特　許　証
(CERTIFICATE OF PATENT)

特許第６９５４６８０号
(PATENT NUMBER)

発明の名称
(TITLE OF THE INVENTION)　話者の確認方法及び話者の確認装置

特許権者
(PATENTEE)　中華人民共和国北京市海淀区清華大学　郵編１００００８４
国籍・地域　中華人民共和国
清華大学

発明者
(INVENTOR)
王　東
李　藍天
鄭　方

出願番号
(APPLICATION NUMBER)　特願２０１９－５５３９１３

出願日
(FILING DATE)　平成２９年１２月　１日(December 1, 2017)

登録日
(REGISTRATION DATE)　令和　３年１０月　４日(October 4, 2021)

この発明は、特許するものと確定し、特許原簿に登録されたことを証する。
(THIS IS TO CERTIFY THAT THE PATENT IS REGISTERED ON THE REGISTER OF THE JAPAN PATENT OFFICE.)

令和　３年１０月　４日(October 4, 2021)

特許庁長官
(COMMISSIONER, JAPAN PATENT OFFICE)

森　清

(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6954680号
(P6954680)
(45)発行日　令和3年10月27日(2021.10.27)　(24)登録日　令和3年10月4日(2021.10.4)

(51)Int.Cl.　　F I
G10L 17/18 (2013.01)　G10L 17/18
G10L 15/06 (2013.01)　G10L 15/06　500P
G06N 3/04 (2006.01)　G06N 3/04　145

請求項の数 10　(全 12 頁)

(21)出願番号　特願2019-553913(P2019-553913)
(86)(22)出願日　平成29年12月1日(2017.12.1)
(65)公表番号　特表2020-515905(P2020-515905A)
(43)公表日　令和2年5月28日(2020.5.28)
(86)国際出願番号　PCT/CN2017/114293
(87)国際公開番号　WO2018/176894
(87)国際公開日　平成30年10月4日(2018.10.4)
審査請求日　令和1年11月29日(2019.11.29)
(31)優先権主張番号　201710214666.0
(32)優先日　平成29年4月1日(2017.4.1)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　502192546
清華大学
Ｔｓｉｎｇｈｕａ　Ｕｎｉｖｅｒｓｉｔｙ
中華人民共和国北京市海淀区清華大学　郵編１０００８４
Ｔｓｉｎｇｈｕａ　Ｕｎｉｖｅｒｓｉｔｙ，Ｈａｉｄｉａｎ　Ｄｉｓｔｒｉｃｔ，Ｂｅｉｊｉｎｇ　１００００８４，Ｐ．Ｒ．Ｃｈｉｎａ
(74)代理人　100108833
弁理士　早川　裕司
(74)代理人　100162156
弁理士　村雨　圭介

最終頁に続く

(54)【発明の名称】話者の確認方法及び話者の確認装置

【特許請求の範囲】
【請求項１】
第２の音声を取得することと、
予め取得した第１の音声と前記第２の音声を、対応する第１の音声スペクトログラムと第２の音声スペクトログラムに変換することと、
畳み込みニューラルネットワークを使用して、前記第１の音声スペクトログラムと前記第２の音声スペクトログラムに対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得することと、
時間遅延ニューラルネットワークを使用して、前記第１の特徴と前記第２の特徴に対して特徴抽出を行い、対応する第３の特徴と第４の特徴を取得することと、
前記第３の特徴と前記第４の特徴により、話者を確認することと、を含み、
前記の、時間遅延ニューラルネットワークを使用して、前記第１の特徴と前記第２の特徴に対して特徴抽出を行い、対応する第３の特徴と第４の特徴を取得することは、
前記第１の音声スペクトログラムの前後のフレームに対応する前記第１の特徴に対してつなぎ合わせを行い、前記第２の音声スペクトログラムの前後のフレームに対応する前記第２の特徴に対してつなぎ合わせを行うことと、
つなぎ合わせた後の前記第１の特徴とつなぎ合わせた後の前記第２の特徴のそれぞれに対して線形変換と次元削減を行い、対応する前記第３の特徴と前記第４の特徴を取得することと、を含むことを特徴とする話者の確認方法。
【請求項２】

前記の、予め取得した第１の音声と前記第２の音声を対応する第１の音声スペクトログラムと第２の音声スペクトログラムに変換することは、具体的に、
前記第１の音声と前記第２の音声のそれぞれにおけるフレームをつなぎ合わせることと、
前記第１の音声におけるフレームをつなぎ合わせた後の音声を、対応する第１の音声スペクトログラムに変換し、前記第２の音声におけるフレームをつなぎ合わせた後の音声を、対応する第２の音声スペクトログラムに変換することと、を含むことを特徴とする請求項１に記載の話者の確認方法。
【請求項３】
前記の、畳み込みニューラルネットワークを使用して、前記第１の特徴と前記第２の特徴に対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得することは、
前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対して畳み込み処理を行うことと、
畳み込んだ後の前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対してプーリング処理を行うことと、
プーリング後の前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対して次元削減を行い、前記対応する第１の特徴と第２の特徴を得ることと、を含むことを特徴とする請求項１または２に記載の話者の確認方法。
【請求項４】
畳み込みニューラルネットワークを使用して前記第１の音声スペクトログラムと前記第２の音声スペクトログラムに対して特徴抽出を行う前に、更に
前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングすることを含むことを特徴とする請求項１から３のいずれか１項に記載の話者の確認方法。
【請求項５】
前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングすることは、
交差エントロピー関数を目的関数として使用して、前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングすることを含むことを特徴とする請求項４に記載の話者の確認方法。
【請求項６】
第２の音声を取得するための取得ユニットと、
予め取得した第１の音声と前記第２の音声を、対応する第１の音声スペクトログラムと第２の音声スペクトログラムに変換するための変換ユニットと、
畳み込みニューラルネットワークを使用して、前記第１の音声スペクトログラムと前記第２の音声スペクトログラムに対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得するための第１の抽出ユニットと、
時間遅延ニューラルネットワークを使用して、前記第１の特徴と前記第２の特徴に対して特徴抽出を行い、対応する第３の特徴と第４の特徴を取得するための第２の抽出ユニットと、
前記第３の特徴と前記第４の特徴により、話者を確認するための確認ユニットと、を含み、
前記第２の抽出ユニットは、具体的に、
前記第１の音声スペクトログラムの前後のフレームに対応する前記第１の特徴に対してつなぎ合わせを行い、前記第２の音声スペクトログラムの前後のフレームと対応する前記第２の特徴に対してつなぎ合わせを行うための第２のつなぎ合わせサブユニットと、
つなぎ合わせた後の前記第１の特徴とつなぎ合わせた後の前記第２の特徴のそれぞれに対して線形変換と次元削減を行い、対応する前記第３の特徴と前記第４の特徴を取得するための第２の変換サブユニットを含むことを特徴とする話者の確認装置。
【請求項７】

前記変換ユニットは、
具体的に、それぞれ前記第１の音声と前記第２の音声におけるフレームをつなぎ合わせるためのつなぎ合わせサブユニットと、
前記第１の音声におけるフレームをつなぎ合わせた後の音声を、対応する第１の音声スペクトログラムに変換し、前記第２の音声におけるフレームをつなぎ合わせた後の音声を、対応する第２の音声スペクトログラムに変換するための変換サブユニットと、を含むことを特徴とする請求項６に記載の話者の確認装置。
【請求項８】
前記第１の抽出ユニットは、
前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対して畳み込み処理を行うための畳み込みサブユニットと、
畳み込んだ後の前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対してプーリング処理を行うためのプーリングサブユニットと、
プーリング後の前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対して次元削減を行い、前記対応する第１の特徴と第２の特徴を得るための次元削減サブユニットと、を含むことを特徴とする請求項６または７に記載の話者の確認装置。
【請求項９】
更に、前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングするためのトレーニングユニットを含むことを特徴とする請求項６から８のいずれか１項に記載の話者の確認装置。
【請求項１０】
前記トレーニングユニットは、具体的に、交差エントロピー関数を目的関数として使用して、前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングするために使われることを特徴とする請求項９に記載の話者の確認装置。
【発明の詳細な説明】
【参照関係】
【０００１】
本願は、２０１７年４月１日に中国に出願された、特許名称が「話者の確認方法及び装置」である中国特許出願２０１７１０２１４６６６０号を引用し、その全体が本出願において参照により引用されている。
【技術分野】
【０００２】
本発明は音声情報処理分野に関し、より具体的には、話者の確認方法及び話者の確認装置に関する。
【背景技術】
【０００３】
話者の確認方法とは、音声に含まれる声紋特徴により、話者に対して身分検証を行う方法である。話者の確認を行う時に、ユーザは、音声を予め保存しておき、そして検証音声を入力する。検証音声とシステムに予め保存した音声を比較すると、検証音声がそのユーザの発した音声であるかを判断することができ、ユーザの身分認証が実現される。
【０００４】
現在、話者の確認方法は、統計モデルを主にしており、性能が良い話者の確認方法は、一般的にｉ－ｖｅｃｔｏｒモデルとＰＬＤＡモデルに基づくものである。ｉ－ｖｅｃｔｏｒモデルは、音声信号に対して、下記の線形モデルを構築する。
【０００５】
Ｘ＝Ｔｗ＋ｖ
【０００６】
ただし、Ｘは音声信号のＭＦＣＣ特徴であり、Ｔは低次元行列であり、ｗはセンテンスベクトル、すなわちｉ－ｖｅｃｔｏｒであり、ｖはガウス雑音である。当該モデルは、本

質的に確率的ＰＣＡモデルである。実際の応用では、一般的に、音声空間は複数の領域に分けられ、それぞれの領域に対して上記線形モデリングが行われ、全ての領域はセンテンスベクトルｗを共有する。ｗは低次元ベクトルであり、その中に話者、音声内容、チャンネル等の情報を含む。話者情報の区別性を高めるために、次のようなＰＬＤＡモデルを取り入れる。

【０００７】

ｗ＝Ｈｕ＋Ｋｃ＋ｎ

【０００８】

ただし、ｕは話者ベクトルであり、ｃは表現ベクトルであり、発音方式、チャンネル等を含み、ｎはガウス雑音である。ＰＬＤＡモデルは、話者特徴と表現特徴を区別する。

【０００９】

上記モデルは、一般に公知のＭＦＣＣ特徴に基づくものであり、モデルにより話者情報を分離する。当該方法は、音声信号の分布状態に基づいてモデリングを行うものであるので、より良い結果を得るにはより多くのデータが必要であり、その計算量が多く、かつチャンネル、ノイズ及び時間的変化等の要因の影響を受けやすい。

【発明の概要】

【発明が解決しようとする課題】

【００１０】

本発明は、上記の必要なデータが多く、計算量が多くかつロバスト性が悪い等の課題又は少なくともその一部の課題を解決するために、話者の確認方法及び話者の確認装置を提供することを目的とする。

【課題を解決するための手段】

【００１１】

本発明のある局面に係る話者の確認方法は、第２の音声を取得することと、予め取得した第１の音声と前記第２の音声を、対応する第１の音声スペクトログラムと第２の音声スペクトログラムに変換することと、畳み込みニューラルネットワークを使用して、前記第１の音声スペクトログラムと前記第２の音声スペクトログラムに対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得することと、時間遅延ニューラルネットワークを使用して、前記第１の特徴と前記第２の特徴に対して特徴抽出を行い、対応する第３の特徴と第４の特徴を取得することと、前記第３の特徴と前記第４の特徴により、話者を確認することと、を含む。

【００１２】

具体的に、前記の、予め取得した第１の音声と前記第２の音声を対応する第１の音声スペクトログラムと第２の音声スペクトログラムに変換することは、それぞれ前記第１の音声と前記第２の音声におけるフレームをつなぎ合わせることと、それぞれ前記第１の音声におけるフレームをつなぎ合わせた後の音声を、対応する第１の音声スペクトログラムに変換し、前記第２の音声におけるフレームをつなぎ合わせた後の音声を、対応する第２の音声スペクトログラムに変換することと、を含む。

【００１３】

具体的に、前記の、畳み込みニューラルネットワークを使用して、前記第１の音声スペクトログラムと前記第２の音声スペクトログラムに対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得することは、前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対して畳み込み処理を行うことと、畳み込んだ後の前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対してプーリング処理を行うことと、プーリング後の前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対して次元削減を行い、前記対応する第１の特徴と第２の特徴を得ることと、を含む。

【００１４】

具体的に、前記の、時間遅延ニューラルネットワークを使用して、前記第１の特徴と前記第２の特徴に対して特徴抽出を行い、対応する第３の特徴と第４の特徴を取得すること

は、前記第１の音声スペクトログラムの前後のフレームに対応する前記第１の特徴に対してつなぎ合わせを行い、前記第２の音声スペクトログラムの前後のフレームに対応する前記第２の特徴に対してつなぎ合わせを行うことと、つなぎ合わせた後の前記第１の特徴とつなぎ合わせた後の前記第２の特徴のそれぞれに対して線形変換と次元削減を行い、対応する前記第３の特徴と前記第４の特徴とを取得することと、を含む。

【００１５】

具体的に、畳み込みニューラルネットワークを使用して前記第１の音声スペクトログラムと前記第２の音声スペクトログラムに対して特徴抽出を行う前に、更に、前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングすることを含む。

【００１６】

具体的に、前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングすることは、交差エントロピー関数を目的関数として使用して、前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングすることを含む。

【００１７】

本発明の他の局面に係る話者の確認装置は、第２の音声を取得するための取得ユニットと、予め取得した第１の音声と前記第２の音声を、対応する第１の音声スペクトログラムと第２の音声スペクトログラムに変換するための変換ユニットと、畳み込みニューラルネットワークを使用して、前記第１の音声スペクトログラムと前記第２の音声スペクトログラムに対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得するための第１の抽出ユニットと、時間遅延ニューラルネットワークを使用して、前記第１の特徴と前記第２の特徴に対して特徴抽出を行い、対応する第３の特徴と第４の特徴を取得するための第２の抽出ユニットと、前記第３の特徴と前記第４の特徴により、話者を確認するための確認ユニットと、を含む。

【００１８】

具体的に、更に、前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングするためのトレーニングユニットを含む。

【００１９】

具体的に、前記変換ユニットは、更に、前記第１の音声と前記第２の音声のそれぞれにおけるフレームをつなぎ合わせるためのつなぎ合わせサブユニットと、前記第１の音声におけるフレームをつなぎ合わせた後の音声を、対応する第１の音声スペクトログラムに変換し、前記第２の音声におけるフレームをつなぎ合わせた後の音声を、対応する第２の音声スペクトログラムに変換するための変換サブユニットと、を含む。

【００２０】

具体的に、前記第１の抽出ユニットは、前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対して畳み込み処理を行うための畳み込みサブユニットと、畳み込んだ後の前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対してプーリング処理を行うためのプーリングサブユニットと、プーリング後の前記第１の音声スペクトログラムと前記第２の音声スペクトログラムのそれぞれに対して次元削減を行うための次元削減サブユニットと、を含む。

【００２１】

具体的に、前記第２の抽出ユニットは、前記第１の音声スペクトログラムの前後のフレームに対応する前記第１の特徴に対してつなぎ合わせを行い、前記第２の音声スペクトログラムの前後のフレームに対応する前記第２の特徴に対してつなぎ合わせを行うための第２のつなぎ合わせサブユニットと、前記第１の特徴をつなぎ合わせた後の特徴と前記第２の特徴をつなぎ合わせた後の特徴のそれぞれに対して線形変換と次元削減を行い、対応する第３の特徴と第４の特徴を取得するための第２の変換サブユニットと、を含む。

【００２２】

具体的に、前記トレーニングユニットは、交差エントロピー関数を目的関数として使用

して、前記畳み込みニューラルネットワークと前記時間遅延ニューラルネットワークとをトレーニングするために使われる。
【発明の効果】
【００２３】
本発明は、話者の確認方法及び話者の確認装置を提供しており、畳み込みニューラルネットワークと時間遅延ニューラルネットワークとを組み合わせることで、第１の音声と第２の音声に対して二回の特徴抽出を行い、最終的に抽出した第３の特徴と第４の特徴とを比較することにより、話者の確認が実現される。本発明は、計算が簡単で、ロバスト性が強く、良好な認識効果を達成することができる。
【図面の簡単な説明】
【００２４】
【図１】本発明の実施例に係る話者の確認方法のフローチャートである。
【図２】畳み込みニューラルネットワークと時間遅延ニューラルネットワークのモデルの構成図である。
【図３】本発明の実施例に係る話者の確認装置の構成図である。
【図４】本発明の他の実施例に係る話者の確認装置の構成図である。
【発明を実施するための形態】
【００２５】
以下、添付の図面と実施例を参照して、本発明を実施するための形態について、より詳細に説明する。以下の実施例は、本発明を説明するためのものに過ぎず、本発明の範囲を限定するものではない。
【００２６】
図１は、本発明の実施例に係る話者の確認方法のフローチャートであり、第２の音声を取得するステップＳ１と、予め取得した第１の音声と第２の音声を、対応する第１の音声スペクトログラムと第２の音声スペクトログラムに変換するステップＳ２と、畳み込みニューラルネットワークを使用して、第１の音声スペクトログラムと第２の音声スペクトログラムに対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得するステップＳ３と、時間遅延ニューラルネットワークを使用して、第１の特徴と第２の特徴に対して特徴抽出を行い、対応する第３の特徴と第４の特徴を取得するステップＳ４と、第３の特徴と第４の特徴により、話者を確認するステップＳ５、を含む。
【００２７】
具体的に、ステップＳ１において、第２の音声を取得する。第２の音声は、話者が新たに入力した音声であり、検証する必要がある音声である。
【００２８】
ステップＳ２において、第１の音声は、話者が予め入力した音声であり、それぞれの第１の音声は、一つの話者ラベルと対応し、話者ラベルにより、話者を一意に確認することができる。第１の音声は、複数の話者の音声を含むことができ、それぞれの話者は、一つ又は複数の第１の音声と対応することができる。第１の音声におけるフレームをつなぎ合わせて、第１の音声スペクトログラムに変換し、第２の音声におけるフレームをつなぎ合わせて、第２の音声スペクトログラムに変換する。第１の音声スペクトログラムと第２の音声スペクトログラムの横軸は時間を表し、縦軸は周波数を表し、色又は輝度は振幅を表す。
【００２９】
ステップＳ３において、畳み込みニューラルネットワークを使用して、第１の音声スペクトログラムと第２の音声スペクトログラムに対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得することは、第１の音声スペクトログラムと第２の音声スペクトログラムのそれぞれに対して畳み込み処理を行うことと、畳み込んだ後の第１の音声スペクトログラムと第２の音スペクトルのそれぞれに対してプーリング処理を行うことと、プーリング後の第１の音声スペクトログラムと第２の音声スペクトログラムのそれぞれに対して次元削減を行い、対応する第１の特徴と第２の特徴を得ることと、を含む。

【００３０】
具体的に、畳み込みネットワークは、複数の畳み込み層とプーリング層を含んでもよい。それぞれの畳み込み層の畳み込みカーネルの数とサイズは必要に応じて調整してもよい。それぞれの畳み込みカーネルを使用して、第１の音声スペクトログラムと第２の音声スペクトログラムに対して畳み込みを行う時に、何れも一つの特徴プレーンを生成することができる。それぞれの畳み込み層の後ろには、一つのプーリング層を接続してもよい。プーリング層は、最大プーリング層又は平均プーリング層であってもよい。プーリング層のウィンドウが重なり合っていてもよく、重なり合っていなくてもよい。プーリング層のウィンドウのサイズは必要に応じて調整してもよい。最後のプーリング層から得られた特徴プレーンに対して線形変換を行い、次元削減することで、対応する第１の特徴と第２の特徴を得る。もちろん、その他の次元削減方法を使用してもよく、本発明は、これを限定しない。
【００３１】
ステップＳ４において、時間遅延ネットワークは、複数の時間遅延層を含んでもよく、それぞれの時間遅延層の入力特徴は、直前の時間遅延層の出力特徴であり、第１の音声と対応する第１の時間遅延層の入力は、第１の特徴であり、第２の音声と対応する第１の時間遅延層の入力は、第２の特徴である。第１の特徴と第２の特徴は、それぞれ第１の音声スペクトログラムと第２の音声スペクトログラムにより得られるので、第１の特徴と第２の特徴は、音声スペクトログラムにおけるフレームと１対１に対応する。第１の音声スペクトログラムについて、第１の時間遅延層を通過する時に、まず、第１の音声スペクトログラムにおける現在フレームの前後のいくつかのフレームに対応する第１の特徴をつなぎ合わせて、つなぎ合わせた後の特徴に対して、線形変換と次元削減を実行して、第１の時間遅延層の出力特徴を得る。第１の時間遅延層の出力特徴を第２の時間遅延層の入力特徴とする。
【００３２】
第２の時間遅延層を通過する時に、まず、第１の音声スペクトログラムにおける現在フレームの前後のいくつかのフレームに対応する第１の時間遅延層の出力特徴をつなぎ合わせて、つなぎ合わせた後の特徴に対して、線形変換と次元削減を実行して、第２の時間遅延層の出力特徴を得る。ただし、本発明は具体的な次元削減の方法を限定しない。
【００３３】
第２の音声スペクトログラムと第１の音声スペクトログラムとは、時間遅延ネットワークにおけるプロセスが同じである。第１の音声スペクトログラムについて、最後の時間遅延層の出力特徴は、第３の特徴である。第２の音声スペクトログラムについて、最後の時間遅延層の出力特徴は、第４の特徴である。
【００３４】
ステップＳ５において、第１の音声と第２の音声に対して、ニューラルネットワークによりフォワード演算を行い、それぞれ第１の音声と第２の音声におけるそれぞれのフレームに対応する第３の特徴と第４の特徴を抽出する。それぞれのフレームの特徴によって、任意の統計モデルを使用して、話者を確認することができる。例えば、第１の音声と第２の音声のそれぞれにおけるそれぞれのフレームの特徴の平均値を算出し、第１の音声におけるそれぞれのフレームの第３の特徴の平均値と、第２の音声におけるそれぞれのフレームの第４の特徴の平均値と、の間の距離を算出する。距離は、コサイン類似度であってもよいが、この距離に限定されない。算出したコサイン類似度がプリセット閾値よりも大きい時に、第２の音声に対応する話者ラベルにより、現在の話者を確認する。
【００３５】
本実施例は、畳み込みニューラルネットワークと時間遅延ニューラルネットワークを組み合わせることで、第１の音声と第２の音声に対して二回の特徴抽出を行い、最終的に抽出した第３の特徴と第４の特徴とを比較することにより、話者の確認が実現される。本発明は、計算が簡単で、ロバスト性が強く、良好な認識効果を達成することができる。
【００３６】

図２は、畳み込みニューラルネットワークと畳み込みニューラルネットワークの構成図であり、図２に示すように、示された畳み込みニューラルネットワークの入力は、スペクトログラムである。畳み込みニューラルネットワークは、二つの畳み込み層があり、一つ目の畳み込み層の畳み込みカーネルは、１２８個であり、それぞれの畳み込みカーネルのサイズは、６×３３であり、一つ目のプーリング層のプーリングウィンドウのサイズは、３×１１である。二つ目の畳み込み層の畳み込みカーネルは、２５６個であり、それぞれの畳み込みカーネルのサイズは、２×８である。二つ目のプーリング層のプーリングウィンドウのサイズは、２×４である。二つ目のプーリング層から得られた２５６個の特徴プレーンに対して、次元削減を行い、５１２個の特徴まで次元削減され、５１２個のニューロンに対応する。時間遅延ネットワークは、二つの時間遅延層があり、それぞれの時間遅延層は、タイミングつなぎ合わせにより前後のコンテキスト情報を拡張する。一つ目の時間遅延層は、前後の各２フレームの音声スペクトログラムに対応する第１の特徴をつなぎ合わせ、二つ目の時間遅延層は、前後の各４フレームの音声スペクトログラムに対応する一つ目の時間遅延層の出力特徴をつなぎ合わせる。それぞれの時間遅延層は、まず、一回線形変換され、そして、更に一つの次元削減層をつなぎ合わせる。それぞれの次元削減層は、時間遅延層に対して次元削減を行い、４００個の特徴を出力する。最後に、二つ目の時間遅延層の次元削減層から出力された４００個の特徴に対して、線形変換を行う。

【００３７】

上記実施例に基づいて、本実施において畳み込みニューラルネットワークを使用して第１の音声スペクトログラムと第２の音声スペクトログラムに対して特徴抽出を行う前に、更に畳み込みニューラルネットワークと時間遅延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。

【００３８】

具体的に、トレーニングする前に、確認する必要がある話者の音声を取得し、確認する必要がある話者の音声を、トレーニングセットとする。トレーニングを行う時に、音声におけるそれぞれのフレームを学習サンプルとして、畳み込みニューラルネットワークと時間遅延ニューラルネットワークにより二回の特徴抽出を行ってから、最後に算出された特徴に対して、線形変換を行い、その後、当該フレームによって予測された話者ラベルは当該フレームと対応する実際の話者ラベルであるかを確認し、話者が確認した誤差情報を使用して、畳み込みニューラルネットワークと時間遅延ニューラルネットワークにおけるパラメータを逆調整する。目的関数は、交差エントロピー関数である。トレーニングする時に使った逆伝播アルゴリズムは、ＮステップＳＧＤ（Ｎａｔｕｒａｌ　ステップＳｔｏｃｈａステップＳｔｉｃ　Ｇｒａｄｉｅｎｔ　ＤｅステップＳｃｅｎｔ、確率的勾配降下法）アルゴリズム又は任意のディープニューラルネットワークトレーニング方法であってもよい。

【００３９】

本実施例において、音声におけるそれぞれのフレームを学習サンプルとして使用して、畳み込みニューラルネットワークと時間遅延ニューラルネットワークに対してトレーニングを行い、畳み込みニューラルネットワークと時間遅延ニューラルネットワークにおけるパラメータを調整する。当該トレーニング方法に必要なデータ量は少なく、最適化した後のパラメータを使用するので、話者の確認の精度を向上させることができる。

【００４０】

図３は、本発明の実施例に係る話者の確認装置の構成図であり、図３に示すように、取得ユニット１、変換ユニット２、第１の抽出ユニット３、第２の抽出ユニット４、及び確認ユニット５を含む。ただし、取得ユニット１は、第２の音声を取得するためのものであり、変換ユニット２は、予め取得した第１の音声と第２の音声を対応する第１の音声スペクトログラムと第２の音声スペクトログラムに変換するためのものであり、第１の抽出ユニット３は、畳み込みニューラルネットワークを使用して、第１の音声スペクトログラムと第２の音声スペクトログラムに対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得するためのものであり、第２の抽出ユニット４は、時間遅延ニューラルネットワ

ークを使用して、第１の特徴と第２の特徴に対して特徴抽出を行い、対応する第３の特徴と第４の特徴を取得するためのものであり、確認ユニット５は、第３の特徴と第４の特徴により、話者を確認するためのものである。

【００４１】

具体的に、取得ユニット１は、第２の音声を取得する。第２の音声は、話者が新たに入力した音声であり、検証する必要な音声である。

【００４２】

変換ユニット２は、具体的に、第１の音声と第２の音声のそれぞれにおけるフレームをつなぎ合わせるためのつなぎ合わせサブユニットと、第１の音声におけるフレームをつなぎ合わせた後の音声を、対応する第１の音声スペクトログラムに変換し、第２の音声におけるフレームをつなぎ合わせた後の音声を、対応する第２の音声スペクトログラムに変換するための変換サブユニットと、を含む。

【００４３】

具体的に、変換ユニット２は、第１の音声におけるフレームをつなぎ合わせてから、第１の音声スペクトログラムに変換し、第２の音声におけるフレームをつなぎ合わせてから、第２の音声スペクトログラムに変換する。第１の音声は、話者が予め入力した音声であり、それぞれの第１の音声は、一つの話者ラベルと対応し、話者ラベルにより、話者を一意に確認することができる。第１の音声は、複数の話者の音声を含むことができ、それぞれの話者は、一つ又は複数の第１の音声と対応することができる。第１の音声スペクトログラムと第２の音声スペクトログラムの横軸は時間を表し、縦軸は周波数を表し、色又は輝度は振幅を表す。

【００４４】

第１の抽出ユニット３は、第１の音声スペクトログラムと第２の音声スペクトログラムのそれぞれに対して畳み込み処理を行うための畳み込みサブユニットと、畳み込んだ後の第１の音声スペクトログラムと第２の音声スペクトログラムのそれぞれに対してプーリング処理を行うためのプーリングサブユニットと、プーリング後の第１の音声スペクトログラムと第２の音声スペクトログラムのそれぞれに対して次元削減を行うための次元削減サブユニットと、を含む。

【００４５】

具体的に、第１の抽出ユニット３は、畳み込みニューラルネットワークを使用して、第１の音声スペクトログラムと第２の音声スペクトログラムに対して特徴抽出を行い、対応する第１の特徴と第２の特徴を取得する。畳み込みネットワークは、複数の畳み込み層とプーリング層を含んでもよい。それぞれの畳み込み層の畳み込みカーネルの数とサイズは必要に応じて調整してもよい。それぞれの畳み込み層の後ろには、一つのプーリング層を接続してもよい。プーリング層は、最大プーリング層又は平均プーリング層であってもよい。プーリング層のウィンドウが重なり合っていてもよく、重なり合っていなくてもよい。プーリング層のウィンドウのサイズは必要に応じて調整してもよい。最後のプーリング層から得られた特徴プレーンに対して線形変換を行い、次元削減することで、対応する第１の特徴と第２の特徴を得る。もちろん、その他の次元削減方法を使用してもよく、本発明は、これを限定しない。

【００４６】

ただし、畳み込みニューラルネットワークを使用して、第１の音声スペクトログラムと第２の音声スペクトログラムに対して特徴抽出を行う時に、それぞれの畳み込みカーネルは、一枚の特徴プレーンを生成する。畳み込みカーネルの数が多いと、複数枚の特徴プレーンを生成することができ、それぞれのプレーンは、多数の特徴があり、それぞれの畳み込み層の後ろに一つのプーリング層をつなぎ合わせるが、特徴の数が依然として多く、算出スピードを大幅に低減させる。そこで、畳み込みニューラルネットワークにおける最後のプーリング層から得られた特徴プレーンに対して次元削減を行う必要がある。低次元行列を使用して次元削減を行うことができるが、本実施例は、次元削減の方法に限定されない。

【００４７】
本実施例は、最後のプーリング層から得られた特徴プレーンに対して次元削減を行うことで、第１の音声に対応する第１の特徴と、第２の音声に対応する第２の特徴が得られ、算出スピードを大幅に向上させる。
【００４８】
更に、第２の抽出ユニット４が使用した時間遅延ネットワークは、例えば複数の全接続の時間遅延層のような複数の時間遅延層を含むことができ、それぞれの時間遅延層における第２の抽出ユニット４は、前後の各フレームをつなぎ合わせることで、前後のコンテキスト情報に対して拡張を行う。前後の各フレームをつなぎ合わせる数は、必要に応じて設置することができる。
【００４９】
具体的に、第２の抽出ユニット４は、第１の音声スペクトログラムの前後フレームに対応する第１の特徴に対してつなぎ合わせを行い、第２の音声スペクトログラムの前後フレームに対応する第２の特徴に対してつなぎ合わせを行うための第２のつなぎ合わせサブユニットと、第１の特徴をつなぎ合わせた後の特徴と、第２の特徴をつなぎ合わせた後の特徴のそれぞれに対して線形変換と次元削減を行い、対応する第３の特徴と第４の特徴を取得するための第２の変換サブユニットと、を含む。
【００５０】
具体的に、それぞれの時間遅延層において、第２のつなぎ合わせサブユニットがつなぎ合わせるフレームの数は同じであってもよいし、異なる時間遅延層において、第２のつなぎ合わせサブユニットがつなぎ合わせるフレームの数は異なってもよい。つなぎ合わせウィンドウは重なり合っていてもよい。
【００５１】
第１の音声スペクトログラムと第２の音声スペクトログラムとは、時間遅延ネットワークにおけるプロセスが同じである。第１の音声スペクトログラムについて、最後の時間遅延層の出力特徴は、第３の特徴である。第２の音声スペクトログラムについて、最後の時間遅延層の出力特徴は、第４の特徴である。
【００５２】
本発明の実施例は、時間遅延ニューラルネットワークを使用して、第１の特徴と第２の特徴に対して、特徴抽出を行い、対応する第３の特徴と第４の特徴を取得する。時間遅延ニューラルネットワークは、特徴に対して強い抽出能力を有し、話者の正確な確認に対して基礎を築く。
【００５３】
確認ユニット５は、第１の音声と第２の音声に対して、畳み込みニューラルネットワークと時間遅延ニューラルネットワークによりフォワード演算を行い、第１の音声と第２の音声のそれぞれにおけるそれぞれのフレームに対応する第３の特徴と第４の特徴を抽出する。それぞれのフレームの特徴によって、任意の統計モデルを使用して、話者を確認することができる。例えば、第１の音声と第２の音声のそれぞれにおけるそれぞれのフレームの特徴の平均値を算出し、第１の音声におけるそれぞれのフレームの第３の特徴の平均値と、第２の音声におけるそれぞれのフレームの第４の特徴の平均値と、の間の距離を算出する。距離は、コサイン類似度であってもよいが、この距離に限定されない。算出したコサイン類似度がプリセット閾値よりも大きい時に、第２の音声に対応する話者ラベルにより、現在の話者を確認する。
【００５４】
本実施例は、畳み込みニューラルネットワークと時間遅延ニューラルネットワークを組み合わせることで、第１の音声と第２の音声に対して二回の特徴抽出を行い、最終的に抽出した第３の特徴と第４の特徴を比較することにより、話者の確認が実現される。本発明は、計算が簡単で、ロバスト性が強く、良好な認識効果を達成することができる。
【００５５】
図４は、本発明の実施例に係る話者の確認装置の構成図であり、図４に示すように、上

記各実施例に基づいて、装置は、更に畳み込みニューラルネットワークと時間遅延ニューラルネットワークとをトレーニングするためのトレーニングユニット６を含む。

【００５６】

具体的に、トレーニングする前に、確認する必要がある話者の音声を取得し、確認する必要がある話者の音声を、トレーニングセットとする。トレーニングを行う時に、トレーニングユニット６は、音声中におけるそれぞれのフレームを学習サンプルとして、畳み込みニューラルネットワークと時間遅延ニューラルネットワークにより二回の特徴抽出を行ってから、最後に算出された特徴に対して、線形変換を行い、その後、当該フレームによって予測された話者ラベルは当該フレームと対応する実際の話者ラベルであるかを確認し、話者が確認した誤差情報を使用して、畳み込みニューラルネットワークと時間遅延ニューラルネットワークにおけるパラメータを逆調整する。目的関数は、交差エントロピー関数である。トレーニングする時に使った逆伝播アルゴリズムは、ＮステップＳＧＤ（Ｎａｔｕｒａｌ　ステップＳｔｏｃｈａステップＳｔｉｃ　Ｇｒａｄｉｅｎｔ　Ｄｅステップ Ｓｃｅｎｔ、確率的勾配降下法）アルゴリズム又は任意のディープニューラルネットワークトレーニング方法であってもよい。

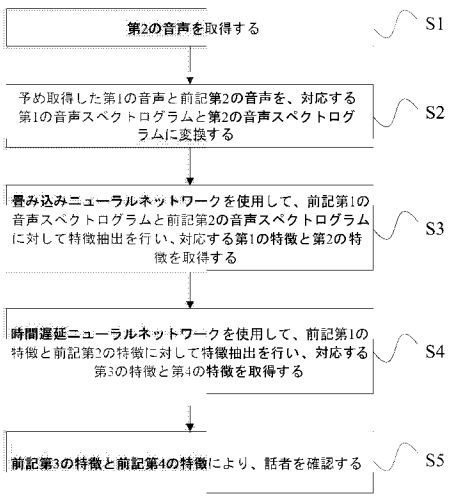
【００５７】

本実施例において、音声におけるそれぞれのフレームを学習サンプルとして使用して、畳み込みニューラルネットワークと時間遅延ニューラルネットワークに対してトレーニングを行い、畳み込みニューラルネットワークと時間遅延ニューラルネットワークにおけるパラメータを調整する。当該トレーニング方法に必要なデータ量が少なく、最適化した後のパラメータを使用するので、話者の確認の精度を向上させることができる。

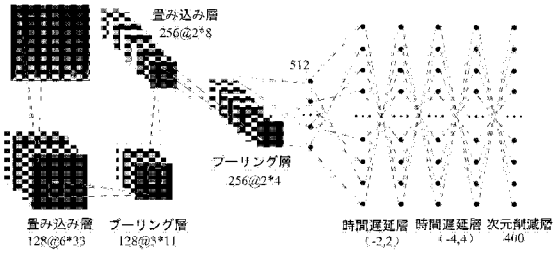
【００５８】

最後に、以上説明した方法は、一つの好適な実施案に過ぎず、本発明の保護範囲を限定するものではない。本発明の要旨と原則を逸脱しない範囲においてなされる様々な修正、等価交換、改善等は、何れも本開示の保護範囲に含まれる。

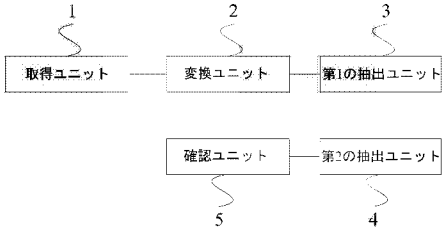
【図１】



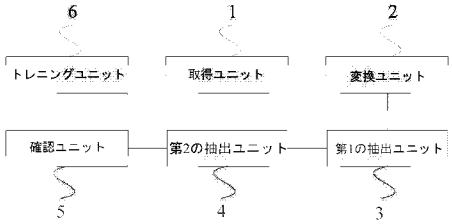
【図２】



【図３】



【図４】



フロントページの続き

発明者　王　東
中華人民共和国１０００８４北京市海淀区双清路３０号清華大学
発明者　李　藍天
中華人民共和国１０００８４北京市海淀区双清路３０号清華大学
発明者　鄭　方
中華人民共和国１０００８４北京市海淀区双清路３０号清華大学

審査官　渡部　幸和

参考文献　中国特許出願公開第１０２０３４４７２（ＣＮ，Ａ）
中国特許出願公開第１０６１２８４６５（ＣＮ，Ａ）
中川　聖一，”再訪：ニューラルネットワークによる音声処理”，電子情報通信学会技術研究報告　Ｖｏｌ．１１３　Ｎｏ．１６１，　年　月　日，　～

調査した分野　　，ＤＢ名
Ｇ１０Ｌ　１５／００－１７／２６
Ｇ０６Ｎ　３／０４