(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6845489号
(P6845489)
(45)発行日　令和3年3月17日(2021.3.17)　(24)登録日　令和3年3月2日(2021.3.2)

(51)Int.Cl.　F I
G10L 17/00 (2013.01)　G10L 17/00 200C

請求項の数 10 (全 16 頁)

(21)出願番号　特願2019-504164(P2019-504164)
(86)(22)出願日　平成29年3月7日(2017.3.7)
(86)国際出願番号　PCT/JP2017/008979
(87)国際公開番号　WO2018/163279
(87)国際公開日　平成30年9月13日(2018.9.13)
審査請求日　令和1年5月21日(2019.5.21)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100077838
弁理士　池田　憲保
(74)代理人　100129023
弁理士　佐々木　敬
(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　大野　弘

最終頁に続く

(54)【発明の名称】音声処理装置、音声処理方法、および音声処理プログラム

(57)【特許請求の範囲】
【請求項１】
音声を表す音声信号に基づき、前記音声信号の部分区間の品質の程度を表す貢献度を算出する貢献度推定手段と、
前記音声信号の前記部分区間の前記貢献度を、前記部分区間の重みとして用いて、前記音声信号から特定の属性情報を認識するための認識特徴量を算出する話者特徴算出手段とを備える、音声処理装置。
【請求項２】
前記音声信号に含まれる音の種類の比率を表す音声統計量を算出する音声統計量算出手段をさらに備え、
前記話者特徴算出手段は、前記音声信号の前記音声統計量と、前記音声信号の前記貢献度とに基づいて、前記認識特徴量を算出する、請求項１に記載の音声処理装置。
【請求項３】
前記貢献度推定手段は、前記音声信号の前記貢献度として、
前記音声信号の一部が音声か否かを識別して算出した音声らしさを表す値、前記音声信号の一部が話者認識に正解する音声か否かを識別して算出した話者認識の正解しやすさを表す値、前記音声信号の一部が話者認識誤りを起こす音声か否かを識別して算出した話者認識の誤りやすさを表す値の少なくともいずれかひとつを算出する、請求項１または２に記載の音声処理装置。
【請求項４】

前記貢献度推定手段は、
ニューラルネットワークを用いて前記音声信号の前記貢献度を算出する、請求項３に記載の音声処理装置。
【請求項５】
前記話者特徴算出手段は、
前記認識特徴量として i-vector を算出する、請求項３または４に記載の音声処理装置。
【請求項６】
前記認識特徴量に基づいて前記属性情報を認識する属性認識手段を備える、請求項１～５のいずれか１項に記載の音声処理装置。
【請求項７】
前記特定の属性情報は、
音声信号を発した話者、音声信号を構成する言語、音声信号に含まれる感情表現、音声信号から推定される話者の性格の少なくともいずれか１つを表す情報である、請求項１～６のいずれか１項に記載の音声処理装置。
【請求項８】
音声を表す音声信号に基づき、前記音声信号の部分区間の品質の程度を表す貢献度を算出し、
前記音声信号の前記部分区間の前記貢献度を、前記部分区間の重みとして用いて、前記音声信号から特定の属性情報を認識するための認識特徴量を算出する、音声処理方法。
【請求項９】
前記音声信号に含まれる音の種類の比率を表す音声統計量をさらに算出し、
前記音声信号の前記音声統計量と、前記音声信号の前記貢献度とに基づいて、前記認識特徴量を算出する、請求項８に記載の音声処理方法。
【請求項１０】
コンピュータに、
音声を表す音声信号に基づき、前記音声信号の部分区間の品質の程度を表す貢献度を算出する処理と、
前記音声信号の前記部分区間の前記貢献度を、前記部分区間の重みとして用いて、前記音声信号から特定の属性情報を認識するための認識特徴量を算出する処理とを実行させる、音声処理プログラム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音声処理装置、音声処理方法、および音声処理プログラムに関する。
【背景技術】
【０００２】
音声信号から、音声を発した話者を特定するための個人性を表す話者特徴を算出する音声処理装置が知られている。また、この話者特徴を用いて、音声信号を発した話者を推定する話者認識装置が知られている。
【０００３】
この種の音声処理装置を用いる話者認識装置は、話者を特定するために、第１の音声信号から抽出した第１の話者特徴と、第２の音声信号から抽出した第２の話者特徴との類似度を評価する。そして、話者認識装置は、類似度の評価結果に基づいて２つの音声信号の話者が同一か否かを判定する。
【０００４】
非特許文献１には、音声信号から話者特徴を抽出する技術が記載されている。非特許文献１に記載の話者特徴抽出技術は、音声モデルを用いて音声信号の音声統計量を算出する

。そして、非特許文献１に記載の話者特徴抽出技術は、因子分析技術に基づいてその音声統計量を処理し、所定の要素数で表現される話者特徴ベクトルとして算出する。すなわち、非特許文献１においては、話者特徴ベクトルを話者の個人性を表す話者特徴として利用する。

【先行技術文献】

【非特許文献】

【０００５】

【非特許文献１】Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, “Front-End Factor Analysis for Speaker Verification,” IEEE Transaction on Audio, Speech and Language Processing, Vol. 19, No. 4, pp. 788-798, 2011.

【発明の概要】

【発明が解決しようとする課題】

【０００６】

しかしながら、非特許文献１に記載の技術には、抽出した話者特徴を用いる話者認識の精度が十分でないという問題があった。

【０００７】

非特許文献１に記載の技術は、話者特徴抽出装置に入力された音声信号に対して所定の統計処理を行い、話者特徴ベクトルを算出する。具体的には、非特許文献１に記載の技術は、話者特徴抽出装置に入力された音声信号の全体に対して一律の統計処理を行うことにより、話者特徴ベクトルを算出している。そのため、非特許文献１に記載の技術は、音声信号の部分区間に、話者の個人性を算出する元として適切ではない信号が含まれている場合であっても、音声信号の全体から話者特徴ベクトルを算出してしまうので、話者認識の精度を損なうおそれがある。具体的には、音声信号の部分区間に、例えば、話者の不明瞭な発声、話者の咳や笑い声などの話し声とは異なる音、雑音などが混入している場合に、話者認識の精度を損なうおそれがある。

【０００８】

本発明は、上記問題に鑑みてなされたものであり、その目的は、話者認識の精度をより高めた音声処理装置、音声処理方法、および音声処理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】

本発明の第１の態様の音声処理装置は、音声を表す音声信号に基づき、前記音声信号の部分区間の品質の程度を表す貢献度を算出する貢献度推定手段と、前記音声信号の前記部分区間の前記貢献度を、前記部分区間の重みとして用いて、前記音声信号から特定の属性情報を認識するための認識特徴量を算出する話者特徴算出手段とを備える。

【００１０】

本発明の第２の態様の音声処理方法は、音声を表す音声信号に基づき、前記音声信号の部分区間の品質の程度を表す貢献度を算出し、前記音声信号の前記部分区間の前記貢献度を、前記部分区間の重みとして用いて、前記音声信号から特定の属性情報を認識するための認識特徴量を算出する。

【００１１】

本発明の第３の態様の音声処理プログラムは、コンピュータに、音声を表す音声信号に基づき、前記音声信号の部分区間の品質の程度を表す貢献度を算出する処理と、前記音声信号の前記部分区間の前記貢献度を、前記部分区間の重みとして用いて、前記音声信号から特定の属性情報を認識するための認識特徴量を算出する処理とを実行させる。

【発明の効果】

【００１２】

本発明によれば、話者認識の精度をより高めた音声処理装置、音声処理方法、およびプログラムを提供することができる。

【図面の簡単な説明】

【００１３】
【図１】本発明の第１の実施形態に係る音声処理装置の構成を示すブロック図である。
【図２】本発明の第１の実施形態に係る音声処理装置の動作の流れを示すフローチャートである。
【図３】本発明の第２の実施形態に係る音声処理装置の構成を示すブロック図である。
【図４】本発明の第２の実施形態に係る音声処理装置の動作の流れを示すフローチャートである。
【図５】本発明の第３の実施形態に係る音声処理装置の構成を示すブロック図である。
【図６】本発明のその他の実施形態に係る音声処理装置の構成を示すブロック図である。
【発明を実施するための形態】
【００１４】
以下、音声処理装置等および話者特徴抽出装置の実施形態について、図面を参照して説明する。なお、実施形態において同じ符号を付した構成要素は同様の動作を行うので、再度の説明を省略する場合がある。
【００１５】
＜第１の実施形態＞
図１は、本発明の第１の実施形態に係る音声処理装置の構成を示すブロック図である。
【００１６】
音声処理装置１００は、貢献度推定部１１と、話者特徴算出部１２とを備える。
【００１７】
貢献度推定部１１は、外部から音声を表す音声信号を受け取る。また、貢献度推定部１１は、受けた音声信号に基づき、その音声信号の部分区間の品質の程度を数値で表した貢献度を算出する。
【００１８】
話者特徴算出部１２は、貢献度推定部１１が算出した音声信号の部分区間の貢献度を、その部分区間の重みとして用いて、音声信号から特定の属性情報を認識するための認識特徴量を算出する。
【００１９】
ここで、特定の属性情報とは、音声信号を発した話者、音声信号を構成する言語、音声信号に含まれる感情表現、および音声信号から推定される話者の性格等を示す情報である。
【００２０】
図２を参照し、音声処理装置１００の動作の流れについて説明する。図２は本発明の第１の実施形態に係る音声処理装置の動作の流れを示すフローチャートである。
【００２１】
まず、貢献度推定部１１は、外部から受けた音声信号に基づいて、音声信号の部分区間の貢献度を算出する（ステップＳ１０１）。次いで、貢献度推定部１１は、算出した音声信号の部分区間の貢献度を話者特徴算出部１２に出力する。
【００２２】
次いで、話者特徴算出部１２は、貢献度推定部１１から受けた貢献度に基づいて、認識特徴量を算出する（ステップＳ１０２）。
【００２３】
＜第２の実施形態＞
図３は、第２の実施形態における音声処理装置２００のブロック図である。音声処理装置２００は、貢献度推定部１１、話者特徴算出部１２、音声区間検出部２１、および音声統計量算出部２２を備える。また、音声処理装置２００は、さらに、貢献度記憶部２３および貢献度学習部２４を備えてもよい。
【００２４】
音声区間検出部２１は、外部から音声信号を受け取る。また、音声区間検出部２１は、受け取った音声信号に含まれる音声区間を検出して区分化する。この時、音声区間検出部

２１は、音声信号を一定の長さに区分化してもよいし、異なる長さに区分化してもよい。例えば、音声区間検出部２１は、音声信号のうち音量が一定時間継続して所定値より小さい区間を無音と判定し、その区間の前後を異なる音声区間と判定して区分化してもよい。そして、音声区間検出部２１は、区分化した結果（音声区間検出部２１の処理結果）である区分化音声信号を、貢献度推定部１１および音声統計量算出部２２に出力する。ここで、音声信号の受け取りとは、例えば、外部の装置または他の処理装置からの音声信号の受信、または他のプログラムからの音声信号処理の処理結果の引き渡しのことである。また、出力とは、例えば、外部の装置や他の処理装置への送信、または他のプログラムへの音声区間検出部２１の処理結果の引き渡しのことである。

【００２５】

音声統計量算出部２２は、音声区間検出部２１から区分化音声信号を受け取る。音声統計量算出部２２は、受け取った区分化音声信号に基づいて、該区分化音声信号に含まれる音の種類を表す音声統計量を算出する。ここで、音の種類とは、例えば、言語により定まる音素や単語、音声信号を類似度に基づいてクラスタリングして得られる音のグループである。そして、音声統計量算出部２２は、音声統計量を話者特徴算出部１２に出力する。以降、ある音声信号に対して算出された音声統計量を、該音声信号の音声統計量と呼ぶ。

【００２６】

音声統計量算出部２２が、音声統計量を算出する方法の一例について説明する。具体的には、音声統計量算出部２２は、音声区間検出部２１から受け取った区分化音声信号に基づいて、該区分化音声信号を周波数分析処理した計算結果で表現される音響特徴を算出し、算出した結果を出力する。例えば、音声統計量算出部２２は、音声区間検出部２１から受け取った区分化音声信号を、短時間フレーム時系列に変換する。そして、音声統計量算出部２２は、短時間フレーム時系列のそれぞれのフレームを周波数分析し、その処理結果を音響特徴として出力する。この場合、音声統計量算出部２２は、例えば、短時間フレーム時系列として、２５ミリ秒区間のフレームを１０ミリ秒ごとに生成する。音声統計量算出部２２は、例えば、周波数分析結果である音響特徴として、高速フーリエ変換処理（Ｆａｓｔ　Ｆｏｕｒｉｅｒ　Ｔｒａｎｓｆｏｒｍ；ＦＦＴ）およびフィルタバンク処理によって得られた周波数フィルタバンク特徴や、さらに加えて離散コサイン変換処理を施して得られたメル周波数ケプストラム係数（Ｍｅｌ－Ｆｒｅｑｕｅｎｃｙ　Ｃｅｐｓｔｒｕｍ　Ｃｏｅｆｆｉｃｉｅｎｔｓ；ＭＦＣＣ）特徴などを算出する。

【００２７】

そして、音声統計量算出部２２は、音響特徴の時系列と、音響特徴と音の種類との対応関係を格納する音声モデルを用いて、音の種類を表す数値情報の時系列を算出する。音声統計量算出部２２は、例えば、音声モデルがガウス混合モデル（Ｇａｕｓｓｉａｎ　Ｍｉｘｔｕｒｅ　Ｍｏｄｅｌ；GMM）である場合、ガウス混合モデルが有する各要素分布の平均、分散、および混合係数に基づいて、各要素分布の事後確率を算出する。ここで、各要素分布の事後確率は、音声信号に含まれる音の種類それぞれの出現度である。また、音声統計量算出部２２は、例えば、音声モデルがニューラルネットワーク（Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋ）である場合、音響特徴と、ニューラルネットワークが有する重み係数に基づいて、音声信号に含まれる音の種類の出現度を算出する。

【００２８】

貢献度記憶部２３は、１つ以上の貢献度推定器を記憶する。貢献度推定器は、音声信号を信号の品質によって複数の種類に仕分けるよう動作するように構成されるものである。貢献度推定器は、例えば、音声信号の品質を表す数値情報を出力する。信号の品質の種類とは、例えば、音声・非音声・無音である。また、信号の品質の種類とは、例えば、話者認識に正解する音声・話者認識に誤りを起こす音声である。

【００２９】

具体的には、貢献度記憶部２３は、貢献度推定器が保有するパラメタを記憶する。貢献度記憶部２３は、例えば、貢献度推定器がニューラルネットワークである場合、それを構成するノードの数やノード間の接続重み係数などの一式をパラメタとして記憶する。

【００３０】
なお、図３では、貢献度記憶部２３が音声処理装置２００内に内蔵されることを例に説明を行ったが、本発明はこれに限定されるものではない。貢献度記憶部２３は、音声処理装置２００の外部に設けられた記憶装置で実現されるものであってもよい。
【００３１】
貢献度推定部１１は、音声区間検出部２１から区分化音声信号を受け取る。貢献度推定部１１は、貢献度記憶部２３に記憶されている貢献度推定器を用いて、区分化音声信号の品質を表す数値情報を算出する。貢献度推定部１１は、音声統計量算出部２２と同様に、区分化音声信号を短時間フレーム時系列に変換し、それぞれのフレームの音響特徴を算出し、音響特徴の時系列を算出する。続いて、貢献度推定部１１は、各フレームの音響特徴と貢献度推定器のパラメタとを用いて、各フレームの品質を表す数値を算出する。以降、ある音声信号に対して算出された信号の品質を表す数値のことを音声信号の貢献度と呼ぶ。
【００３２】
具体的には、貢献度推定部１１は、例えば、貢献度推定器がニューラルネットワークである場合、音響特徴と、ニューラルネットワークが有する重み係数とに基づいて、音響特徴の貢献度を算出する。例えば、貢献度推定器がニューラルネットワークであり、その出力層が、２つの信号の品質の種類「話者認識に正解する信号」と「話者認識誤りを起こす信号」とに相当するものであるとする。このとき、貢献度推定器は、音響特徴が話者認識に正解する信号である確率と、音響特徴が話者認識誤りを起こす信号である確率とを算出し、貢献度として、例えば、「話者認識に正解する信号」である確率を出力する。また、貢献度推定部１１は、話者認識を実行する前に、音声信号の部分区間が音声か否かを識別して音声である確率を算出してもよい。
【００３３】
話者特徴算出部１２は、音声統計量算出部２２が出力した音声統計量および貢献度推定部１１が出力した貢献度を受け取る。話者特徴算出部１２は、音声統計量および貢献度を用いて、音声信号から特定の属性情報を認識するための認識特徴量を算出する。
【００３４】
話者特徴算出部１２が音声信号ｘの認識特徴量としてｉ－ｖｅｃｔｏｒに基づく特徴ベクトルＦ（ｘ）を算出する方法の一例について説明する。なお、話者特徴算出部１２が算出する特徴ベクトルＦ（ｘ）は、音声信号ｘに対して所定の演算を施して算出できるベクトルであればよく、ｉ－ｖｅｃｔｏｒはその一例である。
【００３５】
話者特徴算出部１２は、音声統計量算出部２２から、音声信号ｘの統計量の情報として、例えば、短時間フレームごとに算出された音響事後確率Ｐｔ（ｘ）および音響特徴Ａｔ（ｘ）（ｔ＝｛１…Ｔ｝、Ｔは１以上の自然数）とを受け取る。また、話者特徴算出部１２は、貢献度推定部１１から、音声信号ｘの貢献度の情報として、例えば、短時間フレームごとに算出された貢献度Ｃｔ（ｘ）を受け取る。話者特徴算出部１２は、以下の式（１）のように、音響事後確率Ｐｔ（ｘ）の各要素に対して、貢献度Ｃｔ（ｘ）をかけて、その結果をＱｔ（ｘ）として算出する。
【００３６】
【数１】

$$Q_{t,c}(x) = C_t(x)P_{t,c}(x) \cdots\cdots (1)$$

【００３７】
話者特徴算出部１２は、貢献度によって重みづけされた音響事後確率Ｑｔ（ｘ）および音響特徴Ａｔ（ｘ）を用いて、以下の式（２）に基づいて音声信号ｘの０次統計量Ｓ０（ｘ）を算出し、式（３）に基づいて１次統計量Ｓ１（ｘ）を算出する。
【００３８】

【数２】

$$S_0(x)=\begin{pmatrix} S_{0,1}I_D & \cdots & 0_D \\ \vdots & \ddots & \vdots \\ 0_D & \cdots & S_{0,C}I_D \end{pmatrix},\qquad S_{0,c}=\sum_{t=1}^{T}Q_{t,c}(x)\quad\cdots\cdots(2)$$

【００３９】

【数３】

$$S_1(x)=\left(S_{1,1},S_{1,2},\ldots,S_{1,c}\right)^T,\qquad S_{1,c}=\sum_{t=1}^{T}Q_{t,c}(x)(A_t(x)-m_c)\quad\cdots\cdots(3)$$

【００４０】

話者特徴算出部１２は、続いて、以下の式（４）に基づいて音声信号ｘのｉ－ｖｅｃｔｏｒであるＦ（ｘ）を算出する。

【００４１】

【数４】

$$F(x)=(I+T^T\Sigma^{-1}S_0(x)T)^{-1}T^T\Sigma^{-1}S_1(x)\quad\cdots\cdots(4)$$

【００４２】

式（１）～式（４）において、Ｃは統計量Ｓ０（ｘ）およびＳ１（ｘ）の要素数、Ｄは音響特徴Ａｔ（ｘ）の要素数（次元数）、ｍｃは音響特徴空間におけるｃ番目の領域の音響特徴の平均ベクトル、Ｉは単位行列、０は零行列を表す。Ｔはｉ－ｖｅｃｔｏｒ計算用のパラメタであり、Σは音響特徴空間における音響特徴の共分散行列である。

【００４３】

話者特徴算出部１２が上述の手順で特徴ベクトルＦ（ｘ）を算出する際に、音声信号ｘのすべての時刻ｔ（ｔ＝｛１…Ｔ｝、Ｔは１以上の自然数）において、その貢献度Ｃｔ（ｘ）が１であれば、非特許文献１に記載のｉ－ｖｅｃｔｏｒ算出手順と等価である。本実施形態において、話者特徴算出部１２は、貢献度推定部１１が音声信号ｘの時刻ｔに応じて推定した貢献度Ｃｔ（ｘ）を用いることにより、非特許文献１に記載のｉ－ｖｅｃｔｏｒとは異なる特徴ベクトルＦ（ｘ）を算出できる。

【００４４】

このように、音声処理装置２００において、話者特徴算出部１２が、音声信号ｘに対して、該音声信号の各部分区間の品質に応じた貢献度Ｃｔ（ｘ）を用いて特徴ベクトルＦ（ｘ）を算出することにより、音声信号の品質に応じた特徴ベクトルを出力することができる。

【００４５】

貢献度学習部２４は、訓練用音声信号を用いて貢献度記憶部２３が記憶できる貢献度推定器を学習する。貢献度学習部２４は、例えば、貢献度推定器がニューラルネットワークである場合、それを構成するノード間の接続重み係数などのパラメタを、一般的な最適化基準に従って最適化する。貢献度学習部２４が使用する訓練用音声信号は、複数の音声信号を集めたものであり、それぞれの音声信号は、貢献度推定部１１が出力する信号の品質の種類のいずれかと対応付けられたものである。

【００４６】

以下では、入力が音響特徴であり、出力が「話者認識に正解する音声」および「話者認識に誤りを起こす音声」の２種類の信号の品質である貢献度推定器を貢献度学習部２４が学習する方法の一例を説明する。

【００４７】

（ａ）まず、貢献度学習部２４は、話者ラベルつきの複数の音声信号を用いて、音声信号の話者ラベルを識別することのできる識別器を学習する。（ｂ）次に、貢献度学習部２

４は、話者ラベルつきの複数の音声信号のそれぞれを、短時間フレームごとに算出した音響特徴の時系列に変換し、（ａ）で学習した識別器を用いて、各フレームの話者ラベルを識別する。（ｃ）次に、貢献度学習部２４は、識別された各フレームの話者ラベルのうち、事前に付与された話者ラベルと、識別器が識別した話者ラベルが同一であるフレームを「話者認識に正解する音声」、そうでないフレームを「話者認識に誤りを起こす音声」とする。（ｄ）そして、貢献度学習部２４は、「話者認識に正解する音声」および「話者認識に誤りを起こす音声」を訓練用音声信号として、貢献度推定器を学習する。

【００４８】

以上述べたように、本実施形態に係る音声処理装置２００において、貢献度推定部１１は、音声信号の部分区間に応じた品質を表す指標として、音声信号の貢献度を算出できる。また、話者特徴算出部１２は、音声信号の音響統計量と貢献度とに基づいて特徴ベクトルを算出する。これにより、音声信号に対して、音声信号の各部分区間の品質を考慮した特徴ベクトルを出力できる。すなわち、本実施形態にかかる音声処理装置２００は、話者認識の精度を高めるのに適した話者特徴を算出できる。

【００４９】

なお、本実施形態に係る音声処理装置２００における貢献度記憶部２３は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。また、貢献度記憶部２３に貢献度推定器が記憶される過程は特に限定されない。例えば、記録媒体を介して貢献度推定器が貢献度記憶部２３に記憶されてもよいし、通信回線等を介して送信された貢献度推定器が貢献度記憶部２３に記憶されてもよい。または、入力デバイスを介して入力された貢献度推定器が貢献度記憶部２３で記憶されてもよい。

【００５０】

（第２の実施形態の動作）

次に、第２の実施形態における音声処理装置２００の動作について、図４のフローチャートを用いて説明する。図４は、音声処理装置２００の動作の一例を示すフローチャートである。

【００５１】

音声処理装置２００は、外部から１つ以上の音声信号を受け取り、音声区間検出部２１に提供する。具体的には、音声区間検出部２１は、受け取った音声信号を区分化し、区分化音声信号を貢献度推定部１１および音声統計量算出部２２に出力する（ステップＳ２０１）。

【００５２】

音声統計量算出部２２は、受け取った１つ以上の区分化音声信号それぞれについて、短時間フレーム分析処理を行い、音響特徴と音声統計量の時系列を算出する（ステップＳ２０２）。

【００５３】

貢献度推定部１１は、受け取った１つ以上の区分化音声信号のそれぞれについて、短時間分析フレーム処理を行い、貢献度の時系列を算出する（ステップＳ２０３）。

【００５４】

話者特徴算出部１２は、受け取った１つ以上の音響特徴・音声統計量・貢献度の時系列に基づいて、話者認識特徴量を算出して出力する。（ステップＳ２０４）。音声処理装置２００は、外部からの音声信号の受理が終了したら、一連の処理を終了する。

【００５５】

（第２の実施形態の効果）

以上、説明したように、本実施形態にかかる音声処理装置２００によれば、音声処理装置２００が算出した話者特徴を用いる話者認識の精度を高めることができる。なぜならば、音声処理装置２００は、貢献度推定部１１が音声信号の品質を貢献度として算出し、話者特徴算出部１２が貢献度を考慮した特徴ベクトルを算出することで、音声信号の品質の高い部分区間に重きを置いた特徴ベクトルを出力するからである。

【００５６】

このように、本実施形態に係る音声処理装置２００は、音声信号に対して、各部分区間の品質に応じた貢献度を考慮した特徴ベクトルを算出する。これにより、音声信号の部分区間に、話者の不明瞭な発声、話者の咳や笑い声などの話し声とは異なる音、雑音などが混入している場合にも、話者認識に適した認識特徴量を求めることができる。

【００５７】

＜第３の実施形態＞

図５は、本発明の第３の実施形態に係る、音声処理装置の構成の一例を示すブロック図である。

【００５８】

図５に示すように、音声処理装置３００は、貢献度推定部１１と、話者特徴算出部１２と、属性認識部１３とを備える。音声処理装置３００は、属性情報を認識することのできる音声処理装置である。

【００５９】

貢献度推定部１１および話者特徴算出部１２については、第１および第２の実施形態と同様なので説明は省略する。

【００６０】

属性認識部１３は、話者特徴算出部１２から属性情報を認識するための認識特徴量を受け取る。属性認識部１３は、認識特徴量に基づいて、音声信号を発した話者、音声信号を構成する言語、音声信号に含まれる感情表現、音声信号から推定される発話者の性格等を認識する。具体的には、属性認識部１３は、例えば、認識特徴量を比較するための比較用音声データを格納する記憶装置（図示しない）を参照する。この場合、属性認識部１３は、認識特徴量と、比較用音声データの類似の度合い等を算出することで、属性情報を認識することができる。

【００６１】

＜第３の実施形態の具体例＞

次に、本発明の第３の実施形態に係る音声処理装置３００の具体的な応用例について説明する。

【００６２】

本発明の第３の実施形態に係る音声処理装置３００が算出した話者特徴は、音声信号の話者を推定する話者認識に利用可能である。例えば、第１の音声信号から算出した第１の話者特徴と、第２の音声信号から算出した第２の話者特徴とから、２つの話者特徴の類似性を現す指標として、コサイン類似度を算出する。例えば、話者照合することを目的とする場合は、前記の類似度に基づく照合可否の判定情報を出力してもよい。また、話者識別することを目的とする場合は、第１の音声信号に対して複数の第２の音声信号を用意して各々の類似度を求め、値の大きい組を出力してもよい。

【００６３】

本発明の第３の実施形態に係る音声処理装置３００は、音声信号から特定の属性情報を認識するための認識特徴量を算出する特徴算出装置の一例である。音声処理装置３００は、特定の属性が音声信号を発した話者であるとき、話者特徴抽出装置として利用可能である。また、音声処理装置３００は、例えば文発話の音声信号に対して、当該話者特徴を用いて推定した話者情報に基づいて、当該話者の話し方の特徴に適応化する機構を備える音声認識装置の一部としても利用可能である。また、ここで、話者を示す情報は、話者の性別を示す情報や、話者の年齢あるいは年齢層を示す情報であってもよい。

【００６４】

本発明の第３の実施形態に係る音声処理装置３００は、特定の属性を音声信号が伝える言語（音声信号を構成する言語）を示す情報とするとき、言語特徴算出装置として利用可能である。また、音声処理装置３００は、例えば文発話の音声信号に対して、当該言語特徴を用いて推定した言語情報に基づいて、翻訳する言語を選択する機構を備える音声翻訳装置の一部としても利用可能である。

【００６５】

本発明の第３の実施形態に係る音声処理装置３００は、特定の属性が話者の発話時の感情を示す情報であるとき、感情特徴算出装置として利用可能である。また、音声処理装置３００は、例えば蓄積された多数の発話の音声信号に対して、当該感情特徴を用いて推定した感情情報に基づいて、特定の感情に対応する音声信号を特定する機構を備える音声検索装置や音声表示装置の一部としても利用可能である。この感情情報には、例えば、感情表現を示す情報、発話者の性格を示す情報等が含まれる。

【００６６】

以上のように、本実施形態における特定の属性情報は、音声信号を発した話者、音声信号を構成する言語、音声信号に含まれる感情表現、音声信号から推定される発話者の性格、の少なくともいずれか一つを表す情報である。

【００６７】

（ハードウエア構成についての説明）

以上、実施形態を用いて本発明を説明したが、本発明は、上記実施形態に限定されるものではない。本発明の構成や詳細には、本発明のスコープ内で当業者が理解しうる様々な変更をすることができる。すなわち、本発明は、以上の実施形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

【００６８】

以上のように、本発明の一態様における音声処理装置等は、音声信号の品質を考慮した特徴ベクトルを抽出し話者認識の精度を高めることができるという効果を有しており、音声処理装置等および話者認識装置として有用である。なお、本発明において使用者に関する情報を取得、利用する場合は、これを適法に行うものとする。

【００６９】

＜その他の実施形態＞

音声処理装置は、ハードウエアによって実現してもよいし、ソフトウエアによって実現してもよい。また、音声処理装置は、ハードウエアとソフトウエアの組み合わせによって実現してもよい。

【００７０】

図６は、音声処理装置を構成する情報処理装置（コンピュータ）の一例を示すブロック図である。

【００７１】

図６に示すように、情報処理装置４００は、制御部（ＣＰＵ：Central Processing Unit）４１０と、記憶部４２０と、ＲＯＭ（Read Only Memory）４３０と、ＲＡＭ（Random Access Memory）４４０と、通信インターフェース４５０と、ユーザインターフェース４６０とを備えている。

【００７２】

制御部（ＣＰＵ）４１０は、記憶部４２０またはＲＯＭ４３０に格納されたプログラムをＲＡＭ４４０に展開して実行することで、音声処理装置および話者認識装置の各種の機能を実現することができる。また、制御部（ＣＰＵ）４１０は、データ等を一時的に格納できる内部バッファを備えていてもよい。

【００７３】

記憶部４２０は、各種のデータを保持できる大容量の記憶媒体であって、ＨＤＤ（Hard Disc Drive）、およびＳＳＤ（Solid State Drive）等の記憶媒体で実現することができる。また、記憶部４２０は、情報処理装置４００が通信インターフェース４５０を介して通信ネットワークと接続されている場合には、通信ネットワーク上に存在するクラウドストレージであってもよい。また、記憶部４２０は、制御部（ＣＰＵ）４１０が読み取り可能なプログラムを保持していてもよい。

【００７４】

ＲＯＭ４３０は、記憶部４２０と比べると小容量なフラッシュメモリ等で構成できる不揮発性の記憶装置である。また、ＲＯＭ４３０は、制御部（ＣＰＵ）４１０が読み取り可

能なプログラムを保持していてもよい。なお、制御部（ＣＰＵ）４１０が読み取り可能なプログラムは、記憶部４２０およびＲＯＭ４３０の少なくとも一方が保持していればよい。

【００７５】

なお、制御部（ＣＰＵ）４１０が読み取り可能なプログラムは、コンピュータが読み取り可能な様々な記憶媒体に非一時的に格納した状態で、情報処理装置４００に供給してもよい。このような記憶媒体は、例えば、磁気テープ、磁気ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－Ｒ／Ｗ、半導体メモリである。

【００７６】

ＲＡＭ４４０は、ＤＲＡＭ（Dynamic Random Access Memory）及びＳＲＡＭ（Static Random Access Memory）等の半導体メモリであり、データ等を一時的に格納する内部バッファとして用いることができる。

【００７７】

通信インターフェース４５０は、有線または無線を介して、情報処理装置４００と、通信ネットワークとを接続するインターフェースである。

【００７８】

ユーザインターフェース４６０は、例えば、ディスプレイ等の表示部、およびキーボード、マウス、タッチパネル等の入力部である。

【００７９】

上記の実施の形態の一部又は全部は、以下の付記のようにも記載され得るが以下には限られない。

【００８０】

［付記１］

音声を表す音声信号に基づき、前記音声信号の部分区間の品質の程度を表す貢献度を算出する貢献度推定手段と、

前記音声信号の前記部分区間の前記貢献度を、前記部分区間の重みとして用いて、前記音声信号から特定の属性情報を認識するための認識特徴量を算出する話者特徴算出手段とを備える、音声処理装置。

【００８１】

［付記２］

前記音声信号に含まれる音の種類の比率を表す音声統計量を算出する音声統計量算出手段をさらに備え、

前記話者特徴算出手段は、前記音声信号の前記音声統計量と、前記音声信号の前記貢献度とに基づいて、前記認識特徴量を算出する、付記１に記載の音声処理装置。

【００８２】

［付記３］

前記貢献度推定手段は、前記音声信号の前記貢献度として、

前記音声信号の一部が音声か否かを識別して算出した音声らしさを表す値、前記音声信号の一部が話者認識に正解する音声か否かを識別して算出した話者認識の正解しやすさを表す値、前記音声信号の一部が話者認識誤りを起こす音声か否かを識別して算出した話者認識の誤りやすさを表す値の少なくともいずれかひとつを算出する、付記１または２に記載の音声処理装置。

【００８３】

［付記４］

前記貢献度推定手段は、

ニューラルネットワークを用いて前記音声信号の前記貢献度を算出する、付記３に記載の音声処理装置。

【００８４】

［付記５］

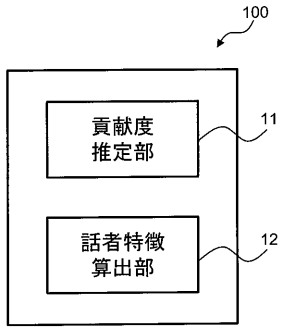
前記話者特徴算出手段は、

前記認識特徴量として i-vector を算出する、付記３または４に記載の音声処理装置。
【００８５】
［付記６］
前記認識特徴量に基づいて前記属性情報を認識する属性認識手段を備える、付記１～５のいずれか１つに記載の音声処理装置。
【００８６】
［付記７］
前記特定の属性情報は、
音声信号を発した話者、音声信号を構成する言語、音声信号に含まれる感情表現、音声信号から推定される話者の性格の少なくともいずれか１つを表す情報である、付記１～６のいずれか１つに記載の音声処理装置。
【００８７】
［付記８］
音声を表す音声信号に基づき、前記音声信号の部分区間の品質の程度を表す貢献度を算出し、
前記音声信号の前記部分区間の前記貢献度を、前記部分区間の重みとして用いて、前記音声信号から特定の属性情報を認識するための認識特徴量を算出する、音声処理方法。
【００８８】
［付記９］
前記音声信号に含まれる音の種類の比率を表す音声統計量をさらに算出し、
前記音声信号の前記音声統計量と、前記音声信号の前記貢献度とに基づいて、前記認識特徴量を算出する、付記８に記載の音声処理方法。
【００８９】
［付記１０］
前記音声信号の前記貢献度として、
前記音声信号の一部が音声か否かを識別して算出した音声らしさを表す値、前記音声信号の一部が話者認識に正解する音声か否かを識別して算出した話者認識の正解しやすさを表す値、前記音声信号の一部が話者認識誤りを起こす音声か否かを識別して算出した話者認識の誤りやすさを表す値の少なくともいずれかひとつを算出する、付記８または９に記載の音声処理方法。
【００９０】
［付記１１］
ニューラルネットワークを用いて前記音声信号の前記貢献度を算出する、付記１０に記載の音声処理方法。
【００９１】
［付記１２］
前記認識特徴量として i-vector を算出する、付記１０または１１に記載の音声処理方法。
【００９２】
［付記１３］
前記認識特徴量に基づいて前記属性情報を認識する、付記８～１２のいずれか１つに記載の音声処理方法。
【００９３】
［付記１４］
前記特定の属性情報は、
音声信号を発した話者、音声信号を構成する言語、音声信号に含まれる感情表現、音声信号から推定される話者の性格の少なくともいずれか１つを表す情報である、付記８～１３のいずれか１つに記載の音声処理方法。
【００９４】
［付記１５］

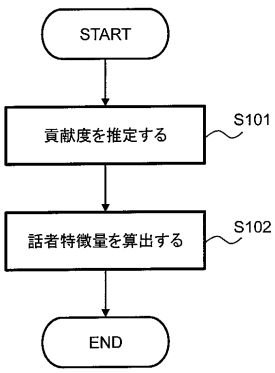
コンピュータに、
音声を表す音声信号に基づき、前記音声信号の部分区間の品質の程度を表す貢献度を算出する処理と、
前記音声信号の前記部分区間の前記貢献度を、前記部分区間の重みとして用いて、前記音声信号から特定の属性情報を認識するための認識特徴量を算出する処理とを実行させる、音声処理プログラム。
【００９５】
［付記１６］
前記コンピュータに、
前記音声信号に含まれる音の種類の比率を表す音声統計量をさらに算出する処理と、
前記音声信号の前記音声統計量と、前記音声信号の前記貢献度とに基づいて、前記認識特徴量を算出する処理とを実行させる、付記１５に記載の音声処理方法。
【００９６】
［付記１７］
前記コンピュータに、
前記音声信号の前記貢献度として、
前記音声信号の一部が音声か否かを識別して算出した音声らしさを表す値、前記音声信号の一部が話者認識に正解する音声か否かを識別して算出した話者認識の正解しやすさを表す値、および前記音声信号の一部が話者認識誤りを起こす音声か否かを識別して算出した話者認識の誤りやすさを表す値の少なくともいずれかひとつを算出する処理を実行させる、付記１５または１６に記載の音声処理プログラム。
【００９７】
［付記１８］
前記コンピュータに、
ニューラルネットワークを用いて前記音声信号の前記貢献度を算出する処理を実行させる、付記１７に記載の音声処理プログラム。
【００９８】
［付記１９］
前記コンピュータに、
前記認識特徴量として i-vector を算出する処理を実行させる、付記１７または１８に記載の音声処理プログラム。
【００９９】
［付記２０］
前記コンピュータに、
前記認識特徴量に基づいて前記属性情報を認識する処理を実行させる、付記１５～１９のいずれか１つに記載の音声処理プログラム。
【０１００】
［付記２１］
前記特定の属性情報は、
音声信号を発した話者、音声信号を構成する言語、音声信号に含まれる感情表現、音声信号から推定される話者の性格の少なくともいずれか１つを表す情報である、付記１５～２０のいずれか１つに記載の音声処理プログラム。
【符号の説明】
【０１０１】
１１・・・貢献度推定部
１２・・・話者特徴算出部
１３・・・属性認識部
２１・・・音声区間検出部
２２・・・音声統計量算出部
２３・・・貢献度記憶部

２４・・・貢献度学習部
１００,２００,３００・・・音声処理装置
４００・・・情報処理装置
４１０・・・制御部（ＣＰＵ）
４２０・・・記憶部
４３０・・・ＲＯＭ
４４０・・・ＲＡＭ
４５０・・・通信インターフェース
４６０・・・ユーザインターフェース

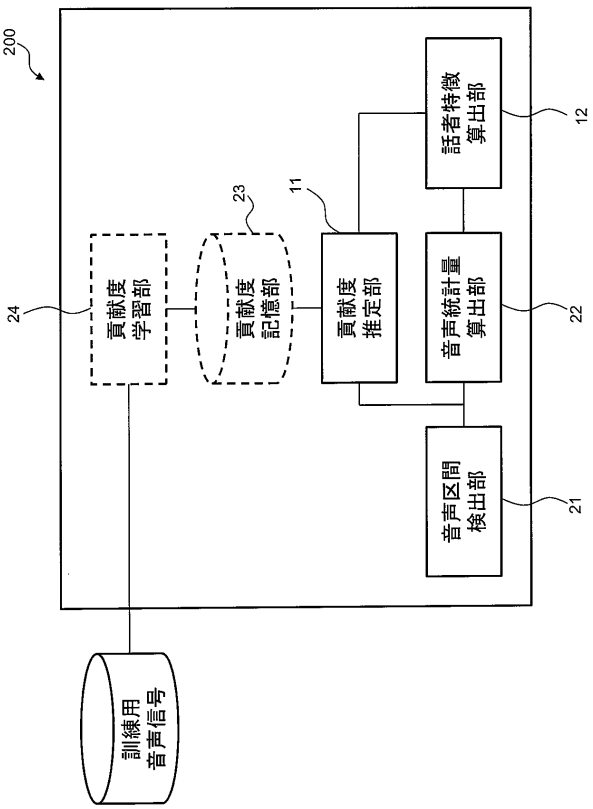
【図１】



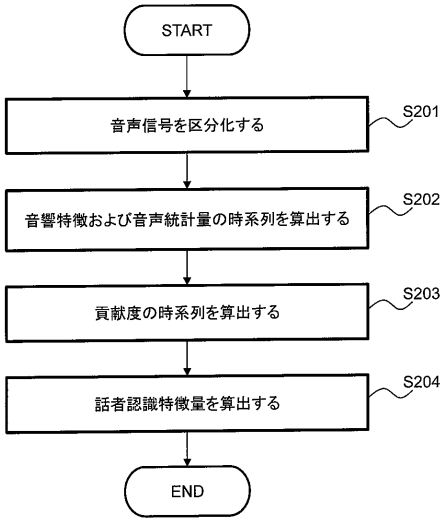
【図２】



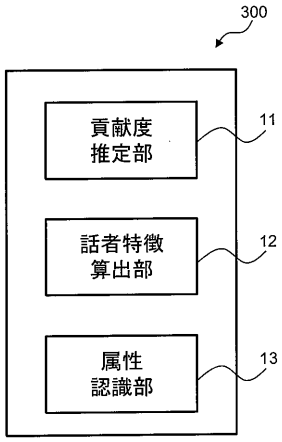
【図３】



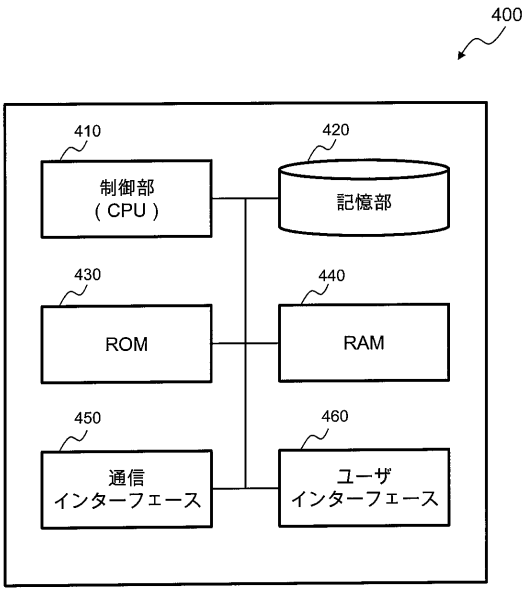
【図４】


START
音声信号を区分化する
S201
音響特徴および音声統計量の時系列を算出する
S202
貢献度の時系列を算出する
S203
話者認識特徴量を算出する
S204
END

【図５】


300
貢献度
推定部
11
話者特徴
算出部
12
属性
認識部
13

【図６】


400
410
制御部
(CPU)
420
記憶部
430
ROM
440
RAM
450
通信
インターフェース
460
ユーザ
インターフェース

フロントページの続き

(56)参考文献　特開２０１５－１８４３７８（ＪＰ，Ａ）
特開２００４－３４１３４０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１７／００