリーミング予測（例えば、１組のＮベスト仮説（Ｎ－ｂｅｓｔ ｈｙｐｏｔｈｅｓｅｓ））を生成し、ＬＡＳデコーダ２３０は、予測を確定する（例えば、１つのベストの再スコアリングされた仮説を識別する）。ここで、具体的には、ＬＡＳデコーダ２３０は、ＲＮＮ－Ｔデコーダ２２０からストリーミングされた仮説y_Rを再スコアリングする。一般に、ＬＡＳデコーダ２３０は、ＲＮＮ－Ｔデコーダ２２０からストリーミングされた仮説y_Rを再スコアリングする再スコアリングモードで機能すると説明されているが、ＬＡＳデコーダ２３０は、設計またはその他の要因（例えば、発話の長さ）に応じてビームサーチモードなどの異なるモードで動作することも可能である。

【００２５】

少なくとも１つのエンコーダ２１０は、ストリーミング音声データ１２に対応する音響フレームを音声入力２０２として受信するように構成される。音響フレームは、音声サブシステム１１６によってパラメータ化された音響フレーム（例えば、メルフレームおよび／またはスペクトルフレーム）に事前に処理され得る。いくつかの実施形態では、パラメータ化された音響フレームは、ログメル特徴（ｌｏｇ－ｍｅｌ ｆｅａｔｕｒｅｓ）を有するログメルフィルタバンクエネルギーに対応する。例えば、音声サブシステム１１６に

よって出力され、エンコーダ２１０に入力されるパラメータ化された入力音響フレームは、$x=(x_1,...,x_T)$として表すことができ、ここで、

【００２６】

【数１】

$$x_t \in \mathbb{R}^d$$

は、ログメルフィルタバンクエネルギーであり、Ｔはｘのフレーム数を示し、ｄはログメル特徴の数を表す。いくつかの例では、各パラメータ化された音響フレームは、短いシフトウィンドウ（例えば、３２ミリ秒、１０ミリ秒ごとにシフト）内で計算された１２８次元のログメル特徴を含む。各特徴は、前のフレーム（例えば、３つ前のフレーム）と重ねられて、より高次元のベクトル（例えば、３つ前のフレームを使用した５１２次元のベクトル）が形成されてもよい。次に、ベクトルを形成する特徴は、（例えば、３０ミリ秒のフレームレートに）ダウンサンプリングされ得る。エンコーダ２１０は、音声入力２０２に基づいて、エンコーディングｅを生成するように構成される。例えば、エンコーダ２１０は、エンコードされた音響フレーム（例えば、エンコードされたメルフレームまたは音響埋め込み（ａｃｏｕｓｔｉｃ　ｅｍｂｅｄｄｉｎｇｓ））を生成する。

【００２７】

エンコーダ２１０の構造は、異なる方法で実施することができるが、いくつかの実施形態では、エンコーダ２１０は、長短期記憶（ＬＳＴＭ：ｌｏｎｇ－ｓｈｏｒｔ　ｔｅｒｍ　ｍｅｍｏｒｙ）ニューラルネットワークである。例えば、エンコーダ２１０は、８個のＬＳＴＭ層を含む。ここで、各層は、２，０４８個の隠れユニットと、それに続く６４０次元の射影層（ｐｒｏｊｅｃｔｉｏｎ　ｌａｙｅｒ）とを含む。いくつかの例では、エンコーダ２１０の第２のＬＳＴＭ層の後に、短縮係数（ｒｅｄｕｃｔｉｏｎ　ｆａｃｔｏｒ）Ｎ＝２を有する時間短縮層（ｔｉｍｅ－ｒｅｄｕｃｔｉｏｎ　ｌａｙｅｒ）が挿入される。

【００２８】

いくつかの構成では、エンコーダ２１０は共有エンコーダネットワークである。言い換えれば、各パスネットワーク２０６、２０８がそれ自体の別個のエンコーダを有する代わりに、各パス２０６、２０８は単一のエンコーダ２１０を共有する。エンコーダを共有することによって、２パスアーキテクチャを使用するＡＳＲ音声認識器２００は、そのモデルサイズおよび／またはその計算コストを削減することができる。ここで、モデルサイズの縮小は、音声認識器２００が完全にオンデバイス（ｏｎ－ｄｅｖｉｃｅ）で良好に機能することを可能にするのに役立ち得る。

【００２９】

いくつかの例では、図２の音声認識器２００は、ＬＡＳデコーダ２３０の第２のパス２０８に適したものとなるようにエンコーダ２１０の出力２１２を適応させるための音響エンコーダ２４０などの追加のエンコーダをも含む。音響エンコーダ２４０は、出力２１２をエンコードされた出力２５２にさらにエンコードするように構成される。いくつかの実施形態では、音響エンコーダ２４０は、エンコーダ２１０からの出力２１２をさらにエンコードするＬＳＴＭエンコーダ（例えば、２層ＬＳＴＭエンコーダ）である。追加のエンコーダを含むことによって、エンコーダ２１０は、パス２０６、２０８の間の共有エンコーダとして依然として保持され得る。

【００３０】

第１のパス２０６を通じて、エンコーダ２１０は、音声入力２０２の各音響フレームを受信して、出力２１２（例えば、音響フレームのエンコーディングｅとして示される）を生成する。ＲＮＮ－Ｔデコーダ２２０は、各フレームの出力２１２を受信して、仮説y_Rとして示される出力２２２を各タイムステップにおいてストリーミング方式で生成する。言い換えれば、ＲＮＮ－Ｔデコーダ２２０は、フレーム毎の埋め込みｅまたは出力２１２を消費して、単語ピース出力２２２を仮説として生成し得る。いくつかの例では、ＲＮＮ－Ｔデコーダ２２０は、受信したエンコードされた音響フレーム２１２に基づいてビーム

サーチを実行することによって、Ｎベスト仮説２２２を生成する。ＲＮＮ－Ｔデコーダ２２０の構造に関して、ＲＮＮ－Ｔデコーダ２２０は、予測ネットワークおよび結合ネットワークを含み得る。ここで、予測ネットワークは、２，０４８個の隠れユニットおよび６４０次元の射影（層ごと）の２つのＬＳＴＭ層、並びに１２８ユニットの埋め込み層を有し得る。エンコーダ２１０および予測ネットワークの出力２１２は、ソフトマックス予測層を含む結合ネットワークに供給され得る。いくつかの例では、ＲＮＮ－Ｔデコーダ２２０の結合ネットワークは、６４０個の隠れユニットと、それに続く４，０９６個の大文字と小文字が混在する単語ピースを予測するソフトマックス層とを含む。

【００３１】

図２の２パスモデルにおいて、第２のパス２０８を通じて、ＬＡＳデコーダ２３０は、各フレームに関してエンコーダ２１０からの出力２１２を受信し、仮説y_Lとして指定された出力２３２を生成する。ＬＡＳデコーダ２３０がビームサーチモードで動作する場合、ＬＡＳデコーダ２３０は、出力２１２のみから出力２３２を生成し、ＲＮＮ－Ｔデコーダ２２０の出力２２２を無視する。ＬＡＳデコーダ２３０が再スコアリングモードで動作する場合、ＬＡＳデコーダ２３０は、ＲＮＮ－Ｔデコーダ２２０から（例えば、ＲＮＮ－Ｔデコーダ２２０によって生成されたＮベスト仮説に対応する）トップＫ仮説２２２、y_Rを取得し、次いで、ＬＡＳデコーダ２３０は、出力２１２をアテンション（ａｔｔｅｎｔｉｏｎ）しつつ、教師強制モードで各シーケンスに対して動作して、スコアを計算する。例えば、スコアは、シーケンスの対数確率とアテンションカバレッジペナルティ（ａｔｔｅｎｔｉｏｎ ｃｏｖｅｒａｇｅ ｐｅｎａｌｔｙ）とを組み合わせたものである。ＬＡＳデコーダ２３０は、最も高いスコアを有するシーケンスを出力２３２として選択する。言い換えれば、ＬＡＳデコーダ２３０は、ＲＮＮ－Ｔデコーダ２２０からの仮説２２２のＮベストリストから、最大尤度（ｍａｘｉｍｕｍ ｌｉｋｅｌｉｈｏｏｄ）を有する単一の仮説y_Rを選択し得る。ここで、再スコアリングモードでは、ＬＡＳデコーダ２３０は、出力２１２をアテンションするために、（例えば、４つのヘッドを有する）マルチヘッドアテンション（ｍｕｌｔｉ－ｈｅａｄｅｄ ａｔｔｅｎｔｉｏｎ）を含み得る。さらに、ＬＡＳデコーダ２３０は、予測のためのソフトマックス層を備えた２層ＬＡＳデコーダ２３０であり得る。例えば、ＬＡＳデコーダ２３０の各層は、２，０４８個の隠れユニットと、それに続く６４０次元の射影とを有する。ソフトマックス層は、ＲＮＮ－Ｔデコーダ２２０のソフトマックス層から同じ大文字と小文字が混在する単語ピースを予測するために、４，０９６次元を含み得る。

【００３２】

ニューラルネットワークは、通常、損失関数（例えば、クロスエントロピー損失関数）を定義するバックプロパゲーションによってトレーニングされる。例えば、損失関数は、ネットワークの実際の出力とネットワークの所望の出力との差として定義される。クロスエントロピー（ＣＥ）損失関数を使用してモデルをトレーニングするために、モデルは、トレーニングデータの対数尤度（ｌｏｇ－ｌｉｋｅｌｉｈｏｏｄ）を最大化することによって、ＣＥ損失関数を最適化するようにトレーニングする。図３Ａ～図３Ｃを参照すると、トレーニング手順３００は、音声認識器２００の各コンポーネントを、対応する組のトレーニングデータ３０２、３０２ａ－ｄでトレーニングすることができる。トレーニング手順３００は、データ処理ハードウェア５１０（図５）と、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５２０（図５）とを含むシステム５００上で実行することができ、メモリハードウェア５２０は命令を格納し、命令は、データ処理ハードウェア５１０上での実行時に、処理ハードウェア５１０に動作を実行させる。例えば、図２の音声認識器２００の２パスモデルアーキテクチャをトレーニングするためのトレーニング手順３００は、３つの段階３１０、３２０、３３０で行われ得る。第１段階３１０の間、トレーニング手順３００は、（例えば、ＣＥ損失関数を使用して）エンコーダ２１０およびＲＮＮ－Ｔデコーダ２２０をトレーニングする。いくつかの例では、トレーニング手順３００は、$P(y_R = y \mid x)$を最大化するようにエンコーダ２１０およびＲＮＮ－Ｔデコーダ２２０をトレーニングする。第２段階３２０の間、トレーニング手順３００は、エンコーダ

２１０またはＲＮＮ－Ｔデコーダ２２０のパラメータを更新することなく、ＬＡＳデコーダ２３０をトレーニングする。いくつかの実施形態では、トレーニング手順３００は、損失を強制することを教示するクロスエントロピーを使用してＬＡＳデコーダ２３０をトレーニングする。例えば、トレーニング手順３００は、Ｐ（$y_L = y \mid x$）を最大化するようにＬＡＳデコーダ２３０をトレーニングする。第３段階３３０の間、トレーニング手順３００はさらに、ｎベスト仮説を使用することによって期待される単語誤り率を最適化するために、最小ＷＥＲ（ＭＷＥＲ）損失を用いてＬＡＳデコーダ２３０をトレーニングする。例えば、ＷＥＲ目的関数は、Ｎベストビームの仮説２２２における単語誤りの加重平均として損失をモデル化する。この第３段階３３０の間、ＬＡＳデコーダ２３０は、以下の式によって表されるＷＥＲ目的関数に従って微調整され得る。

【００３３】

【数２】

$$L_{MWER}(x, y^*) = \sum_{y \in B_{LAS}} P(y|x)\widehat{W}(y|y^*) \quad (1)$$

ここで、ｙ＊は、グラウンドトゥルースであり、B_{LAS}は、ビームサーチ中のＬＡＳデコーダ２３０からの仮説のＮベストリストであり、Ｐ（ｙ｜ｘ）は、仮説ｙの正規化された事後確率（ｎｏｒｍａｌｉｚｅｄ　ｐｏｓｔｅｒｉｏｒ）であり、

【００３４】

【数３】

$$\widehat{W}(y|y^*)$$

は、仮説ｙにおける単語誤り数とビーム全体の単語誤りの平均数との間の差を表す。いくつかの実施形態では、ＬＡＳデコーダ２３０が再スコアリング器として機能する場合、ＬＡＳデコーダ２３０は、ＲＮＮ－Ｔデコーダ２２０からの最良の仮説y_Rに高い可能性を割り当てることを最適化するようにトレーニングする。ここで、この損失最適化関数は次の式で表すことができる。

【００３５】

【数４】

$$L_{MWER}(x, y^*) = \sum_{y \in B_{RNN-T}} P(y|x)\widehat{W}(y|y^*) \quad (2)$$

ここで、B_{RNN-T}は、ＲＮＮ－Ｔデコーダ２２０でのビームサーチから取得される。ここで、これらの最適化モデルの各々は、損失基準が、音声認識器２００またはその一部が確率質量を割り当てるように学習すべき分布を表していることを示している。

【００３６】

図３Ｂを参照すると、いくつかの実施形態では、第３トレーニング段階３３０または微調整段階の間、トレーニング手順３００は、ＭＷＥＲ損失を使用してトレーニングを実行するが、修正損失関数$MWER_{AUG}$を使用する。ここで、修正損失関数$MWER_{AUG}$は、固有名詞の損失拡張（ｌｏｓｓ　ａｕｇｍｅｎｔａｔｉｏｎ）の一形態である。このトレーニング手法では、損失は、トレーニングにおける固有名詞の性能を強調するように構成される。いくつかの例では、モデルが固有名詞を正確に外している仮説ｙに高い確率を割り当てるときに、モデル（例えば、ＬＡＳデコーダ２３０）に適用されるペナルティ３３２を増加させることによって、損失は、固有名詞の性能を強調する。例示すると、図３Ｂは、トレーニング手順３００の第３段階３３０の間に、ＬＡＳデコーダ２３０が、入力

３０２ｄを予測する１組の可能性のある仮説ｙＬ、ｙＬ１－３を生成することを示している。ここで、入力３０２ｄは、固有名詞Ｐｎを含んでいるが、ＬＡＳデコーダ２３０は、固有名詞Ｐｎを実際には含んでいないにも関わらず、入力３０２ｄに対する最高確率の仮説ｙＬである第１の仮説ｙＬ１を識別する。この例では、ＬＡＳデコーダ２３０が固有名詞Ｐｎを誤って識別した仮説ｙＬに最高確率の仮説ｙＬを割り当てたため、修正損失関数ＭＷＥＲＡＵＧは、ペナルティ３３２を適用する。いくつかの構成では、トレーニング手順３００は、モデル（例えば、ＬＡＳデコーダ２３０）がペナルティ基準を満たす仮説ｙに確率を割り当てたと判断する。ペナルティ基準は、モデルが、確率しきい値を満たす（例えば、確率しきい値に割り当てられた値を超える）、固有名詞に対する誤った仮説に確率を割り当てたことを含み得る。ここで、確率しきい値は、誤った仮説に対する許容可能なレベルまたは値を示す事前構成された値であり得る。これらの例では、トレーニング手順３００が、モデル（例えば、ＬＡＳデコーダ２３０）がペナルティ基準を満たす仮説ｙに確率を割り当てたと判断した場合、トレーニング手順３００は、修正損失関数にペナルティ３３２を適用する。いくつかの例では、固有名詞の損失拡張に対する修正損失関数は、次の式で表される。

【００３７】

【数５】

$$L_{AUG}(x,y^*) = \sum_{y \in B_{RNN-T}} P(y|x)\widehat{W}(y|y^*) \cdot C_\lambda(y|y^*) \quad (3)$$

ここで

$$C_\lambda(y,y^*) = \begin{cases} \lambda & y^*がyにない固有名詞を含む場合 \\ 1, & その他の場合 \end{cases} \quad (4)$$

定数λ＞１である。ここで、λは、一般的な発話１２に対する音声認識器２００の性能に関して、固有名詞の認識の有効性をバランスさせるために選択されたハイパーパラメータを指す。例えば、ハイパーパラメータの設定は、他の誤りのタイプとのトレードオフで、固有名詞の誤りに起因する勾配の増加を回避しようとするものである。いくつかの構成では、各グラウンドトゥルース文字起こし（例えば、トレーニングデータ３０２ｄ）に対する固有名詞Ｐｎは、固有名詞識別システム３４０によるトレーニングの前に識別される。仮説ｙが固有名詞Ｐｎを含むことを保証するために、仮説ｙが固有名詞Ｐｎの単語列全体を適切な順序で含む場合に、仮説ｙが固有名詞Ｐｎを含むものとして定義される。例えば、固有名詞Ｐｎ「シーダー・ラピッズ（Ｃｅｄａｒ　Ｒａｐｉｄｓ）」は、仮説「シーダー・ラピッズの人口（Ｐｏｐｕｌａｔｉｏｎ　ｏｆ　Ｃｅｄａｒ　Ｒａｐｉｄｓ）」には含まれているが、仮説「シーダー・ツリーの高さ（Ｃｅｄａｒ　ｔｒｅｅ　ｈｅｉｇｈｔ）」または「シーダー・ラピッドエスエスエス（Ｃｅｄａｒ　Ｒａｐｉｄｓｓｓ）」には含まれていない。

【００３８】

図３Ｃは、ファズトレーニング（ｆｕｚｚ　ｔｒａｉｎｉｎｇ）を適用して、音声認識器２００が固有名詞を区別する能力を最適化するトレーニング手順３００の別の例を示す。この手法では、ファズトレーニングは、固有名詞と、音声学的に類似した誤った代替語とを区別する方法を音声認識器２００に教えることを目的としている。言い換えれば、ファズトレーニングにおいて、音声認識器２００のようなモデルは、モデルが起こり得る間違いおよび同音異綴り（ａｌｔｅｒｎａｔｉｖｅ　ｓｐｅｌｌｉｎｇｓ）の知識を得ることを可能にする。トレーニング中に、モデル（例えば、ＬＡＳデコーダ２３０）が固有名詞の間違いに高い可能性を割り当てると、トレーニング手順３００は、モデルにペナルティ３３２を課す。ペナルティ３３２を課すことによって、トレーニングは、将来同様の誤

りの可能性を減少させることを意図している。

【００３９】

これらの潜在的な間違いに関して音声認識器２００（例えば、音声認識器２００のＬＡＳデコーダ２３０）をトレーニングするために、ファズトレーニングは、ビーム修正を実行し得る。一般に、ビームサーチは、最適なポテンシャル解（例えば、仮説または候補）を評価する数を指定するビームサイズまたはビーム幅パラメータＢを含む。ファズトレーニングは、ビームサーチからの仮説ｙを置換するか、ビームサーチからの仮説ｙの数を拡張するかのいずれかによって、ビームサイズＢを活用することができる。例示すると、図３Ｃは、５個の仮説y_L、y_{L1-5}に対応する５のビームサイズ、または５個の仮説y_L、y_{L1-5}に拡張された３のビームサイズを有するビームサーチを示す例である。この例では、ビームサイズが５である場合、ファジングシステム３５０は、仮説のうちの２個を誤った固有名詞の代替語３５２、３５２ａ－ｂに置換し得る。同様に、ビームサイズが３である場合、ファジングシステム３５０は、誤った固有名詞の代替語３５２、３５２ａ－ｂを含む追加の仮説ｙを生成し得る。いくつかの実施形態では、ファジングシステム３５０は、トレーニングデータ３０２に含まれる固有名詞Ｐｎに音声学的に類似する代替語３５２を生成する音声ファジングと呼ばれる技法を使用して、固有名詞の代替語３５２を生成する。音声ファジングにより、ファジングシステム３５０は、トレーニングデータ３０２の従来型のコーパスでは強調されていないか、または含まれていない可能性がある新たな単語または同音異綴りを生成し得る。$y \in B_{RNN-T}$であり、かつグラウンドトゥルースy^*に対応する仮説に関して、ファズ処理は次の式で表すことができる。

【００４０】

【数６】

$$Fuzz(y, y^*) = \begin{cases} y^{fuzz} & y^*\text{および}y\text{が固有名詞を共有している場合} \\ y, & \text{その他の場合} \end{cases} \quad (5)$$

いくつかの構成では、ファズ仮説y^{fuzz}は、ｙをコピーし、固有名詞Ｐｎの発生を音声学的に類似した代替語３５２に置換することによって形成される。ファズトレーニングでは、損失関数は、ＲＮＮ－Ｔデコーダ２２０からの元のビームを代替語３５２（ファズまたはファズ仮説とも呼ばれる）と組み合わせることによって定義される。以下の式は、ファズトレーニングによるトレーニング手順３００間の損失関数を表し得る。

【００４１】

【数７】

$$L_{Fuzz}(x, y^*) = \sum_{y \in B_{RNN-T} \cup Fuzz_{(B_{RNN-T})}} P(y|x)\widehat{W}(y|y^*) \quad (6)$$

ここで、Ｐ（ｙ｜ｘ）は、（例えば、付加項「Ｆｕｚｚ（B_{RNN-T}）」で表されるような）修正されたビームサイズを考慮した再正規化された事後確率に対応する。いくつかの実施形態では、ファズトレーニングの損失関数Ｌは、ハイパーパラメータτ（０≦τ≦１）をも含み、ハイパーパラメータは、ファズトレーニング損失関数L_{Fuzz}を使用する確率を規定するようになっている。これらの実施形態において、トレーニング手順３００がファズトレーニング損失関数L_{Fuzz}を使用しない場合、トレーニング手順３００は式（２）によって表される損失関数を使用する。ハイパーパラメータτは任意の確率に設定され得るが、いくつかの構成では、ハイパーパラメータは、トレーニング手順３００が常にファズトレーニング損失関数L_{Fuzz}を組み込むように、１に設定される。

【００４２】

いくつかの構成では、トレーニング手順３００は、ファズトレーニングの前に、トレー

ニングデータセット３０２に含まれる各固有名詞Ｐｎに対して一定数の代替語３５２（例えば、２５個の代替語３５２）を決定する。ここで、ファズトレーニングの前に生成される代替語３５２の数は、計算コストを最小限に抑えつつ、代替語３５２の多様性を確保するように構成することができる。トレーニング手順３００が、ファズトレーニングの前に一定数の代替語３５２を生成する場合、ファズトレーニング中に、トレーニング手順３００は、必要に応じて、既に生成されたランダムな代替語３５２を選択し得る。

【００４３】

引き続き図３Ｃを参照すると、第３段階３３０の間に、トレーニング手順３００は、ファズトレーニングを使用してＬＡＳデコーダ２３０をトレーニングする。ここで、ＬＡＳデコーダ２３０は、固有名詞Ｐｎを含むトレーニングデータ３０２、３０２ｄを受信し、トレーニングデータ３０２ｄの固有名詞Ｐｎに対応する５個の仮説y_L、y_{L1-5}（例えば、ビーム幅Ｂ＝５）を生成する。また、ＬＡＳデコーダ２３０は、各仮説y_Lに、特定の仮説が入力（例えば、トレーニングデータ３０２）を正しく識別するとＬＡＳデコーダ２３０が考える可能性を示す確率（たとえば、０．２、０．２、０．１、０．４、０．１として示される）を割り当てる。この例では、ファジングシステム３５０は、（例えば、代替語３５２がファズトレーニングの前に生成された場合）１組の潜在的な仮説y_L、「ベルモント（Ｂｅｌｍｏｎｔ）」および「ブームント（Ｂｏｏｍｕｎｄｔ）」を含むように２個のファズ仮説３５２ａ－ｂを生成または選択する。この例に示されるように、ＬＡＳデコーダ２３０は、誤った代替語「ベルモント」３５２ａに最も高い可能性（例えば、０．４として示される）を割り当てる。ＬＡＳデコーダ２３０は、誤った代替語３５２に最も高い可能性を割り当てているので、トレーニング手順３００はファズトレーニング損失関数L_{Fuzz}にペナルティ３３２を適用する。ここで、ペナルティ３３２などのペナルティは、ニューラルネットワークの重みまたはパラメータを調整するためにトレーニング中にフィードバックを提供する。一般に、ペナルティは、特定の入力に適用される重みを操作して、不所望の出力または不正確な出力ではなく、意図された出力に近づくか、または示すように機能する。言い換えれば、ペナルティ３３２は、ＬＡＳデコーダ２３０が将来、誤った代替語３５２が最良の仮説ｙである可能性が高いことを示す可能性を低減するように機能する。

【００４４】

図４は、音声認識モデル（例えば、音声認識器２００）をトレーニングする方法４００の例示的な動作の構成のフローチャートである。方法４００は、動作４０２－４０８によって、最小単語誤り率（ＭＷＥＲ）損失関数を用いて音声認識モデルをトレーニングする。動作４０２において、方法４００は、固有名詞Ｐｎを含むトレーニング例３０２を受信する。動作４０４において、方法４００は、トレーニング例３０２に対応する複数の仮説ｙを生成する。ここで、複数の仮説の各仮説ｙは、固有名詞Ｐｎを表し、各仮説には、個々の仮説ｙに対する可能性を示す確率が割り当てられる。動作４０６において、方法４００は、仮説ｙに関連付けられた確率がペナルティ基準を満たすことを決定する。ペナルティ基準は、（ｉ）確率が確率しきい値を満たしていること、および（ｉｉ）仮説が固有名詞を誤って表していることを示す。動作４０８において、方法４００は、ペナルティ３３２を最小単語誤り率損失関数に適用する。

【００４５】

図５は、本明細書で説明されるシステム（例えば、音声認識器２００）および方法（例えば、方法４００）を実施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、パーソナルデジタルアシスタント、サーバ、ブレードサーバ、メインフレーム、および他の適切なコンピュータなどの様々な形態のデジタルコンピュータを代表することが意図されている。本明細書に示された構成要素、それらの接続および関係、およびそれらの機能は、例示的なものに過ぎず、本明細書に記載および／または特許請求の範囲に記載される本発明の実施形態を限定するものではない。

【００４６】

コンピューティングデバイス５００は、プロセッサ５１０（例えば、データ処理ハードウェア）、メモリ５２０（例えば、メモリハードウェア）、ストレージデバイス５３０、メモリ５２０および高速拡張ポート５４０に接続する高速インタフェース／コントローラ５４０、および低速バス５７０およびストレージデバイス５３０に接続する低速インタフェース／コントローラ５６０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、かつ共通のマザーボード上に、または適切な他の方法で搭載され得る。プロセッサ５１０は、メモリ５２０またはストレージデバイス５３０に格納された命令を含むコンピューティングデバイス５００内での実行のための命令を処理して、高速インタフェース５４０に接続されたディスプレイ５８０などの外部入力／出力デバイス上にグラフィカルユーザインタフェース（ＧＵＩ）用のグラフィカル情報を表示する。他の実施形態では、複数のメモリおよび複数のタイプのメモリと共に、複数のプロセッサおよび／または複数のバスが適宜使用されてもよい。また、複数のコンピューティングデバイス５００が接続され、各デバイスが（例えば、サーババンク、ブレードサーバのグループ、またはマルチプロセッサシステムとして）必要な処理の一部を提供してもよい。

【００４７】

メモリ５２０は、コンピューティングデバイス５００内に非一時的に情報を記憶する。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もよい。非一時的メモリ５２０は、コンピューティングデバイス５００による使用のための一時的または永久的な基準でプログラム（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）を格納するために使用される物理的デバイスであってもよい。不揮発性メモリの例には、これらに限定されないが、フラッシュメモリおよび読み出し専用メモリ（ＲＯＭ）／プログラム可能読み出し専用メモリ（ＰＲＯＭ）／消去可能プログラム可能読み出し専用メモリ（ＥＰＲＯＭ）／電子消去可能プログラム可能読み出し専用メモリ（ＥＥＰＲＯＭ）（例えば、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）が含まれる。揮発性メモリの例には、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）、およびディスクまたはテープが含まれる。

【００４８】

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の大容量ストレージデバイスを提供することができる。いくつかの実施形態では、ストレージデバイス５３０は、コンピュータ可読媒体である。種々の異なる実施形態では、ストレージデバイス５３０は、フロッピーディスク（登録商標）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同様のソリッドステートメモリデバイス、またはストレージエリアネットワークまたは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。追加の実施形態では、コンピュータプログラム製品は、情報媒体に有形的に具体化される。コンピュータプログラム製品は、実行時に、上記したような１つまたは複数の方法を実行する命令を含む。情報媒体は、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。

【００４９】

高速コントローラ５４０は、コンピューティングデバイス５００の帯域幅を大量に使用する処理を管理し、低速コントローラ５６０は、より低い帯域幅を大量に使用する処理を管理する。このような役割の配分は、例示的なものに過ぎない。いくつかの実施形態では、高速コントローラ５４０は、メモリ５２０、ディスプレイ５８０（例えば、グラフィックプロセッサまたはアクセラレータを介する）、および各種拡張カード（図示せず）を受け入れる高速拡張ポート５５０に接続される。いくつかの実施形態では、低速コントローラ５６０は、ストレージデバイス５３０および低速拡張ポート５９０に接続される。様々な通信ポート（例えば、ＵＳＢ、ブルートゥース（登録商標）、イーサネット（登録商標

）、無線イーサネット（登録商標））を含む低速拡張ポート５９０は、キーボード、ポインティングデバイス、スキャナ、または例えばネットワークアダプターを介する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／出力デバイスに接続され得る。

【００５０】

コンピューティングデバイス５００は、図面に示されるように、いくつかの異なる形態で実施することができる。例えば、標準サーバ５００ａとして、またはそのようなサーバ５００ａのグループ内で複数回、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の一部として実施することができる。

【００５１】

本明細書に記載のシステムおよび技術の様々な実施形態は、デジタル電子回路および／または光回路、集積回路、特別に設計されたＡＳＩＣ（特定用途向け集積回路）、コンピュータハードウェア、ファームウェア、ソフトウェア、および／またはそれらの組み合わせにおいて実現することができる。これらの様々な実施形態は、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ように接続された、特別または一般的な目的であってもよい、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施形態を含むことができる。

【００５２】

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ている）は、プログラマブルプロセッサ用の機械命令を含み、高水準の手続き型言語および／またはオブジェクト指向のプログラミング言語、および／またはアセンブリ言語／機械語で実施することができる。本明細書で使用する場合、「機械可読媒体」および「コンピュータ可読媒体」という用語は、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、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、プログラマブル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装置および／またはデバイス（例えば、磁気ディスク、光ディスク、メモリ、プログラマブルロジックデバイス（ＰＬＤ））を指す。「機械可読信号」という用語は、機械命令および／またはデータをプログラマブルプロセッサに提供するために使用される任意の信号を指す。

【００５３】

本明細書で説明するプロセスおよび論理フローは、入力データを処理して出力を生成することによって機能を実行する１つまたは複数のコンピュータプログラムを実行する１つまたは複数のプログラマブルプロセッサによって実行することができる。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）などの特定用途論理回路によっても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は、一例として、汎用マイクロプロセッサおよび専用マイクロプロセッサの両方、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、読み出し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。コンピュータの必須要素は、命令を実行するプロセッサと、命令および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とである。一般に、コンピュータは、データを格納するための１つまたは複数の大容量ストレージデバイス（例えば、磁気ディスク、光磁気ディスク、または光ディスク）からのデータを受信するか、またはデータを転送するか、あるいはその両方を行うように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を有する必要はない。コンピュータプログラム命令およびデータを格納するのに適したコンピュータ可読媒体には、半導体メモリデバイス（例えば、ＥＰＲＯＭ、ＥＥＰＲＯＭ、およびフラッシュメモリデバイス）、磁気ディスク（例えば、内蔵ハードディスクまたはリムーバブルディスク）、光磁気ディスク、およびＣＤＲＯＭおよびＤＶＤ－

ＲＯＭディスクを含む全ての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る。プロセッサおよびメモリは、特定用途論理回路によって補完または特定用途論理回路に組み込むことができる。

【００５４】

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、例えば、ＣＲＴ（陰極線管）、ＬＤＣ（液晶ディスプレイ）モニタ、またはタッチスクリーンなどのユーザに情報を表示するためのディスプレイデバイスと、任意選択でユーザがコンピュータに入力を提供するキーボードおよびポインティングデバイス（例えば、マウスやトラックボール）とを有するコンピュータ上で実施することができる。他の種類の装置を使用して、例えば、任意の形態の感覚フィードバック（例えば、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）であり得るユーザに提供されるフィードバックとともにユーザとのインタラクションを提供することもでき、ユーザからの入力は、音響、音声、または触覚入力を含む任意の形態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（例えば、ウェブブラウザから受信し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によって）、ユーザとインタラクションすることができる。

【００５５】

いくつかの実施形態が説明されている。それにもかかわらず、本開示の技術思想および範囲から逸脱することなく、様々な変更がなされ得ることが理解されるであろう。従って、他の実施形態も以下の特許請求の範囲内にある。

【図面】

【図１Ａ】

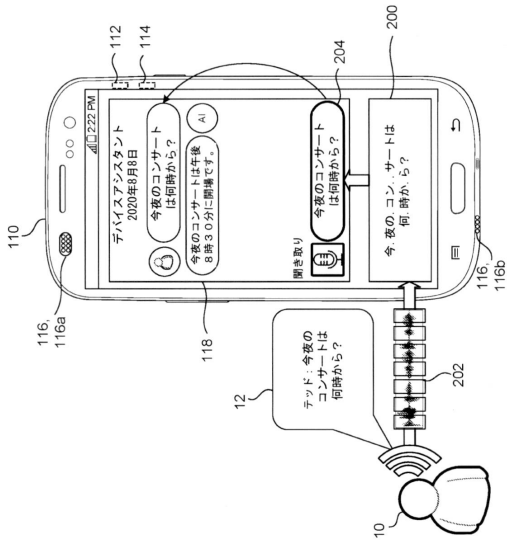

FIG. 1A

【図１Ｂ】

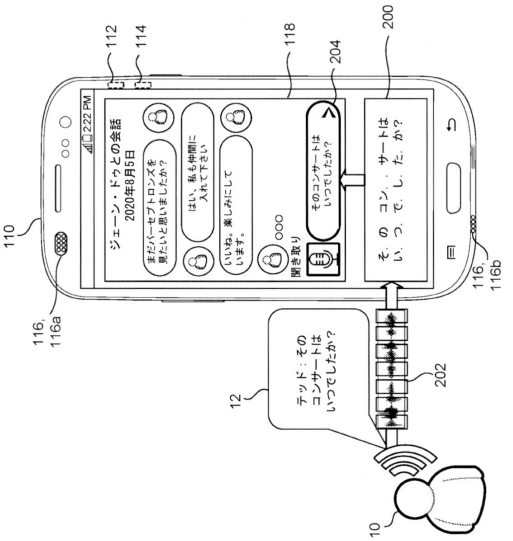

FIG. 1B

【図２】

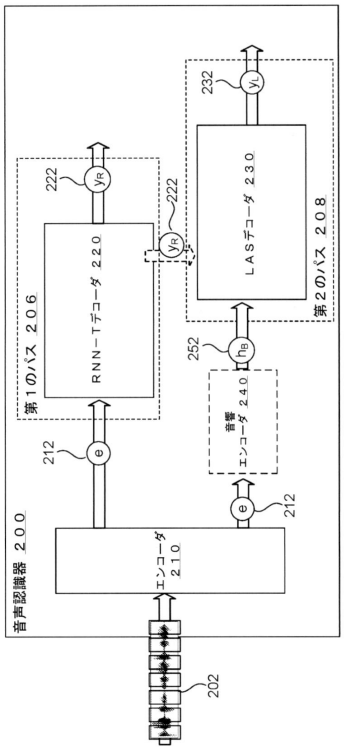

FIG. 2

【図３Ａ】

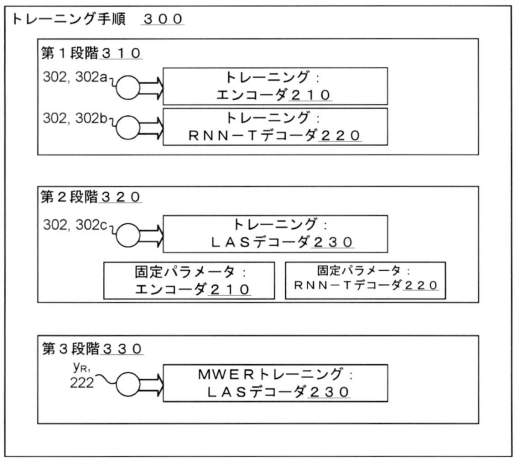

FIG. 3A

【図３Ｂ】

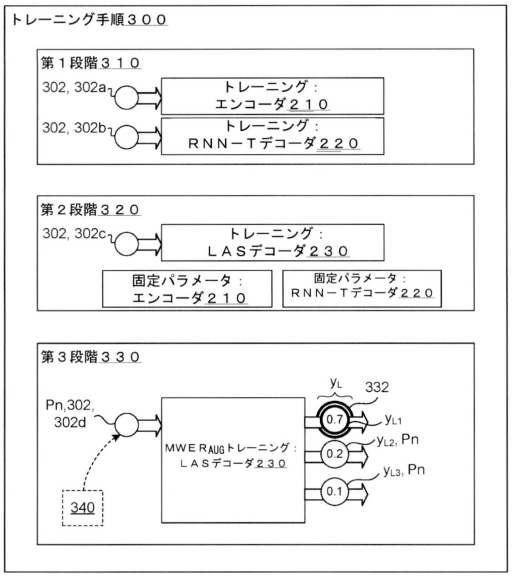

FIG. 3B

【図３Ｃ】

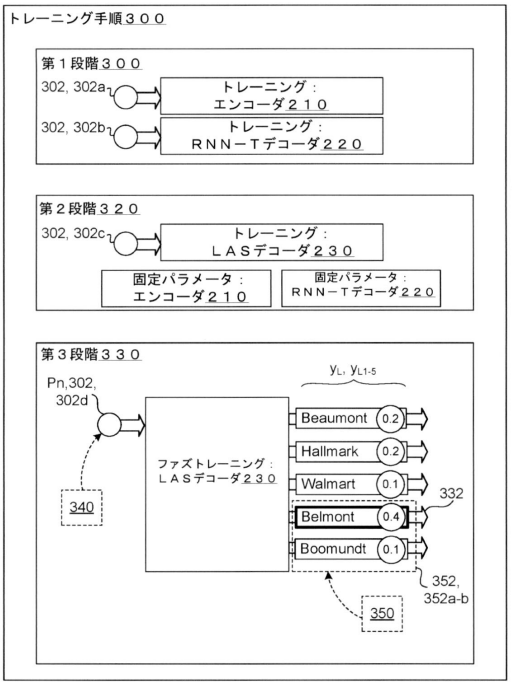

FIG. 3C

【図４】

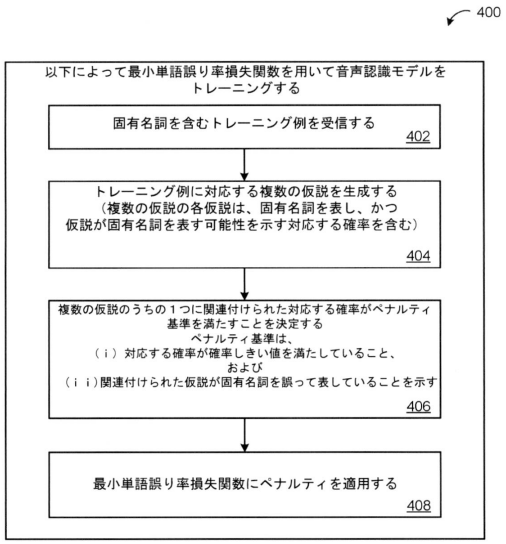

FIG. 4

【図５】

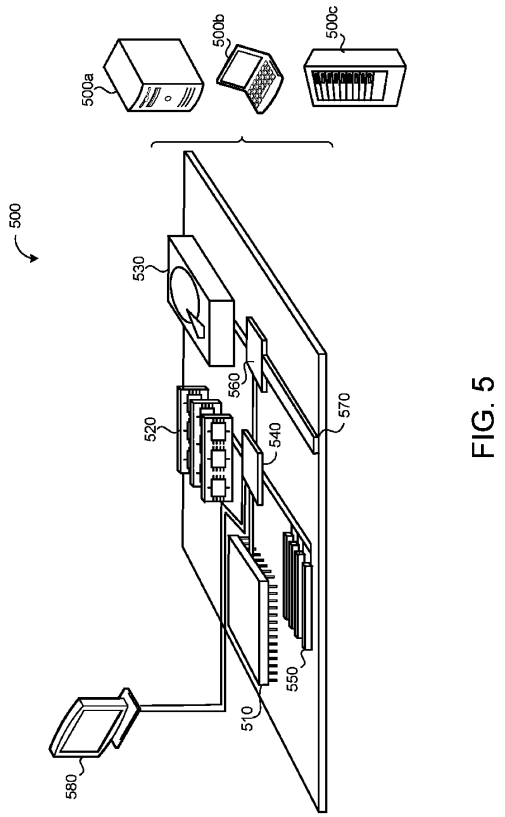

FIG. 5

フロントページの続き

ｅ．ｉｅｅｅ．ｏｒｇ／ｄｏｃｕｍｅｎｔ／９０５４２３５にて発表
早期審査対象出願
(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　特開２０２０－１９４４９４（ＪＰ，Ａ）
国際公開第２０１２／１６５５２９（WO，Ａ１）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１５／３４
ＩＥＥＥ Ｘｐｌｏｒｅ