

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7205546号
(P7205546)

(45)発行日 令和5年1月17日(2023.1.17)

(24)登録日 令和5年1月6日(2023.1.6)

(51)国際特許分類

F I

G 1 0 L 17/04 (2013.01)

G 1 0 L 17/04

G 1 0 L 17/02 (2013.01)

G 1 0 L 17/02

請求項の数 21 (全20頁)

(21)出願番号	特願2020-552456(P2020-552456)	(73)特許権者	000004237
(86)(22)出願日	平成30年10月25日(2018.10.25)		日本電気株式会社
(86)国際出願番号	PCT/JP2018/039735		東京都港区芝五丁目7番1号
(87)国際公開番号	WO2020/084741	(74)代理人	110002044
(87)国際公開日	令和2年4月30日(2020.4.30)		弁理士法人プライタス
審査請求日	令和3年4月20日(2021.4.20)	(72)発明者	山本 仁
			東京都港区芝五丁目7番1号 日本電気株式会社内
		(72)発明者	越仲 孝文
			東京都港区芝五丁目7番1号 日本電気株式会社内
		審査官	菊池 智紀

最終頁に続く

(54)【発明の名称】 音声処理装置、音声処理方法、及びプログラム

(57)【特許請求の範囲】

【請求項1】

第1の音声信号をサンプルデータとして取得し、

取得した前記サンプルデータに対して、信号処理を実行し、前記サンプルデータとの類似度が設定範囲内となる新たな第2の音声信号を、前記訓練データとして生成する、
音声処理装置と、

生成された前記第2の音声信号を前記第1の音声信号の話者とは異なる話者として学習する、話者認識装置と、

を備えている、

ことを特徴とするシステム。

【請求項2】

請求項1に記載のシステムであって、

前記音声処理装置が、前記信号処理として、前記サンプルデータを、時間軸又は周波数軸において伸張又は収縮させる処理を実行する、

ことを特徴とするシステム。

【請求項3】

請求項1に記載のシステムであって、

前記音声処理装置が、前記信号処理として、前記サンプルデータに対する符号化処理と、符号化処理によって得られた潜在変数に対する演算処理と、演算処理された前記潜在変数に対する復号処理とを実行する、

ことを特徴とするシステム。

【請求項 4】

請求項 1 ~ 3 のいずれかに記載のシステムであって、

前記音声処理装置が、前記信号処理の実行後に、前記類似度として、前記サンプルデータから抽出される話者特徴と前記新たな音声信号から抽出される話者特徴との類似度を求め、求めた類似度が設定範囲内にない場合は、再度、前記信号処理を実行する、ことを特徴とするシステム。

【請求項 5】

請求項 1 ~ 4 のいずれかに記載のシステムであって、

前記音声処理装置が、前記信号処理の実行後に、前記新たな音声信号の評価を行い、得られた評価結果が設定範囲内とならない場合は、再度、前記信号処理を実行する、ことを特徴とするシステム。

10

【請求項 6】

請求項 3 に記載のシステムであって、

前記音声処理装置が、
前記信号処理の前に、前記サンプルデータの発話者の別の音声信号、及び前記サンプルデータの発話者とは異なる発話者の音声信号、それぞれに対して、符号化処理を行って潜在変数を生成し、更に、生成した潜在変数間の差分を算出し、
前記信号処理において、算出した前記差分を用いて、前記演算処理を実行する、ことを特徴とするシステム。

20

【請求項 7】

請求項 3 に記載のシステムであって、

前記音声処理装置が、前記類似度が所定の値より大きい場合に、前記潜在変数に対する演算処理として、前記潜在変数に乱数を加算する、ことを特徴とするシステム。

【請求項 8】

(a) 第 1 の音声信号をサンプルデータとして取得する、ステップと、

(b) 取得した前記サンプルデータに対して、信号処理を実行し、前記サンプルデータとの類似度が設定範囲内となる新たな第 2 の音声信号を、前記訓練データとして生成する、ステップと、

30

(c) 生成された前記第 2 の音声信号を前記第 1 の音声信号の話者とは異なる話者として学習する、ステップと、

を有する、

ことを特徴とする音声処理方法。

【請求項 9】

請求項 8 に記載の音声処理方法であって、

前記 (b) のステップにおいて、前記信号処理として、前記サンプルデータを、時間軸又は周波数軸において伸張又は収縮させる処理を実行する、ことを特徴とする音声処理方法。

【請求項 10】

40

請求項 8 に記載の音声処理方法であって、

前記 (b) のステップにおいて、前記信号処理として、前記サンプルデータに対する符号化処理と、符号化処理によって得られた潜在変数に対する演算処理と、演算処理された前記潜在変数に対する復号処理とを実行する、ことを特徴とする音声処理方法。

【請求項 11】

請求項 8 ~ 10 のいずれかに記載の音声処理方法であって、

前記 (b) のステップにおいて、前記信号処理の実行後に、前記類似度として、前記サンプルデータから抽出される話者特徴と前記新たな音声信号から抽出される話者特徴との類似度を求め、求めた類似度が設定範囲内にない場合は、再度、前記信号処理を実行する、

50

ことを特徴とする音声処理方法。

【請求項 1 2】

請求項 8 ~ 1 1 のいずれかに記載の音声処理方法であって、

前記 (b) のステップにおいて、前記信号処理の実行後に、前記新たな音声信号の評価を行い、得られた評価結果が設定範囲内とならない場合は、再度、前記信号処理を実行する、

ことを特徴とする音声処理方法。

【請求項 1 3】

請求項 1 0 に記載の音声処理方法であって、

前記 (b) のステップにおいて、

前記信号処理の前に、前記サンプルデータの発話者の別の音声信号、及び前記サンプルデータの発話者とは異なる発話者の音声信号、それぞれに対して、符号化処理を行って潜在変数を生成し、更に、生成した潜在変数間の差分を算出し、

前記信号処理において、算出した前記差分を用いて、前記演算処理を実行する、

ことを特徴とする音声処理方法。

【請求項 1 4】

請求項 1 0 に記載の音声処理方法であって、

前記 (b) のステップにおいて、前記類似度が所定の値より大きい場合に、前記潜在変数に対する演算処理として、前記潜在変数に乱数を加算する、

ことを特徴とする音声処理方法。

【請求項 1 5】

コンピュータに、

(a) 第 1 の音声信号をサンプルデータとして取得する、ステップと、

(b) 取得した前記サンプルデータに対して、信号処理を実行し、前記サンプルデータとの類似度が設定範囲内となる新たな第 2 の音声信号を、前記訓練データとして生成する、ステップと、

(c) 生成された前記第 2 の音声信号を前記第 1 の音声信号の話者とは異なる話者として学習する、ステップと、

を実行させる、プログラム。

【請求項 1 6】

請求項 1 5 に記載のプログラムであって、

前記 (b) のステップにおいて、前記信号処理として、前記サンプルデータを、時間軸又は周波数軸において伸張又は収縮させる処理を実行する、

ことを特徴とするプログラム。

【請求項 1 7】

請求項 1 5 に記載のプログラムであって、

前記 (b) のステップにおいて、前記信号処理として、前記サンプルデータに対する符号化処理と、符号化処理によって得られた潜在変数に対する演算処理と、演算処理された前記潜在変数に対する復号処理とを実行する、

ことを特徴とするプログラム。

【請求項 1 8】

請求項 1 5 ~ 1 7 のいずれかに記載のプログラムであって、

前記 (b) のステップにおいて、前記信号処理の実行後に、前記類似度として、前記サンプルデータから抽出される話者特徴と前記新たな音声信号から抽出される話者特徴との類似度を求め、求めた類似度が設定範囲内にない場合は、再度、前記信号処理を実行する、ことを特徴とするプログラム。

【請求項 1 9】

請求項 1 5 ~ 1 8 のいずれかに記載のプログラムであって、

前記 (b) のステップにおいて、前記信号処理の実行後に、前記新たな音声信号の評価を行い、得られた評価結果が設定範囲内とならない場合は、再度、前記信号処理を実行す

10

20

30

40

50

る、
ことを特徴とするプログラム。

【請求項 20】

請求項 17 に記載のプログラムであって、

前記 (b) のステップにおいて、

前記信号処理の前に、前記サンプルデータの発話者の別の音声信号、及び前記サンプルデータの発話者とは異なる発話者の音声信号、それぞれに対して、符号化処理を行って潜在変数を生成し、更に、生成した潜在変数間の差分を算出し、
前記信号処理において、算出した前記差分を用いて、前記演算処理を実行する、
ことを特徴とするプログラム。

10

【請求項 21】

請求項 17 に記載のプログラムであって、

前記 (b) のステップにおいて、前記類似度が所定の値より大きい場合に、前記潜在変数に対する演算処理として、前記潜在変数に乱数を加算する、
ことを特徴とするプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、話者認識に必要となる訓練データを生成するための、音声処理装置及び音声処理方法に関し、更には、これらを実現するためのプログラムに関する。

20

【背景技術】

【0002】

従来から、音声認識の分野では、音声信号をテキストデータに変換する処理だけでなく、音声信号から音声の特徴を抽出し、抽出した特徴に基づいて話者を識別する処理（話者認識）も行われている。

【0003】

ここで、話者認識について説明する。特許文献 1 は、話者認識を行うシステムを開示している。特許文献 1 に開示されたシステムは、まず、音声信号が入力されてくると、入力された音声信号から、人の発話の特徴を抽出する。続いて、特許文献 1 に開示されたシステムは、抽出した特徴を、予め登録されている特徴に照合し、照合結果に基づいて話者を識別する。

30

【0004】

また、特許文献 1 に開示されたシステムにおいて、音声信号からの特徴の抽出は、特徴抽出器によって行われている。具体的には、特徴抽出器は、機械学習によって構築されたモデルを用いて、音声信号から、発話した人の特徴を抽出する。また、モデルは、例えば、多数の人から得られた訓練データを用いて、ニューラルネットワークのパラメータを最適化することによって構築される。

【先行技術文献】

【特許文献】

【0005】

【文献】国際公開第 2016/092807 号

【発明の概要】

【発明が解決しようとする課題】

【0006】

ところで、特許文献 1 に開示されたシステムにおいて、話者の識別精度を高めるためには、特徴抽出器における抽出精度を高める必要がある。そして、特徴抽出器における抽出精度を高めるためには、できるだけ多くの人から訓練データを収集する必要がある。

【0007】

しかしながら、訓練データの収集は、個々の発話を録音することによって行われるので、多くの人から訓練データを収集するためには、多大なコストがかかってしまう、という

50

問題がある。また、コストは、収集元の人が多くなるほど増加してしまう。このため、従来においては、訓練データの収集には、限界がある。

【0008】

本発明の目的の一例は、上記問題を解消し、話者認識に必要となる訓練データの収集にかかるコストの上昇を抑制しつつ、特徴抽出器の抽出精度の向上を図り得る、音声処理装置、音声処理方法、及びプログラムを提供することにある。

【課題を解決するための手段】

【0009】

上記目的を達成するため、本発明の一側面における音声処理装置は、話者認識における訓練データを生成するための装置であって、

前記訓練データの元になる音声信号をサンプルデータとして取得する、データ取得部と、
取得された前記サンプルデータに対して、信号処理を実行し、前記サンプルデータとの類似度が設定範囲内となる新たな音声信号を、前記訓練データとして生成する、データ生成部と、
を備えている、
ことを特徴とする。

【0010】

また、上記目的を達成するため、本発明の一側面における音声処理方法は、話者認識における訓練データを生成するための方法であって、

(a) 前記訓練データの元になる音声信号をサンプルデータとして取得する、ステップと、
(b) 取得された前記サンプルデータに対して、信号処理を実行し、前記サンプルデータとの類似度が設定範囲内となる新たな音声信号を、前記訓練データとして生成する、ステップと、
を有する、
ことを特徴とする。

【0011】

更に、上記目的を達成するため、本発明の一側面におけるプログラムは、コンピュータによって、話者認識における訓練データを生成するためのプログラムであって、前記コンピュータに、

(a) 前記訓練データの元になる音声信号をサンプルデータとして取得する、ステップと、
(b) 取得された前記サンプルデータに対して、信号処理を実行し、前記サンプルデータとの類似度が設定範囲内となる新たな音声信号を、前記訓練データとして生成する、ステップと、
を実行させる、プログラム。

【発明の効果】

【0012】

以上のように本発明によれば、話者認識に必要となる訓練データの収集にかかるコストの上昇を抑制しつつ、特徴抽出器の抽出精度の向上を図ることができる。

【図面の簡単な説明】

【0013】

【図1】図1は、本発明の実施の形態1における音声処理装置の概略構成を示すブロック図である。

【図2】図2は、本発明の実施の形態1における音声処理装置の構成をより具体的に示すブロック図である。

【図3】図3は、本発明の実施の形態1における音声処理装置の動作を示すフロー図である。

【図4】図4は、本発明の実施の形態1の変形例1における音声処理装置の構成を示すブロック図である。

【図5】図5は、本発明の実施の形態1の変形例2における音声処理装置の構成を示すブロック図である。

10

20

30

40

50

【図 6】図 6 は、本発明の実施の形態 2 における音声処理装置の構成を示すブロック図である。

【図 7】図 7 は、本発明の実施の形態 2 における音声処理装置の動作を示すフロー図である。

【図 8】図 8 は、本発明の実施の形態 2 の変形例 1 における音声処理装置の構成を示すブロック図である。

【図 9】図 9 は、本発明の実施の形態 2 の変形例 2 における音声処理装置の構成を示すブロック図である。

【図 10】図 10 は、本発明の実施の形態 2 の変形例 3 における音声処理装置の構成を示すブロック図である。

10

【図 11】図 11 は、本発明の実施の形態 2 の変形例 3 におけるデータ生成部の処理を具体的に示す図である。

【図 12】図 12 は、本発明の実施の形態 1 及び 2 における音声処理装置を実現するコンピュータの一例を示すブロック図である。

【発明を実施するための形態】

【0014】

(実施の形態 1)

以下、本発明の実施の形態 1 における、音声処理装置、音声処理方法、及びプログラムについて、図 1～図 5 を参照しながら説明する。

【0015】

20

[装置構成]

最初に、図 1 を用いて、本実施の形態 1 における音声処理装置の構成について説明する。図 1 は、本発明の実施の形態 1 における音声処理装置の概略構成を示すブロック図である。

【0016】

図 1 に示す本実施の形態 1 における音声処理装置 100 は、話者認識における訓練データを生成するための装置である。図 1 に示すように、音声処理装置 100 は、データ取得部 10 と、データ生成部 20 とを備えている。

【0017】

データ取得部 10 は、訓練データの元になる音声信号をサンプルデータとして取得する。データ生成部 20 は、取得されたサンプルデータに対して、信号処理を実行し、サンプルデータとの類似度が設定範囲内となる新たな音声信号を、訓練データとして生成する。

30

【0018】

このように、本実施の形態 1 では、既存の音声信号から、話者認識に必要となる訓練データを生成できるので、訓練データの収集にかかるコストの上昇を抑制できる。また、本実施の形態 1 によれば、簡単に訓練データの量を増加させることができるので、話者認識における特徴抽出器の抽出精度の向上を図ることもできる。

【0019】

続いて、図 2 を用いて、本実施の形態 1 における音声処理装置のより具体的な構成について説明する。図 2 は、本発明の実施の形態 1 における音声処理装置の構成をより具体的に示すブロック図である。

40

【0020】

図 2 に示すように、本実施の形態では、音声処理装置 100 は、外部の話者データベース 200 に接続されている。話者データベース 200 は、録音された話者の音声信号を格納している。データ取得部 10 は、本実施の形態では、話者データベース 200 から、サンプルとなる音声信号を取得する。

【0021】

図 2 に示すように、本実施の形態では、データ生成部 20 は、信号処理を実行する音声変換部 21 を備えている。音声変換部 21 は、信号処理として、サンプルデータを、時間軸又は周波数軸において伸張又は収縮させる処理を実行する。

50

【 0 0 2 2 】

具体的には、音声変換部 2 1 は、例えば、サンプルデータである音声信号に対して、時間軸上の伸縮処理を施し、この音声信号を、声の高さが異なる人物を模した音声信号に変換する。また、音声変換部 2 1 は、サンプルデータである音声信号に対して、周波数軸上の伸縮処理を施し、この音声信号を声道長が異なる人物を模した音声信号に変換することもできる。

【 0 0 2 3 】

また、データ生成部 2 0 は、変換後の音声信号を、訓練データとして、外部の話者認識装置 3 0 0 へと出力する。この場合、話者認識装置 3 0 0 において、例えば、話者性の特徴を算出する特徴抽出器は、出力されてきた訓練データを用いて、話者間の差異を学習する。また、類似度を評価してスコアを算出する話者照合器、及び話者による類似度の値域を揃える類似度正規化器も、この訓練データを用いて学習することができる。

10

【 0 0 2 4 】

[装置動作]

次に、本実施の形態 1 における音声処理装置 1 0 0 の動作について図 3 を用いて説明する。図 3 は、本発明の実施の形態 1 における音声処理装置の動作を示すフロー図である。以下の説明においては、適宜図 1 を参酌する。また、本実施の形態 1 では、音声処理装置 1 0 0 を動作させることによって、音声処理方法が実施される。よって、本実施の形態 1 における音声処理方法の説明は、以下の音声処理装置 1 0 0 の動作説明に代える。

【 0 0 2 5 】

図 3 に示すように、最初に、データ取得部 1 0 は、話者データベース 2 0 0 から、サンプルとなる音声信号を取得する（ステップ A 1 ）。

20

【 0 0 2 6 】

次に、データ生成部 2 0 において、音声変換部 2 1 が、サンプルデータである音声信号を、時間軸又は周波数軸において伸張又は収縮させる処理を実行して、訓練データとなる新たな音声信号を生成する（ステップ A 2 ）。

【 0 0 2 7 】

ステップ A 2 の実行後、データ生成部 2 0 は、話者認識装置 3 0 0 に、ステップ A 2 で生成した訓練データを出力する（ステップ A 3 ）。ステップ A 3 の実行によって、音声処理装置 1 0 0 における処理は、一旦終了するが、上述のステップ A 1 ~ A 3 は、サンプルとなる音声信号を変えて、必要な訓練データが揃うまで、繰り返し実行される。

30

【 0 0 2 8 】

[実施の形態における効果]

以上のように、本実施の形態 1 では、元の音声信号から、声の高さが異なる人物を模した音声信号、又は声道長が異なる人物を模した音声信号が得られる。本実施の形態 1 によれば、訓練データの収集にかかるコストの上昇を抑制しつつ、話者認識における特徴抽出器の抽出精度の向上を図ることができる。

【 0 0 2 9 】

[プログラム]

本実施の形態 1 におけるプログラムは、コンピュータに、図 3 に示すステップ A 1 ~ A 3 を実行させるプログラムであれば良い。このプログラムをコンピュータにインストールし、実行することによって、本実施の形態における音声処理装置 1 0 0 と音声処理方法とを実現することができる。この場合、コンピュータのプロセッサは、データ取得部 1 0 、及びデータ生成部 2 0 として機能し、処理を行なう。

40

【 0 0 3 0 】

また、本実施の形態 1 におけるプログラムは、複数のコンピュータによって構築されたコンピュータシステムによって実行されても良い。この場合は、例えば、各コンピュータが、それぞれ、データ取得部 1 0 、及びデータ生成部 2 0 のいずれかとして機能しても良い。

【 0 0 3 1 】

50

[変形例 1]

ここで、図 4 を用いて、本実施の形態 1 における音声処理装置 100 の変形例 1 について説明する。図 4 は、本発明の実施の形態 1 の変形例 1 における音声処理装置の構成を示すブロック図である。

【 0032 】

図 4 に示すように、本変形例 1 では、データ生成部 20 は、音声変換部 21 に加えて、類似度判定部 22 を備え、これにより、既存話者の音声信号と、変換後の音声信号との類似度を評価する。

【 0033 】

類似度判定部 22 は、音声変換部 21 による信号処理の実行後に、類似度として、サンプルデータから抽出される話者特徴と新たな音声信号から抽出される話者特徴との類似度を求める。そして、類似度判定部 22 は、求めた類似度が設定範囲内でない場合は、音声変換部 21 に対して、再度、信号処理を実行させる。

10

【 0034 】

具体的には、類似度判定部 22 は、音声信号から、話者特徴として、例えば、既存の手法を用いて i-vector を抽出する。また、類似度判定部 22 は、類似度として、例えば、コサイン類似度を算出する。

【 0035 】

音声変換部 21 は、本変形例 1 では、求められた類似度を取得し、類似度が設定範囲内となるように、再度変換処理を行う。例えば、類似度が所定の値より大きい場合、即ち、サンプルデータと新たな音声信号とが似ている場合は、音声変換部 21 は、話者特徴の差異が大きくなるように変換処理を実行する。

20

【 0036 】

本変形例 1 によれば、声質が既存話者と異なる話者の音声信号を、確実に生成することができるので、話者認識における特徴抽出器の抽出精度をよりいっそう向上させることができる。

【 0037 】

[変形例 2]

続いて、図 5 を用いて、本実施の形態 1 における音声処理装置 100 の変形例 2 について説明する。図 5 は、本発明の実施の形態 1 の変形例 2 における音声処理装置の構成を示すブロック図である。

30

【 0038 】

図 5 に示すように、本変形例 2 では、データ生成部 20 は、音声変換部 21 に加えて、評価確認部 23 を備え、これにより、信号処理後の新たな音声信号の音声らしさを評価する。

【 0039 】

評価確認部 23 は、信号処理の実行後に、新たな音声信号の評価を実行する。そして、評価確認部 23 は、得られた評価結果が設定範囲内とならない場合は、音声変換部 21 に対して、再度、信号処理を実行させる。

【 0040 】

40

具体的には、評価確認部 23 は、既存の手法を用いて、変換処理後の新たな音声信号について、音声らしさを評価する。既存の手法としては、例えば、VAD (Voice Activity Detection) 等が挙げられる。また、音声変換部 21 は、本変形例 2 では、評価結果を取得し、評価結果が低く、音声らしさが不足している場合は、評価結果が高くなるように変換処理を実行する。

【 0041 】

本変形例 2 によれば、人の音声らしくない音声信号は除外されるので、この場合も、話者認識における特徴抽出器の抽出精度をよりいっそう向上させることができる。

【 0042 】

また、本実施の形態 1 は、上述した変形例 1 と変形例 2 とを合わせた態様であっても良

50

い。この場合は、データ生成部 20 は、音声変換部 21 に加えて、類似度判定部 22 と、評価確認部 23 との両方を備えることになる。

【0043】

(実施の形態 2)

次に、本発明の実施の形態 2 における、音声処理装置、音声処理方法、及びプログラムについて、図 6 ~ 図 10 を参照しながら説明する。

【0044】

[装置構成]

最初に、図 6 を用いて、本実施の形態 2 における音声処理装置の構成について説明する。図 6 は、本発明の実施の形態 2 における音声処理装置の構成を示すブロック図である。

10

【0045】

図 6 に示す本実施の形態 2 における音声処理装置 101 も、図 1 及び図 2 に示した実施の形態 1 における音声処理装置 100 と同様に、話者認識における訓練データを生成するための装置である。但し、本実施の形態 2 では、音声処理装置 101 は、データ生成部 20 の構成及び機能において、実施の形態 1 と異なっている。以下、相違点を中心に説明する。

【0046】

本実施の形態 2 では、データ生成部 20 は、符号化処理部 24 と、演算処理部 25 と、復号処理部 26 とを備えている。符号化処理部 24 は、サンプルデータに対する符号化処理を行う。演算処理部 25 は、符号化処理によって得られた潜在変数に対する演算処理を行う。復号処理部 26 は、演算処理された潜在変数に対する復号処理を実行する。

20

【0047】

具体的には、符号化処理部 24 は、例えば、自己符号化器 (オートエンコーダ) の符号化部 (エンコーダ) を用いて、音声信号を符号化して、潜在変数、即ち、圧縮された特徴を生成する。演算処理部 25 は、演算処理として、例えば、潜在変数に乱数を加算する。復号処理部 26 は、同じ自己符号化器の復号部 (デコーダ) を用いて、演算処理後の潜在変数に対して、復号を実行する。この結果、新たな音声信号が生成される。本実施の形態 2 では、自己符号化器として、変分自己符号化器 (Variational Autoencoder) を用いてもよい。

【0048】

本実施の形態 2 では、このように、データ生成部 20 は、信号処理として、符号化処理と、演算処理と、復号処理とを実行する。そして、符号化によって得られた潜在変数に対して、演算処理が行われるので、復号された音声信号は、元のサンプルデータと異なった音声信号となる。なお、演算処理は、上述した乱数の加算処理以外の処理であっても良い。

30

【0049】

[装置動作]

次に、本実施の形態 2 における音声処理装置 101 の動作について図 7 を用いて説明する。図 7 は、本発明の実施の形態 2 における音声処理装置の動作を示すフロー図である。以下の説明においては、適宜図 6 を参酌する。また、本実施の形態 2 では、音声処理装置 101 を動作させることによって、音声処理方法が実施される。よって、本実施の形態 2 における音声処理方法の説明は、以下の音声処理装置 101 の動作説明に代える。

40

【0050】

図 7 に示すように、最初に、データ取得部 10 は、話者データベース 200 から、サンプルとなる音声信号を取得する (ステップ B1)。

【0051】

次に、データ生成部 20 において、符号化処理部 24 が、サンプルデータに対して、符号化処理を実行する (ステップ B2)。続いて、演算処理部 25 は、ステップ B2 の符号化処理によって得られた潜在変数に対して、演算処理を実行する (ステップ B3)。更に、復号処理部 26 は、ステップ B3 の演算処理された潜在変数に対して、復号処理を実行して、新たな音声信号を生成する (ステップ B4)。

50

【 0 0 5 2 】

ステップ B 4 の実行後、データ生成部 2 0 は、話者認識装置 3 0 0 に、ステップ B 4 で生成した訓練データを出力する（ステップ B 5）。ステップ B 5 の実行によって、音声処理装置 1 0 1 における処理は、一旦終了するが、上述のステップ B 1 ~ B 5 は、サンプルとなる音声信号を変えて、必要な訓練データが揃うまで、繰り返し実行される。

【 0 0 5 3 】

[実施の形態 2 における効果]

以上のように、本実施の形態 2 においても、実施の形態 1 と同様に、元の音声信号から、それとは異なる新たな音声信号が得られる。本実施の形態 2 によっても、訓練データの収集にかかるコストの上昇を抑制しつつ、話者認識における特徴抽出器の抽出精度の向上を図ることができる。

10

【 0 0 5 4 】

[プログラム]

本実施の形態 2 におけるプログラムは、コンピュータに、図 7 に示すステップ B 1 ~ B 5 を実行させるプログラムであれば良い。このプログラムをコンピュータにインストールし、実行することによって、本実施の形態 2 における音声処理装置 1 0 1 と音声処理方法とを実現することができる。この場合、コンピュータのプロセッサは、データ取得部 1 0、及びデータ生成部 2 0 として機能し、処理を行なう。

【 0 0 5 5 】

また、本実施の形態 2 におけるプログラムは、複数のコンピュータによって構築されたコンピュータシステムによって実行されても良い。この場合は、例えば、各コンピュータが、それぞれ、データ取得部 1 0、及びデータ生成部 2 0 のいずれかとして機能しても良い。

20

【 0 0 5 6 】

[変形例 1]

ここで、図 8 を用いて、本実施の形態 2 における音声処理装置 1 0 1 の変形例 1 について説明する。図 8 は、本発明の実施の形態 2 の変形例 1 における音声処理装置の構成を示すブロック図である。

【 0 0 5 7 】

図 4 に示すように、本変形例 1 では、データ生成部 2 0 は、符号化処理部 2 4、演算処理部 2 5、及び復号処理部 2 6 に加えて、類似度判定部 2 2 を備え、これにより、既存話者の音声信号と、変換後の音声信号との類似度を評価する。

30

【 0 0 5 8 】

類似度判定部 2 2 は、実施の形態 1 の変形例 1 と同様に、信号処理の実行後に、類似度として、サンプルデータから抽出される話者特徴と新たな音声信号から抽出される話者特徴との類似度を求める。そして、類似度判定部 2 2 は、求めた類似度が設定範囲内でない場合は、符号化処理部 2 4、演算処理部 2 5、及び復号処理部 2 6 に対して、再度、信号処理を実行させる。

【 0 0 5 9 】

具体的には、本変形例 1 でも、類似度判定部 2 2 は、音声信号から、話者特徴として、例えば、既存の手法を用いて i-vector を抽出する。また、類似度判定部 2 2 は、類似度として、例えば、コサイン類似度を算出する。

40

【 0 0 6 0 】

本変形例 1 では、演算処理部 2 5 が、求められた類似度を取得し、類似度が設定範囲内となるように、演算処理を行う。例えば、類似度が所定の値より大きい場合、即ち、サンプルデータと新たな音声信号とが似ている場合は、演算処理部 2 5 は、加算する乱数の値を増加させて演算処理を実行する。

【 0 0 6 1 】

本変形例 1 によれば、実施の形態 1 の変形例 1 と同様に、声質が既存話者と異なる話者の音声信号を、確実に生成することができるので、話者認識における特徴抽出器の抽出精

50

度をよりいっそう向上させることができる。

【 0 0 6 2 】

[変形例 2]

続いて、図 9 を用いて、本実施の形態 2 における音声処理装置 1 0 1 の変形例 2 について説明する。図 9 は、本発明の実施の形態 2 の変形例 2 における音声処理装置の構成を示すブロック図である。

【 0 0 6 3 】

図 9 に示すように、本変形例 2 では、データ生成部 2 0 は、符号化処理部 2 4、演算処理部 2 5、及び復号処理部 2 6 に加えて、評価確認部 2 3 を備え、これにより、信号処理後の新たな音声信号の音声らしさを評価する。

10

【 0 0 6 4 】

評価確認部 2 3 は、実施の形態 1 の変形例 2 と同様に、信号処理の実行後に、新たな音声信号の評価を実行する。そして、評価確認部 2 3 は、得られた評価結果が設定範囲内とまらない場合は、符号化処理部 2 4、演算処理部 2 5、及び復号処理部 2 6 に対して、再度、信号処理を実行させる。

【 0 0 6 5 】

具体的には、本変形例 2 でも、評価確認部 2 3 は、既存の手法を用いて、変換処理後の新たな音声信号について、音声らしさを評価する。既存の手法としては、例えば、VAD (Voice Activity Detection) 等が挙げられる。また、本変形例 2 では、演算処理部 2 5 が、評価結果を取得し、評価結果が低く、音声らしさが不足している場合は、評価結果が高くなるように演算処理を実行する

20

【 0 0 6 6 】

本変形例 2 によれば、人の音声らしくない音声信号は除外されるので、この場合も、実施の形態 1 の変形例 2 と同様に、話者認識における特徴抽出器の抽出精度をよりいっそう向上させることができる。

【 0 0 6 7 】

また、本実施の形態 2 も、実施の形態 1 と同様に、上述した変形例 1 と変形例 2 とを合わせた態様であっても良い。この場合は、データ生成部 2 0 は、符号化処理部 2 4、演算処理部 2 5、及び復号処理部 2 6 に加えて、類似度判定部 2 2 と、評価確認部 2 3 との両方を備えることになる。

30

【 0 0 6 8 】

[変形例 3]

ここで、図 1 0 及び図 1 1 を用いて、本実施の形態 2 における音声処理装置 1 0 1 の変形例 3 について説明する。図 1 0 は、本発明の実施の形態 2 の変形例 3 における音声処理装置の構成を示すブロック図である。

【 0 0 6 9 】

図 1 0 に示すように、本変形例 3 では、データ生成部 2 0 は、符号化処理部 2 4、演算処理部 2 5、及び復号処理部 2 6 に加えて、第 2 の符号化処理部 2 7 と、差分算出部 2 8 とを備えている。

【 0 0 7 0 】

40

第 2 の符号化処理部 2 7 は、信号処理の前に、まず、データ取得部 1 0 を介して、サンプルデータの発話者の別の音声信号、及びサンプルデータの発話者とは異なる発話者の音声信号を取得する。そして、第 2 の符号化処理部 2 7 は、サンプルデータの発話者の別の音声信号、及びサンプルデータの発話者とは異なる発話者の音声信号、それぞれに対して、符号化処理を行って潜在変数を生成する。

【 0 0 7 1 】

差分算出部 2 8 は、第 2 の符号化処理部 2 7 で生成された各潜在変数間の差分を算出する。その後、演算処理部 2 5 は、差分算出部 2 8 で算出された差分を用いて、演算処理を実行する。

【 0 0 7 2 】

50

続いて、図 11 を用いて、本変形例 3 でのデータ生成部 20 の処理を具体的に説明する。図 11 は、本発明の実施の形態 2 の変形例 3 におけるデータ生成部の処理を具体的に示す図である。

【0073】

図 11 に示すように、まず、サンプルデータは音声信号 E1 の音声信号である。また、サンプルデータの発話者は、識別番号 (ID) が 123 の発話者である。この場合において、第 2 の符号化処理部 27 には、データ取得部 10 を介して、ID 123 の発話者のサンプルデータとは異なる音声信号 E3 と、ID 456 の発話者の音声信号 E4 とが、入力される。

【0074】

よって、第 2 の符号化処理部 27 は、音声信号 E3 の潜在変数と、音声信号 E4 の潜在変数とを生成し、これらを差分算出部 28 に入力する。差分算出部 28 は、入力された 2 つの潜在変数について、両者の差分 D を算出し、算出した差分 D を演算処理部 25 に入力する。

【0075】

そして、符号化処理部 24 は、サンプルデータである音声信号 E1 の潜在変数を生成するので、演算処理部 25 は、入力された差分 D を用いて、音声信号 E1 の潜在変数に対して演算処理を実行する。この場合の演算処理としては、例えば、音声信号 E1 の潜在変数への差分 D の加算処理が挙げられる。また、この場合においては、差分 D に対して、所定の係数 α が乗算されていても良い。その後、符号処理部 16 は、演算処理後の潜在変数に対して、符号化処理を実行して、新たな音声信号 E2 を生成する。

【0076】

本変形例 3 によれば、既存の話者間の相違に基づいて、新たな音声信号を生成できるので、話者認識における特徴抽出器の抽出精度をよりいっそう向上させることができる。

【0077】

(物理構成)

ここで、実施の形態 1 及び 2 におけるプログラムを実行することによって、音声処理装置を実現するコンピュータについて図 12 を用いて説明する。図 12 は、本発明の実施の形態 1 及び 2 における音声処理装置を実現するコンピュータの一例を示すブロック図である。

【0078】

図 12 に示すように、コンピュータ 110 は、CPU (Central Processing Unit) 111 と、メインメモリ 112 と、記憶装置 113 と、入力インターフェイス 114 と、表示コントローラ 115 と、データリーダ/ライタ 116 と、通信インターフェイス 117 とを備える。これらの各部は、バス 121 を介して、互いにデータ通信可能に接続される。なお、コンピュータ 110 は、CPU 111 に加えて、又は CPU 111 に代えて、GPU (Graphics Processing Unit)、又は FPGA (Field-Programmable Gate Array) を備えていても良い。

【0079】

CPU 111 は、記憶装置 113 に格納された、本実施の形態におけるプログラム (コード) をメインメモリ 112 に展開し、これらを所定順序で実行することにより、各種の演算を実施する。メインメモリ 112 は、典型的には、DRAM (Dynamic Random Access Memory) 等の揮発性の記憶装置である。また、本実施の形態におけるプログラムは、コンピュータ読み取り可能な記録媒体 120 に格納された状態で提供される。なお、本実施の形態におけるプログラムは、通信インターフェイス 117 を介して接続されたインターネット上で流通するものであっても良い。

【0080】

また、記憶装置 113 の具体例としては、ハードディスクドライブの他、フラッシュメモリ等の半導体記憶装置が挙げられる。入力インターフェイス 114 は、CPU 111 と、キーボード及びマウスといった入力機器 118 との間のデータ伝送を仲介する。表示コ

10

20

30

40

50

ントローラ 115 は、ディスプレイ装置 119 と接続され、ディスプレイ装置 119 での表示を制御する。

【0081】

データリーダ/ライタ 116 は、CPU 111 と記録媒体 120 との間のデータ伝送を仲介し、記録媒体 120 からのプログラムの読み出し、及びコンピュータ 110 における処理結果の記録媒体 120 への書き込みを実行する。通信インターフェイス 117 は、CPU 111 と、他のコンピュータとの間のデータ伝送を仲介する。

【0082】

また、記録媒体 120 の具体例としては、CF (Compact Flash (登録商標)) 及び SD (Secure Digital) 等の汎用的な半導体記憶デバイス、フレキシブルディスク (Flexible Disk) 等の磁気記録媒体、又は CD-ROM (Compact Disk Read Only Memory) などの光学記録媒体が挙げられる。

10

【0083】

なお、本実施の形態における音声処理装置 100 は、プログラムがインストールされたコンピュータではなく、各部に対応したハードウェアを用いることによっても実現可能である。更に、音声処理装置 100 は、一部がプログラムで実現され、残りの部分がハードウェアで実現されていてもよい。

【0084】

上述した実施の形態の一部又は全部は、以下に記載する(付記 1) ~ (付記 18) によって表現することができるが、以下の記載に限定されるものではない。

20

【0085】

(付記 1)

話者認識における訓練データを生成するための装置であって、
前記訓練データの元になる音声信号をサンプルデータとして取得する、データ取得部と、
取得された前記サンプルデータに対して、信号処理を実行し、前記サンプルデータとの類似度が設定範囲内となる新たな音声信号を、前記訓練データとして生成する、データ生成部と、
を備えている、
ことを特徴とする音声処理装置。

【0086】

30

(付記 2)

付記 1 に記載の音声処理装置であって、
前記データ生成部が、前記信号処理として、前記サンプルデータを、時間軸又は周波数軸において伸張又は収縮させる処理を実行する、
ことを特徴とする音声処理装置。

【0087】

(付記 3)

付記 1 に記載の音声処理装置であって、
前記データ生成部が、前記信号処理として、前記サンプルデータに対する符号化処理と、符号化処理によって得られた潜在変数に対する演算処理と、演算処理された前記潜在変数に対する復号処理とを実行する、
ことを特徴とする音声処理装置。

40

【0088】

(付記 4)

付記 1 ~ 3 のいずれかに記載の音声処理装置であって、
前記データ生成部が、前記信号処理の実行後に、前記類似度として、前記サンプルデータから抽出される話者特徴と前記新たな音声信号から抽出される話者特徴との類似度を求め、求めた類似度が設定範囲内でない場合は、再度、前記信号処理を実行する、
ことを特徴とする音声処理装置。

【0089】

50

(付記 5)

付記 1 ~ 4 のいずれかに記載の音声処理装置であって、

前記データ生成部が、前記信号処理の実行後に、前記新たな音声信号の評価を行い、得られた評価結果が設定範囲内とならない場合は、再度、前記信号処理を実行する、ことを特徴とする音声処理装置。

【 0 0 9 0 】

(付記 6)

付記 3 に記載の音声処理装置であって、

前記データ生成部が、
前記信号処理の前に、前記サンプルデータの発話者の別の音声信号、及び前記サンプルデータの発話者とは異なる発話者の音声信号、それぞれに対して、符号化処理を行って潜在変数を生成し、更に、生成した潜在変数間の差分を算出し、
前記信号処理において、算出した前記差分を用いて、前記演算処理を実行する、ことを特徴とする音声処理装置。

【 0 0 9 1 】

(付記 7)

話者認識における訓練データを生成するための方法であって、

(a) 前記訓練データの元になる音声信号をサンプルデータとして取得する、ステップと、
(b) 取得された前記サンプルデータに対して、信号処理を実行し、前記サンプルデータとの類似度が設定範囲内となる新たな音声信号を、前記訓練データとして生成する、ステップと、
を有する、
ことを特徴とする音声処理方法。

【 0 0 9 2 】

(付記 8)

付記 7 に記載の音声処理方法であって、

前記 (b) のステップにおいて、前記信号処理として、前記サンプルデータを、時間軸又は周波数軸において伸張又は収縮させる処理を実行する、
ことを特徴とする音声処理方法。

【 0 0 9 3 】

(付記 9)

付記 7 に記載の音声処理方法であって、

前記 (b) のステップにおいて、前記信号処理として、前記サンプルデータに対する符号化処理と、符号化処理によって得られた潜在変数に対する演算処理と、演算処理された前記潜在変数に対する復号処理とを実行する、
ことを特徴とする音声処理方法。

【 0 0 9 4 】

(付記 1 0)

付記 7 ~ 9 のいずれかに記載の音声処理方法であって、

前記 (b) のステップにおいて、前記信号処理の実行後に、前記類似度として、前記サンプルデータから抽出される話者特徴と前記新たな音声信号から抽出される話者特徴との類似度を求め、求めた類似度が設定範囲内にならない場合は、再度、前記信号処理を実行する、ことを特徴とする音声処理方法。

【 0 0 9 5 】

(付記 1 1)

付記 7 ~ 1 0 のいずれかに記載の音声処理方法であって、

前記 (b) のステップにおいて、前記信号処理の実行後に、前記新たな音声信号の評価を行い、得られた評価結果が設定範囲内とならない場合は、再度、前記信号処理を実行する、
ことを特徴とする音声処理方法。

10

20

30

40

50

【 0 0 9 6 】

(付記 1 2)

付記 9 に記載の音声処理方法であって、

前記 (b) のステップにおいて、

前記信号処理の前に、前記サンプルデータの発話者の別の音声信号、及び前記サンプルデータの発話者とは異なる発話者の音声信号、それぞれに対して、符号化処理を行って潜在変数を生成し、更に、生成した潜在変数間の差分を算出し、

前記信号処理において、算出した前記差分を用いて、前記演算処理を実行する、

ことを特徴とする音声処理方法。

【 0 0 9 7 】

(付記 1 3)

コンピュータによって、話者認識における訓練データを生成するためのプログラムであって、

前記コンピュータに、

(a) 前記訓練データの元になる音声信号をサンプルデータとして取得する、ステップと、

(b) 取得された前記サンプルデータに対して、信号処理を実行し、前記サンプルデータとの類似度が設定範囲内となる新たな音声信号を、前記訓練データとして生成する、ステップと、

を実行させる、プログラム。

【 0 0 9 8 】

(付記 1 4)

付記 1 3 に記載のプログラムであって、

前記 (b) のステップにおいて、前記信号処理として、前記サンプルデータを、時間軸又は周波数軸において伸張又は収縮させる処理を実行する、

ことを特徴とするプログラム。

【 0 0 9 9 】

(付記 1 5)

付記 1 3 に記載のプログラムであって、

前記 (b) のステップにおいて、前記信号処理として、前記サンプルデータに対する符号化処理と、符号化処理によって得られた潜在変数に対する演算処理と、演算処理された前記潜在変数に対する復号処理とを実行する、

ことを特徴とするプログラム。

【 0 1 0 0 】

(付記 1 6)

付記 1 3 ~ 1 5 のいずれかに記載のプログラムであって、

前記 (b) のステップにおいて、前記信号処理の実行後に、前記類似度として、前記サンプルデータから抽出される話者特徴と前記新たな音声信号から抽出される話者特徴との類似度を求め、求めた類似度が設定範囲内でない場合は、再度、前記信号処理を実行する、

ことを特徴とするプログラム。

【 0 1 0 1 】

(付記 1 7)

付記 1 3 ~ 1 6 のいずれかに記載のプログラムであって、

前記 (b) のステップにおいて、前記信号処理の実行後に、前記新たな音声信号の評価を行い、得られた評価結果が設定範囲内とならない場合は、再度、前記信号処理を実行する、

ことを特徴とするプログラム。

【 0 1 0 2 】

(付記 1 8)

付記 1 5 に記載のプログラムであって、

前記 (b) のステップにおいて、

10

20

30

40

50

前記信号処理の前に、前記サンプルデータの発話者の別の音声信号、及び前記サンプルデータの発話者とは異なる発話者の音声信号、それぞれに対して、符号化処理を行って潜在変数を生成し、更に、生成した潜在変数間の差分を算出し、
前記信号処理において、算出した前記差分を用いて、前記演算処理を実行する、
ことを特徴とするプログラム。

【 0 1 0 3 】

以上、実施の形態を参照して本願発明を説明したが、本願発明は上記実施の形態に限定されるものではない。本願発明の構成や詳細には、本願発明の範囲内で当業者が理解し得る様々な変更をすることができる。

【産業上の利用可能性】

10

【 0 1 0 4 】

以上のように本発明によれば、話者認識に必要となる訓練データの収集にかかるコストの上昇を抑制しつつ、特徴抽出器の抽出精度の向上を図ることができる。本発明は、話者認識が求められる各種分野に有用である。

【符号の説明】

【 0 1 0 5 】

1 0 データ取得部

2 0 データ生成部

2 1 音声変換部

2 2 類似度判定部

20

2 3 評価確認部

2 4 符号化処理部

2 5 演算処理部

2 6 復号処理部

2 7 第2の符号化処理部

2 8 差分算出部

1 0 0 音声処理装置（実施の形態1）

1 0 1 音声処理装置（実施の形態2）

1 1 0 コンピュータ

1 1 1 C P U

30

1 1 2 メインメモリ

1 1 3 記憶装置

1 1 4 入力インターフェイス

1 1 5 表示コントローラ

1 1 6 データリーダー/ライター

1 1 7 通信インターフェイス

1 1 8 入力機器

1 1 9 ディスプレイ装置

1 2 0 記録媒体

1 2 1 バス

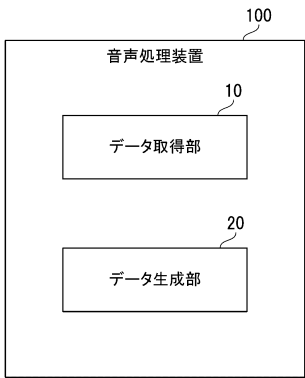
40

2 0 0 話者データベース

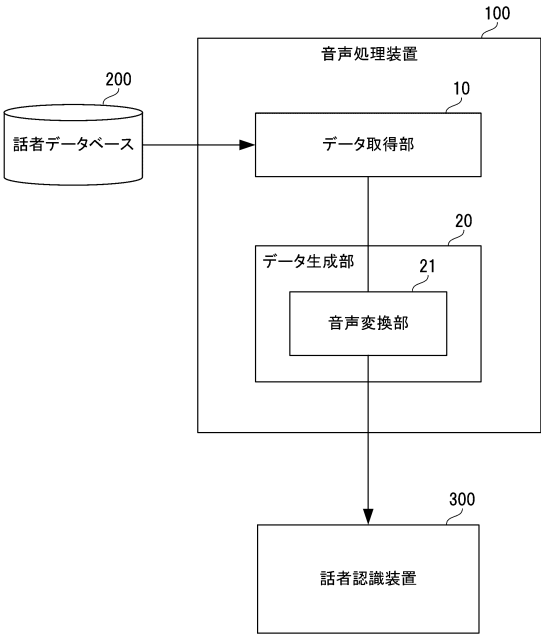
3 0 0 話者認識装置

【図面】

【図 1】



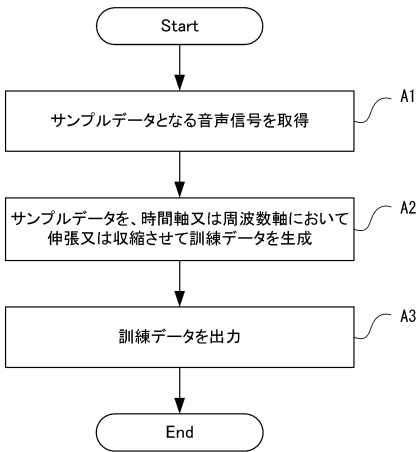
【図 2】



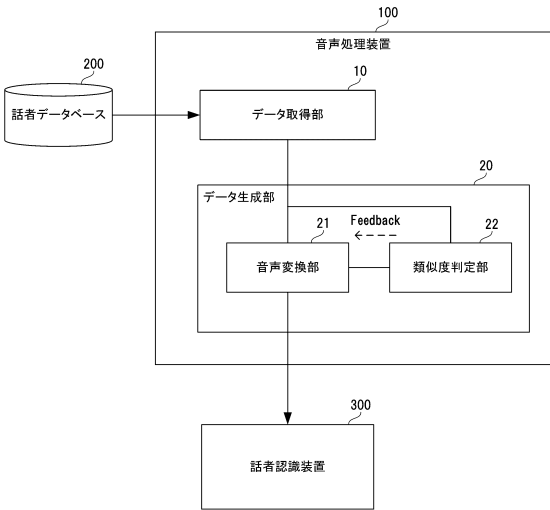
10

20

【図 3】



【図 4】

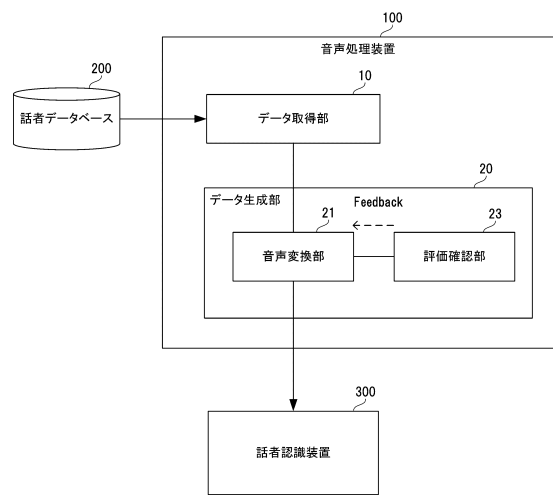


30

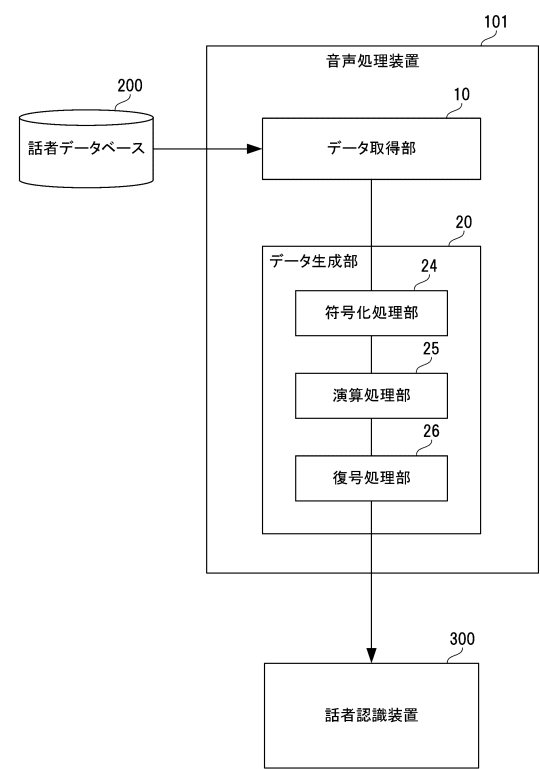
40

50

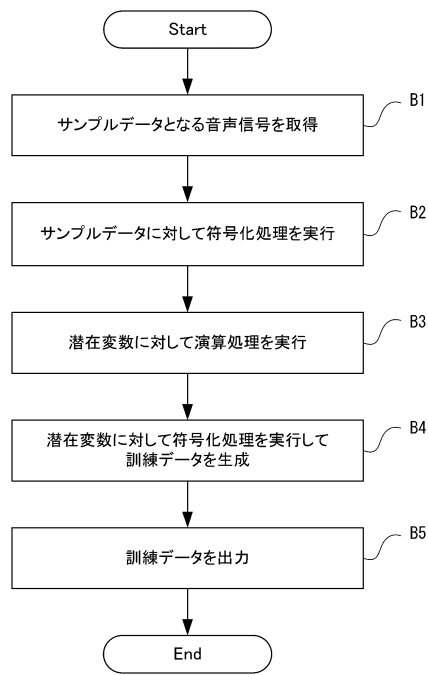
【図 5】



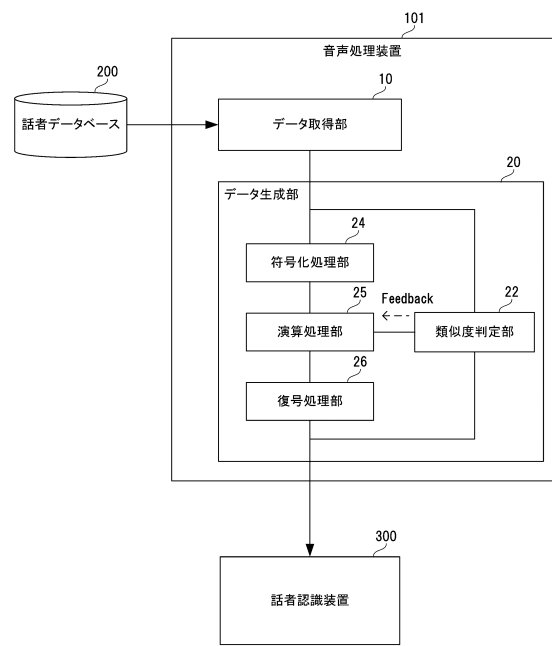
【図 6】



【図 7】



【図 8】



10

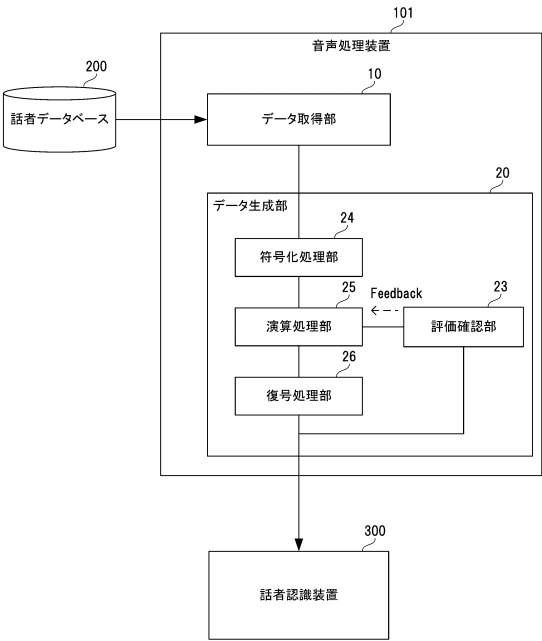
20

30

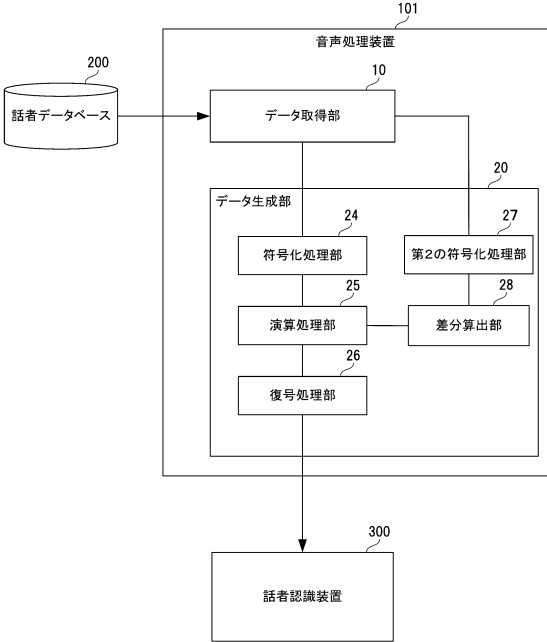
40

50

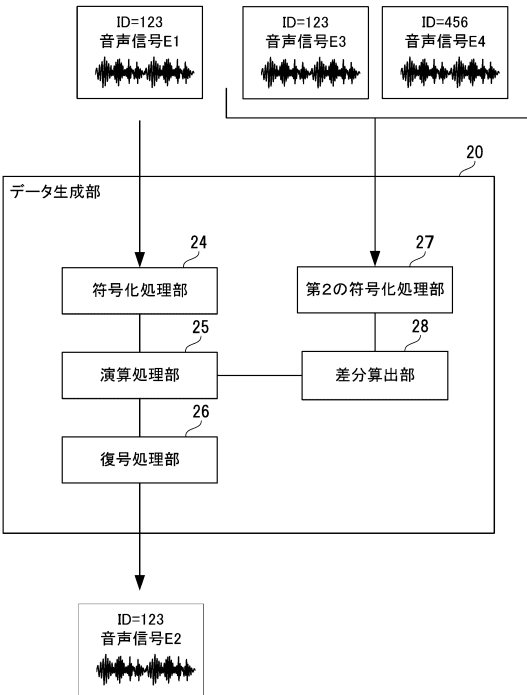
【図 9】



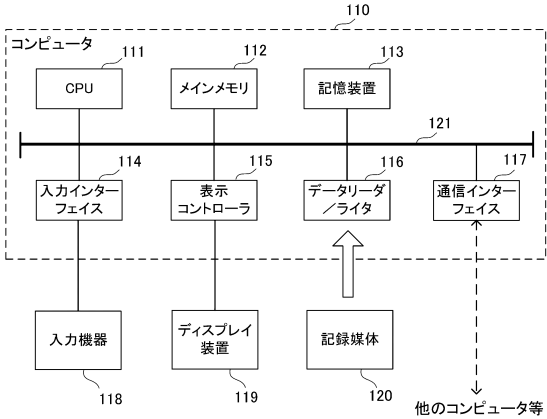
【図 10】



【図 11】



【図 12】



10

20

30

40

50

フロントページの続き

- (56)参考文献 特開 2 0 0 3 - 0 2 2 0 8 8 (J P , A)
特開平 0 4 - 2 9 5 9 0 0 (J P , A)
特開昭 6 3 - 0 7 8 2 0 0 (J P , A)
特開昭 6 1 - 0 8 4 6 9 4 (J P , A)
特開 2 0 1 8 - 1 3 9 0 7 1 (J P , A)
西崎博光 他 , " 音声コーデックと変分オートエンコーダを利用した音響モデル学習データの
拡張 " , 日本音響学会 2017 年秋季研究発表会講演論文集 CD-ROM , 2017 年 09 月 11 日 , pp.8
7-90
- (58)調査した分野 (Int.Cl. , D B 名)
G 1 0 L 1 5 / 0 0 - 1 7 / 2 6
G 0 6 N 3 / 0 0 - 9 9 / 0 0
I E E E X p l o r e