(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2018-517927
(P2018-517927A)
(43)公表日　平成30年7月5日(2018.7.5)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 17/18	(2013.01)	G10L	17/18	
G10L 17/00	(2013.01)	G10L	17/00	200C

審査請求 有　予備審査請求 未請求　（全 36 頁）

(21)出願番号	特願2017-556908 (P2017-556908)
(86)(22)出願日	平成28年7月27日 (2016.7.27)
(85)翻訳文提出日	平成29年12月20日 (2017.12.20)
(86)国際出願番号	PCT/US2016/044181
(87)国際公開番号	WO2017/039884
(87)国際公開日	平成29年3月9日 (2017.3.9)
(31)優先権主張番号	14/846,187
(32)優先日	平成27年9月4日 (2015.9.4)
(33)優先権主張国	米国 (US)

(71)出願人　502208397
グーグル　エルエルシー
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
(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
(72)発明者　ゲオルク・ハイゴルト
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ヴュー・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】話者検証のためのニューラルネットワーク

(57)【要約】
本明細書は一般に、（ｉ）話者検証モデルに対するニューラルネットワークをトレーニングするステップと、（ｉｉ）クライアントデバイスでユーザを加入させるステップと、（ｉｉｉ）ユーザのアイデンティティを前記ユーザの音声の特性に基づいて検証するステップとを含む、話者検証に関連するシステム、方法、デバイス、および他の技術を説明する。幾つかの実装はコンピュータ実行型の方法を含む。前記方法は、コンピューティングデバイスで、前記コンピューティングデバイスのユーザの発声を特徴付けるデータを受信するステップを含むことができる。話者表現を、前記コンピューティングデバイスで、前記コンピューティングデバイス上のニューラルネットワークを用いて前記発声に対して生成することができる。前記ニューラルネットワークは、それぞれが、（ｉ）第１の発声を特徴付けるデータおよび１つまたは複数の第２の発声を特徴付けるデータを含み、（ｉｉ）マッチング話者サンプルまたは非マッチング話者サンプルとしてラベル付けされる、複数のトレーニング・サンプルに基づいてトレーニングされることができる。

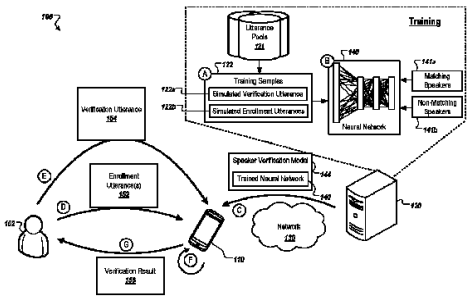


FIG. 1

【特許請求の範囲】
【請求項１】
コンピューティングデバイスで、前記コンピューティングデバイスのユーザの発声を特徴付けるデータを受信するステップと、
前記コンピューティングデバイスで、前記コンピューティングデバイス上のニューラルネットワークを用いて前記発声に対する話者表現を生成するステップであって、前記ニューラルネットワークが、それぞれ、
（ｉ）第１の発声を特徴付けるデータおよび１つまたは複数の第２の発声を特徴付けるデータを含み、
（ｉｉ）前記第１の発声の話者が前記１つまたは複数の第２の発声の話者と同一であるかどうかに従って、マッチングサンプルまたは非マッチングサンプルとしてラベル付けされる、
複数のトレーニング・サンプルに基づいてトレーニングされている、ステップと、
前記コンピューティングデバイスで、前記コンピューティングデバイスの認証されたユーザに対する話者モデルにアクセスするステップと、
前記コンピューティングデバイスで、前記話者モデルに関する前記発声に対する前記話者表現を評価して、前記発声が前記コンピューティングデバイスの前記認証されたユーザにより話された可能性があったかどうかを判定するステップと、
を含む、コンピュータ実行型の方法。
【請求項２】
前記複数のトレーニング・サンプルの各々は、各発声グループが前記各発声グループに対する前記対応する話者の発声のみから構成されるように、前記第１の発声および前記１つまたは複数の第２の発声を異なる話者に対応する発声グループから選択することにより生成された、請求項１に記載のコンピュータ実行型の方法。
【請求項３】
前記コンピューティングデバイスの前記認証されたユーザの１組の発声を取得するステップと、
前記発声に対する各話者表現を生成するために、各発声を前記１組の発声から前記ニューラルネットワークに入力するステップと、
前記認証されたユーザの前記１組の発声における前記発声に対する前記各話者表現の平均に基づいて前記コンピューティングデバイスの前記認証されたユーザに対する前記話者モデルを生成するステップと、
をさらに含む、請求項１または２に記載のコンピュータ実行型の方法。
【請求項４】
前記ニューラルネットワークがトレーニングされている前記複数のトレーニング・サンプルの何れも、前記コンピューティングデバイスの前記ユーザの前記発声を特徴付けるデータを含まない、請求項１乃至３の何れか１項に記載のコンピュータ実行型の方法。
【請求項５】
前記コンピューティングデバイスで、前記発声に対する前記話者表現を生成するステップは、前記発声の全体を特徴付けるデータを前記ニューラルネットワークで単一のパスで前記ニューラルネットワークを通じて処理するステップを含む、請求項１乃至４の何れか１項に記載のコンピュータ実行型の方法。
【請求項６】
前記ユーザの前記発声が前記コンピューティングデバイスの前記認証されたユーザにより話された可能性があったと判定したことに応答して、機能を前記コンピューティングデバイス上で実施するステップをさらに含む、請求項１乃至５の何れか１項に記載のコンピュータ実行型の方法。
【請求項７】
第１の１組の発声に対するニューラルネットワークの出力に基づいて特定の話者に対する話者モデルを決定するステップであって、前記第１の１組の発声は前記特定の話者の複

数の異なる発声を含む、ステップと、
前記第１の１組の発声内にない特定の発声に対する前記ニューラルネットワークの出力に基づいて話者表現を決定するステップと、
前記話者表現を前記特定の話者に対する前記話者モデルと比較して、前記特定の発声を前記特定の話者の発声としてまたは前記特定の話者と異なる話者の発声として分類するステップと、
前記発声の前記特定の話者の発声としてまたは前記特定の話者と異なる話者の発声としての前記分類が正しかったかどうかに基づいて前記ニューラルネットワークを調節するステップと、
を含む、コンピュータ実行型の方法。
【請求項８】
コンピューティングシステムで、複数の異なる発声のセットを、前記ニューラルネットワークをトレーニングするためのトレーニングデータとして選択するステップであって、各１組の発声は、
（ｉ）前記各１組の発声に対する第１の第１の話者の複数の異なる発声と、
（ｉｉ）前記各１組の発声に対する前記第１の話者または前記第１の話者と異なる前記各１組の発声に対する第２の話者の何れかの第２の発声と、
を含む、ステップと、
前記複数の異なる発声のセットを用いて、前記ニューラルネットワークのトレーニングを反復するステップであって、前記複数の異なる発声のセットからの各１組の発声は前記ニューラルネットワークをトレーニングするために異なるトレーニングの反復で使用される、ステップと、
を含み、
前記第１の１組の発声は前記複数の異なる発声のセットから選択され、前記特定の話者は前記第１の１組の発声に対する前記第１の話者である、
請求項７に記載のコンピュータ実行型の方法。
【請求項９】
コンピューティングシステムで、ニューラルネットワークをトレーニングするためのトレーニングデータの複数の異なるサブセットを選択するステップであって、トレーニングデータの各サブセットは、第１の話者の各発声を分類する複数の第１のコンポーネントおよび前記第１の話者または第２の話者の発声を特徴付ける第２のコンポーネントを含む、ステップと、
トレーニングデータの前記選択されたサブセットの各々に対して、
前記第１のコンポーネントの各々に対応する各第１の話者表現を生成するために、前記第１のコンポーネントの各々を前記ニューラルネットワークに入力するステップと、
前記第２のコンポーネントに対応する第２の話者表現を生成するために、前記第２のコンポーネントを前記ニューラルネットワークに入力するステップと、
前記複数の第１のコンポーネントに対する前記各第１の話者表現の平均に基づいて前記第１の話者に対するシミュレートされた話者モデルを決定するステップと、
前記第２の話者表現を前記シミュレートされた話者モデルと比較して、前記第２のコンポーネントにより特徴付けられた前記発声を前記第１の話者の発声としてまたは前記第１の話者と異なる話者の発声として分類するステップと、
前記第２のコンポーネントにより特徴付けられた前記発声が前記第１の話者の発声としてまたは前記第１の話者と異なる話者の発声として正しく分類されたかどうかに基づいて、前記ニューラルネットワークを調節するステップと、
を含む、コンピュータ実行型の方法。
【請求項１０】
前記第２のコンポーネントを前記ニューラルネットワークに入力するステップに応答して、単一のパスで前記ニューラルネットワークを通じて、前記第２のコンポーネントにより特徴付けられた前記発声の全体に対するデータを処理することで、前記第２の話者表現

を前記ニューラルネットワークで生成するステップをさらに含む、請求項９に記載のコンピュータ実行型の方法。
【請求項１１】
トレーニングデータの前記複数の異なるサブセットからトレーニングデータの第１のサブセットを選択するステップは、
各発声グループが前記各発声グループに対する前記対応する話者の発声のみを特徴づけるデータから構成されるように、それぞれ異なる話者に対応する複数の発声グループからの前記第１の話者に対応する第１の発声グループを選択するステップと、
前記第１の話者または前記第２の話者に対応する第２の発声グループを前記複数の発声グループから選択するステップと、
前記複数の第１のコンポーネントを前記第１の発声グループ内の発声を特徴付ける前記データから決定するステップと、
前記第２のコンポーネントを前記第２の発声グループ内の発声を特徴付ける前記データから決定するステップと、
を含む、請求項９または１０に記載のコンピュータ実行型の方法。
【請求項１２】
前記第１の発声グループおよび前記第２の発声グループの少なくとも１つは前記複数の発声グループからランダムに選択される、請求項１１に記載のコンピュータ実行型の方法。
【請求項１３】
前記第１の話者はトレーニングデータの前記複数の異なるサブセットの少なくとも幾つかの間で互いに異なり、
前記第２の話者はトレーニングデータの前記複数の異なるサブセットの少なくとも幾つかの間で互いに異なる、
請求項９乃至１２の何れか１項に記載のコンピュータ実行型の方法。
【請求項１４】
トレーニングデータの前記複数の異なるサブセットのうちトレーニングデータの第１のサブセット内の第１のコンポーネントの総数は、トレーニングデータの前記複数の異なるサブセットのうちトレーニングデータの第２のサブセット内の第１のコンポーネントの総数と異なる、請求項９乃至１３の何れか１項に記載のコンピュータ実行型の方法。
【請求項１５】
記第２の話者表現を前記シミュレートされた話者モデルと比較して、前記第２のコンポーネントにより特徴付けられた前記発声を前記第１の話者の発声としてまたは前記第１の話者と異なる話者の発声として分類するステップは、前記第２の話者表現からの値および前記シミュレートされた話者モデルからの値の間の距離を決定するステップと、ロジスティック回帰機能を前記距離に適用するステップとを含む、請求項９乃至１４の何れか１項に記載のコンピュータ実行型の方法。
【請求項１６】
前記ニューラルネットワークは複数の隠蔽されたレイヤを含み、前記ニューラルネットワークはソフトマックス出力レイヤを有さない、請求項９乃至１５の何れか１項に記載のコンピュータ実行型の方法。
【請求項１７】
前記ニューラルネットワークは、ローカルに接続された隠蔽されたレイヤと、前記レイヤに続く複数の完全に接続された隠蔽されたレイヤを有するディープ・ニューラルネットワークを含む、請求項９乃至１６の何れか１項に記載のコンピュータ実行型の方法。
【請求項１８】
前記複数の第１のコンポーネントにより特徴付けられた前記発声と前記第２のコンポーネントにより特徴付けられた前記発声とは全て固定長を有する、請求項１７に記載のコンピュータ実行型の方法。
【請求項１９】

前記ニューラルネットワークは、可変長を有する発声を特徴付けるデータでトレーニングされるように構成された長短期メモリ再帰型ニューラルネットワークを含む、請求項９乃至１８の何れか１項に記載のコンピュータ実行型の方法。

【請求項２０】

トレーニングデータの前記複数の異なるサブセットにわたって前記各複数の第１のコンポーネントにより特徴付けられた前記発声と、トレーニングデータの前記複数の異なるサブセットにわたって前記各第２のコンポーネントにより特徴付けられた前記発声とは全て、同一の単語またはフレーズの発声である、請求項９乃至１９の何れか１項に記載のコンピュータ実行型の方法。

【請求項２１】

トレーニングデータの前記複数の異なるサブセットにわたって前記各複数の第１のコンポーネントにより特徴付けられた前記発声の少なくとも幾つかと、トレーニングデータの前記複数の異なるサブセットにわたって前記各第２のコンポーネントにより特徴付けられた前記発声の少なくとも幾つかとは、異なる単語または異なるフレーズの発声である、請求項９乃至１９の何れか１項に記載のコンピュータ実行型の方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本明細書の主題は一般に、話者検証タスクで使用されるニューラルネットワークおよび他のモデルに関する。

【背景技術】

【０００２】

話者検証は一般に人のアイデンティティを、当該人の音声の特性に基づいて検証することに関する。幾つかのコンピューティングデバイスは、ユーザに、ユーザにより話された会話の１つまたは複数のサンプルを当該デバイスに提供することで当該デバイスに「加入」させることができる。当該サンプルから、ユーザの音声を表す話者モデルが決定される。当該デバイスで受信された後続の会話サンプルをついで、当該話者モデルに関して処理し評価して、ユーザのアイデンティティを検証してもよい。

【発明の概要】

【課題を解決するための手段】

【０００３】

本明細書では一般に、話者検証のためのニューラルネットワーク、または他のタイプのモデルをトレーニングし使用するためのシステム、方法、デバイス、および他の技術を説明する。幾つかの実装では、ニューラルネットワークは、話者検証を実施するコンピューティングデバイスによりアクセス可能な話者検証モデルのコンポーネントであってもよい。一般に、ニューラルネットワークは、それぞれ話者の加入および発声の検証をシミュレートする反復でトレーニングされてもよい。例えば、各トレーニングの反復において、所与の発声に対してニューラルネットワークにより生成された話者表現を話者モデルに関して評価してもよい。シミュレートされた検証発声に対する話者表現の、１つまたは複数のシミュレートされた加入発声に対する話者表現の組合せ（例えば、平均）との比較に基づいて、ニューラルネットワークのパラメータを、所与の発声を同一の人または加入した人とは異なる人により話されているとして分類するように当該話者検証モデルの能力を最適化するように更新してもよい。これが、当該システムの信頼性を増大する点で利点を有することは理解される。ニューラルネットワークはさらに、当該発声のフレームを独立にまたは逐次的に処理するのではなく、単一のパスでニューラルネットワークを通じて、発声全体を特徴づけるデータを処理するように構成されてもよい。これらのおよび他の実装は下記でより完全に説明され、図面で示される。

【０００４】

本明細書で説明する主題の幾つかの実装はコンピュータ実行型の方法を含む。当該方法は、コンピューティングシステムで、ニューラルネットワークをトレーニングするための

トレーニングデータの複数の異なるサブセットを選択するステップを含むことができる。トレーニングデータの各サブセットは、第１の話者の各発声を分類する複数の第１のコンポーネントと、第１の話者または第２の話者の発声を特徴付ける第２のコンポーネントとを含むことができる。トレーニングデータの当該選択されたサブセットごとに、当該方法は、第１のコンポーネントの各々に対応する各第１の話者表現を生成するために、第１のコンポーネントの各々をニューラルネットワークに入力するステップと、第２のコンポーネントに対応する第２の話者表現を生成するために、第２のコンポーネントをニューラルネットワークに入力するステップと、当該複数の第１のコンポーネントに対する当該各第１の話者表現の平均に基づいて第１の話者に対するシミュレートされた話者モデルを決定するステップと、第２の話者表現を当該シミュレートされた話者モデルと比較して、第２のコンポーネントにより特徴付けられた発声を第１の話者の発声としてまたは第１の話者と異なる話者の発声として分類するステップと、第２のコンポーネントにより特徴付けられた発声が第１の話者の発声としてまたは第１の話者と異なる話者の発声として正しく分類されたかどうかに基づいて、ニューラルネットワークを調節するステップとを含むことができる。

【０００５】

これらのおよび他の実装は以下の特徴のうち１つまたは複数を含むことができる。

【０００６】

第２のコンポーネントをニューラルネットワークに入力するステップに応答して、第２の話者表現を、単一のパスでニューラルネットワークを通じて、第２のコンポーネントにより特徴付けられた当該発声の全体に対するデータを処理することで、ニューラルネットワークで生成することができる。

【０００７】

トレーニングデータの当該複数の異なるサブセットからトレーニングデータの第１のサブセットを選択するステップは、各発声グループが当該各発声グループに対する当該対応する話者の発声のみを特徴づけるデータから構成されるように、それぞれ異なる話者に対応する複数の発声グループからの第１の話者に対応する第１の発声グループを選択するステップと、第１の話者または第２の話者に対応する第２の発声グループを発声の当該複数のグループから選択するステップと、当該複数の第１のコンポーネントを第１の発声グループ内の発声を特徴付ける当該データから決定するステップと、第２のコンポーネントを第２の発声グループ内の発声を特徴付ける当該データから決定するステップとを含むことができる。

【０００８】

第１の発声グループおよび第２の発声グループの少なくとも１つを発声の当該複数のグループからランダムに選択することができる。

【０００９】

第１の話者は、互いにトレーニングデータの当該複数の異なるサブセットのうちの少なくとも幾つかと異なることができる。第２の話者は、互いにトレーニングデータの当該複数の異なるサブセットのうちの少なくとも幾つかと異なることができる。

【００１０】

トレーニングデータの当該複数の異なるサブセットのうちトレーニングデータの第１のサブセット内の第１のコンポーネントの総数は、トレーニングデータの当該複数の異なるサブセットのうちトレーニングデータの第２のサブセット内の第１のコンポーネントの総数と異なることができる。

【００１１】

第２の話者表現を当該シミュレートされた話者モデルと比較して、第２のコンポーネントにより特徴付けられた当該発声を第１の話者の発声としてまたは第１の話者と異なる話者の発声として分類するステップは、第２の話者表現からの値および当該シミュレートされた話者モデルからの値の間の距離を決定するステップと、ロジスティック回帰機能を当該距離に適用するステップとを含むことができる。

【００１２】
ニューラルネットワークは複数の隠蔽されたレイヤを含むことができる。ニューラルネットワークはソフトマックス出力レイヤを有さなくてもよい。
【００１３】
ニューラルネットワークは、ローカルに接続された隠蔽されたレイヤと、当該レイヤに続く複数の完全に接続された隠蔽されたレイヤを有するディープ・ニューラルネットワークを含むことができる。
【００１４】
当該複数の第１のコンポーネントにより特徴付けられた当該発声、および第２のコンポーネントにより特徴付けられた当該発声は全て、固定長を有することができる。
【００１５】
ニューラルネットワークは、可変長を有する発声を特徴付けるデータでトレーニングされるように構成される長短期メモリ再帰型ニューラルネットワークであることができる。
【００１６】
トレーニングデータの当該複数の異なるサブセットにわたって各複数の第１のコンポーネントにより特徴付けられた発声、およびトレーニングデータの当該複数の異なるサブセットにわたって各第２のコンポーネントにより特徴付けられた発声は全て、同一の単語またはフレーズの発声であることができる。
【００１７】
トレーニングデータの当該複数の異なるサブセットにわたって各複数の第１のコンポーネントにより特徴付けられた発声の少なくとも幾つか、およびトレーニングデータの当該複数の異なるサブセットにわたって各第２のコンポーネントにより特徴付けられた当該発声の少なくとも幾つかは異なる単語または異なるフレーズから成る発声であることができる。
【００１８】
当該トレーニングされたニューラルネットワークを、話者検証を当該１つまたは複数のコンピューティングデバイスで実施する際に使用するための当該コンピューティングシステムとは別の１つまたは複数のコンピューティングデバイスに提供することができる。
【００１９】
第１の話者表現の各々を、第１の話者表現に対応するニューラルネットワークに入力された当該各第１のコンポーネントに基づいてニューラルネットワーク生成することができる。第２の話者表現を、第２の話者表現に対応するニューラルネットワークに入力された第２のコンポーネントに基づいてニューラルネットワークにより生成することができる。
【００２０】
本明細書で説明する主題の幾つかの実装はコンピューティングデバイスを含むことができる。コンピューティングデバイスは、１つまたは複数のコンピュータプロセッサと、当該１つまたは複数のプロセッサにより実行されたとき、動作を実施させる命令を格納した１つまたは複数のコンピュータ可読媒体とを備えることができる。当該動作は、コンピューティングデバイスで、コンピューティングデバイスのユーザの発声を特徴付けるデータを受信するステップと、コンピューティングデバイスで、コンピューティングデバイス上のニューラルネットワークを用いて当該発声に対する話者表現を生成するステップであって、ニューラルネットワークが、それぞれ（ｉ）第１の発声を特徴付けるデータおよび１つまたは複数の第２の発声を特徴付けるデータを含み、（ｉｉ）第１の発声の話者が当該１つまたは複数の第２の発声の話者と同一であるかどうかに従って、マッチングサンプルまたは非マッチングサンプルとしてラベル付けされる、複数のトレーニング・サンプルに基づいてトレーニングされている、ステップと、コンピューティングデバイスで、コンピューティングデバイスの認証されたユーザに対する話者モデルにアクセスするステップと、コンピューティングデバイスで、当該話者モデルに関する当該発声に対する話者表現を評価して、当該発声がコンピューティングデバイスの認証されたユーザにより話された可能性があったかどうかを判定するステップとを含むことができる。

【００２１】
各発声グループが当該各発声グループに対する対応する話者の発声のみから構成されるように、第１の発声および当該１つまたは複数の第２の発声を異なる話者に対応する発声グループから選択することによって、当該複数のトレーニング・サンプルの各々を生成することができる。
【００２２】
当該動作はさらに、コンピューティングデバイスの認証されたユーザの１組の発声を取得するステップと、当該発声に対する各話者表現を生成するために、各発声を当該１組の発声からニューラルネットワークに入力するステップと、当該認証されたユーザの当該１組の発声における当該発声に対する各話者表現の平均に基づいてコンピューティングデバイスの認証されたユーザに対する話者モデルを生成するステップとを含むことができる。
【００２３】
ニューラルネットワークがトレーニングされている当該複数のトレーニング・サンプルの何れも、コンピューティングデバイスのユーザの発声を特徴付けるデータを含まなくてもよい。
【００２４】
コンピューティングデバイスで、当該発声に対する話者表現を生成するステップは、当該発声の全体をニューラルネットワークで単一のパスでニューラルネットワークを通じて特徴付けるデータを処理するステップを含むことができる。
【００２５】
ニューラルネットワークは再帰型ニューラルネットワークであることができる。ユーザの発声は第１の時間長を有することができる。当該発声に対する話者表現は、当該発声の第１の時間長の全体に対して当該発声を特徴付けるデータを処理するステップを含むことができる。当該動作はさらに、コンピューティングデバイスのユーザの別の発声を受信するステップであって、当該他の発声は、第１の時間長と異なる第２の時間長を有する、ステップと、当該他の発声の第２の時間長の全体に対して当該他の発声を特徴付けるデータを処理することでユーザの当該他の発声に対する第２の話者表現を生成するステップとを含むことができる。
【００２６】
当該動作はさらに、ユーザの発声がコンピューティングデバイスの認証されたユーザにより話された可能性があったと判定したことに応答して、機能をコンピューティングデバイスで実施するステップを含むことができる。当該機能は、コンピューティングデバイスの状態をロックされた状態からロックされていない状態に変更するステップであって、コンピューティングデバイスは、当該ロックされた状態のコンピューティングデバイスの１つまたは複数の能力へのアクセスをブロックするように構成され、コンピューティングデバイスは当該ロックされていない状態のコンピューティングデバイスの当該１つまたは複数の能力へのアクセスを許可するように構成される、ステップを含むことができる。
当該話者表現は、決定された発声に基づいて決定されるユーザの音声の区別的な特徴を示すニューラルネットワークの出力を含むことができる。
【００２７】
本明細書で説明する主題の幾つかの実装はコンピュータ実行型の方法を含むことができる。当該方法は、コンピューティングデバイスで、コンピューティングデバイスのユーザの発声を特徴付けるデータを受信するステップを含むことができる。話者表現を、コンピューティングデバイスで、コンピューティングデバイス上のニューラルネットワークを用いて当該発声に対して生成することができる。ニューラルネットワークを、それぞれが（ｉ）第１の発声を特徴付けるデータおよび１つまたは複数の第２の発声を特徴付けるデータを含み、（ｉｉ）第１の発声の話者が当該１つまたは複数の第２の発声の話者と同一であるかどうかに従って、マッチングサンプルまたは非マッチングサンプルとしてラベル付けされる、複数のトレーニング・サンプルに基づいてトレーニングすることができる。話者モデルに、コンピューティングデバイスで、コンピューティングデバイスの認証された

ユーザに対してアクセスすることができる。当該発声に対する話者表現を、コンピューティングデバイスで、当該話者モデルに関して評価して、当該発声がコンピューティングデバイスの認証されたユーザにより話された可能性があったかどうかを判定することができる。

【００２８】

本明細書で説明する主題の幾つかの実装はコンピュータ実行型の方法を含むことができる。当該コンピュータ実行型の方法は、第１の１組の発声に対するニューラルネットワークの出力に基づいて特定の話者に対する話者モデルを決定するステップであって、第１のセットは複数の異なる特定の話者の発声を含む、ステップと、第１の１組の発声内にない特定の発声に対するニューラルネットワークの出力に基づいて話者表現を決定するステップと、当該話者表現を当該特定の話者に対する当該話者モデルと比較して、当該特定の発声を当該特定の話者の発声としてまたは当該特定の話者と異なる話者の発声として分類するステップと、当該発声の当該特定の話者の発声としてまたは当該特定の話者と異なる話者の発声としての当該分類が正しかったかどうかに基づいてニューラルネットワークを調節するステップとを含むことができる。

【００２９】

これらのおよび他の実装は以下の特徴のうち１つまたは複数を含むことができる。複数の異なる発声のセットを、ニューラルネットワークをトレーニングするためのトレーニングデータとして選択することができる。各１組の発声は、（ｉ）当該各１組の発声に対する第１の第１の話者の複数の異なる発声、および（ｉｉ）当該各１組の発声に対する第１の話者、または、第１の話者と異なる当該各１組の発声に対する第２の話者の何れかの第２の発声を含むことができる。当該複数の異なる発声のセットを、ニューラルネットワークを繰り返しトレーニングするために使用することができる。当該複数の異なる発声のセットからの各１組の発声を、ニューラルネットワークをトレーニングするために異なるトレーニングの反復において使用することができる。第１の１組の発声を当該複数の異なる発声のセットから選択することができる。当該特定の話者は第１の１組の発声に対する第１の話者である。

【００３０】

本明細書で説明する主題のさらなる実装は、当該方法のアクションを実施するように構成され、コンピュータ記憶デバイスで符号化された対応するシステム、装置、およびコンピュータプログラムを含むことができる。１つまたは複数のコンピュータのシステムを、動作中に当該システムに当該アクションを実施させるソフトウェア、ファームウェア、ハードウェア、または当該システムにインストールされたそれらの組合せにより構成することができる。１つまたは複数のコンピュータプログラムを、１つまたは複数のデータ処理装置により実行されたとき、当該装置に当該アクションを実施させる命令を有することによりそのように構成することができる。

【００３１】

本明細書で説明する主題の幾つかの実装は以下の利点の１つまたは複数を実現しうる。ニューラルネットワークを、ユーザの音声の特性に基づいて話者のアイデンティティをより正確に検証できる話者検証モデルで使用するための話者表現を生成するために、トレーニングすることができる。ターゲット性能レベルを、有限のコンピューティングリソースを有するモバイルコンピューティングデバイスに格納しそこで使用しうるコンパクトなニューラルネットワークで実現することができる。さらに、ニューラルネットワークを、話者検証プロセスの検証および加入段階をシミュレートする方式でトレーニングしてもよい。したがって、ニューラルネットワークは、当該話者検証プロセスのトレーニング段階と検証および加入段階との間の対称性のため良好な性能を実現しうる。本明細書で説明するアプローチに従うニューラルネットワークをトレーニングする利益は、事前選択された数の話者の間の特定の話者に属するとして入力を分類するためにニューラルネットワークをトレーニングすることを含む他のアプローチと対照的に、より多くの数の様々な話者を、当該ネットワークをトレーニングするために使用しうるということである。さらに、高信

頼のトレーニングを保証するためのトレーニング話者ごとに要求された最小数のトレーニング発声がなくてもよい。さらに、ニューラルネットワークは、独立なパス内のニューラルネットワークを通じて当該発声のフレームを処理する必要なく、発声全体を特徴づけるデータを単一のパスでニューラルネットワークを通じて処理するように構成されてもよい。上で参照した態様は、発声がコンピューティングデバイスの認証されたユーザにより話された可能性があったかどうかを評価することに寄与しうる。当該態様はかかる評価を特に高信頼としうる。これを、少なくとも、コンピューティングデバイスの認証されたユーザのような、発声が特定の人により話されたかどうかを評価する際のエラーを減らすことで、より有効な音声認識システムとすることができる。当該高信頼性は、上で参照した態様に関連付けられた広範囲のシステムのセキュリティを増大させうる。例えば、当該態様が認証されたユーザを認識するために使用され、応答して、コンピューティングデバイスの状態をロックされた状態からロックされていない状態に変更する場合、上で参照したエラーの減少は、コンピューティングデバイスのアンロックをより安全にする。実際、当該システムの高信頼性は、コンピューティングデバイスを、非認証されたユーザからの詐欺的なアンロック試行に対してあまり脆弱でなくしうる。当該エラーの減少はまた、エラー訂正の必要性を削減でき、これはコンピューティングリソースをエラー訂正に割り当てる必要性を削減できる。これは、コンピュータリソースがより制限されうるモバイルコンピューティングデバイスにおいて特に有利である。

【図面の簡単な説明】

【００３２】

【図１】（ｉ）ニューラルネットワークをトレーニングし、（ｉｉ）ユーザをコンピューティングデバイスで加入させ、（ｉｉｉ）ユーザの音声の区別的な特徴に基づいてコンピューティングデバイスのユーザの発声を検証する動作を実行する、例示的なクライアントデバイスおよびコンピューティングシステムの略図である。

【図２】話者検証タスクで使用するための話者表現を決定するためにニューラルネットワークをトレーニングするためのブロック図である。

【図３】音声の区別的な特性を示す話者表現を決定するためにニューラルネットワークをトレーニングするための例示的なプロセスの流れ図である。

【図４Ａ】発声プール内の異なる話者に対する発声のグループからのトレーニング発声のサブセットの例示的な選択を示す概念図である。

【図４Ｂ】ニューラルネットワークをトレーニングするための発声プールからのトレーニングデータのバッチの例示的な選択を示す概念図である。

【図５Ａ】発声の少なくとも一部を特徴づけるデータを処理し、当該発声の当該少なくとも一部を特徴づけるデータに基づいて話者表現を生成するように構成された例示的なディープ・ニューラルネットワークのブロック図である。

【図５Ｂ】話者検証モデルで使用するために構成される長短期メモリレイヤを有する例示的な再帰型ニューラルネットワークのブロック図である。

【図６】コンピューティングデバイス上のニューラルネットワークを用いて、ユーザの発声から決定されたユーザの音声の特性に基づいてユーザのアイデンティティを検証するための例示的なプロセスの流れ図である。

【図７】本明細書で説明するコンピュータ実行型の方法および他の技術を実行する際に使用できるコンピューティングデバイスおよびモバイルコンピューティングデバイスの１例を示す図である。

【発明を実施するための形態】

【００３３】

図１は、話者検証モデルに対するニューラルネットワークをトレーニングし、当該モデルを用いて話者検証のプロセスを実行するための例示的なシステム１００の略図である。一般に、話者検証とは、当該話者の１つまたは複数の発声から決定される当該話者の音声の特性に基づいて話者のアイデンティティ要求を受理または拒否するタスクである。図１に示すように、話者検証は一般に３つの段階、即ち（ｉ）当該話者検証モデルに対するニ

ューラルネットワークのトレーニング、（ｉｉ）新たな話者の加入、および（ｉｉｉ）当該加入した話者の検証を含むことができる。

【００３４】

システム１００はクライアントデバイス１１０、コンピューティングシステム１２０、およびネットワーク１３０を含む。幾つかの実装では、コンピューティングシステム１２０は、トレーニングされたニューラルネットワーク１４０に基づいて話者検証モデル１４４をクライアントデバイス１１０に提供してもよい。幾つかの実装では、話者検証モデル１４４は、例えば、オペレーティング・システムまたはアプリケーションのコンポーネントとしてクライアントデバイス１１０に事前インストールされてもよい。他の実装では、話者検証モデル１４４がネットワーク１３０上で受信されてもよい。クライアントデバイス１１０は、話者検証モデル１４４を使用してユーザ１０２を当該話者検証プロセスに加入させてもよい。後の時点でユーザ１０２のアイデンティティを検証する必要があるとき、クライアントデバイス１１０は、話者検証モデル１４４を用いてユーザ１０２のアイデンティティを検証するためにユーザ１０２の会話発声を受信してもよい。話者検証モデル１４４をクライアントデバイス１１０にローカルに格納してもよいので、クライアントデバイス１１０はネットワーク１３０上の通信なしに話者検証判定を行うことができうる。

【００３５】

図１には示していないが、幾つかの実装では、コンピューティングシステム１２０は、クライアントデバイス１１０に格納されるニューラルネットワーク１４０ではなくまたはそれに加えて、当該トレーニングされたニューラルネットワーク１４０に基づいて話者検証モデル１４４を格納してもよい。これらの実装では、クライアントデバイス１１０は、ネットワーク１３０を介してコンピューティングシステム１２０と通信して話者検証モデル１４４にリモートにアクセスし、それをユーザ１０２の加入のために使用してもよい。後の時点でユーザ１０２のアイデンティティを検証する必要があるとき、クライアントデバイス１１０はユーザ１０２の会話発声を受信してもよく、ネットワーク１３０を介してコンピューティングシステム１２０と通信して、当該リモートに配置された話者検証モデル１４４を用いてユーザ１０２のアイデンティティを検証してもよい。コンピューティングシステム１２０およびコンピューティング・デバイス１１０は、互いと異なってもよく、物理的に別々であってもよい。

【００３６】

システム１００では、クライアントデバイス１１０は、例えば、デスクトップコンピュータ、ラップトップコンピュータ、タブレットコンピュータ、時計、ウェラブルコンピュータ、セルラ電話、スマートフォン、音楽プレイヤ、ｅブックリーダ、ナビゲーションシステム、またはユーザが対話しうる任意の他の適切なコンピューティングデバイスであることができる。幾つかの実装では、クライアントデバイス１１０はモバイルコンピューティングデバイスであってもよい。コンピューティングシステム１２０は１つまたは複数のコンピュータを含むことができ、当該コンピュータのうち独立なものに対して機能を実施してもよく、または当該機能は、複数のコンピュータにわたって実施するために分散されてもよい。ネットワーク１３０は、有線またはワイヤレスまたはその両方の組合せであることができ、インターネットを含むことができる。

【００３７】

幾つかの実装では、ユーザの電話のようなクライアントデバイス１１０は、話者検証モデル１４４をクライアントデバイス１１０にローカルに格納して、クライアントデバイス１１０が、加入または検証処理のためにリモートサーバ（例えば、コンピューティングシステム１２０）でのモデルに頼ることなくユーザのアイデンティティを検証できるようにしてもよく、したがって通信帯域幅および時間を節約することができる。さらに、幾つかの実装では、１つまたは複数の新たなユーザが加入するとき、本明細書で説明した話者検証モデル１４４は当該新たなユーザを用いた話者検証モデル１４４の任意の再トレーニングを必要とせず、これはまた、計算的に効率的でありうる。他の実装では、ニューラルネットワーク（およびしたがって当該話者検証モデル）を新たに収集されたトレーニングデ

ータの使用に基づいて定常的に更新できるように、加入、検証、またはその両方のために提供された所与のユーザの発声をコンピューティングシステム１２０に提供し、当該トレーニングデータに追加してもよい。

【００３８】

当該トレーニングされたニューラルネットワーク１４０を含む話者検証モデル１４４のサイズは、クライアントデバイス１１０の記憶空間およびメモリ空間が制限されうるので、コンパクトでありうるのが望ましい。後述のように、話者検証モデル１４４はトレーニングされたニューラルネットワーク１４０に基づく。話者検証モデル１４４は、発声を特徴付けるデータに基づいて、当該発声の話者の音声の区別的な特徴を示す話者表現を生成するためのニューラルネットワーク１４０を含んでもよい。話者検証モデル１４４は、当該発声の話者のアイデンティティ要求を検証できるように、当該話者表現を処理し、当該発声の話者の音声が加入したユーザの音声と十分に同様であるかかどうかを判定するためのさらなるコンポーネントを含んでもよい。

【００３９】

幾つかの実装では、ニューラルネットワークは大規模な１組のトレーニングデータを用いてトレーニングされてもよい。様々な技術を当該トレーニングデータの前処理の間に、トレーニング自体の間に、または後トレーニング段階の間に適用して、ニューラルネットワークのサイズを強制および／または削減し、コンパクトなモデルサイズを実現してもよい。例えば、話者検証モデル１４４を、ニューラルネットワーク１４０の特定のレイヤのみを選択することで構築してもよい。これは、クライアントデバイス１１０に格納するのに適したコンパクトな話者検証モデルをもたらしうる。当該話者モデルに対する話者表現を生成する際にソフトマックスまたは他の分類レイヤなしに、加入を実施してもよい。

【００４０】

図１はまた、段階（Ａ）乃至（Ｆ）で示す、データの例示的なフローを示す。段階（Ａ）乃至（Ｆ）を、当該示されたシーケンスで行ってもよく、または、当該示されたシーケンスと異なるシーケンスで行ってもよい。幾つかの実装では、段階（Ａ）乃至（Ｆ）の１つまたは複数をオフラインで行ってもよい。コンピューティングシステム１２０は、クライアントデバイス１１０がネットワーク１３０に接続されないときに計算を実施してもよい。段階（Ａ）および（Ｂ）は一般に上で参照した当該トレーニング段階の間に発生する。段階（Ｄ）は一般に加入段階の間に発生する。段階（Ｅ）乃至（Ｇ）は一般に検証段階の間に発生する。

【００４１】

段階（Ａ）で、コンピューティングシステム１２０は、ニューラルネットワーク１４０の監視されたトレーニングをニューラルネットワーク１４０に提供するためのトレーニング発声のサンプルを選択する。幾つかの実装では、トレーニング・サンプル１２２における発声がそれぞれ、多数の異なるトレーニング話者により話された１つまたは複数の所定の単語から構成されてもよい。当該発声は以前に記録されておりコンピューティングシステム１２０により使用するためにアクセス可能とされている。各トレーニング話者は所定の発声をコンピューティングデバイスに話してもよく、コンピューティングデバイスは当該発声を含むオーディオ信号を記録してもよい。例えば、各トレーニング話者を、当該トレーニングフレーズ「Hello Phone」を話すように促してもよい。幾つかの実装では、各トレーニング話者を、同一のトレーニングフレーズ複数回を話すように促してもよい。各トレーニング話者の当該記録されたオーディオ信号がコンピューティングシステム１２０に送信されてもよく、コンピューティングシステム１２０は当該記録されたオーディオ信号を多数の異なるコンピューティングデバイスおよび多数の異なるトレーニング話者から収集してもよい。幾つかの実装では、ユーザのアイデンティティが当該予め定義されたトレーニングフレーズの発声から決定されたユーザの音声の特性に基づいて検証されてもよいという点で、ニューラルネットワーク１４０をテキスト依存の話者検証に対して最適化してもよい。かかる実装では、ニューラルネットワーク１４０は、全ての、または実質的に全ての、当該予め定義されたトレーニングフレーズを含む発声でトレーニングされても

よい。他の実装では、ニューラルネットワーク１４０は、ユーザのアイデンティティが多種多様な単語またはフレーズの発声から決定されたユーザの音声の特性に基づいて検証されてもよいという点で、テキスト独立な話者検証を可能とするためにトレーニングされてもよく、当該多種多様な単語またはフレーズの発声は予め定義されていなくてもよい。例えば、ユーザは、自分のアイデンティティを検証するために、どの単語またはフレーズを彼または彼女が話したいかを独立に判定でき、当該トレーニングされたニューラルネットワーク１４０に基づく話者検証モデルは次いで、当該話された単語またはフレーズが与えられた場合にユーザを認証することができる。テキスト独立な話者検証を可能とするために、ニューラルネットワーク１４０は、多数の異なるトレーニング話者により話された多種多様な単語またはフレーズの発声でトレーニングされてもよい。

【００４２】

段階（Ｂ）で、ニューラルネットワーク１４０は、クライアントデバイスでユーザの加入および検証と平行する方式でトレーニングされてもよい。したがって、コンピューティングシステム１２０は、各トレーニング・サンプル１２２において、１組のシミュレートされた加入発声１２２ｂおよびシミュレートされた検証発声１２２ａを選択することができる。シミュレートされた加入発声１２２ｂは、シミュレートされた話者モデルをトレーニング・サンプル１２２ごとに決定できるように、同一のトレーニング話者の全ての発声であってもよい。シミュレートされた検証発声１２２ａは、シミュレートされた加入発声１２２ｂの話者と同一の話者の発声であってもよく、または異なる話者の発声であってもよい。トレーニング・サンプル１２２を次いでニューラルネットワーク１４０に提供でき、シミュレートされた検証発声１２２ａがシミュレートされた加入発声１２２ｂの話者と同一の話者により話されたか、またはシミュレートされた加入発声１２２ｂの話者と異なる話者により話されたかどうかに関して、ニューラルネットワーク１４０の出力に基づいて分類を行ってもよい。ニューラルネットワーク１４０は次いで、当該話者判定が正しかったかどうかに基づいて更新することができる。幾つかの実装では、各トレーニング・サンプル１２２を、２つのクラス、即ち、（当該シミュレートされた検証発声および加入発声の話者が同一であるサンプルに関する）マッチング話者クラス１４１ａおよび（当該シミュレートされた検証発声および加入発声の話者が異なるサンプルに関する）非マッチング話者クラス１４１ｂのうち１つに属するとしてラベル付けしてもよい。これらのラベルは、同一の話者が発声１２２ａおよび発声１２２ｂを話したかどうかのグランド・トルースを示す。トレーニング・サンプル１２２の分類の正確性を当該サンプルのラベルに基づいて決定することができる。幾つかの実装では、ニューラルネットワークに対する調節は、入力サンプルの分類の正確性に厳密に基づかなくてもよいが、一般にシミュレートされた検証発声１２２ａおよびシミュレートされた加入発声１２２ｂに対するニューラルネットワークにより生成された話者表現の比較から決定された１つまたは複数のメトリックに基づいてもよい。幾つかの実装では、トレーニング・サンプル１２２をトレーニングデータのリポジトリから選択してもよい。当該トレーニングデータのリポジトリを、発声プール１２１に編成してもよい。発声プール１２１の各々は当該発声のトレーニング話者によりグループ化されたトレーニング発声を含んでもよい。

【００４３】

ニューラルネットワーク１４０は、トレーニング・サンプル１２２における発声に関する情報を入力するための入力レイヤ、およびサンプル１２２を処理するための幾つかの隠蔽されたレイヤを含んでもよい。当該トレーニングされたニューラルネットワーク１４０が、サンプル１２２のシミュレートされた検証発声および加入発声のうちマッチングまたは非マッチング話者の何れかを有するとのトレーニング・サンプル１２２の所望の分類を生成させる出力を話者検証モデル１４４が生成するように、１つまたは複数の隠蔽されたレイヤの重みまたは他のパラメータを調節してもよい。幾つかの実装では、ニューラルネットワーク１４０のパラメータをコンピューティングシステム１２０により自動的に調節してもよい。幾つかの他の実装では、ニューラルネットワーク１４０のパラメータをコンピューティングシステム１２０のオペレータにより手動で調節してもよい。ニューラルネ

ットワークのトレーニング段階を、例えば図２、３、４Ａ－Ｂ、および５Ａ－Ｂの説明において以下でより詳細に説明する。

【００４４】

段階（Ｃ）で、ニューラルネットワーク１４０がトレーニングされると、当該トレーニングされたニューラルネットワーク１４０に基づく話者検証モデル１４４が、例えば、ネットワーク１３０を通じてコンピューティングシステム１２０からクライアントデバイス１１０に送信される。幾つかの実装では、当該トレーニングされたニューラルネットワーク１４０、またはその一部は、話者検証モデル１４４のコンポーネントであってもよい。話者検証モデル１４４を、１つまたは複数のユーザ１０２の発声から決定されたユーザの音声の特性に基づいてユーザ１０２のアイデンティティを検証するように構成することができる。モデル１４４を、ユーザの音声の区別的な特徴を示すユーザ１０２に対する話者表現を生成するために、ユーザ１０２の発声を特徴づけるデータを当該トレーニングされたニューラルネットワーク１４０への入力として提供するように構成してもよい。当該話者表現を次いで、以前に決定されたユーザの音声のモデルと比較することができる。当該話者表現がユーザの話者モデルと十分に同様である場合、話者検証モデル１４４は、ユーザ１０２のアイデンティティが正当であるという指示を出力することができる。対照的に、当該話者表現がユーザの話者モデルと十分に同様でない場合、話者検証モデル１４４は、ユーザ１０２のアイデンティティが無効である（検証されない）という指示を出力することができる。

【００４５】

段階（Ｄ）で、クライアントデバイス１１０で自分の音声を加入させたいと望むユーザ１０２が、加入段階で１つまたは複数の加入発声１５２をクライアントデバイス１１０に提供する。一般に、ユーザ１０２は、音声がニューラルネットワーク１４０のトレーニングで使用されなかったトレーニング話者の１つではない。幾つかの実装では、クライアントデバイス１１０は、トレーニング・サンプル１２２の発声において当該１組のトレーニング話者により話された同一のフレーズである加入フレーズを話すことをユーザ１０２に促してもよい。幾つかの実装では、クライアントデバイス１１０は、加入フレーズを何回か話すようにユーザに促してもよく、当該話された加入発声に対するオーディオ信号を加入発声１５２として記録してもよい。

【００４６】

クライアントデバイス１１０は、加入発声１５２を使用して、クライアントデバイス１１０の話者検証システムにおいてユーザ１０２を加入させる。一般に、ユーザ１０２の加入はニューラルネットワーク１４０の再トレーニングなしに行われる。同一の話者検証モデル１４４の夫々のインスタンスは、多数の異なる話者を加入させるために、ニューラルネットワーク１４０における重み値または他のパラメータを変更するのを必要とせずに、多数の異なるクライアントデバイスで使用されてもよい。話者検証モデル１４４を、再トレーニングニューラルネットワーク１４０なしに任意のユーザを加入させるために使用できるので、制限された処理要件でクライアントデバイス１１０で加入を実施してもよい。

【００４７】

幾つかの実装では、加入発声１５２に関する情報は話者検証モデル１４４へ入力され、話者検証モデル１４４は、ユーザ１０２に対応する基準ベクトルまたは他の１組の値を出力してもよい。当該基準ベクトルまたは他の１組の値は、ユーザの音声の区別的な特徴を特徴付ける話者モデルを構成してもよい。当該話者モデルをクライアントデバイス１１０に、またはクライアントデバイス１１０から離れたコンピューティングシステムに格納してもよく、その結果、後にクライアントデバイス１１０により受信された発声に基づいて生成された話者表現を当該話者モデルと比較して、後に受信された発声の各話者がユーザ１０２であるかまたは他の話者であるかどうかを検証してもよい。

【００４８】

段階（Ｅ）で、ユーザ１０２が、音声認証を用いてクライアントデバイス１１０へのアクセスを得ようと試みる。ユーザ１０２は、検証段階で検証発声１５４をクライアントデ

バイス１１０に提供する。幾つかの実装では、検証発声１５４は、加入発声１５２と同一の、話されたフレーズの発声である。検証発声１５４は話者検証モデル１４４への入力として使用される。

【００４９】

段階（Ｆ）で、クライアントデバイス１１０は、ユーザの音声が加入したユーザの当該音声にマッチするかどうかを判定する。幾つかの実装では、ニューラルネットワーク１４０は検証発声１５４を特徴付けるデータを処理してもよく、検証発声１５４に基づいてユーザ１０２に対する話者表現を出力してもよい。幾つかの実装では、クライアントデバイス１１０は、ユーザ１０２に対する話者表現を加入したユーザに対する当該話者モデルと比較して、検証発声１５４が加入したユーザにより話されたかどうかを判定してもよい。ニューラルネットワークの検証段階を、例えば図６に関して以下でより詳細に説明する。

【００５０】

段階（Ｇ）で、クライアントデバイス１１０が、検証結果１５６を表す指示をユーザ１０２に提供する。幾つかの実装では、クライアントデバイス１１０がユーザ１０２のアイデンティティを受理した場合、クライアントデバイス１１０は、検証が成功したという視覚的指示またはオーディオ指示をユーザ１０２に送信してもよい。幾つかの他の実装では、クライアントデバイス１１０がユーザ１０２のアイデンティティを受理した場合、クライアントデバイス１１０はユーザ１０２に次の入力を促してもよい。例えば、クライアントデバイス１１０は、メッセージ「Device enabled. Please enter your search」を当該ディスプレイに出力してもよい。幾つかの他の実装では、クライアントデバイス１１０がユーザ１０２のアイデンティティを受理した場合、クライアントデバイス１１０は、さらなる入力をユーザ１０２から待つことなく後続のアクションを実施してもよい。例えば、ユーザ１０２は検証段階の間に、クライアントデバイス１１０に対して「Hello Phone , search the nearest coffee shop」を話してもよい。クライアントデバイス１１０は、検証フレーズ「Hello Phone」を用いてユーザ１０２のアイデンティティを検証してもよい。ユーザ１０２のアイデンティティが受理された場合、クライアントデバイス１１０は、ユーザ１０２にさらなる入力を求めることなく、最も近いコーヒー店の検索を実施してもよい。一般に、幾つかの実装では、クライアントデバイス１１０がユーザ１０２のアイデンティティを受理した場合、クライアントデバイス１１０はロックされた状態からロックされていない状態に遷移することで応答してもよい。当該ロックされた状態において、クライアントデバイス１１０の１つまたは複数の能力は無効化またはブロックであり、当該ロックされていない状態では、当該能力は有効であり、または、ユーザ１０２がアクセスするのに利用可能とされる。同様に、クライアントデバイス１１０は、成功した検証に応答してより完全に特徴付けられた状態に「活性化」または遷移してもよい。

【００５１】

幾つかの実装では、クライアントデバイス１１０がユーザ１０２のアイデンティティを拒絶した場合、クライアントデバイス１１０は、検証が拒絶されたという視覚的指示またはオーディオ指示をユーザ１０２に送信してもよい。幾つかの実装では、クライアントデバイス１１０がユーザ１０２のアイデンティティを拒絶した場合、クライアントデバイス１１０はユーザ１０２に別の発声試行を促してもよい。幾つかの実装では、試行の数が閾値を超過した場合、クライアントデバイス１１０は、ユーザ１０２が自分のアイデンティティをさらに検証するのを試みるのをブロックしてもよい。

【００５２】

図２を参照すると、ニューラルネットワーク２０６をトレーニングするための例示的なシステム２００のブロック図が示されている。図２により示されたトレーニング段階が完了すると、当該トレーニングされたニューラルネットワーク２０６は、話者の発声を特徴付けるデータを処理し、当該話者の音声の区別的な特徴を示す当該話者に対する話者表現を生成することができる。当該話者表現はついで、加入段階の間に当該話者に対する話者モデルを生成するか、または検証段階の間に当該話者のアイデンティティを検証する際に話者検証モデルにより使用されてもよい。

【００５３】
一般に、図２は、ニューラルネットワーク２０６が、後に話者検証タスクを実施するクライアントデバイスで発生する加入および検証段階と平行する方式でトレーニングされてもよいことを示す。トレーニング発声を有限数の話者から当該話者の各々に対する対応するクラスに分類するためにニューラルネットワーク２０６をトレーニングする幾つかのアプローチとは異なり、図２のニューラルネットワーク２０６は、所与の発声の特定の話者を決定するためにトレーニングされない。その代わり、ニューラルネットワーク２０６は、当該発声の何れかを特定の話者イデンティティと必ずしもマッチすることなく、区別的かつ所与の発声の話者が別の１組の発声の話者と同一であるか否かを判定するために使用可能である話者表現を生成するためにトレーニングされる。このように、トレーニングの間に最適化された損失関数は、検証段階の間に当該話者検証モデルにより利用される同一の関数である。換言すれば、検証の間に、検証発声に基づく話者表現は加入したユーザに対する話者モデルと比較される。当該話者表現が当該話者モデルと十分に同様である場合、検証発声の話者のアイデンティティが検証される。図２に示す当該アプローチはトレーニングの間に同様な技術を使用する。即ち、シミュレートされた話者モデル２１４が１つまたは複数の加入発声に対する話者表現に基づいて生成され、話者表現２０８はまた、シミュレートされた検証発声２０２に対して生成される。ニューラルネットワーク２０６の重み値および他のパラメータは、シミュレートされた検証発声２０２をシミュレートされた加入発声２０４ａ乃至ｎと同一または異なる話者により話されているとして分類する際のエラーを最小化するために、トレーニングの間に調節される。

【００５４】
図２は、シミュレートされた検証発声２０２を特徴づけるデータおよび１つまたは複数のシミュレートされた加入発声２０４ａ乃至ｎを特徴づけるデータを含むトレーニングデータのサンプルに基づく単一のトレーニングの反復の前方パスを示す。実際に、ニューラルネットワーク２０６は、多数の反復およびトレーニングデータの多数の異なるサンプルにわたってトレーニングされる。各反復により、ニューラルネットワーク２０６を、当該各反復に対するトレーニングデータの対応するサンプルを処理した結果に基づいて調節してもよい。図４Ａおよび４Ｂは、さらに以下で説明するように、シミュレートされた検証発声２０２およびシミュレートされた加入発声２０４ａ乃至ｎが選択されうる例示的な技術を示す。特定のサンプルに対するシミュレートされた加入発声２０４ａ乃至ｎは一般に、同一のトレーニング話者により話された全ての発声である。シミュレートされた加入発声２０４ａ乃至ｎの話者は異なるトレーニングの反復に対するトレーニングデータの異なるサンプルの間で異なってもよいが、所与のトレーニング反復に対する所与のトレーニング・サンプルにおいて、加入発声２０４ａ乃至ｎの全ては一般に同一のトレーニング話者により話される。シミュレートされた検証発声２０２はシミュレートされた加入発声２０４ａ乃至ｎの話者と同一のトレーニング話者により話されているかもしれず、または、シミュレートされた加入発声２０４ａ乃至ｎの話者と異なるトレーニング話者により話されているかもしれない。当該話者がシミュレートされた検証発声２０２とシミュレートされた加入発声２０４ａ乃至ｎの両方の間で同一であるトレーニングデータのサンプルに対して、当該サンプルを「マッチング」サンプルとしてラベル付けしてもよい。当該話者がシミュレートされた検証発声２０２およびシミュレートされた加入発声２０４ａ乃至ｎの間で異なるトレーニングデータのサンプルに対して、当該サンプルを「非マッチング」サンプルとしてラベル付けしてもよい。当該ラベルは当該トレーニング・サンプルの真の分類を表してもよく、トレーニングの前の前処理段階で決定してもよい。幾つかの実装では、ニューラルネットワーク２０６の出力に基づくトレーニング・サンプルの当該推定された分類は、当該トレーニング・サンプルに対するラベルにより示される真の分類と比較して、ニューラルネットワーク２０６を調節するかどうかを判定してもよい。

【００５５】
幾つかの実装では、当該トレーニング・サンプルにおけるデータは、当該シミュレートされた検証発声および加入発声２０２、２０４ａ乃至ｎに対する生のオーディオ信号はで

なくてもよい。その代わり、発声２０２、２０４ａ乃至ｎが、ニューラルネットワーク２０６により処理するための適切なフォーマットに処理および変換されていてもよい。例えば、当該トレーニング・サンプルにおけるデータは、生のオーディオ信号自体ではなく、当該シミュレートされた検証発声および加入発声２０２、２０４ａ乃至ｎの各特徴を特徴付けてもよい。幾つかの実装では、当該トレーニング・サンプル内のシミュレートされた発声２０２、２０４ａ乃至ｎの各々を表すデータは各発声に対する１つまたは複数のログフィルタバンクを含んでもよい。幾つかの実装では、各発声を当該発声に対する複数のフレームに分割してもよく、別々のログフィルタバンクを当該発声のフレームごとに生成することができる。例えば、当該発声の各フレームは例えば４０個のログフィルタバンクにより表されてもよい。

【００５６】

幾つかの実装では、シミュレートされた検証発声２０２を特徴づけるデータおよびシミュレートされた加入発声２０４ａ乃至ｎの各々を特徴づけるデータをニューラルネットワーク２０６を通じて一度（即ち、単一のパスで）処理することができる。したがって、所与の発声に対するトレーニングデータが各々各１組のログフィルタバンクにより表された複数のフレームに分割されているが、当該発声の全体に対するフレームの全てを特徴づけるデータを、ニューラルネットワークを通じて単一のパスで処理するために、ニューラルネットワーク２０６に（例えば、それぞれ４０個のログフィルタバンクを有する８０個のフレームに対する８０ｘ４０個の特徴ベクトルとして）入力することができる。これは、当該フレームを別々に処理するために、当該発声のフレームごとにデータをニューラルネットワーク２０６に独立して入力するのとは対照的である。他の実装では、各発声２０２、２０４ａ乃至ｎの全体を特徴づけるデータを単一のパスでニューラルネットワーク２０６を通じて処理するためにニューラルネットワーク２０６をトレーニングするのではなく、発声２０２、２０４ａ乃至ｎの独立なフレームを特徴づけるデータを、ニューラルネットワーク２０６への入力として提供することができる。

【００５７】

幾つかの実装では、シミュレートされた検証発声および加入発声２０２、２０４ａ乃至ｎを１つまたは複数の追加の技術に従って事前に処理してもよい。例えば、ニューラルネットワーク２０６の構造は、トレーニング発声が全て固定された長さ（例えば、０．８秒のオーディオ）を有するのを要求してもよい。少なくとも幾つかの発声２０２、２０４ａ乃至ｎはしたがって、長い発声を均一な長さに刈り込み、かつ／または幾つかの短い発声をパディングして長いクリップを作成した結果であってもよい。他の実装では、しかし、ニューラルネットワーク２０６は可変長さ発声を処理できてもよく、この場合、トレーニングデータ内の発声２０２、２０４ａ乃至ｎを固定された長さに刈り込むかまたはパディングしてもよい。発声２０２、２０４ａ乃至ｎに対するオーディオはまた均一化されていてもよく、雑音の存在において堅牢に実施されるようにニューラルネットワークがトレーニングされることを保証するために、トレーニング発声２０２、２０４ａ乃至ｎに雑音が追加されているかまたはそこから雑音が除去されていてもよい。

【００５８】

点線ボックス２１５内のシステム２００の部分は、複数のシミュレートされた加入発声２０４ａ乃至ｎを特徴づけるデータがシミュレートされた加入発声２０４ａ乃至ｎの特定のトレーニング話者に対するシミュレートされた話者モデル２１４を生成するために使用されるという点で、話者検証プロセスの加入段階をシミュレートする。シミュレートされた加入発声２０４ａ乃至ｎの各々を特徴づけるそれぞれのデータはニューラルネットワーク２０６ニューラルネットワーク２０６の入力レイヤに別々に入力される。ニューラルネットワーク２０６は１つまたは複数の隠蔽されたレイヤを通じて当該データを処理し、シミュレートされた加入発声２０４ａ乃至ｎの各々に対する各話者表現２１０ａ乃至ｎを生成する。例えば、図２に示すように、話者表現１（２１０ａ）はシミュレートされた加入発声１に基づいてニューラルネットワーク２０６により生成される（２０４ａ）。同様に、話者表現２（２１０ｂ）はシミュレートされた加入発声２に基づいてニューラルネット

ワーク２０６により生成される（２０４ｂ）。話者表現はしたがって、シミュレートされた加入発声２０４ａ乃至ｎの各々に対するニューラルネットワーク２０６により生成されてもよい。幾つかの実装では、話者表現２１０ａ乃至ｎを、ニューラルネットワーク２０６を通じてシミュレートされた加入発声２０４ａ乃至ｎの各々を逐次的に処理することで生成してもよい。幾つかの実装では、話者表現２１０ａ乃至ｎを、発声２０４ａ乃至ｎを特徴付けるデータを、シミュレートされた加入発声２０４ａ乃至ｎの各々に対するニューラルネットワーク２０６の各インスタンスと並列に処理することで並列に生成することができる。話者表現２１０ａ乃至ｎは一般にそれぞれ、シミュレートされた加入発声２０４ａ乃至ｎの対応する１つに基づいてニューラルネットワーク２０６により決定された、当該シミュレートされた加入トレーニング話者の音声の区別的な特性を表す値の集合を含む。幾つかの実装では、話者表現２１０ａ乃至ｎはニューラルネットワーク２０６の最後の隠蔽されたレイヤの重み値または他のパラメータを示してもよい。幾つかの実装では、話者表現２１０ａ乃至ｎは、ニューラルネットワーク２０６がソフトマックス出力レイヤなしに構成されたときの、ニューラルネットワーク２０６の最終的な出力であってもよい。

【００５９】

シミュレートされた話者モデル２１４を生成するために、話者表現２１０ａ乃至ｎを図２のボックス２１２に示すように平均化することができる。したがって、シミュレートされた話者モデル２１４は、シミュレートされた加入発声２０４ａ乃至ｎのトレーニング話者の音声の区別的な特性を表す値の集合を定義してもよい。シミュレートされた話者モデル２１４を決定するために複数の話者表現２１０ａ乃至ｎを平均化することで、当該異なるシミュレートされた加入発声２０４ａ乃至ｎの間の話者の音声の変形を平滑化することができる。シミュレートされた話者モデル２１４は、独立な話者表現２１０ａ乃至ｎの何れかよりも高信頼な話者の音声の表現であってもよい。これは、所与のシミュレートされた加入発声２０４ａ乃至ｎの特質を独立に反映してもよい。

【００６０】

幾つかの実装では、トレーニングデータの各サンプル内のシミュレートされた加入発声２０４ａ乃至ｎの総数はトレーニングの反復ごとに変化してもよい。例えば、第１のトレーニングの反復に対する第１のトレーニング・サンプルは９個のシミュレートされた加入発声２０４ａ乃至ｎを含んでもよい。第２のトレーニングの反復に対する第２のトレーニング・サンプルは、しかし、４個のシミュレートされた加入発声２０４ａ乃至ｎのみを含んでもよい。他の実装では、トレーニングデータの各サンプル内のシミュレートされた加入発声２０４ａ乃至ｎの総数はトレーニングの反復ごとに固定されていてもよい。例えば、ニューラルネットワーク２０６は、当該１組のトレーニングデータが反復ごとに全体で５個のシミュレートされた加入発声２０４ａ乃至ｎを含む一連の反復でトレーニングされてもよい。幾つかの実装では、当該トレーニングの反復の１つ、一部または全部を、単一のシミュレートされた加入発声２０４ａ乃至ｎのみを含むトレーニング・サンプルで実施してもよい。

【００６１】

話者表現２１０ａ乃至ｎが、シミュレートされた加入発声２０４ａ乃至ｎを特徴付けるデータから生成されたのと同一の方式で、話者表現２０８を、シミュレートされた検証発声２０２を特徴付けるデータから生成することができる。当該シミュレートされた検証発声２０２を特徴付けるデータ（例えば、検証発声２０２の特徴を特徴づけるログフィルタバンク値）をニューラルネットワーク２０６の入力レイヤに提供することができる。ニューラルネットワーク２０６次いで、当該ネットワークの１つまたは複数の隠蔽されたレイヤを通じて入力を処理する。ニューラルネットワーク２０６の出力は、シミュレートされた検証発声２０２を話した話者の音声の区別的な特性を示す値の集合を定義する話者表現２０８である。

【００６２】

ニューラルネットワーク２０６のトレーニングの間の検証段階とさらに並行するために、シミュレートされた検証発声２０２に基づく話者表現２０８は、例えば、検証段階の間

に話者検証モデルによりクライアントデバイスで行われるのと同一の方式でシミュレートされた話者モデル２１４と比較されることができる。幾つかの実装では、当該比較を、（１）シミュレートされた話者表現２０８に対して値の集合を定義する第１のベクトルおよび（２）シミュレートされた話者モデル２１４に対する値の集合を定義する第２のベクトルの余弦距離を（ブロック２１６に示すように）取得することにより実施することができる。ロジスティック回帰２１８を次いで当該距離に適用して、シミュレートされた検証発声２０２を話したトレーニング話者が、シミュレートされた加入発声２０４ａ乃至ｎを話したトレーニング話者と同一であるかまたは異なるかどうかを推定することができる。これは、マッチング話者クラスに対する第１のブロック２２０ａ、および非マッチング話者クラスに対する第２のブロック２２０ｂにより図２で表されている。幾つかの実装では、ロジスティック回帰２１８と異なる分類技術を適用して、シミュレートされた検証発声２０２を話したトレーニング話者が、シミュレートされた加入発声２０４ａ乃至ｎを話したトレーニング話者と同一であるかまたは異なるかどうかに関する判定を行ってもよい。例えば、ヒンジレイヤまたはソフトマックスレイヤを幾つかの代替的な分類に対して使用してもよい。図２に示すような２つのクラスモデルでは、当該そフトマックスおよびロジスティック回帰技術は同一のまたは同様な最適化機能を使用してもよい。

【００６３】

ニューラルネットワーク２０６の重み値または他のパラメータを次いで、ブロック２２２により表されるように、シミュレートされた検証発声２０２に対する話者表現２０８のシミュレートされた話者モデル２１４との比較の結果に基づいて、調節することができる。例えば、当該トレーニング・サンプルが真に非マッチング話者を有するとしてラベル付けされ、不正確に分類されたトレーニング・サンプルがマッチング話者を有するとして分類された場合、ニューラルネットワーク２０６はエラーを補正するように自動的に調節されてもよい。より一般に、ニューラルネットワーク２０６を最適化して、話者サンプルをマッチングするための類似性スコアを最大化するか、または、ロジスティック回帰によるスコア出力を最適化してもよく、ニューラルネットワーク２０６をまた最適化して、非マッチング話者サンプルに対する類似性スコアを最小化するか、またはロジスティック回帰によるスコア出力を最適化してもよい。幾つかの実装では、ニューラルネットワーク２０６に対する調節を、トレーニングの反復ごとに各トレーニング・サンプルの結果に応答して行うことができ、またはニューラルネットワーク２０６をトレーニングの反復の幾つかのみの結果に基づいて調節してもよい。幾つかの実装では、ニューラルネットワーク２０６を、話者表現２０８および非マッチング話者に対するシミュレートされた話者モデル２１４の間の距離（即ち、差異の最大化）を最大化し、話者表現２０８およびマッチング話者に対するシミュレートされた話者モデル２１４の間の距離を最小化（即ち、差異を最小化）するように調節してもよく。幾つかの実装では、トレーニング・サンプルをマッチング話者クラス２２０ａまたは非マッチング話者クラス２２０ｂの何れかに属するとして分類するための硬判定をトレーニング段階の間に行わなくてもよいことに留意されたい。寧ろ、ニューラルネットワーク２０６を、ロジスティック回帰レイヤ２１８により出力される当該スコアを最適化するか、または、１つまたは複数の他のメトリックを最適化する方式で調節してもよい。

【００６４】

次に図３を参照すると、話者検証モデルで使用されうるニューラルネットワークをトレーニングするための例示的なプロセス３００の流れ図が示されている。幾つかの実装では、プロセス３００を、図１からのコンピューティングシステム１２０および図２からのコンピューティングシステム２００のような本明細書で説明するコンピューティングシステムにより実行してもよい。

【００６５】

プロセス３００は段階３０２で開始し、第１の１組のトレーニングデータが選択される（即ち、第１のトレーニング・サンプル）。第１の１組のトレーニングデータは、シミュレートされた検証発声を特徴づけるデータおよび１つまたは複数のシミュレートされた加

入発声を特徴づけるデータを含むことができる。当該トレーニングセットにおける発声は、それらがトレーニング段階の間に話者検証の加入段階および検証段階を並行させるかまたは「シミュレート」する方式でトレーニングプロセスで使用されるという点で、「シミュレートされる」。しかし、当該発声自体は一般に、トレーニング話者により話された記録された会話の実際のスニペットである。当該トレーニング話者は一般に、当該話者検証プロセスの実際の加入および検証段階の間に発声を提供した同一の話者ではない。下記でさらに説明する図４Ａおよび４Ｂは、当該シミュレートされた検証発声および加入発声を選択するための例示的な技術を示す。

【００６６】

当該選択された１組のトレーニングデータ（即ち、選択されたサンプル）を、それがマッチング話者の会話または非マッチング話者に対するサンプルを表すかどうかに従ってラベル付けしてもよい。当該シミュレートされた検証発声の話者が当該シミュレートされた加入発声の話者と同じである場合、当該１組のトレーニングデータはマッチング話者サンプルとしてラベル付けされる。当該シミュレートされた検証発声の話者が当該シミュレートされた加入発声の話者と異なる場合、当該１組のトレーニングデータが非マッチング話者サンプルとしてラベル付けされる。幾つかの実装では、当該ラベルを、マッチングまたは非マッチングサンプルの何れかであるとして当該１組のトレーニングデータの推定された分類が正確であるか否かを判定するために、後にトレーニングプロセス３００において使用することができる。

【００６７】

幾つかの実装では、選択された１組のトレーニングデータは、シミュレートされた検証発声および加入発声に対する生のオーディオ信号ではなく、その代わり当該発声の特徴を特徴付けるデータを含んでもよい。例えば、当該１組のトレーニングデータで表された各発声は、当該発声の固定長のフレームに対して決定された１組のログフィルタバンクにより特徴付けられることができる。当該発声のフレームごとの当該ログフィルタバンクをついで、ニューラルネットワークへの入力として提供され当該発声の全体を分類する単一の１組の入力値に連結してもよい。

【００６８】

プロセス３００の段階３０４および３０６で、話者表現は、第１の１組のトレーニングデータで分類される発声の各々に対して決定される。当該話者表現はそれぞれ、当該各話者表現に対する対応する発声を話した当該トレーニング話者の音声の区別的な特徴を示す値の集合であることができる。例えば、第１の話者表現が当該シミュレートされた検証発声に基づいて生成されてもよく、各第２の話者表現が当該シミュレートされた加入発声の各々に基づいて生成されてもよい。当該話者表現を生成するために、発声を特徴づけるデータが、トレーニングされているニューラルネットワークの入力レイヤに提供される。ニューラルネットワークは次いで、当該ネットワークの１つまたは複数の隠蔽されたレイヤを通じて当該入力データを処理する。当該話者表現は次いでニューラルネットワークの出力である。幾つかの実装では、当該出力は、ソフトマックスレイヤではない出力レイヤで出力される。当該出力を提供する最終的なレイヤは完全に接続された線形レイヤであってもよい。幾つかの実装では、当該話者表現は、ソフマックス出力レイヤの出力ではなく、ニューラルネットワークの最後の隠蔽されたレイヤで生成された値またはその活性化を含んでもよい。幾つかの実装では、ニューラルネットワークをソフトマックス出力レイヤなしで構成してもよい。

【００６９】

段階３０８で、当該シミュレートされた加入発声に対応する話者表現が、シミュレートされた話者モデルを生成するために結合される。当該シミュレートされた話者モデルは、当該シミュレートされた加入発声に対する話者表現の平均であることができる。当該話者表現を平均化することで、当該トレーニング話者の音声を特徴づける高信頼のモデルを決定することができる。例えば、当該話者が当該シミュレートされた加入発声の各々を話した方式の変形を、当該シミュレートされた検証発声に対する話者表現が比較される堅牢な

ベースラインで当該話者モデルを使用できるように、平滑化してもよい。幾つかの実装では、プロセス３００は、当該シミュレートされた加入発声に対する話者表現のサブセットのみを選択して、当該シミュレートされた話者モデルを生成する際に結合してもよい。例えば、当該シミュレートされた加入発声の各々または当該対応するシミュレートされた加入発声の品質の測定値を決定してもよい。プロセス３００はついで、当該シミュレートされた話者モデルを生成するために使用される１組の表現に含めるために、閾値品質スコアを満たすこれらの話者表現のみ、または対応する発声が閾値品質スコアを満たすこれらの話者表現を選択してもよい。

【００７０】

段階３１０で、シミュレートされた検証発声に対する話者表現がシミュレートされた話者モデルと比較される。幾つかの実装では、二進クラシファイアが、マッチング話者を表すかまたは表さないとしてデータサンプルを分類するために使用される。幾つかの実装では、当該比較は、当該シミュレートされた検証発声に対する話者表現および当該シミュレートされた話者モデルの間の類似性の測定値を決定するステップを含むことができる。例えば、類似性の測定値は、当該話者表現に対する値のベクトルと当該シミュレートされた話者モデルに対する値のベクトルの間の余弦距離であってもよい。類似性の測定値はついで、マッチング話者サンプルまたは非マッチング話者サンプルの何れかとして第１の１組のトレーニングデータの分類を推定するために使用されてもよい。例えば、類似性の測定値が十分に高い（例えば、閾値類似性スコアを満たす）場合、ロジスティック回帰を、当該１組のトレーニングデータをマッチング話者のクラスにマップするために使用してもよい。他方、類似性の測定値が低すぎる（例えば、閾値類似性スコアを満たさない）場合、ロジスティック回帰を、当該１組のトレーニングデータを非マッチング話者のクラスにマップするために使用してもよい。

【００７１】

次に、段階３１２で、ニューラルネットワークの１つまたは複数のパラメータを、シミュレートされた検証発声に対する話者表現およびシミュレートされた話者モデルの間の段階３１０での比較の結果に基づいて調節してもよい。例えば、当該トレーニングデータが非マッチング話者サンプルとしてラベル付けされた場合に、隠蔽されたレイヤにおける様々なノードの重み、またはニューラルネットワークの他のパラメータを調節して、当該話者表現および当該シミュレートされた話者モデルの間の距離を増大（類似性スコアを減少）させてもよい。さらに、当該トレーニングデータがマッチング話者サンプルとしてラベル化された場合に、ニューラルネットワークの重みまたは他のパラメータを調節して、当該話者表現および当該シミュレートされた話者モデルの間の距離を削減（当該類似性スコアを増大）させてもよい。一般に、トレーニングプロセス３００の各反復は各加入段階および各検証段階をシミュレートすることを意図しているので、ニューラルネットワークを調節して、話者検証の間に実際の加入および検証段階に適用されるたものと同一の損失関数を最適化してもよい。このアプローチの１つの利益は、ニューラルネットワークが、より正確な検証話者のアイデンティティのための話者検証モデルにおいて使用できる話者表現をより良く生成するためにトレーニングされるということである。例えば、幾つかの実装では、ニューラルネットワークをトレーニングするときに考慮されない発声の実際の検証の間に追加の後処理ステップが行われない。これらの技術を、ニューラルネットワークをトレーニングするための「エンド・ツー・エンド」のアプローチと考えてもよい。

【００７２】

最後に、段階３１４で、次に１組のトレーニングデータが、トレーニングニューラルネットワークの別の反復に対して選択される。再度、この段階で選択された当該１組のトレーニングデータはシミュレートされた検証発声を特徴付けるデータおよび１つまたは複数のシミュレートされた加入発声を特徴付けるデータを含んでもよい。プロセス３００はついで、段階３０４乃至３１２を反復し、追加のトレーニングの反復に対するトレーニングデータの追加のセットを限界に達するまで選択し続けてもよい。幾つかの実装では、当該限界が、利用可能なトレーニングデータの全てが期限切れになることから生じてもよい。

幾つかの実装では、プロセス３００をターゲット性能レベルに到達するまで続けてもよい。例えば、何回かのトレーニングの反復の後、ニューラルネットワークを、トレーニングプロセス３００の間に使用されなかった差し出された１組のデータに対してテストしてもよい。トレーニングを、当該差し出されたセット上の試験が、ニューラルネットワークが少なくともターゲット性能レベルを達成したことを示すまで、続けてもよい。

【００７３】

次に図４Ａおよび４Ｂを参照すると、話者検証モデルに対するニューラルネットワークをトレーニングする際に使用するためのトレーニングデータのセットを選択するための例示的な技術を示す略図が示されている。幾つかの実装では、図４Ａおよび４Ｂに関して説明された技術は、多数のトレーニングの反復にわたって選択されたトレーニング発声の多様性を保証することができる。これは、所与の数のトレーニング発声に対してニューラルネットワークをより良く実施することをもたらしうる。

【００７４】

幾つかの実装では、当該利用可能なトレーニング発声の全部のまたは一部を複数のグループ４１０ａ乃至ｎにクラスタ化してもよい。グループ４１０ａ乃至ｎを、トレーニング発声のグループの集合を含む発声プール４０８にさらに構成してもよい。当該トレーニング発声を幾つかの実装では話者によりグループ化してもよい。例えば、グループ４１０ａは、全て第１の話者により話された複数の発声を含み、グループ４１０ｎは全て別の話者により話された複数の発声を含む。したがって、グループ４１０ａ乃至ｎの各々は異なる話者に対応してもよい。グループ４１０ａ乃至ｎは全て同一のトレーニング発声の数を含んでもよく、またはトレーニング発声の数はグループ４１０ａ乃至ｎの異なるものの間で変化してもよい。

【００７５】

トレーニングの反復ごとに、発声プール４０８にアクセスしてもよく、特定の発声を、各トレーニングの反復における入力として適用されるトレーニングデータのサンプルに対して選択してもよい。例えば、図４Ａは、入力サンプル４０２としてトレーニングの反復のために発声プール４０８からランダムに選択された１つの１組のトレーニングデータを示す。第１の話者に対応する第１の発声グループは、当該シミュレートされた話者モデルを生成する際に使用するための発声プール４０８内のグループ４１０ａ乃至ｎから選択することができる。当該グループをランダムにまたは別の方式で選択してもよい。当該選択されたグループ、例えば、図４Ａのグループ４１０ａから、当該第１の話者の発声のサブセットが、入力サンプル４０２内のシミュレートされた加入発声４０６として選択される。このサブセットは一般に複数の発声を含み、或るトレーニングの反復から別の反復へと、同一数または異なる数の発声を含んでもよい。当該選択されたグループ、例えば、グループ４１０ａからの発声をランダムに選択してもよく、その結果、当該発声の異なる組合せが、異なるトレーニングの反復において第１の話者に対する異なるシミュレートされた話者モデルを生成するために使用される。

【００７６】

発声４０４もシミュレートされた検証発声として選択される。発声４０４は、当該トレーニングの反復が加入発声４０６とのマッチまたは非マッチの１例であるかどうかに依存して、第１の話者または異なる話者の発声であってもよい。マッチングおよび非マッチングの例の両方がトレーニングで使用される。結果として、幾つかのトレーニングの反復に対して、発声４０４は、第１の話者の発声、例えば、グループ４１０ａからの発声である。他のトレーニングの反復に対して、発声４０４は、図４Ａに示すように、第１の話者と異なる第２の話者の発声であり、その結果入力サンプル４０２はシミュレートされた検証発声４０４とシミュレートされた加入発声４０６の間のマッチを表さない。

【００７７】

図４Ａの例では、特定の発声が、シミュレートされた検証発声４０４として第２の発声グループ４１０ｎから選択される（例えば、ランダムに選択される）。幾つかの実装では、（発声４０４がそこから選択される）第２の発声グループを、発声プール４０８内のグ

ループ４１０ａ乃至ｎからランダムに、または、グループ４１０ａ乃至ｎの変化する選択のパターンに従って選択してもよい。他の実装では、当該シミュレートされた加入発声の話者と同一の話者からの別の発声が当該シミュレートされた検証発声として適用されるべきかどうかに関してランダムな選択を行ってもよい。したがって、おそらく当該ランダムな選択がバイアスされ、その結果、シミュレートされた検証発声４０４がシミュレートされた加入発声の話者４０６と同一の話者の発声であるという５０パーセントの可能性が存在する。当該ランダムな選択の結果が、入力サンプル４０２がマッチング話者サンプルであるというものである場合、シミュレートされた検証発声４０４を、シミュレートされた加入発声４０６が選択された発声のグループと同一の発声のグループ４１０から選択することができる。しかし、当該ランダムな選択の結果が、入力サンプル４０２が非マッチング話者サンプルであるというものである場合、シミュレートされた検証発声４０４を、シミュレートされた加入発声４０６がそこから選択された発声のグループと異なる話者に対応する発声４１０の異なるグループから選択することができる。

【００７８】

一般に、図４Ａにより示される選択技術は、話者の異なる組合せからの発声を異なるトレーニングの反復で適用させることができる。例えば、第１のトレーニングの反復において、当該シミュレートされた加入発声は第１の話者により話されているかもしれず、当該シミュレートされた検証発声はまた第１の話者により話されているかもしれない。第２のトレーニングの反復において、当該シミュレートされた加入発声は第２の話者により話されているかもしれず、当該シミュレートされた検証発声は第３の話者により話されているかもしれない。次に第３のトレーニングの反復において、当該シミュレートされた加入発声は第１の話者により話されているかもしれず、当該シミュレートされた検証発声は第２の話者により話されているかもしれない。幾つかの実装では、異なる順列を生成するかまたは当該シミュレートされた検証発声と加入発声の話者の間の入力サンプル４０２における順列数を最大化する方式で発声４１０ａ乃至ｎのグループをランダムに選択せずその代わり決定的に発声４１０ａ乃至ｎのグループを選択する、選択アルゴリズムを使用してもよい。単純な例として、３つの異なるトレーニング話者からの発声の３つのグループＡ、Ｂ、およびＣが発声プール４０８で利用可能であった場合、９個の異なる入力サンプル４０２が９個のトレーニングの反復、即ち、（Ａ、Ａ）、（Ａ、Ｂ）、（Ａ、Ｃ）、（Ｂ、Ａ）、（Ｂ、Ｂ）、（Ｂ、Ｃ）、（Ｃ、Ａ）、（Ｃ、Ｂ）、および（Ｃ、Ｃ）に関して生成されてもよい。トレーニングの反復はまた、これらの同一のグループのペアとともに生じうるが、当該グループ内では異なる発声が選択される。

【００７９】

本明細書で説明する当該トレーニングアプローチの１つの履歴は、事前に選択された数の話者のうち特定の話者に属するとして入力を分類するためにニューラルネットワークをトレーニングすることを含む他のアプローチと対照的に、より多くの様々な話者を当該ネットワークをトレーニングするために使用してもよい。さらに、高信頼のトレーニングを保証するためにトレーニング話者ごとに要求される（トレーニング話者ごとに実際に使用される当該１つまたは複数の発声以外の）最小数のトレーニング発声はない。なぜならば、当該ネットワークは特定の話者に対してトレーニングされず、その代わり所与の入力サンプル４０２が当該シミュレートされた検証発声および加入発声の中にマッチング話者または非マッチング話者があるかどうかに基づいてトレーニングされるからである。

【００８０】

図４Ｂは、ニューラルネットワークのトレーニングの間に入力サンプルに関する発声を選択するためのシャッフル技術の略図４００ｂを示す。本図に示すように、トレーニング・サンプルのバッチにおけるサンプルは全て、当該バッチ内のトレーニング・サンプルの間の発声の良好なシャッフルおよび多様性を取得するために異なるプールから来ることができる。当該シャッフル技術は、より堅牢なおよび高信頼のニューラルネットワークのトレーニングをもたらしうる。

【００８１】

図５Ａおよび５Ｂを参照すると、話者検証モデルで使用されうる例示的なニューラルネットワーク５０２、５１２のブロック図が示されている。幾つかの実装ではニューラルネットワーク５０２、５１２の何れかを、図１乃至４Ｂおよび６に関して説明した技術を実装するために使用してもよく、図２乃至４Ｂに関して説明したトレーニング技術を含む。

【００８２】

図５Ａのディープ・ニューラルネットワーク５０２のアーキテクチャは、ローカルに接続されたレイヤ５０４を含み、それに続いて１つまたは複数の完全に接続された隠蔽されたレイヤ５０６ａ乃至ｎを含む。ローカルに接続されたレイヤ５０４および完全に接続されたレイヤ５０６ａ乃至ｎは整流線形ユニット（ＲｅＬＵ）を有してもよい。ネットワーク５０２の最後のレイヤは完全に接続された線形レイヤ５０８であり、これは、入力発声（または発声のフレーム）５０３ａに基づいて話者表現５１０ａを出力する。表現５１０ａの前の最後のレイヤ５０２は、幾つかの実装では、非負の活性化を全空間にマップし、投影を決定するための線形レイヤである。当該全空間は、ＲｅＬｕ活性化がｙ＝ｍａｘ（ｘ、０）のような関数でありうるという概念を指す。したがって、当該話者表現を形成する活性化（ｙ）は常に正のベクトルであってもよい。かかる活性化関数が線形活性化関数ｙ＝ｘにより変更される場合、当該話者表現を潜在的に正負の値を有するベクトルとして生成することができる。後者は、例えば、それに余弦距離比較関数が続くとき、当該話者のより適切な表現であることができる。

【００８３】

ニューラルネットワーク５０２の構成は一般に、固定された長さのトレーニング発声、または固定された数の発声のフレームを処理することができる。ニューラルネットワーク５０２がトレーニングされ、後に加入および検証段階における実行時の間に使用されるとき、発声を適切に刈り込みまたはパディングして、当該発声がニューラルネットワーク５０２により処理するために要求される固定長を有することを保証してもよい。結果として、ニューラルネットワーク５０２は、単一のパス、例えば、ディープ・ニューラルネットワーク５０２を通じた単一の前方伝播で話者表現を計算することができる。これにより、当該話者表現を、発声の異なる部分の逐次的処理を含む技術より低いレイテンシで生成することができる。

【００８４】

次に、図５Ｂに示すニューラルネットワーク５１２は、再帰型ニューラルネットワークである。ニューラルネットワーク５０２のアーキテクチャと異なり、ニューラルネットワーク５１２は可変長の入力発声を処理することができる。例えば、発声５０３ｂは、ニューラルネットワーク５１２が使用されているコンテキストに依存して、トレーニング発声、加入発声、または検証発声であってもよい。発声５０３ｂを複数のフレームに分割してもよく、当該複数のフレームは固定長を有してもよい。ニューラルネットワーク５１２に入力されたフレームの数は、発声５０３ｂの全体の長さの関数であってもよい。換言すれば、長い発声はより多くのフレームを有してもよく、短い発声はより少ないフレームを有してもよい。発声５０３ｂのフレームは長短期メモリ（ＬＳＴＭ）レイヤ５１６に入力される。１つまたは複数の追加の隠蔽されたレイヤがＬＳＴＭレイヤ５１６に続いてもよい。ネットワーク５１２の最後のレイヤは再度、完全に接続された線形レイヤ５１８である。幾つかのケースでは、完全に接続された線形レイヤ５１８は、非負の活性化を当該全空間にマッピングし、投影を決定することで話者表現５１０ｂを出力してもよい。ニューラルネットワーク５１２は可変長さ発声を扱うことができるので、これは発声の単語またはフレーズが予め定義されず異なる発声の間で変化しうるテキスト独立な話者検証に良く適しうる。

【００８５】

図５Ａおよび５Ｂに示すニューラルネットワーク５０２および５１２は特定の構成を有するとして示されているが、本明細書で説明する当該技術で使用されうるニューラルネットワークはこれらの例により限定されない。例えば、ニューラルネットワークの隠蔽されたトポロジは異なる数および配置のレイヤを有してもよく、当該レイヤは、完全に接続さ

れたレイヤ、ローカルに接続されたレイヤ、または長短期メモリレイヤのような任意の回帰レイヤを含んでも含まなくてもよい。幾つかの実装では、ニューラルネットワークは従来型のニューラルネットワークであってもよい。

【００８６】

図６は、本明細書で説明する技術に従ってトレーニングされている話者検証モデルおよびニューラルネットワークを用いて発声を検証するための例示的なプロセス６００の流れ図である。プロセス６００は一般に図１に示す検証段階（段階Ｅ乃至Ｇ）に対応する。図６で参照されるニューラルネットワークは、幾つかの実装では、図２乃至４Ｂに関して説明した技術に従ってトレーニングされてもよく、図５Ａまたは５Ｂに示すような構造を有してもよい。

【００８７】

段階６０２で、発声はコンピューティングデバイスのユーザから受信されることができる。例えば、ユーザは、自分のスマートフォンをアンロックするか、または、幾つかの他の機能をコンピューティングデバイスで実施したいかもしれない。しかし、スマートフォンは、それがロックされない前に、または所望の機能が実施される前にユーザを認証するようにユーザに要求してもよい。幾つかの実装では、当該認証を、スマーフォンの話者検証モデルを用いてユーザの音声の特性に基づいて実施してもよい。スマーフォンは、検証発声を話すようにユーザに促してもよく、検証発声を段階６０２でスマーフォンにより受信し記録してもよい。

【００８８】

段階６０４で、スマーフォンは、当該受信された発声に基づいて話者表現を生成するためにニューラルネットワークにアクセスする。ニューラルネットワークは、ローカルにスマーフォンに格納されてもよく、または例えば、アプリケーションプログラミングインタフェース（ＡＰＩ）を介してリモートコンピューティングシステム上でアクセスされてもよい。ニューラルネットワークは本明細書で説明する技術に従ってトレーニングされてもよく、それぞれシミュレートされた検証発声および複数のシミュレートされた加入発声を含むデータのサンプルに基づいてトレーニングされているかもしれない。ニューラルネットワークは、単一のパス内でニューラルネットワークを通じて、発声の全体を特徴付けるデータを処理するように構成されてもよい。段階６０６で、当該受信された発声を特徴付けるデータがニューラルネットワークへの入力として提供される。ニューラルネットワークは、当該入力を処理し、ユーザの音声の区別的な特性を示す話者表現を生成する。

【００８９】

段階６０８で、話者モデルがスマーフォンでアクセスされる。当該話者モデルは加入したユーザの音声の区別的な特徴を示してもよい。幾つかの実装では、当該話者モデルが、加入したユーザの各発声を特徴付けるデータからニューラルネットワークにより生成された複数の話者表現の平均に基づいてもよい。段階６１０で、検証発声に基づいて段階６０６で生成された話者表現が、当該話者モデルと比較され、または、そうでない場合は当該話者モデルに関して評価される。幾つかの実装では、当該比較または他の評価がユーザのスマーフォン上で話者検証モデルにより実施される。当該話者検証モデルは、検証発声に対する話者モデルと話者表現の間の類似性の距離または他の測定値を決定してもよい。類似性の距離または他の測定値に基づいて、当該話者検証モデルは、ユーザの音声が加入したユーザの音声と十分に同様である場合に、ユーザを認証してもよい。そうでない場合、ユーザの音声の類似性が加入したユーザの音声に関して少なくとも閾値類似性スコアを満たさない場合に、当該話者検証モデルはユーザが認証されないという指示を生成してもよい。

【００９０】

幾つかの実装では、当該話者検証モデルが、検証発声が加入した話者により話されたことを十分な確信度で判定した場合、加入したユーザに対する話者モデルをついで、検証発声に基づいて更新してもよい。例えば、当該デバイスが以下の３つの検証発声にどのように応答しうるかを考える。当該話者検証モデルが第１の検証発声を話したユーザのアイデ

ンティティを拒否する（例えば、デバイスは第１の検証発声に応答してアンロックを拒否してもよい）ように、３つの検証発声のうち最初のものに対する類似性スコアは第１の閾値より小さい。当該３つの検証発声のうち２番目のものに対する類似性スコアは、第２の検証発声を話したユーザのアイデンティティが受理されるように、第１の閾値を満たしてもよい。しかし、第２の検証発声に対する類似性スコアは、加入したユーザの話者モデルを第２の検証発声に基づいて更新するために十分に高い。最後に、検証発声のうち第３のものに対する類似性スコアが、第３の検証発声を話したユーザのアイデンティティが受理される（例えば、デバイスのアンロックのような第１の１組のアクションを実施してもよい）ように、第１の閾値を満たし、また、加入したユーザに対する当該話者モデルが第３の検証発声に基づいて更新されうるように、より高い第２の閾値を満たす。当該話者モデルが、第３の検証発声に対してニューラルネットワークにより生成された話者表現を第１のインスタンス内の話者モデルを生成するために使用されたユーザの加入発声からの他の話者表現と結合（例えば、平均化）することで更新されてもよい。

【００９１】

段階６１２で、スマーフォンは次いで、ユーザが認証されたかどうかに基づいて動作を行うことができる。例えば、当該発声を提供したユーザが加入したユーザであるという判定に応答して、スマーフォンを起動またはアンロックしてもよい。しかし、当該発声を提供したユーザがが加入したユーザでないか、または、複数の加入したユーザの１つでないと判定された場合、スマーフォンはロックされたままであってもよく、または、そうでない場合はユーザが実施するために選択された１つまたは複数の機能の実施をブロックしてもよい。別のアプリケーションでは、本明細書で説明する話者検証技術をユーザデバイス（例えば、スマートフォン、ノートブックコンピュータ、ウェラブルデバイス）で使用して、当該デバイスにより非認証されたユーザ（例えば、音声当該デバイスで加入されていないユーザ）から検出された会話入力を拒否してもよい。例えば、当該デバイスがロックされていない状態であるとき、当該デバイスは、ユーザが当該デバイスに実施してほしいアクション（例えば「Navigate to the football game」または「Play my music collection」）を示すデバイスの認証されたユーザにより話された音声コマンドをリッスンしてもよい。幾つかの実装では、当該音声コマンドが当該認証されたユーザにより話されたと判定できる場合には、当該デバイスは当該音声コマンドにより示される当該要求されたアクションをのみを実施してもよい。このように、他の非認証されたユーザからの雑談を、例えば、拒絶してもよい。

【００９２】

図７は、本明細書で説明する技術を実装するために使用できるコンピューティングデバイス７００およびモバイルコンピューティングデバイスの１例を示す。コンピューティングデバイス７００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のような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。当該モバイルコンピューティングデバイスは、携帯情報端末、セルラスマーフォン、スマートフォン、および他の同様なコンピューティングデバイスのような様々な形態のモバイルデバイスを表すことを意図している。ここで示したコンポーネント、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すぎず、本明細書で説明および／またはクレームした発明の実装を限定しようとするものではない。

【００９３】

コンピューティングデバイス７００は、プロセッサ７０２、メモリ７０４、記憶デバイス７０６、メモリ７０４および複数の高速拡張ポート７１０に接続する高速インタフェース７０８、および低速拡張ポート７１４および記憶デバイス７０６に接続する低速インタフェース７１２を含む。プロセッサ７０２、メモリ７０４、記憶デバイス７０６、高速インタフェース７０８、高速拡張ポート７１０、および低速インタフェース７１２の各々は様々なバスを用いて相互接続され、必要に応じて共通のマザーボード上でまたは他の方式でマウントされてもよい。プロセッサ７０２は、高速インタフェース７０８に接続される

ディスプレイ７１６のような外部入力／出力デバイスにＧＵＩに関するグラフィカル情報を表示するためのメモリ７０４または記憶デバイス７０６に格納された命令を含む、コンピューティングデバイス７００内で実行するための命令を処理することができる。他の実装では、複数のプロセッサおよび／または複数のバスは、必要に応じて、複数のメモリおよびメモリのタイプに沿って使用されてもよい。また、複数のコンピューティングデバイスは、（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）必要な動作の部分を提供する各デバイスと接続されてもよい。

【００９４】

メモリ７０４は情報をコンピューティングデバイス７００内に格納する。幾つかの実装では、メモリ７０４は揮発性メモリユニットまたはユニットである。幾つかの実装では、メモリ７０４は非揮発性メモリユニットまたはユニットである。メモリ７０４はまた、磁気または光ディスクのような別の形態のコンピュータ可読媒体であってもよい。

【００９５】

記憶デバイス７０６は大容量記憶をコンピューティングデバイス７００に提供することができる。幾つかの実装では、記憶デバイス７０６は、記憶領域ネットワークまたは他の構成でのデバイスを含む、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・メモリまたは他の同様な固体状態メモリデバイス、またはデバイスのアレイのようなコンピュータ可読媒体であってもよいかまたはそれを含んでもよい。当該コンピュータプログラム製品はまた、実行されたとき上述したような１つまたは複数の方法を実施する命令を含んでもよい。当該コンピュータプログラム製品をまた、プロセッサ７０２上のメモリ７０４、記憶デバイス７０６、またはメモリのようなコンピュータまたはマシン可読媒体で有形に具体化することができる。

【００９６】

高速インタフェース７０８はコンピューティングデバイス７００に対する帯域幅集約的な動作を管理し、低速インタフェース７１２はより低い帯域幅集約的な動作を管理する。かかる機能の割当ては例示的なものにすぎない。幾つかの実装では、高速インタフェース７０８は、（例えば、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）メモリ７０４、ディスプレイ７１６に接続され、高速拡張ポート７１０に接続される。高速拡張ポート７１０は様々な拡張カード（図示せず）を受けつけてもよい。当該実装では、低速インタフェース７１２は記憶デバイス７０６および低速拡張ポート７１４に接続される。低速拡張ポート７１４は、様々な通信ポート（例えば、ＵＳＢ、Ｂｌｕｅｔｏｏｔｈ（登録商標）、イーサネット（登録商標）、ワイヤレスイーサネット（登録商標））を含んでもよく、キーボード、ポインティングデバイス、スキャナのような１つまたは複数の入力／出力デバイス、またはスイッチまたはルータのようなネットワーキングデバイスに、例えば、ネットワークアダプタを通じて接続してもよい。

【００９７】

コンピューティングデバイス７００を本図に示すように幾つかの異なる形態で実装してもよい。例えば、それを標準サーバ７２０として、またはかかるサーバのグループにおいて複数回、実装してもよい。さらに、ラップトップコンピュータ７２２のようなパーソナルコンピュータで実装してもよい。また、ラックサーバシステム７２４の一部として実装してもよい。代替的に、コンピューティングデバイス７００からのコンポーネントは、モバイルコンピューティングデバイス７５０のようなモバイルデバイス（図示せず）内の他のコンポーネントと結合されてもよい。かかるデバイスの各々はコンピューティングデバイス７００およびモバイルコンピューティングデバイス７５０の１つまたは複数を含んでもよく、システム全体は互いと通信する複数のコンピューティングデバイスで構成されてもよい。

【００９８】

モバイルコンピューティングデバイス７５０は、プロセッサ７５２、メモリ７６４、他のコンポーネントのうちディスプレイ７５４、通信インタフェース７６６、および送受信機７６８のような入力／出力デバイスを含む。モバイルコンピューティングデバイス７５

０にまた、追加の記憶を提供するための、マイクロドライブまたは他のデバイスのような記憶デバイスが提供されてもよい。プロセッサ７５２、メモリ７６４、ディスプレイ７５４、通信インタフェース７６６、および送受信機７６８の各々は様々なバスを用いて相互接続され、当該コンポーネントの幾つかは必要に応じて共通のマザーボード上でまたは他の方式でマウントされてもよい。

【００９９】

プロセッサ７５２は、メモリ７６４に格納された命令を含む、モバイルコンピューティングデバイス７５０内の命令を実行することができる。プロセッサ７５２を、別々のおよび複数のアナログおよびデジタルプロセッサを含むチップから成るチップ・セットとして実装してもよい。プロセッサ７５２は、例えば、モバイルコンピューティングデバイス７５０により実行されるユーザインタフェース、アプリケーション、およびモバイルコンピューティングデバイス７５０によるワイヤレス通信の制御のような、モバイルコンピューティングデバイス７５０の他のコンポーネントの協調を提供してもよい。

【０１００】

プロセッサ７５２は、ディスプレイ７５４に接続される制御インタフェース７５８およびディスプレイインタフェース７５６を通じてユーザと通信してもよい。ディスプレイ７５４は、例えば、ＴＦＴ（薄膜トランジスタ液晶ディスプレイ）ディスプレイまたはＯＬＥＤ（有機発光ダイオード）ディスプレイ、または他の適切なディスプレイ技術であってもよい。ディスプレイインタフェース７５６は、ディスプレイ７５４を駆動してグラフィカルおよび他の情報をユーザに提供するための適切な回路を備えてもよい。制御インタフェース７５８は、ユーザからコマンドを受信し、プロセッサ７５２に送信するために当該コマンドを変換してもよい。さらに、外部インタフェース７６２は、他のデバイスとのモバイルコンピューティングデバイス７５０の近領域通信を可能するために、プロセッサ７５２との通信を提供してもよい。外部インタフェース７６２は、例えば、幾つかの実装では有線通信を提供し、または他の実装ではワイヤレス通信を提供してもよく、複数のインタフェースをまた使用してもよい。

【０１０１】

メモリ７６４はモバイルコンピューティングデバイス７５０内に格納する。メモリ７６４を、コンピュータ可読媒体または媒体、揮発性メモリユニットまたはユニット、または非揮発性メモリユニットまたはユニットの１つまたは複数として実装することができる。拡張メモリ７７４はまた、拡張インタフェース７７２を通じてモバイルコンピューティングデバイス７５０に提供され接続されてもよい。拡張インタフェース７７２は、例えば、ＳＩＭＭ（Single In Line Memory Module）カードインタフェースを含んでもよい。拡張メモリ７７４はモバイルコンピューティングデバイス７５０に対する追加の記憶空間を提供してもよく、または、モバイルコンピューティングデバイス７５０に対するアプリケーションまたは他の情報を格納してもよい。特に、拡張メモリ７７４は、上述したプロセスを実行または補完する命令を含んでもよく、セキュア情報を含んでもよい。したがって、例えば、拡張メモリ７７４を、モバイルコンピューティングデバイス７５０に対するセキュリティモジュールとして提供してもよく、モバイルコンピューティングデバイス７５０の安全な使用を許可する命令でプログラムされてもよい。さらに、セキュアアプリケーションを、ハック不能な方式でＳＩＭＭカードに識別情報を配置するといった、当該ＳＩＭＭカードを介して追加の情報とともに提供してもよい。

【０１０２】

当該メモリは、以下で説明するように、例えば、フラッシュ・メモリおよび／またはＮＶＲＡＭメモリ（非揮発性ランダム・アクセスメモリ）を含んでもよい。当該コンピュータプログラム製品は、実行されたとき、上述したもののような１つまたは複数の方法を実施する命令を含む。当該コンピュータプログラム製品は、メモリ７６４、拡張メモリ７７４、またはプロセッサ７５２上のメモリのようなコンピュータまたはマシン可読媒体であることができる。幾つかの実装では、当該コンピュータプログラム製品を、伝播信号で、例えば、送受信機７６８または外部インタフェース７６２上で受信することができる。

【０１０３】
モバイルコンピューティングデバイス７５０は通信インタフェース７６６を通じて無線で通信してもよい。通信インタフェース７６６は、必要な場合はデジタル信号処理回路を含んでもよい。通信インタフェース７６６は、とりわけＧＳＭ（登録商標）通話（Global System for Mobile communications）、ＳＭＳ（Short Message Service）、ＥＭＳ（Enhanced Messaging Service）、またはＭＭＳメッセージング（Multimedia Messaging Service）、ＣＤＭＡ（code division multiple access、ＴＤＭＡ（time division multiple access）、ＰＤＣ（Personal Digital Cellular）、ＷＣＤＭＡ（登録商標）（Wideband Code Division Multiple Access）、ＣＤＭＡ２０００、またはＧＰＲＳ（General Packet Radio Service）のような、様々なモードまたはプロトコルの下での通信を提供してもよい。かかる通信を、例えば、送受信機７６８を通じて無線周波数を用いて行ってもよい。さらに、短波通信を、例えばＢｌｕｅｔｏｏｔｈ（登録商標）、ＷｉＦｉ、または他のかかる送受信機（図示せず）を用いて行ってもよい。さらに、ＧＰＳ（Global Positioning System）受信器モジュール７７０は追加のナビゲーション位置関連のワイヤレスデータをモバイルコンピューティングデバイス７５０に提供してもよい。モバイルコンピューティングデバイス７５０は、モバイルコンピューティングデバイス７５０で実行されているアプリケーションにより必要に応じて使用されてもよい。

【０１０４】
モバイルコンピューティングデバイス７５０はまた、オーディオコーデック７６０を用いて可聴的に通信してもよく、オーディオコーデック７６０は話された情報をユーザから受信し、それを使用可能なデジタル情報に変換してもよい。オーディオコーデック７６０は、話者を通じて、例えば、モバイルコンピューティングデバイス７５０のハンドセットでユーザに対する可聴音を同様に生成してもよい。かかる音は音声通話からの音を含んでもよく、記録された音（例えば、音声メッセージ、音楽ファイル等）を含んでもよく、また、モバイルコンピューティングデバイス７５０で動作しているアプリケーションにより生成された音を含んでもよい。

【０１０５】
モバイルコンピューティングデバイス７５０を、本図で示すように幾つかの異なる形態で実装してもよい。例えば、セルラスマーフォン７８０として実装してもよい。また、スマートフォン７８２、携帯情報端末、または他の同様なモバイルデバイスの一部として実装してもよい。

【０１０６】
本明細書で説明したシステムおよび技術の様々な実装を、デジタル電子回路、集積回路、特別に設計されたＡＳＩＣ（特殊用途向け集積回路）、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／またはその組合せで実現することができる。これらの様々な実装は、少なくとも１つのプログラム可能プロセッサを含むプログラム可能システムで実行可能および／または解釈可能である１つまたは複数のコンピュータプログラムでの実装を含むことができる。当該少なくとも１つのプログラム可能プロセッサは、特殊目的または一般的な目的であってもよく、記憶システム、少なくとも１つの入力デバイス、および少なくとも１つの出力デバイスとデータおよび命令を送受信するために接続されてもよい。

【０１０７】
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）はプログラム可能プロセッサに対するマシン命令を含み、高レベル手続型および／またはオブジェクト指向プログラミング言語で、および／またはアセンブリ／マシン言語で実装することができる。本明細書で使用する際、当該用語マシン可読媒体およびコンピュータ可読媒体は、マシン命令をマシン可読信号として受信するマシン可読媒体を含む、マシン命令および／またはデータをプログラム可能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、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、プログラム可能論理デ

バイス（ＰＬＤ））を指す。マシン可読信号という用語は、マシン命令および／またはデータをプログラム可能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。

【０１０８】

ユーザとの対話を提供するために、本明細書で説明したシステムおよび技術を、ユーザが当該コンピュータへの入力を提供できるユーザおよびキーボードおよびポインティングデバイス（例えば、マウスまたはトラックボール）に情報を表示するための、ディスプレイデバイス（例えば、ＣＲＴ（カソード・レイ・チューブ）またはＬＣＤ（液晶ディスプレイ）モニタ）を有するコンピュータで実装することができる。他の種類のデバイスを、ユーザとの対話を提供するために使用することができる。例えば、ユーザに提供されるフィードバックは任意の形態のセンサフィードバック（例えば、視覚フィードバック、可聴フィードバック、または触覚フィードバック）であることができ、ユーザからの入力を音響、会話、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。

【０１０９】

本明細書で説明したシステムおよび技術を、（例えば、データサーバとして）バックエンドコンポーネントを含む、またはミドルウェアコンポーネント（例えば、アプリケーションサーバ）を含む、またはフロントエンドコンポーネント（例えば、ユーザがそれを通じて本明細書で説明したシステムおよび技術の実装と対話できるグラフィカルユーザインタフェースまたはウェブ・ブラウザを有するクライアントコンピュータ）を含む、またはかかるバックエンド、ミドルウェア、またはフロントエンドコンポーネント任意の組合せを含む、コンピューティングシステムで実装することができる。当該システムの当該コンポーネントは、デジタルデータ通信（例えば、通信ネットワーク）の任意の形態または媒体により相互接続されることができる。通信ネットワークの例はローカル・エリア・ネットワーク（ＬＡＮ）、広帯域ネットワーク（ＷＡＮ）、およびインターネットを含む。

【０１１０】

当該コンピューティングシステムはクライアントおよびサーバを含むことができる。クライアントおよびサーバは一般に互いから離れており、一般に通信ネットワークを通じて対話する。クライアントおよびサーバの関係は、当該各コンピュータで実行され互いにクライアントサーバ関係を有するコンピュータプログラムにより生ずる。

【０１１１】

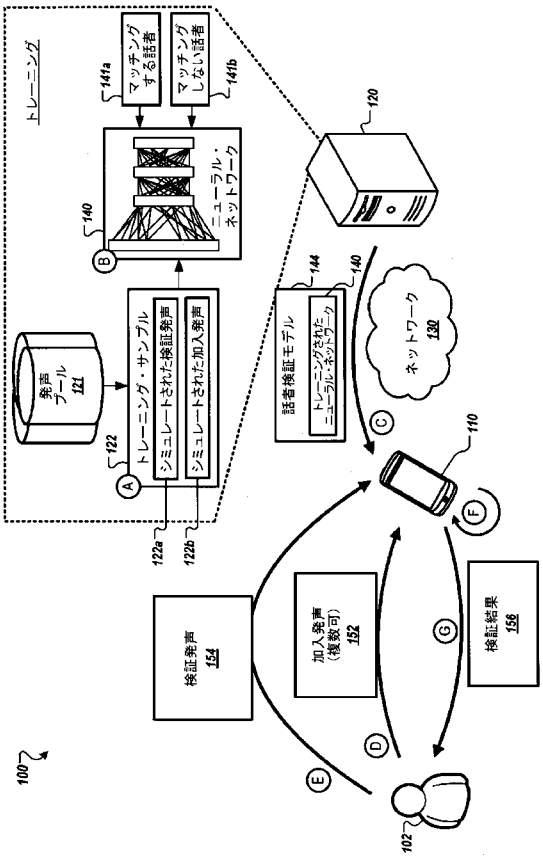
様々な実装を上で詳細に説明したが、他の修正が可能である。さらに、本図で示した論理フローは、所望の結果を実現するために、示した特定の順序、または逐次的な順序を必要としない。さらに、他のステップを提供してもよく、またはステップを当該説明したフローから削除してもよく、他のコンポーネントを当該説明したシステムに追加し、または、そこから削除してもよい。したがって、他の実装は添付の特許請求の範囲内にある。

【符号の説明】

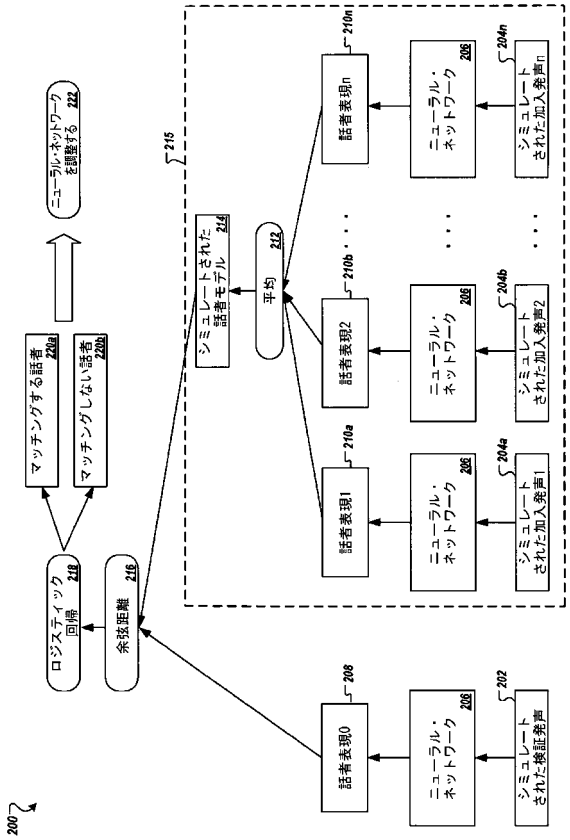
【０１１２】

１２１　発声プール
１２２　トレーニング・サンプル
１２２ａ　シミュレートされた検証発声
１２２ｂ　シミュレートされた加入発声
１３０　ネットワーク
１４０　トレーニングされたニューラル・ネットワーク
１４１ａ　マッチングする話者
１４１ｂ　マッチングしない話者
１４４　話者検証モデル
１５２　加入発声（複数可）
１５４　検証発声
１５６　検証結果

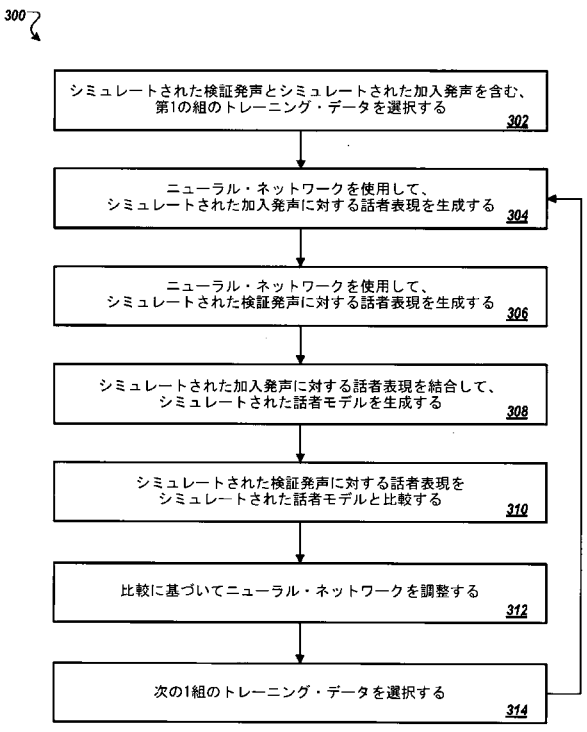
【図１】



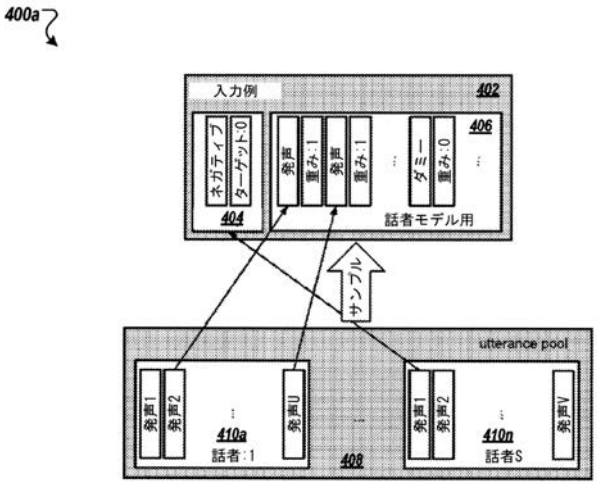
【図２】



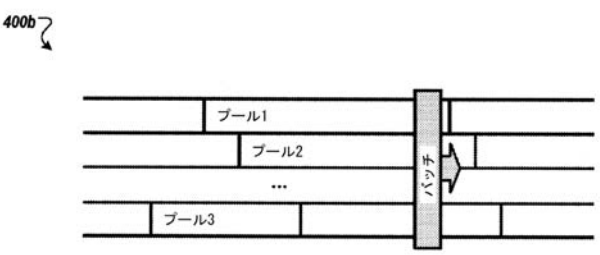
【図３】



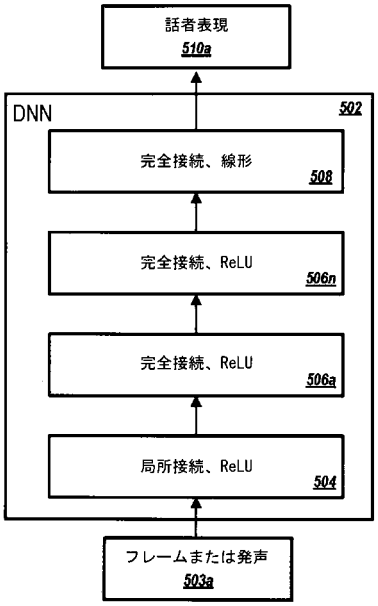
【図４Ａ】



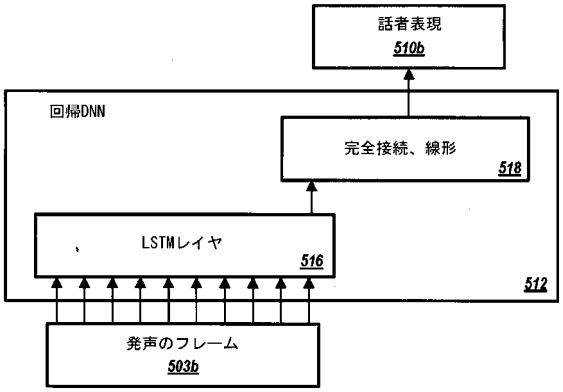
【図４Ｂ】



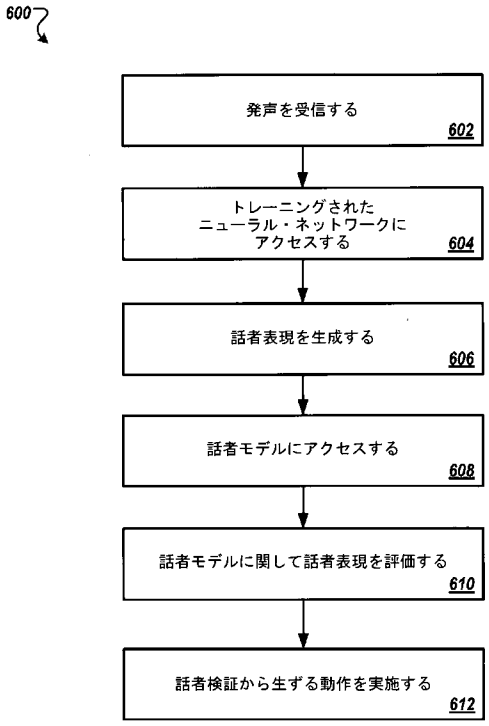
【図５Ａ】


話者表現
510a
DNN
502
完全接続、線形
508
完全接続、ReLU
506n
完全接続、ReLU
506a
局所接続、ReLU
504
フレームまたは発声
503a

【図５Ｂ】


話者表現
510b
回帰DNN
完全接続、線形
518
LSTMレイヤ
516
512
発声のフレーム
503b

【図６】


600
発声を受信する
602
トレーニングされた
ニューラル・ネットワークに
アクセスする
604
話者表現を生成する
606
話者モデルにアクセスする
608
話者モデルに関して話者表現を評価する
610
話者検証から生ずる動作を実施する
612

【図７】

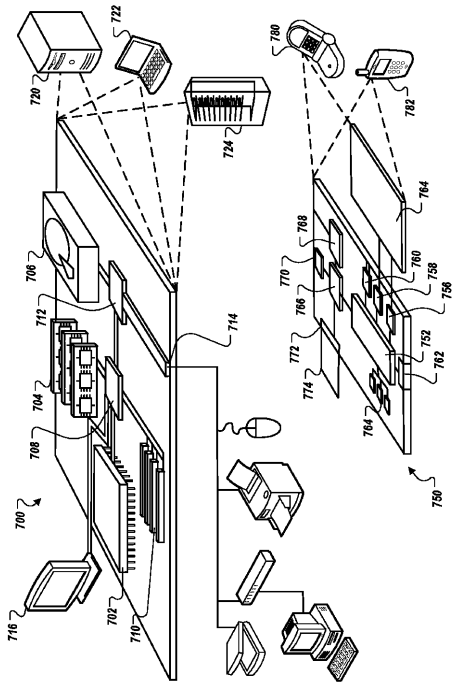


FIG. 7

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2016/044181

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L17/04 G10L17/18 G10L17/02
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2015/127336 A1 (LEI XIN [US] ET AL) 7 May 2015 (2015-05-07)	1-6
A	paragraph [0030] paragraphs [0035] - [0040] paragraphs [0042], [0043] paragraph [0045] paragraphs [0060], [0061] paragraph [0068] paragraphs [0076] - [0079]	7,9
	-/--	

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
6 October 2016	19/10/2016

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Ramos Sánchez, U

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2016/044181

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X,P	HEIGOLD GEORG ET AL: "End-to-end text-dependent speaker verification", 2016 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), IEEE, 20 March 2016 (2016-03-20), pages 5115-5119, XP032901578, DOI: 10.1109/ICASSP.2016.7472652 [retrieved on 2016-05-18] section '6. Summary & Conclusion', first 4 lines; page 5118 section '3. D-Vector Baseline Approach'; page 5116, left-hand column section '4. End-To-End Speaker Verification'; page 5116, right-hand column, paragraph 3 - page 5117, left-hand column, paragraph 4 page 5117, right-hand column, paragraphs 2,3 -----	1-21

1

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2016/044181

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2015127336 A1	07-05-2015	NONE	

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JP,KE,KG,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US

(72)発明者 サミー・ベンジオ
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ヴュー・アンフィシアター・パークウェイ・１６００
(72)発明者 イグナシオ・ロペス・モレーノ
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ヴュー・アンフィシアター・パークウェイ・１６００