(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7023934号
(P7023934)

(45)発行日　令和4年2月22日(2022.2.22)

(24)登録日　令和4年2月14日(2022.2.14)

(51)国際特許分類

Ｇ１０Ｌ	*15/16*	*(2006.01)*	Ｇ１０Ｌ　15/16
Ｇ１０Ｌ	*17/18*	*(2013.01)*	Ｇ１０Ｌ　17/18

ＦＩ

請求項の数　19　（全25頁）

(21)出願番号　特願2019-510589(P2019-510589)
(86)(22)出願日　平成29年8月24日(2017.8.24)
(65)公表番号　特表2019-528476(P2019-528476A)
(43)公表日　令和1年10月10日(2019.10.10)
(86)国際出願番号　PCT/US2017/048499
(87)国際公開番号　WO2018/039500
(87)国際公開日　平成30年3月1日(2018.3.1)
審査請求日　令和2年7月6日(2020.7.6)
(31)優先権主張番号　201610741622.9
(32)優先日　平成28年8月26日(2016.8.26)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　511050697
アリババ グループ ホウルディング リミテッド
英国領ケイマン諸島 グランド ケイマン ジョージ タウン ピーオーボックス ８４７ ワン キャピタル プレイス フォース フロア
(74)代理人　100079108
弁理士　稲葉 良幸
(74)代理人　100109346
弁理士　大貫 敏史
(74)代理人　100117189
弁理士　江口 昭彦
(74)代理人　100134120
弁理士　内藤 和彦

最終頁に続く

(54)【発明の名称】　音声認識方法及び装置

(57)【特許請求の範囲】
【請求項１】
音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出することと、
前記話者認識特徴を含む前記ベクトルに従って第２のニューラルネットワーク内のバイアスを補償することと、
前記第２のニューラルネットワークに基づく音響モデルを介し前記音声データ内の音声を認識することと、を含む音声認識方法。
【請求項２】
前記話者認識特徴を含む前記ベクトルに従って前記第２のニューラルネットワーク内のバイアスを補償することは、前記話者認識特徴を含む前記ベクトルに前記第２のニューラルネットワークのバイアス項となるべき重み行列を掛けることを含む、請求項１に記載の音声認識方法。
【請求項３】
前記第１のニューラルネットワーク、前記第２のニューラルネットワーク、及び前記重み行列は、前記第１のニューラルネットワーク及び前記第２のニューラルネットワークをそれぞれトレーニングし、次に前記トレーニングされた第１のニューラルネットワーク、前記重み行列、及び前記トレーニングされた第２のニューラルネットワークを一括してトレーニングすることによりトレーニングされる、請求項２に記載の音声認識方法。
【請求項４】

前記第１のニューラルネットワーク、前記第２のニューラルネットワーク、及び前記重み行列を初期化することと、
所定客観的判定基準に従って逆伝搬アルゴリズムを使用することにより前記重み行列を更新することと、
所定客観的判定基準に従って前記逆伝搬アルゴリズムを使用することにより前記第２のニューラルネットワーク及び接続行列を更新することと、をさらに含む請求項３に記載の音声認識方法。
【請求項５】
前記話者認識特徴は少なくとも話者声紋情報を含む、請求項１乃至４のいずれか一項に記載の音声認識方法。
【請求項６】
前記話者認識特徴を含む前記ベクトルに従って前記第２のニューラルネットワーク内のバイアスを補償することは、前記話者認識特徴を含む前記ベクトルに従って前記第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償することを含み、
前記話者認識特徴を含む前記ベクトルは前記第１のニューラルネットワーク内の最後の隠れ層の出力ベクトルである、請求項１に記載の音声認識方法。
【請求項７】
前記話者認識特徴を含む前記ベクトルに従って前記第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償することは、前記第１のニューラルネットワークの前記最後の隠れ層においてノードにより出力された前記話者認識特徴を含む前記ベクトルを、前記第２のニューラルネットワーク内の前記入力層を除く層の前記すべて又は一部に対応するバイアスノードへ送信することを含む、請求項６に記載の音声認識方法。
【請求項８】
前記音声データは、収集された元音声データ又は前記収集された元音声データから抽出された音声特徴である、請求項１に記載の音声認識方法。
【請求項９】
前記話者認識特徴は、様々なユーザ又は様々なユーザのクラスタに対応する、請求項１に記載の音声認識方法。
【請求項１０】
一組の命令を格納する非一時的コンピュータ可読媒体であって、前記一組の命令は、装置の１つ又は複数のプロセッサによって、前記装置に音声認識の方法を行わせるように、実行可能であり、前記方法は、
音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出することと、
前記話者認識特徴を含む前記ベクトルに従って第２のニューラルネットワーク内のバイアスを補償することと、
前記第２のニューラルネットワークに基づく音響モデルを介し前記音声データ内の音声を認識することと、を含む、非一時的コンピュータ可読媒体。
【請求項１１】
前記話者認識特徴を含む前記ベクトルに従って前記第２のニューラルネットワーク内のバイアスを補償することは、前記話者認識特徴を含む前記ベクトルに前記第２のニューラルネットワークのバイアス項となるべき重み行列を掛けることを含む、請求項１０に記載の非一時的コンピュータ可読媒体。
【請求項１２】
前記第１のニューラルネットワーク、前記第２のニューラルネットワーク、及び前記重み行列は、前記第１のニューラルネットワーク及び前記第２のニューラルネットワークをそれぞれトレーニングし、次に前記トレーニングされた第１のニューラルネットワーク、前記重み行列、及び前記トレーニングされた第２のニューラルネットワークを一括してトレ

ーニングすることによりトレーニングされる、請求項１１に記載の非一時的コンピュータ可読媒体。
【請求項１３】
前記一組の命令は、前記装置の前記１つ又は複数のプロセッサにより、前記装置に、
前記第１のニューラルネットワーク、前記第２のニューラルネットワーク、及び前記重み行列を初期化することと、
所定客観的判定基準に従って逆伝搬アルゴリズムを使用することにより前記重み行列を更新することと、
所定客観的判定基準に従って前記逆伝搬アルゴリズムを使用することにより前記第２のニューラルネットワーク及び接続行列を更新することと、をさらに行わせるように、実行可能である、請求項１２に記載の非一時的コンピュータ可読媒体。
【請求項１４】
前記話者認識特徴は少なくとも話者声紋情報を含む、請求項１０に記載の非一時的コンピュータ可読媒体。
【請求項１５】
前記話者認識特徴を含む前記ベクトルに従って前記第２のニューラルネットワーク内のバイアスを補償することは、前記話者認識特徴を含む前記ベクトルに従って前記第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償することを含み、
前記話者認識特徴を含む前記ベクトルは前記第１のニューラルネットワーク内の最後の隠れ層の出力ベクトルである、請求項１０に記載の非一時的コンピュータ可読媒体。
【請求項１６】
前記話者認識特徴を含む前記ベクトルに従って前記第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償することは、前記第１のニューラルネットワークの前記最後の隠れ層においてノードにより出力された前記話者認識特徴を含む前記ベクトルを、前記第２のニューラルネットワーク内の前記入力層を除く層の前記すべて又は一部に対応するバイアスノードへ送信することを含む、請求項１５に記載の非一時的コンピュータ可読媒体。
【請求項１７】
前記音声データは、収集された元音声データ又は前記収集された元音声データから抽出された音声特徴である、請求項１０に記載の非一時的コンピュータ可読媒体。
【請求項１８】
前記話者認識特徴は、様々なユーザ又は様々なユーザのクラスタに対応する、請求項１０に記載の非一時的コンピュータ可読媒体。
【請求項１９】
音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出するように構成された抽出ユニットと、
前記話者認識特徴を含む前記ベクトルに従って第２のニューラルネットワーク内のバイアスを補償し、前記第２のニューラルネットワークに基づく音響モデルを介し前記音声データ内の音声を認識するように構成された認識ユニットと、を含む音声認識装置。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願の相互参照
[001] 本出願は、参照のためその全体を本明細書に援用する２０１６年８月２６日申請の中国特許出願第２０１６１０７４１６２２．９号への優先権の恩恵を主張する。
【０００２】
技術分野
[002] 本出願は、音声認識に関し、より具体的には音声認識方法及び装置に関する。
【背景技術】

【０００３】
背景
[003] 現在のところ、大きな進歩が話者非依存（ＳＩ：speaker independent）音声認識システムに対しなされてきた。しかし、様々なユーザ間の差異が特定ユーザの音声認識システムの性能劣化を生じさせ得る。
【０００４】
[004] 話者依存（ＳＤ：speaker dependent）音声認識システムはＳＩ音声認識システムの性能劣化の問題を解決し得る。しかし、ＳＤ音声認識システムはトレーニングのための大量のユーザ音声データの入力を必要とし、これはユーザの大きな不都合と高コストとを生じる。
【０００５】
[005] 話者適応化（speaker adaptation）技術はＳＩ及びＳＤ音声認識システムの欠点をある程度まで補い得る。話者適応化技術により、ＳＤ音声特徴はＳＩ音声特徴へ変換され得、ＳＩ音声特徴は次に認識のためにＳＩ音響モデルへ提供される。代替的に、ＳＩ音響システムはＳＤ音響システムへ変換され得る。次に、ＳＤ音声特徴が認識される。
【０００６】
[006] ＳＩ音声認識システムと比較して、話者適応化技術は、ユーザ個人差を有する音声特徴を考慮し、したがってより良好な認識性能を有し得る。ＳＤ認識システムと比較して、話者適応化技術は、ＳＩシステムの事前情報を導入し、したがって必要とされるユーザ音声データの量は著しく低減される。
【０００７】
[007] 話者適応化技術は、ユーザ音声データが予め取得されるかどうかに依存してオフライン話者適応化技術とオンライン話者適応化技術とに分割され得る。オンライン話者適応化技術により、音声認識システムのパラメータは、現在のユーザ音声入力に従って等間隔（例えば６００ｍｓ）で調整され得、これにより話者適応化を実現する。
【０００８】
[008] 現時点で、オンライン話者適応化方法の解決策が図１に示される。この解決策は、ユーザの音声特徴とユーザに関して抽出されたｉベクトル（すなわち識別可能ベクトル）とを繋ぐことを含み得る。この解決策はまた、繋がれた特徴を音声認識のためにディープニューラルネットワーク（ＤＮＮ：deep neural network）内へ入力することを含み得る。ｉベクトルの抽出プロセスは、平均スーパーベクトルを取得するために音声の音響特性をガウス混合モデルに入力することと、ｉベクトルを取得するために平均スーパーベクトルにＴ行列を掛けることとを含み得る。ユーザが話している時、この解決策によると、ｉベクトルがユーザの音声の始めの部分から抽出され得る。抽出されたｉベクトルは、ユーザの音声の残りの音声認識のために使用され、こうしてオンライン話者適応化を実現する。
【０００９】
[009] この解決策は主として以下の問題を有する。オンライン話者適応化技術では、ｉベクトル抽出プロセスは、複雑であり、一定時間長の音声データを必要とするので、ｉベクトルを抽出するための音声データと、音声認識のための音声データとは、互いに異なる。音声認識では、ｉベクトルを抽出するための音声データは、認識されるべきそれらの音声データの予備的（preliminary）音声データである。このため、ｉベクトルは、認識される必要がある音声データと整合しなく、したがって音声認識の性能に影響を与える。
【発明の概要】
【課題を解決するための手段】
【００１０】
概要
[010] 本開示の実施形態は、余りに大きな計算複雑性を導入することなくオンライン話者適応化における音声認識の性能を効果的に改善し得る音声認識方法及び装置を提供する。
【００１１】

[011] これらの実施形態は音声認識方法を含む。本方法は、音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出することを含み得る。本方法はまた、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することを含み得る。本方法はさらに、第２のニューラルネットワークに基づく音響モデルを介し音声データ内の音声を認識することを含み得る。話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することは、話者認識特徴を含むベクトルに第２のニューラルネットワークのバイアス項となるべき重み行列を掛けることを含み得る。

【００１２】

[012] 第１のニューラルネットワーク、第２のニューラルネットワーク、及び重み行列は、第１のニューラルネットワーク及び第２のニューラルネットワークをそれぞれトレーニングし、次にトレーニングされた第１のニューラルネットワーク、重み行列、及びトレーニングされた第２のニューラルネットワークを一括してトレーニングすることによりトレーニングされ得る。

【００１３】

[013] 加えて、本方法は、第１のニューラルネットワーク、第２のニューラルネットワーク、及び重み行列を初期化することを含み得る。本方法はまた、所定客観的判定基準に従って逆伝搬（back propagation）アルゴリズムを使用することにより重み行列を更新することを含み得る。本方法はさらに、所定客観的判定基準に従って誤差逆伝搬アルゴリズムを使用することにより第２のニューラルネットワーク及び接続行列を更新することを含み得る。話者認識特徴は少なくとも話者声紋情報を含み得る。

【００１４】

[014] 話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することは、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償することを含み得る。話者認識特徴を含むベクトルは、第１のニューラルネットワーク内の最後の隠れ層の出力ベクトルであり得る。

【００１５】

[015] 話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償することは、第１のニューラルネットワークの最後の隠れ層においてニューロンノードにより出力された話者認識特徴を含むベクトルを、第２のニューラルネットワーク内の入力層を除く層のすべて又は一部に対応するバイアスノードへ送信することを含み得る。第１のニューラルネットワークは再帰型（recursive）ニューラルネットワークであり得る。音声データは、収集された元音声データ又は収集された元音声データから抽出される音声特徴であり得る。話者認識特徴は様々なユーザに対応してもよいし、様々なユーザのクラスタに対応してもよい。

【００１６】

[016] これらの実施形態はまた、音声認識方法を含む。本方法は音声データを収集することを含み得る。本方法はまた、収集された音声データを第１のニューラルネットワークに入力することにより、話者認識特徴を含むベクトルを抽出することを含み得る。本方法はさらに、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することを含み得る。加えて、本方法は収集された音声データを第２のニューラルネットワークに入力することにより音声を認識することを含み得る。

【００１７】

[017] 話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することは、話者認識特徴を含むベクトルに第２のニューラルネットワークのバイアス項となるべき重み行列を掛けることを含み得る。話者認識特徴は少なくとも話者声紋情報を含み得る。第１の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り得る。

【００１８】

[018] 話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することは、第１のニューラルネットワークの最後の隠れ層においてニューロンノードにより出力された話者認識特徴を含むベクトルを、第２のニューラルネットワーク内の入力層を除く層のすべて又は一部に対応するバイアスノードへ送信することを含み得る。

【００１９】

[019] さらに、これらの実施形態は音声認識装置を含む。音声認識装置は、音声認識のプログラムを格納するように構成されたメモリを含み得る。音声認識装置はまた、音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出するために音声認識のプログラムを実行するように構成されたプロセッサを含み得る。プロセッサはまた、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成され得る。プロセッサはさらに、音声データ内の音声を第２のニューラルネットワークに基づく音響モデルを介し認識するように構成され得る。

【００２０】

[020] 話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成されたプロセッサは、話者認識特徴を含むベクトルに、第２のニューラルネットワークのバイアス項となるべき重み行列を掛けるように構成され得ることを含み得る。話者認識特徴は少なくとも話者声紋情報を含み得る。第１の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り得る。

【００２１】

[021] 話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成されたプロセッサは、第１のニューラルネットワークの最後の隠れ層においてニューロンノードにより出力された話者認識特徴を含むベクトルを、第２のニューラルネットワーク内の入力層を除く層のすべて又は一部に対応するバイアスノードへ送信するように構成されることを含み得る。

【００２２】

[022] さらに、これらの実施形態は音声認識装置を含む。音声認識装置は、音声認識のプログラムを格納するように構成されたメモリを含み得る。音声認識装置はまた、音声データを収集するために音声認識のプログラムを実行するように構成されたプロセッサを含み得る。プロセッサはまた、収集された音声データを第１のニューラルネットワークに入力することにより、話者認識特徴を含むベクトルを抽出するように構成され得る。プロセッサはさらに、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成され得る。加えて、プロセッサは、収集された音声データを第２のニューラルネットワークに入力することにより音声を認識するように構成され得る。

【００２３】

[023] 話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成されたプロセッサは、話者認識特徴を含むベクトルに、第２のニューラルネットワークのバイアス項となるべき重み行列を掛けるように構成されることを含む。話者認識特徴は少なくとも話者声紋情報を含み得る。第１の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り得る。

【００２４】

[024] 話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成されたプロセッサは、第１のニューラルネットワークの最後の隠れ層においてニューロンノードにより出力された話者認識特徴を含むベクトルを、第２のニューラルネットワーク内の入力層を除く層のすべて又は一部に対応するバイアスノードへ送信するように構成されることを含み得る。

【００２５】

[025] これらの実施形態はまた、音声認識装置を含む。音声認識装置は、音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出するように構成された抽出ユニットを含み得る。音声認識装置はまた、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償し、第２のニューラルネットワー

クに基づく音響モデルを介し音声データ内の音声を認識するように構成された認識ユニットを含み得る。

【００２６】

[026] これらの実施形態はさらに音声認識装置を含む。音声認識装置は音声データを収集するように構成された収集ユニットを含み得る。音声認識装置はまた、収集された音声データを第１のニューラルネットワークに入力することにより話者認識特徴を含むベクトルを抽出し、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成された抽出及び補償ユニットを含み得る。音声認識装置はさらに、収集された音声データを第２のニューラルネットワークに入力することにより音声を認識するように構成された認識ユニットを含み得る。

【００２７】

図面の簡単な説明

[027] 本明細書の一部を構成する添付図面は、いくつかの実施形態を示し、開示された原理について本明細書と共に説明する役目を果たす。

【図面の簡単な説明】

【００２８】

【図１】[028]例示的ｉベクトルベースオンライン話者適応化解決策の概要図である。

【図２】[029]本開示のいくつかの実施形態による例示的音声認識方法のフローチャートである。

【図３】[030]本開示のいくつかの実施形態による音声認識のための例示的システムアーキテクチャの概要図である。

【図４】[031]本開示のいくつかの実施形態による例示的ニューラルネットワークの概要図を示す。

【図５】[032]本開示のいくつかの実施形態による例示的システムアーキテクチャの概要図である。

【図６】[033]本開示のいくつかの実施形態による例示的音声認識装置の概要図である。

【図７】[034]本開示のいくつかの実施形態による例示的音声認識方法のフローチャートである。

【図８】[035]本開示のいくつかの実施形態による音声認識方法の例示的実施形態プロセスの概要図である。

【図９】[036]本開示のいくつかの実施形態による例示的音声認識装置の概要図である。

【発明を実施するための形態】

【００２９】

詳細な説明

[037] 多くの詳細が、本開示の包括的理解を容易にするために以下の明細書に示される。本開示における方法及び装置は本明細書で説明されるものとは異なる多くの他のやり方で実現され得る。当業者は、本開示の暗示するものから逸脱することなく同様な拡張をなし得る。したがって、本開示は以下に開示される特定実施形態に限定されない。

【００３０】

[038] 本出願の技術的解決策は添付図面及び実施形態を参照して詳細に説明される。本出願の保護範囲内に入るすべての本出願の実施形態及び実施形態における様々な特徴は相反しない限り互いに組み合わせさられ得るということに注意すべきである。加えて、論理的順番がフローチャート内に示されるが、いくつかのケースでは、示される又は説明される工程は本明細書のものとは異なる順番で行われ得る。

【００３１】

[039] いくつかの実施形態では、音声認識方法を実行するコンピュータデバイスは１つ又は複数のプロセッサ（ＣＰＵ）、入出力インターフェース、ネットワークインターフェース、及びメモリを含み得る。

【００３２】

[040] メモリは、非恒久的メモリ、ランダムアクセスメモリ（ＲＡＭ：random access

memory）、及び／又はコンピュータ可読媒体内の読み取り専用メモリ（ＲＯＭ：read-only memory）又はフラッシュメモリ（フラッシュＲＡＭ）などの非揮発性メモリを含み得る。メモリはコンピュータ可読媒体の一例である。メモリはモジュール１、モジュール２、．．．、モジュールＮを含み得る、ここでＮは２より大きい整数である。

【００３３】

[041] コンピュータ可読媒体は、恒久的及び非恒久的ストレージ媒体、着脱可能及び着脱不能ストレージ媒体を含む。ストレージ媒体は、任意の方法又は技術により情報格納を実現し得る。情報はコンピュータ可読命令、データ構造、プログラムモジュール、又は他のデータであり得る。コンピュータ記憶媒体の例は、限定しないが、相転移メモリ（ＰＲＡＭ：phase change memory）、スタティックランダムアクセスメモリ（ＳＲＡＭ：static random access memory）、ダイナミックランダムアクセスメモリ（ＤＲＡＭ：dynamic random access memory）、他のタイプのランダムアクセスメモリ（ＲＡＭ）、読み取り専用メモリ（ＲＯＭ）、電気的消去可能ＰＲＯＭ（ＥＥＰＲＯＭ：electrically erasable programmable read-only memory）、フラッシュメモリ又は他のメモリ技術、コンパクトディスク読み出し専用メモリ（ＣＤ－ＲＯＭ：compact disk read-only memory）、デジタルバーサタイルディスク（ＤＶＤ：digital versatile disc）又は他の光学的ストレージ、磁気カセットテープ、磁気ディスクストレージ又は他の磁気ストレージデバイス、又はコンピュータデバイスによりアクセスされ得る情報を格納するために使用し得る任意の他の非伝送媒体を含む。本明細書で規定されるように、コンピュータ可読媒体は変調データ信号及び搬送波などの過渡的媒体を含まない。

【００３４】

[042] 本開示の実施形態は多くの利点を提供する。これらの実施形態うちののいくつかは、音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出し、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することを含む。この結果、音声認識のためのニューラルネットワークをＳＩ音響システムからＳＤ音響システムへ変換し、これにより認識性能を改善する。話者認識特徴を含むベクトルが抽出される音声データは、音声認識のための音声データと同じであるので、認識性能は著しく改善され得る。さらに、話者認識特徴を含むベクトルが話者認識のためにニューラルネットワークを介し抽出される場合、抽出はニューラルネットワークの順方向プロセスを介し実現し得る。

【００３５】

[043] 図２は本開示のいくつかの実施形態による例示的音声認識方法のフローチャートである。音声認識方法は工程Ｓ１１０、Ｓ１２０を含み得る。

【００３６】

[042] 工程Ｓ１１０では、音声データからの話者認識特徴を含むベクトルが第１のニューラルネットワークを介し抽出される。いくつかの実施形態では、工程Ｓ１１０後、本方法はさらに、話者認識特徴を含む抽出されたベクトルの長さを正規化することを含み得る。いくつかの実施形態では、話者認識特徴を含む抽出されたベクトルは、長さ正規化無しに直接使用され得る。

【００３７】

[045] 工程Ｓ１２０では、バイアスは、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内で補償され、音声データ内の音声は、第２のニューラルネットワークに基づく音響モデルを介し認識される。

【００３８】

[046] 用語「第１」及び「第２」は、例えば様々なニューラルネットワークを識別するためだけに使用される。例示的実施形態の範囲から逸脱することなく、第１のニューラルネットワークは第２のニューラルネットワークと呼ばれ得る。同様に、第２のニューラルネットワークは第１のニューラルネットワークと呼ばれ得る。

【００３９】

[047] 第１のニューラルネットワークは話者を類別するためのニューラルネットワークで

あり得る。第１のニューラルネットワークは、限定しないが話者声紋情報などの入力音声データに従って話者認識特徴を抽出し得る。第２のニューラルネットワークは音声認識のためのニューラルネットワークであり得る。第２のニューラルネットワークは入力音声データに従ってテキスト情報を認識し得る。

【００４０】

[048] 本方法は、第１のニューラルネットワーク及び第２のニューラルネットワークを含むシステムへ適用され得る。音声データは、認識のために第１のニューラルネットワーク及び第２のニューラルネットワーク内に入力され得る。本方法は、音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出することを含み得る。本方法はまた、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することを含み得る。本方法はまた、音声データからテキスト情報を取得するために、第２のニューラルネットワークに基づく音響モデルを介し音声を認識することを含み得る。

【００４１】

[049] 話者認識特徴は、話者の個人差を効果的に特徴付け得る特徴を指す。話者の個人差は声道差により引き起こされ得る。話者の個人差はまた、環境又はチャネルにより引き起こされ得る。話者適応化技術は、声道補償と、雑音環境及びオフィス環境などの様々な話す環境の補償とのために使用され得る補償技術である。話者適応化技術はまた、電話チャネル及びマイクロホンチャネルなどの様々なチャネルのための補償に使用され得る。異なる環境内で又は異なるチャネルを介して同じユーザから収集された音声データは、異なる話者認識特徴のために、異なる話者の音声データと見なされ得る。

【００４２】

[050] 本明細書の実施形態では、ニューラルネットワークは、互いに接続された複数のニューロンノードを含み得る。１つのニューロンノードの出力は別のニューロンノードの入力であり得る。ニューラルネットワークは複数のニューラル層を含み得る。それらの機能及び性質に従って、ニューラルネットワーク内のニューラル層は入力層、隠れ層、及び出力層に分割され得る。隠れ層はユーザに見えない層を指す。入力層は入力を受信し隠れ層へ分配する責任を負う。１つ又は複数の隠れ層が存在し得る。最後の隠れ層の出力結果が出力層へ提供される。ユーザは出力層から出力される最終結果を見ることができる。ニューラルネットワーク及びバイアス補償については以下に詳細に説明される。

【００４３】

[051] 本方法では、第２のニューラルネットワークに基づく音響モデルは、話者非依存であり得、バイアス補償のための話者認識特徴を含むベクトルを導入することによりＳＤ音響モデルに変換され得る。したがって、音声認識の性能が改善され得る。

【００４４】

[052] ニューラルネットワークは、多次元の音響特性を並列に計算し得る一種のグローバルモデルである。別の態様では、ｉベクトルを抽出するために使用されるガウスモデルがローカルモデルであり、次元のそれぞれを別々に計算する必要がある。したがって、第１のニューラルネットワークが本方法における話者認識特徴を含むベクトルを抽出するために使用される場合、抽出がより良好なリアルタイム性能を実現するとともに実製品において実現可能となるように短い音声データが使用され得る。

【００４５】

[053] 加えて、ガウス混合モデルを使用するシステムはニューラルネットワークベースシステムとは異なるので、これらの２つのシステムの同時最適化は容易ではないかもしれない。それにもかかわらず、本出願の実施形態では、第１のニューラルネットワーク及び第２のニューラルネットワークが全体として最適化され得る。さらに、話者認識特徴を含むベクトルが第１のニューラルネットワークを介し抽出されるプロセスは単純であり、少ない演算量を含む。同プロセスはオンライン話者適応化認識のリアルタイム要件を満たし得る。その上、短時間データが抽出のために使用され得る。

【００４６】

[054] 短期的データが抽出のために使用され得る場合、話者認識特徴を含むベクトルが抽出される音声データは、認識されるべき音声データであり得る。換言すれば、話者認識特徴を含む抽出されたベクトルは、認識されるべき音声データに良く整合し得る。したがって、音声認識の性能は著しく改善され得る。

【００４７】

[055] 図３は、本開示のいくつかの実施形態による音声認識のための例示的システムアーキテクチャの概要図である。本システムは、音声収集デバイス１１、音声認識デバイス１２、及び出力デバイス１３を含む。

【００４８】

[056] 音声認識デバイス１２は、上記工程Ｓ１１０を実行するように構成された話者認識ユニット１２１と、上記工程Ｓ１２０を実行するように構成された音声認識ユニット１２２とを含む。換言すれば、話者認識ユニット１２１は、話者認識特徴を含む抽出されたベクトルを音声認識ユニット１２２へ送信するように構成され得る。代替的に、音声認識ユニット１２２は、話者認識特徴を含むベクトルを話者認識ユニット１２１から取得するように構成され得る。

【００４９】

[057] 音声収集デバイス１１は、元音声データを収集し、元音声データ又は元音声データから抽出された音声特徴を、話者認識ユニット１２１及び音声認識ユニット１２２それぞれへ出力するように構成され得る。

【００５０】

[058] 出力デバイス１３は、音声認識ユニット１２２の認識結果を出力するように構成される。出力デバイス１３の出力方式は、限定しないが、認識結果をデータベース内に格納すること、認識結果を所定デバイスへ送信すること、又は認識結果を所定デバイス上に表示することのうちの１つ又は複数を含み得る。

【００５１】

[059] いくつかの実施形態では、音声収集デバイス１１及び音声認識デバイス１２は１つのデバイスに一体化され得る。代替的に、音声収集デバイス１１は、元音声データ又は抽出された音声特徴を接続線、無線接続などを介し音声認識デバイス１２へ送信し得る。いくつかの実施形態では、音声認識デバイス１２がネットワーク側に配置される場合、音声収集デバイス１１は、元音声データ又は抽出された音声特徴をインターネットを介し音声認識デバイス１２へ送信し得る。

【００５２】

[060] 出力デバイス１３及び音声認識デバイス１２は１つのデバイスに一体化され得る。代替的に、出力デバイス１３は、認識結果を接続線、無線接続などを介し音声認識デバイス１２から受信又は取得するように構成され得る。いくつかの実施形態では、音声認識デバイス１２がネットワーク側に配置される場合、出力デバイス１３は、認識結果をインターネットを介し音声認識デバイス１２から受信又は取得するように構成され得る。

【００５３】

[061] 音声認識デバイス１２はさらに、話者認識特徴を含むベクトルに重み行列を掛けるための計算ユニットを含み得る。話者認識特徴を含むベクトルは話者認識ユニット１２１により抽出される。音声認識デバイス１２は乗算の積を音声認識ユニット１２２へ提供するように構成され得る。代替的に、話者認識ユニット１２１又は音声認識ユニット１２２は、話者認識特徴を含むベクトルに重み行列を掛けるように構成され得る。

【００５４】

[062] 音声認識デバイス１２は、単独のデバイスでなくてもよい。例えば、話者認識ユニット１２１及び音声認識ユニット１２２は２つのデバイスに分散され得る。話者認識ユニット１２１又は音声認識ユニット１２２はまた１つ又は複数の分散デバイスにより実現され得る。

【００５５】

[063] 図４は、本開示のいくつかの実施形態による例示的ニューラルネットワークの概要

図を示す。図４に示すように、ニューラルネットワークは、入力層Ｌ１、隠れ層Ｌ２、及び出力層Ｌ３を含む。入力層Ｌ１は３つのニューロンノードＸ１、Ｘ２及びＸ３を含む。隠れ層Ｌ２は３つのニューロンノードＹ１、Ｙ２及びＹ３を含む。出力層Ｌ３は１つのニューロンノードＺを含む。図４に示すニューラルネットワークは、ニューラルネットワークの原理を単に示すために使用されており、上述の第１のニューラルネットワーク及び第２のニューラルネットワークを規定するようには意図されていない。

【００５６】

[064] 図４において、バイアスノードＢ１は、隠れ層Ｌ２に対応しており、隠れ層Ｌ２におけるバイアス補償のためのバイアス項を格納するために使用される。バイアスノードＢ１におけるバイアス項及び入力層Ｌ１内の各ニューロンノードの出力は、隠れ層Ｌ２内の各ニューロンノードの入力を提供する。バイアスノードＢ２は、出力層Ｌ３に対応しており、出力層Ｌ３におけるバイアス補償のためのバイアス項を格納するために使用される。バイアスノードＢ２におけるバイアス項と隠れ層Ｌ２における各ニューロンノードの出力は、出力層Ｌ３内の各ニューロンノードの入力を提供する。バイアス項は、事前設定され得るか、又は外部デバイスからニューラルネットワーク内へ入力され得るかのいずれかである。

【００５７】

[065] バイアス項は、バイアス補償のために使用されるベクトルを指す。ある層におけるバイアス補償は、同層のニューロンノード毎に、同層に対応するバイアスノードにより提供されるバイアス項内の、ニューロンノードに対応する値を加えた、前の層のすべてのニューロンノードの出力値の加重和の結果に基づく計算を指す。

【００５８】

[066] 例えば、入力層Ｌ１内のニューロンノードＸ１、Ｘ２及びＸ３の出力値がそれぞれｘ１、ｘ２及びｘ３であると仮定する。隠れ層Ｌ２内のニューロンノードＹ１に関しては、出力値は次のようになる：

【数１】

$$\mathrm{f}\left(\mathrm{W}_{11}^{L1}\,x1+\mathrm{W}_{12}^{L1}\,x2+\mathrm{W}_{13}^{L1}\,x3+\mathrm{b}_{1}^{B1}\right)$$

ここで、ｆは、括弧内の内容に対しニューロンノードＹ１によりなされた計算を表し、括弧内の内容は、ニューロンノードＹ１により受信された入力値を表す。

【数２】

$$W_{ij}^{L1}$$

は、例えばＹ１、ｉ＝１、ｊ＝１、２、３に関しては、層Ｌ１内の第ｊニューロンノードと次層（すなわち層Ｌ２）内の第ｉニューロンノード間の重み付けを指し、

【数３】

$$b_{s}^{B1}$$

は、隠れ層Ｌ２内のｓ番目ニューロンノード（ｓ＝１、２、３）に対応するバイアスノードＢ１内のバイアス項の値を指し、例えば、ニューロンノードＹ１に対応するバイアスノードＢ１内のバイアス項の値は

【数４】

$$b_{s}^{B1}$$

である。

【００５９】

[067] 図２に戻って参照すると、工程Ｓ１２０において、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することは、話者認識特徴を含むベクトルを線型変換し、これを第２のニューラルネットワークにおける入力層以外のある層又はいくつかの層のバイアス項として採用することを指し得る。線形変換は、限定しないが重み行列による乗算のやり方で行われ得る。

【００６０】

[068] 第１のニューラルネットワークは３つの隠れ層を含み得る。いくつかの実施形態では、第１のニューラルネットワークは１つ又は２つの隠れ層を含んでもよいし、４以上の隠れ層を含んでもよい。いくつかの実施形態では、第２のニューラルネットワークは３つの隠れ層を含み得る。いくつかの実施形態では、第２のニューラルネットワークは１つ又は２つの隠れ層を含んでもよいし、４以上の隠れ層を含んでもよい。

【００６１】

[069] いくつかの実施形態では、話者認識特徴は少なくとも話者声紋情報を含み得る。話者声紋情報は異なるユーザの音声データを識別するために使用され得る。換言すれば、異なるユーザの音声データから抽出される話者声紋情報は異なる。いくつかの実施形態では、話者認識特徴は、話者声紋情報、環境情報、及びチャネル情報のうちの１つ又は複数を含み得る。環境情報は、音声データが収集される環境の特徴を特徴付けるために使用され得る。チャネル情報は、音声データが収集されるチャネルの特徴を特徴付けるために使用され得る。

【００６２】

[070] いくつかの実施形態では、第１の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り得る。再帰型ニューラルネットワークは、１つ又は複数のフィードバックループを有するニューラルネットワークを指し、非線形システムのリアルな動的モデリングを実現し得る。再帰型ニューラルネットワークが、話者認識特徴を含むベクトルを抽出するために使用される場合、抽出は短期データに対し行われ得る。再帰型ニューラルネットワークは、限定しないが、ＬＳＴＭ（long-short term memory）再帰型ニューラルネットワークであり得る。

【００６３】

[071] いくつかの実施形態では、工程Ｓ１２０における話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することは、話者認識特徴を含むベクトルに、第２のニューラルネットワークのバイアス項となるべき重み行列を掛けることを含み得る。

【００６４】

[072] いくつかの実施形態では、重み行列が単位行列である場合、重み行列が掛けられた後、話者認識特徴を含むベクトルは変化しなくてもよい。話者認識特徴を含むベクトルは、第２のニューラルネットワークのバイアス項として直接採用され得る。

【００６５】

[073] いくつかの実施形態では、第１のニューラルネットワーク、第２のニューラルネットワーク、及び重み行列は、第１のニューラルネットワーク及び第２のニューラルネットワークをそれぞれトレーニングし、次にトレーニングされた第１のニューラルネットワーク、重み行列、及びトレーニングされた第２のニューラルネットワークを一括してトレーニングすることによりトレーニングされ得る。一括的にトレーニングすることは、トレーニングするための音声データを、第１のニューラルネットワーク及び第２のニューラルネットワーク内にそれぞれ入力することと、第１のニューラルネットワークにより抽出された話者認識特徴を含むベクトルに重み行列を掛けた後に、第２のニューラルネットワーク上のバイアスを補償することとを指し得る。トレーニングは、限定しないがグラフィック処理ユニット（ＧＰＵ：graphics processing unit）により行われ得る。

【００６６】

[074] いくつかの実施形態では、トレーニングされた第１のニューラルネットワーク、重み行列、及びトレーニングされた第２のニューラルネットワークを一括してトレーニング

した後、本方法はさらに、第１のニューラルネットワーク、第２のニューラルネットワーク、及び重み行列を初期化することを含み得る。本方法はまた、所定客観的判定基準に従って逆伝搬アルゴリズムを使用することにより重み行列を更新することを含み得る。加えて、本方法は、所定客観的判定基準に従って誤差逆伝搬アルゴリズムを使用することにより第２のニューラルネットワーク及び接続行列を更新することを含み得る。重み行列に関する初期化は、ガウス分布による乱数的初期化であり得る。上記所定客観的判定基準は、限定しないが、標的最小二乗平均誤差（ＬＭＳ）、再帰型最小二乗（ＲＬＳ：recursive least square）、及び正規化最小二乗平均誤差（ＮＬＭＳ：normalized least mean square error）を含み得る。

【００６７】

[075] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することは、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償することを含む。話者認識特徴を含むベクトルは、第１のニューラルネットワーク内の最後の隠れ層の出力ベクトルであり得る。例えば、第２のニューラルネットワークが入力層、３つの隠れ層、及び１つの出力層を含むと仮定すると、入力層を除くすべての層は、出力層及び３つの隠れ層を指し得る。入力層を除くいくつかの層は、４つの層（すなわち、出力層及び３つの隠れ層）のうちの１つ又は複数を指し得る。

【００６８】

[076] 話者認識特徴を含むベクトルに基づく第２のニューラルネットワーク内のある層上のバイアス補償は、話者認識特徴を含むベクトルに重み行列を掛けることにより取得されるベクトルを、同層のバイアス項として採用することを指し得る。例えば、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内の入力層以外のすべての層上のバイアス補償は、話者認識特徴を含むベクトルに重み行列を掛けることにより取得されたベクトルを、第２のニューラルネットワーク内の出力層及び３つの隠れ層のそれぞれのバイアス項として採用することを指し得る。

【００６９】

[077] いくつかの実施形態では、話者認識特徴を含むベクトルは、第１のニューラルネットワーク内の最後の隠れ層の出力ベクトルであり得る。最後の隠れ層の出力ベクトルは出力層の出力ベクトルより少ない次元を有し、これにより過剰フィッティング（overfitting）を回避する。

【００７０】

[078] いくつかの実施形態では、話者認識特徴を含むベクトルは、第１のニューラルネットワーク内の最後の隠れ層以外の隠れ層の出力ベクトルであってもよいし、出力層の出力ベクトルであってもよい。

【００７１】

[079] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償することは、第１のニューラルネットワークの最後の隠れ層においてニューロンノードにより出力された話者認識特徴を含むベクトルを、第２のニューラルネットワーク内の入力層を除く層のすべて又は一部に対応するバイアスノードへ送信することを含み得る。ある層に対応するバイアスノードは、同層の上のバイアス補償のために使用されるバイアス項を格納し得る。話者認識特徴を含むベクトルは、第１のニューラルネットワーク内の最後の隠れ層内の複数のニューロンノードのそれぞれの出力値からなるベクトルであり得る。

【００７２】

[080] 話者認識特徴を含むベクトルをバイアスノードへ送信することは、話者認識特徴を含むベクトルをバイアスノードへ直接送信することを指してもよいし、話者認識特徴を含むベクトルを線型変換し、次にこれをバイアスノードへ送信することを指してもよい。

【００７３】

[081] 第２のニューラルネットワーク内の複数層が同じベクトルによりバイアス補償を受

ける場合、複数層は個別のバイアスノードに対応してもよいし、同じバイアスノードに対応してもよい。例えば、第１のニューラルネットワークにより抽出された話者認識特徴を含むベクトルは、複数のバイアスノードそれぞれへ送信され得る。複数のバイアスノードは、第２のニューラルネットワーク内のバイアス補償を必要とする複数の層に１対１ベースで対応する。別の例として、第１のニューラルネットワークにより抽出された話者認識特徴を含むベクトルはまた、１つのバイアスノードへ送信され得る。当該バイアスノードは、第２のニューラルネットワーク内のバイアス補償を必要とする複数の層に対応する。

【００７４】

[082] いくつかの実施形態では、音声データは、収集された元音声データであるか又は収集された元音声データから抽出された音声特徴である。音声特徴は、限定しないが、メル周波数ケプストラム係数（ＭＦＣＣ：Mel frequency cepstral coefficient）、知覚線形予測係数（ＰＬＰ：perceived linear prediction coefficient）、フィルタバンク特徴、又はそれらの任意の組み合わせを含み得る。

【００７５】

[083] いくつかの実施形態では、話者認識特徴は、様々なユーザに１対１ベースで対応してもよいし、様々なユーザのクラスタに１対１ベースで対応してもよい。様々なユーザに１対１ベースで対応する話者認識特徴は、第１のニューラルネットワークの出力層がユーザの識別子を出力することを意味する。様々なユーザのクラスタに１対１ベースで対応する話者認識特徴は、ユーザがクラスタ化された後、第１のニューラルネットワークの出力層がカテゴリ識別子を出力することを意味する。

【００７６】

[084] クラスタは１つ又は複数のパターンを含み得、パターンは測定のベクトルを指してもよいし、多次元空間内の点であってもよい。クラスタ化操作は類似性に基づいており、同じクラスタ内のパターンは、異なるクラスタ内のパターンより高い類似性を有する。クラスタ化のためのアルゴリズムは分割方法、階層的方法、密度アルゴリズム、グラフ理論クラスタリング法、グリッドアルゴリズム、及びモデルアルゴリズムに分割され得る。例えば、これらのアルゴリズムはＫ平均法（K-MEANS）、Ｋ－ＭＥＤＯＩＤＳ、Ｃｌａｒａ又はＣｌａｒａｎｓであり得る。

【００７７】

[085] ユーザをクラスタ化することは、トレーニング中に様々なユーザの話者認識特徴間の類似性に従って複数のユーザの話者認識特徴を複数のクラスタに類別することと、クラスタに対応する話者認識特徴を含むベクトルを取得するために、１つのクラスタに分類された複数の話者認識特徴を計算（例えば重み付け平均化）することとを指し得る。カテゴリ識別子は、１つのクラスタを表すために使用される識別子であり得る。カテゴリ識別子はクラスタに１対１ベースで対応する。

【００７８】

[086] 非常に多くのユーザの音声認識が必要とされる場合、クラスタ化操作が行われれば、一組の出力結果は、話者認識特徴を含む限定数のベクトルであり得る。例えば、何百万のユーザが存在する場合、ユーザが何千ものクラスタに分類されれば、話者認識特徴を含む何千ものベクトルだけが存在し、これにより実装の複雑性を著しく低減する。

【００７９】

[087] 話者認識特徴が、話者認識特徴間の類似性に従って複数のクラスタに分類される場合、様々な次元の類似性（例えば、声紋情報、環境情報、チャネル情報などの様々なタイプの話者認識特徴）に従って、様々なクラスタ化結果が取得され得る。例えば、同様な声紋を有する話者認識特徴は１つのクラスタと見なされ得る。別の例として、同じ又は同様な環境に対応する話者認識特徴が１つのクラスタと見なされ得る。代替的に、同様なチャネルに対応する話者認識特徴が１つのクラスタと見なされ得る。

【００８０】

[088] 図５は、本開示のいくつかの実施形態による音声認識のための例示的システムアーキテクチャの概要図である。図５に示すように、本システムは話者分類器２１と、音声認

識システム２３とを含み得る。本システムにおける話者認識特徴は話者声紋情報である。話者分類器２１は上記工程Ｓ１１０を実行するように構成される。音声認識システム２３は上記工程Ｓ１２０を実行するように構成される。

【００８１】

[089] 話者声紋情報を含むベクトルが接続行列２２により線形に変換され得る。接続行列は限定しないが重み行列であり得る。

【００８２】

[090] 話者認識特徴を含むベクトルを抽出するための第１のニューラルネットワークを利用する話者分類器２１は、入力層２１１、１つ又は複数の隠れ層２１２、及び出力層２１３を含み得る。いくつかの実施形態では、隠れ層２１２の数は３であり得る。代替的に、１つ又は複数の隠れ層２１２が存在し得る。

【００８３】

[091] 音声を認識するための第２のニューラルネットワークを利用する音声認識システム２３は、入力層２３１、１つ又は複数の隠れ層２３２、及び出力層２３３を含み得る。いくつかの実施形態では、隠れ層２１２の数は３であり得る。いくつかの実施形態では、１つ又は複数の隠れ層２１２が存在し得る。

【００８４】

[092] 話者分類器２１内の第１のニューラルネットワークの入力層２１１により受信される音声データは、音声認識システム２３内の第２のニューラルネットワークの入力層２３１により受信されるものと同じであり得る。音声データは、収集された元音声データであり得る。代替的に、音声データは、元音声データから抽出された音声特徴であり得る。

【００８５】

[093] したがって、話者分類器２１内の第１のニューラルネットワークは、音声認識システム２３内の第２のニューラルネットワークと同じ入力を有し得る。すなわち、話者声紋情報を含むベクトルが取得される音声データは、音声認識のための音声データと同じであり得る。したがって、話者声紋情報を含むベクトルによる第２のニューラルネットワーク上のバイアス補償は、認識されるべき音声データと完全に整合し得る。その結果、音声認識の性能は効果的に改善され得る。第１のニューラルネットワーク及び第２のニューラルネットワークはそれぞれ、全結合ニューラルネットワーク（ＤＮＮ：fully connected neural network）、畳み込みニューラルネットワーク（ＣＮＮ：convolution neural network）、及び再帰型ニューラルネットワーク（ＲＮＮ：recurrent neural network）の任意の１つ、又はそのいくつかの組み合わせを含み得る。

【００８６】

[094] 話者声紋情報を含むベクトル表現は、話者分類器２１内の最後の隠れ層の出力ベクトルであり得る。

【００８７】

[095] 音声認識システム２３において、出力層２３３及び１つ又は複数の隠れ層２３２のそれぞれは、話者声紋情報を含む線形変換されたベクトル表現をバイアス項として採用し得る。いくつかの実施形態では、出力層２３３及び１つ又は複数の隠れ層２３２において、少なくとも１つ又は複数の層は、話者声紋情報を含む線型変換されたベクトル表現をバイアス項として採用し得る。

【００８８】

[096] 接続行列２２はまた、話者声紋情報を含むベクトルに対し長さ正規化を行うように構成され得る。いくつかの実施形態では、話者分類器により出力された話者声紋情報を含むベクトルは、長さ正規化を受けることなく、重み付けを掛けられた後、音声認識システム２３へ直接提供され得る。

【００８９】

[097] 話者分類器２１の出力層２１３によるデータ出力は、様々なユーザのタグＩＤであってもよいし、ユーザがクラスタ化された後のクラスタのタグＩＤであってもよい。出力層の出力データはトレーニングのためにだけ使用され得る。音声認識システム２３の出力

層２３３から出力される認識結果は、状態レベル、音素レベル、又は単語レベルタグＩＤであり得る。

【００９０】

[098] 図５に示す例示的システムアーキテクチャはさらに、以下の機能を実行し得る。

【００９１】

[099] トレーニングデータを使用することにより、話者分類器の第１のニューラルネットワークと、第２のニューラルネットワークに基づく音響モデル（例えば図２において参照される音響モデル）とをトレーニングすること。第１及び第２のニューラルネットワークは所望の音声認識性能又は話者認識性能をそれぞれ実現し得る。さらに、このトレーニングは、第１のニューラルネットワーク、接続行列、及び第２のニューラルネットワークを一括してトレーニングすることを含み得る。ＧＰＵがこれらのトレーニングを加速するために使用され得る。

【００９２】

[0100] システムアーキテクチャは、ネットワーク初期化のための音響モデル及び話者分類器として、トレーニングされた音響モデル及び話者分類器を使用し得る。いくつかの実施形態では、ネットワーク初期化はまた図５の接続行列をランダムに初期化することを含み得る。

【００９３】

[0101] 所定客観的判定基準に従って、システムアーキテクチャは、収束状態に到達するように接続行列を更新するために、逆伝播（ＢＰ）アルゴリズムを使用し得る。

【００９４】

[0102] 所定客観的判定基準に従って、システムアーキテクチャは、収束状態に到達するように音響モデル及び接続行列を更新するためにＢＰアルゴリズムを使用し得る。所定客観的判定基準は、実アプリケーションにおけるニーズに従って設定され得る。

【００９５】

[0103] さらに、システムアーキテクチャは、収集された元音声データから音声特徴を抽出し得る。抽出された音声特徴は話者分類器により処理され、その結果、音声特徴に対応する話者声紋情報を含むベクトルが取得される。このベクトルは接続行列により線形変換され、音声認識システムへ送信される。抽出された音声特徴は、音声認識システムにおける第２内のニューラルネットワークに基づき音響モデルにより復号化される。最終的に、音声認識結果が取得され得る。音声認識システムでは、第２のニューラルネットワークの出力層及び３つの隠れ層のバイアス項は、話者声紋情報を含む形型変換されたベクトルであり得る。

【００９６】

[0104] 本出願はまた、音声認識装置に関する。音声認識装置は、音声認識のプログラムを格納するように構成されたメモリを含む。音声認識装置はまた、音声認識のプログラムを実行するように構成されたプロセッサを含む。プロセッサは、音声認識のプログラムを実行する際、第１のニューラルネットワークを介し、音声データから話者認識特徴を含むベクトルを抽出するように構成され得る。

【００９７】

[0105] プロセッサはまた、音声認識のプログラムを実行する際、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成され得る。プロセッサはさらに、音声認識のプログラムを実行する際、第２のニューラルネットワークに基づく音響モデルを介し音声データ内の音声を認識するように構成される。

【００９８】

[0106] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成されたプロセッサは、話者認識特徴を含むベクトルに、第２のニューラルネットワークのバイアス項となるべき重み行列を掛けるように構成されることを含み得る。

【００９９】

[0107] いくつかの実施形態では、話者認識特徴は少なくとも話者声紋情報を含み得る。
【０１００】
[0108] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成されたプロセッサは、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償するように構成されることを含み得る。話者認識特徴を含むベクトルは、第１のニューラルネットワーク内の最後の隠れ層の出力ベクトルであり得る。
【０１０１】
[0109] いくつかの実施形態では、話者認識特徴は、様々なユーザに１対１ベースで対応してもよいし、様々なユーザのクラスタに１対１ベースで対応してもよい。様々なユーザに１対１ベースで対応する話者認識特徴は、第１のニューラルネットワークの出力層がユーザの識別子を出力することを意味する。様々なユーザのクラスタに１対１ベースで対応する話者認識特徴は、ユーザがクラスタ化された後に、第１のニューラルネットワークの出力層がカテゴリ識別子を出力することを意味する。
【０１０２】
[0110] いくつかの実施形態では、第１の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り得る。
【０１０３】
[0111] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成されたプロセッサは、第１のニューラルネットワークの最後の隠れ層においてニューロンノードにより出力された話者認識特徴を含むベクトルを、第２のニューラルネットワーク内の入力層を除く層のすべて又は一部に対応するバイアスノードへ送信するように構成されることを含む。
【０１０４】
[0112] さらに、プロセッサは、音声認識のプログラムを実行する際、上記工程Ｓ１１０、Ｓ１２０を実行するように構成され得る。音声認識のプログラムを実行する際にプロセッサにより実行される動作のさらなる詳細は上に見出され得る。
【０１０５】
[0113] 本出願はさらに音声認識装置に関する。図６は、本開示のいくつかの実施形態による例示的音声認識装置の概要図である。音声認識装置は、抽出ユニット３１及び認識ユニット３２を含む。
【０１０６】
[0114] 一般的に、これらのユニット（そして任意の副ユニット）は、他の部品（例えば集積回路の一部）と共に及び／又は関連機能の特定機能を実行するプログラム（コンピュータ可読媒体上に格納された）の一部と共に使用するために設計されたパッケージ化機能ハードウェアユニットであり得る。このユニットは入口点及び出口点を有し得、例えばJava（登録商標）、Lua、C、又はC++などのプログラミング言語で書かれ得る。ソフトウェアユニットは、コンパイルされ、実行可能プログラム内へリンクされ、動的リンクライブラリ内にインストールされてもよいし、例えばBASIC、Perl、又はPythonなどのインタープリート型プログラミング言語で書かれてもよい。ソフトウェアユニットは他のユニット又は自身から呼出し可能であり得る及び／又は検出された事象又は割り込みに応答して呼び出され得るということが理解される。コンピュータデバイス上で実行するように構成されたソフトウェアユニットは、コンパクトディスク、ディジタルビデオディスク、フラッシュドライブ、磁気ディスク、又は任意の他の非一時的媒体などのコンピュータ可読媒体上に提供されてもよいし、ディジタルダウンロードとして提供されてもよい（そして実行に先立って、インストール、圧縮解除、又は解読を必要とする圧縮された又はインストール可能なフォーマットで元々格納され得る）。このようなソフトウェアコードは、コンピュータデバイスによる実行のために実行コンピュータデバイスのメモリデバイス上に部分的又は完全に格納され得る。ソフトウェア命令はＥＰＲＯＭなどのファームウェアで埋め込まれ得る。ハードウェアユニットはゲート及びフリップフロップなどの接続された論

理ユニットで構成され得る及び／又はプログラマブルゲートアレイ又はプロセッサなどのプログラム可能ユニットで構成され得るということがさらに理解される。本明細書で説明されたユニット又はコンピュータデバイス機能は好適にはソフトウェアユニットとして実現されるが、ハードウェア又はファームウェアで表され得る。一般的に、本明細書で説明されたユニットは、他のユニットと組み合わせられ得る又は物理的編成又はストレージにもかかわらず副ユニットに分割され得る論理ユニットを指す。

【０１０７】

[0115] 抽出ユニット３１は、音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出するように構成され得る。抽出ユニット３１は、上記装置内の話者認識特徴を含むベクトルを抽出するためのものと同様な動作を実行するように構成され得る。

【０１０８】

[0116] 認識ユニット３２は、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償し、第２のニューラルネットワークに基づく音響モデルを介し音声データ内の音声を認識するように構成され得る。認識ユニット３２は上記装置内の音声を認識するためのものと同様な動作を実行するように構成され得る。

【０１０９】

[0117] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成された認識ユニット３２は、話者認識特徴を含むベクトルに、第２のニューラルネットワークのバイアス項となるべき重み行列を掛けるように構成されることを含み得る。

【０１１０】

[0118] いくつかの実施形態では、話者認識特徴は少なくとも話者声紋情報を含み得る。

【０１１１】

[0119] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成された認識ユニット３２は、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内の入力層を除く層のすべて又は一部においてバイアスを補償するように構成されることを含み得る。代替的に、話者認識特徴を含むベクトルは、第１のニューラルネットワーク内の最後の隠れ層の出力ベクトルであり得る。

【０１１２】

[0120] いくつかの実施形態では、第１の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り得る。

【０１１３】

[0121] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成された認識ユニット３２は、第１のニューラルネットワークの最後の隠れ層においてニューロンノードにより出力された話者認識特徴を含むベクトルを、第２のニューラルネットワーク内の入力層を除く層のすべて又は一部に対応するバイアスノードへ送信するように構成されることを含み得る。

【０１１４】

[0122] いくつかの実施形態では、抽出ユニット３１は、図３に示すシステムアーキテクチャ内の話者認識ユニット１２１として構成され得る。認識ユニット３２は、図３に示すシステムアーキテクチャ内の音声認識ユニット１２２として構成され得る。図６の装置は、図３に示すシステムアーキテクチャの音声認識装置として構成され得る。図６の装置のさらなる詳細動作は音声認識図３に示す装置に関して上に説明したものが参照され得る。

【０１１５】

[0123] さらに、抽出ユニット３１及び認識ユニット３２により実行される動作は、上記音声認識方法における工程Ｓ１１０、Ｓ１２０と同様であり得る。抽出ユニット３１及び認識ユニット３２により実行される動作のさらなる詳細も上に見出され得る。

【０１１６】

[0124] 本出願はまた音声認識方法に向けられる。図７は、本開示のいくつかの実施形態による例示的音声認識方法のフローチャートである。この方法は、図３の音声認識デバイス及び／又は図５のシステムアーキテクチャにより行われ得る。図７に示すように、音声認識方法は、以下の工程Ｓ４１０、Ｓ４２０、及び工程Ｓ４３０を含む。

【０１１７】

[0125] 工程Ｓ４１０では、システムアーキテクチャは音声データを収集する。

【０１１８】

[0126] 工程Ｓ４２０では、システムアーキテクチャは、収集された音声データを第１のニューラルネットワークに入力することにより、話者認識特徴を含むベクトルを抽出し、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償する。

【０１１９】

[0127] 工程Ｓ４３０では、システムアーキテクチャは、収集された音声データを第２のニューラルネットワークに入力することにより音声を認識する。

【０１２０】

[0128] 工程Ｓ４１０、Ｓ４２０及び工程Ｓ４３０は、収集プロセス中に連続的に行われ得る。一団の音声データが収集されるたびに、工程Ｓ４２０、Ｓ４３０が、その一団の音声データの音声認識の結果を取得するために、当該一団の音声データに対し行われ得る。一団の音声データのサイズは限定しないが１つ又は複数のフレームであり得る。

【０１２１】

[0129] 図８は、本開示のいくつかの実施形態による図７の音声認識方法の例示的実施形態プロセスの概要図である。

【０１２２】

[0130] この実施形態プロセスはユーザの音声を収集することを含む。この実施形態プロセスはまた、収集された音声データ又はそれから抽出された音声特徴を、第１のニューラルネットワーク及び第２のニューラルネットワークへ直接入力することを含む。この実施形態プロセスはさらに、話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出し、このベクトルを第２のニューラルネットワークへバイアス項として送信することを含む。この実施形態プロセスはまた、第２のニューラルネットワークから音声データの認識結果を出力することを含む。

【０１２３】

[0131] 収集された元音声データは、第１のニューラルネットワーク及び第２のニューラルネットワークへ直接提供され得る。代替的に、音声特徴は、収集された元音声データから抽出され得、抽出された音声特徴は次に第１のニューラルネットワーク及び第２のニューラルネットワークへ提供される。

【０１２４】

[0132] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することは、話者認識特徴を含むベクトルに、第２のニューラルネットワークのバイアス項となるべきに重み行列を掛けることを含み得る。

【０１２５】

[0133] いくつかの実施形態では、話者認識特徴は少なくとも話者声紋情報を含み得る。

【０１２６】

[0134] いくつかの実施形態では、第１の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り得る。

【０１２７】

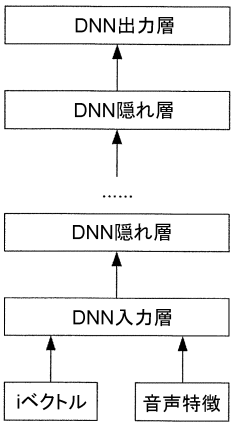
[0135] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償することは、第１のニューラルネットワークの最後の隠れ層においてニューロンノードにより出力された話者認識特徴を含むベクトルを、第２のニューラルネットワーク内の入力層を除く層のすべて又は一部に対応するバイアスノードへ送信することを含む。

【０１２８】
[0136] 第１のニューラルネットワーク、第２のニューラルネットワーク、話者認識特徴を含むベクトルの抽出、話者認識特徴を含むベクトルに従って第２のニューラルネットワークにおけるバイアス補償、及び第２のニューラルネットワークに基づく音声認識のさらなる詳細は、音声認識方法について上で説明したものと同様である。
【０１２９】
[0137] 本出願に開示された実施形態はさらに音声認識装置に関する。音声認識装置は音声認識のプログラムを格納するように構成されたメモリを含む。音声認識装置はまた、音声データから話者認識特徴を含むベクトルを第１のニューラルネットワークを介し抽出するために、音声認識のプログラムを実行するように構成されたプロセッサを含む。プロセッサはまた、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するために、音声認識のプログラムを実行するように構成される。プロセッサはさらに、音声データ内の音声を第２のニューラルネットワークに基づく音響モデルを介し認識するために、音声認識のプログラムを実行するように構成される。
【０１３０】
[0138] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するために音声認識のプログラムを実行するように構成されたプロセッサは、話者認識特徴を含むベクトルに、第２のニューラルネットワークのバイアス項となるべき重み行列を掛けるように構成されることを含み得る。
【０１３１】
[0139] いくつかの実施形態では、話者認識特徴は少なくとも話者声紋情報を含み得る。
【０１３２】
[0140] いくつかの実施形態では、第１の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り得る。
【０１３３】
[0141] いくつかの実施形態では、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するために音声認識のプログラムを実行するように構成されたプロセッサは、第１のニューラルネットワークの最後の隠れ層においてニューロンノードにより出力された話者認識特徴を含むベクトルを、第２のニューラルネットワーク内の入力層を除く層のすべて又は一部に対応するバイアスノードへ送信するように構成されることを含み得る。
【０１３４】
[0142] プロセッサが音声認識のプログラムを読み実行するように構成される場合、音声データを収集するように構成されたプロセッサは、図３の音声収集装置の動作を参照し得る。動作のさらなる詳細は図３の説明に見出され得る。話者認識特徴を含むベクトルを抽出すること、話者認識特徴を含むベクトルに従って第２のニューラルネットワークにおいてバイアスを補償すること、及び第２のニューラルネットワークに基づき音声を認識することのさらなる動作詳細もまた、本音声認識方法における上記それらの説明を参照することができる。
【０１３５】
[0143] 本出願はさらに音声認識装置に関する。図９は、本開示のいくつかの実施形態による例示的音声認識装置の概要図である。図９に示すように、音声認識装置は、音声データを収集するように構成された収集ユニット６１を含む。音声認識装置はまた、収集された音声データを第１のニューラルネットワークに入力することにより、話者認識特徴を含むベクトルを抽出し、話者認識特徴を含むベクトルに従って第２のニューラルネットワーク内のバイアスを補償するように構成された抽出及び補償ユニット６２を含む。音声認識装置はさらに、収集された音声データを第２のニューラルネットワークに入力することにより音声を認識するように構成された認識ユニット６３を含む。これらのユニット（そして任意の副ユニット）は、他の部品（例えば集積回路の一部）と共に及び／又は関連機能の特定機能を実行するプログラム（コンピュータ可読媒体上に格納された）の一部と共に

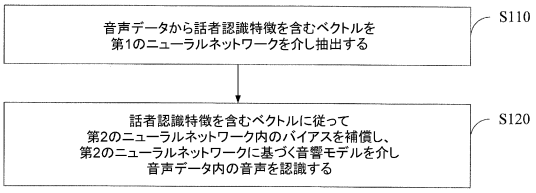
使用するために設計されたパッケージ化機能ハードウェアユニットであり得る。
【０１３６】
[0144] 収集ユニット６１は、上記装置内の音声データを収集するためのものと同様な動作を実行するように構成され得る。
【０１３７】
[0145] 抽出及び補償ユニット６２は、上記装置内の話者認識特徴を含むベクトルを抽出するためのもの、及び第２のニューラルネットワークにおいてバイアスを補償するためのものと同様な動作を実行するように構成され得る。
【０１３８】
[0146] 認識ユニット６３は、音声を認識するためのものと同様な動作を実行するように構成され得る。
【０１３９】
[0147] 収集ユニット６１は、独立したデバイス内に装備され得る。代替的に、収集ユニット６１は、抽出及び補償ユニット６２、認識ユニット６３と共に同じデバイス内に装備され得る。
【０１４０】
[0148] 収集ユニット６１は、図３に示す音声収集装置を参照して実現され得る。抽出及び補償ユニット６２並びに認識ユニット６３による、第１のニューラルネットワークによる話者認識特徴を含むベクトルを抽出すること、話者認識特徴を含むベクトルに従って第２のニューラルネットワークにおいてバイアスを補償すること、及び音声を認識することのさらなる実施詳細は、上記音声認識方法におけるそれらの説明として参照され得る。
【０１４１】
[0149] 上に示すように、上記方法のすべて又はいくつかの工程は、プログラムを介した命令下で当該ハードウェアにより完了され得るということが理解される。プログラムは、読み取り専用メモリ、磁気ディスク又はコンパクトディスクなどのコンピュータ可読記憶媒体内に格納され得る。任意選択的に、上述の実施形態のすべて又はいくつかの工程はまた、１つ又は複数の集積回路を使用して実現され得る。したがって、上記実施形態における様々なモジュール／ユニットはハードウェアの形式で実装されてもよいし、ソフトウェア機能モジュールの形式で実装されてもよい。本出願はハードウェア及びソフトウェアの組み合わせのいかなる特定形式にも限定されない。
【０１４２】
[0150] 確かに、本出願の様々な他の実施形態が存在し得る。当業者は本出願の精神及び本質から逸脱することなく本出願に従って様々な変更及び変形をなすことができるだろう。すべてのこれらの対応する変更及び変形はすべて本出願の特許請求の範囲に入るべきである。

【図面】

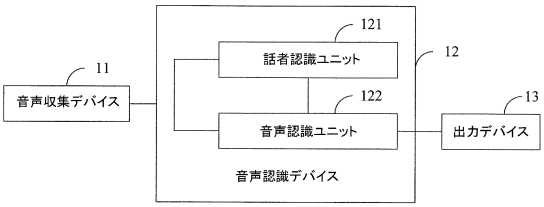
【図１】



【図２】



【図３】



【図４】

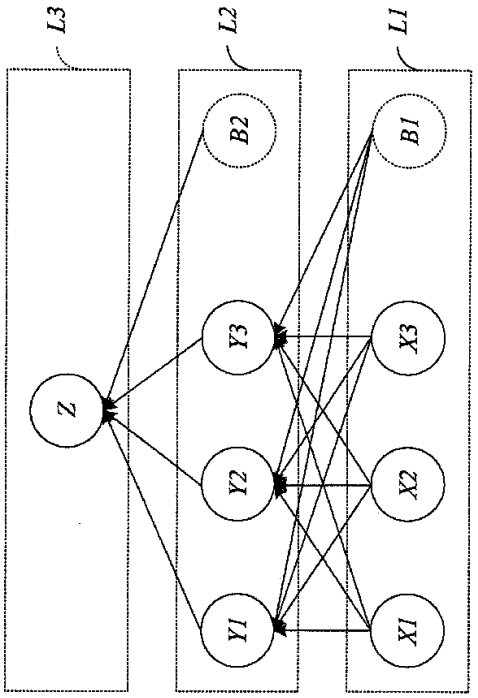
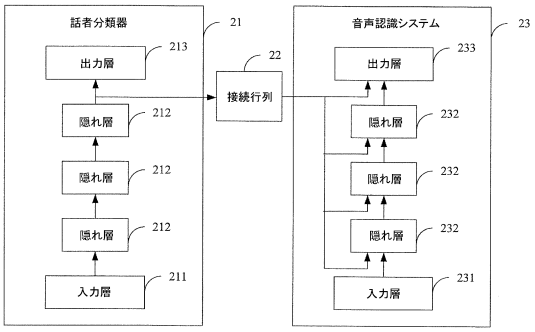
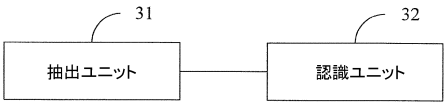


Fig. 4

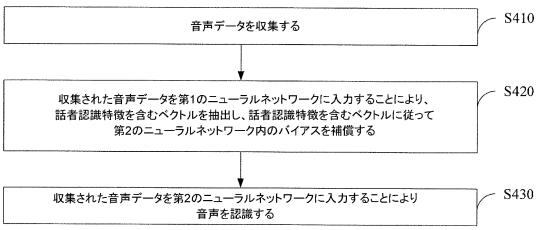
【図５】


話者分類器
21
出力層
213
22
接続行列
隠れ層
212
隠れ層
212
隠れ層
212
入力層
211
音声認識システム
23
出力層
233
隠れ層
232
隠れ層
232
隠れ層
232
入力層
231

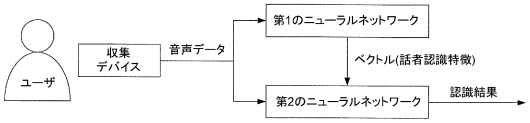
【図６】


31
抽出ユニット
32
認識ユニット

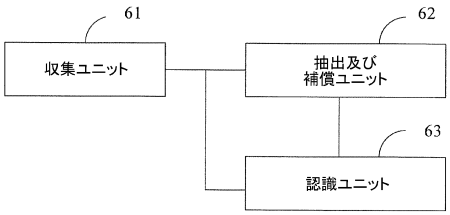
【図７】


音声データを収集する
S410
収集された音声データを第1のニューラルネットワークに入力することにより、
話者認識特徴を含むベクトルを抽出し、話者認識特徴を含むベクトルに従って
第2のニューラルネットワーク内のバイアスを補償する
S420
収集された音声データを第2のニューラルネットワークに入力することにより
音声を認識する
S430

【図８】


ユーザ
収集
デバイス
音声データ
第1のニューラルネットワーク
ベクトル(話者認識特徴)
第2のニューラルネットワーク
認識結果

【図９】



フロントページの続き

(72)発明者　ファン，ジーイン
中華人民共和国，３１１１２１，ハンチョウ，ユ ハン ディストリクト，ウェスト ウェン イ ロード ナンバー ９６９，ビルディング ３，５／エフ，アリババ グループ リーガル デパートメント
(72)発明者　シュエ，シャオフェイ
中華人民共和国，３１１１２１，ハンチョウ，ユ ハン ディストリクト，ウェスト ウェン イ ロード ナンバー ９６９，ビルディング ３，５／エフ，アリババ グループ リーガル デパートメント
(72)発明者　ヤン，ジージエ
中華人民共和国，３１１１２１，ハンチョウ，ユ ハン ディストリクト，ウェスト ウェン イ ロード ナンバー ９６９，ビルディング ３，５／エフ，アリババ グループ リーガル デパートメント

審査官　山下 剛史

(56)参考文献　特開２０１５－１０２８０６（ＪＰ，Ａ）
柏木陽佑他，話者コードに基づく話者正規化学習を利用したニューラルネット音響モデルの適応，電子情報通信学会技術研究報告，2014年12月，Vol.114，No.365，pp.105-110
柏木陽佑他，制約付き話者コードの同時推定によるニューラルネット音響モデルの話者正規化学習，日本音響学会２０１４年秋季研究発表会講演論文集，2014年09月，pp.7-10
Shaofei XUE et al.，Fast Adaptation of Deep Neural Network Based on Discriminant Codes for Speech Recognition，IEEE/ACM Transactions on Audio, Speech and Language Processing，2014年12月，Vol.22, No.12，pp.1713-1725
Zhiying HUANG et al.，Speaker adaptation of RNN-BLSTM for speech recognition based on speaker code，2016 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2016)，2016年03月，pp.5305-5309

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ １５／００－１７／２６
ＩＥＥＥ Ｘｐｌｏｒｅ