(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5672155号
(P5672155)

(45)発行日　平成27年2月18日(2015.2.18)　(24)登録日　平成27年1月9日(2015.1.9)

(51)Int.Cl.		F I		
G10L 15/04	(2013.01)	G10L	15/04	300A
G10L 15/02	(2006.01)	G10L	15/02	200D
G10L 25/93	(2013.01)	G10L	25/93	

請求項の数 4　(全 20 頁)

(21)出願番号　特願2011-122808 (P2011-122808)
(22)出願日　平成23年5月31日 (2011.5.31)
(65)公開番号　特開2012-252060 (P2012-252060A)
(43)公開日　平成24年12月20日 (2012.12.20)
審査請求日　平成26年3月4日 (2014.3.4)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100089118
弁理士　酒井　宏明
(72)発明者　張　霓
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

審査官　上田　雄

最終頁に続く

(54)【発明の名称】話者判別装置、話者判別プログラム及び話者判別方法

(57)【特許請求の範囲】

【請求項１】

各々の話者に配置される複数のマイクから各々の音声データを取得する取得部と、

前記取得部によって取得された音声データを所定の区間のフレームにフレーム化するフレーム化部と、

第１の確率モデルに基づいて、前記フレーム化部によってフレーム化されたフレームが有声音領域または無声音領域のいずれであるかを識別する第１の識別部と、

各音声データにおける同一区間のフレームで有声音領域が重複して識別された場合に、当該同一区間で有声音領域と識別されたフレームのうち最大のエネルギーを持つフレームの識別結果を有効化する有効化部と、

第２の確率モデルに基づいて、前記有効化部によって有効化された後のフレームの識別結果から各々の音声データにおける発話領域および沈黙領域を識別する第２の識別部と

を有することを特徴とする話者判別装置。

【請求項２】

前記有効化部は、当該同一区間で有声音領域と識別されたフレームのうち最大のエネルギーを持つフレーム以外の識別結果を無声音領域に置き換えることを特徴とする請求項１に記載の話者判別装置。

【請求項３】

コンピュータに、

各々の話者に配置される複数のマイクから各々の音声データを取得し、

取得された音声データを所定の区間のフレームにフレーム化し、
第１の確率モデルに基づいて、前記フレームが有声音領域または無声音領域のいずれであるかを識別し、
各音声データにおける同一区間のフレームで有声音領域が重複して識別された場合に、当該同一区間で有声音領域と識別されたフレームのうち最大のエネルギーを持つフレームの識別結果を有効化し、
第２の確率モデルに基づいて、有効化された後のフレームの識別結果から各々の音声データにおける発話領域および沈黙領域を識別する
各処理を実行させることを特徴とする話者判別プログラム。
【請求項４】
コンピュータが、
各々の話者に配置される複数のマイクから各々の音声データを取得し、
取得された音声データを所定の区間のフレームにフレーム化し、
第１の確率モデルに基づいて、前記フレームが有声音領域または無声音領域のいずれであるかを識別し、
各音声データにおける同一区間のフレームで有声音領域が重複して識別された場合に、当該同一区間で有声音領域と識別されたフレームのうち最大のエネルギーを持つフレームの識別結果を有効化し、
第２の確率モデルに基づいて、有効化された後のフレームの識別結果から各々の音声データにおける発話領域および沈黙領域を識別する
各処理を実行することを特徴とする話者判別方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、話者判別装置、話者判別プログラム及び話者判別方法に関する。
【背景技術】
【０００２】
複数の話者によってなされる会話の各場面において各話者のうち誰が発話しているのかを判別する技術が知られている。
【０００３】
かかる話者の判別を閾値判定により実現する技術の一例として、音声認識装置が挙げられる。この音声認識装置には、各参加者に対応してマイクロホンが接続される。このような構成の下、音声認識装置は、マイクロホンによって出力される音声信号のパワーがパワー閾値を超えてから下回るまでの区間の音声信号を音声認識の対象として記憶部の所定のエリアへ記録する。その上で、音声認識装置は、記憶部に記録した音声信号を音声認識した後に、発言者を特定するためのデータとしてマイクロホンの識別情報を紐付けて音声認識の結果を記憶部の議事録エリアへ記録する。
【０００４】
また、話者の判別を音源定位により実現する技術の一例としては、発話イベント分離システムが挙げられる。この発話イベント分離システムでは、それぞれ異なる方向に放射状に向けた複数のマイクロホンを有するマイクロホンアレイが用いられる。発話イベント分離システムは、音源定位のアルゴリズムを用いて、マイクロホンアレイによって収録された多チャネルの音声データを解析して時刻毎に音の到来方向を推定する。また、発話イベント分離システムは、音源となる話者の存在範囲を推定する。その上で、発話イベント分離システムは、音源定位の結果と、話者の存在範囲の推定結果から、時刻毎にどの話者が発話しているかを同定する。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２００８－３０９８５６号公報

【特許文献２】特開２００７－２３３２３９号公報
【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、上記の従来技術では、以下に説明するように、話者の判別を簡易かつ正確に行うことができないという問題がある。
【０００７】
例えば、上記の音声認識装置は、音声信号のパワーがパワー閾値を超過するか否かによって話者が発話しているか否かを判定するものである。このため、上記の音声認識装置では、話者を判別する精度はパワー閾値に依存するが、人間が発話する音声には個人差があるので、パワー閾値に適切な値を設定することは困難である。それゆえ、上記の音声認識装置では、話者の判別を正確に行うことができない。
【０００８】
また、上記の発話イベント分離システムでは、音源定位により音の到来方向を推定するのに複雑なアルゴリズムを使用する必要がある。さらに、上記の発話イベント分離システムでは、話者の存在範囲を推定するために、会議に参加する人数等を予め学習させておく必要もある。よって、上記の発話イベント分離システムでは、話者の判別を簡易に行うことはできない。
【０００９】
開示の技術は、上記に鑑みてなされたものであって、話者の判別を簡易かつ正確に行うことができる話者判別装置、話者判別プログラム及び話者判別方法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】
【００１０】
本願の開示する話者判別装置は、各々の話者に配置される複数のマイクから各々の音声データを取得する取得部を有する。さらに、前記話者判別装置は、前記取得部によって取得された音声データを所定の区間のフレームにフレーム化するフレーム化部を有する。さらに、前記話者判別装置は、第１の確率モデルに基づいて、前記フレーム化部によってフレーム化されたフレームが有声音領域または無声音領域のいずれであるかを識別する第１の識別部を有する。さらに、前記話者判別装置は、各音声データにおける同一区間のフレームで有声音領域が重複して識別された場合に、当該同一区間で有声音領域と識別されたフレームのうち最大のエネルギーを持つフレームの識別結果を有効化する有効化部を有する。さらに、前記話者判別装置は、第２の確率モデルに基づいて、前記有効化部によって有効化された後のフレームの識別結果から各々の音声データにおける発話領域および沈黙領域を識別する第２の識別部を有する。
【発明の効果】
【００１１】
本願の開示する話者判別装置の一つの態様によれば、話者の判別を簡易かつ正確に行うことができるという効果を奏する。
【図面の簡単な説明】
【００１２】
【図１】図１は、実施例１に係る会話分析装置の機能的構成を示すブロック図である。
【図２】図２は、有声音および無声音の一例を示す図である。
【図３】図３は、発話領域および沈黙領域の一例を示す図である。
【図４】図４は、話者判別方法を説明するための図である。
【図５】図５は、隠れマルコフモデルにおける状態遷移図の一例を示す図である。
【図６】図６は、有声音領域および無声音領域の識別結果の一例を示す図である。
【図７】図７は、図６に示した識別結果の置換結果の一例を示す図である。
【図８】図８は、隠れマルコフモデルにおける状態遷移図の一例を示す図である。
【図９】図９は、実施例１に係る会話分析処理の手順を示すフローチャートである。

【図１０】図１０は、実施例１に係る会話分析処理の手順を示すフローチャートである。
【図１１】図１１は、実施例１に係る有効化処理の手順を示すフローチャートである。
【図１２】図１２は、実施例１及び実施例２に係る話者判別プログラムを実行するコンピュータの一例について説明するための図である。
【発明を実施するための形態】
【００１３】
以下に、本願の開示する話者判別装置、話者判別プログラム及び話者判別方法の実施例を図面に基づいて詳細に説明する。なお、この実施例は開示の技術を限定するものではない。そして、各実施例は、処理内容を矛盾させない範囲で適宜組み合わせることが可能である。
【実施例１】
【００１４】
まず、本実施例に係る話者判別装置を含む会話分析装置の機能的構成について説明する。図１は、実施例１に係る会話分析装置の機能的構成を示すブロック図である。図１に示す会話分析装置１０は、話者Ａ、話者Ｂ及び話者Ｃにそれぞれ対応して設けられた接話マイク３０Ａ～３０Ｃを介して集音した複数の音声データから、話者Ａ～話者Ｃの会話に関する特性を抽出して会話スタイルを分析するものである。
【００１５】
この会話分析装置１０には、接話マイク３０Ａ～３０Ｃの３つのマイクが接続される。これら接話マイク３０Ａ～３０Ｃは、話者によって装着される接話型マイクロホン（close‐talking microphone）である。かかる接話マイクの一態様としては、ラペルマイクやヘッドセットマイクなどが挙げられる。以下では、接話マイク３０Ａ～３０Ｃのことを区別なく総称する場合には「接話マイク３０」と記載する場合がある。
【００１６】
なお、図１の例では、接話型マイクロホンを用いる場合を例示したが、必ずしも接話型マイクロホンを用いる必要はなく、各々の話者に他の話者よりも接近して配置するのであれば任意のマイクを採用できる。また、図１の例では、３つのマイクを用いて話者Ａ～話者Ｃの３人の会話を集音する場合を例示するが、２つのマイクを用いて２人の会話を集音することとしてもよいし、また、４つ以上のマイクを用いて４人以上の会話を集音することとしてもかまわない。
【００１７】
登録部３１は、接話マイク３０によって集音された音声信号を会話分析装置１０の記憶部１１へ登録する処理部である。一態様としては、登録部３１は、接話マイク３０から音声入力されたアナログ信号にＡ／Ｄ（Analog/Digital）変換を実行することによりデジタル信号に変換した上で音声記憶部１１へ登録する。なお、以下では、接話マイク３０Ａから音声入力されたアナログ信号がＡ／Ｄ変換されたデジタル信号のことを「第１の音声データ」と記載する場合がある。また、接話マイク３０Ｂから音声入力されたアナログ信号がＡ／Ｄ変換されたデジタル信号のことを「第２の音声データ」と記載する場合がある。さらに、接話マイク３０Ｃから音声入力されたアナログ信号がＡ／Ｄ変換されたデジタル信号のことを「第３の音声データ」と記載する場合がある。
【００１８】
図１に示すように、会話分析装置１０は、音声記憶部１１と、抽出部１３と、分析部１４とを有する。なお、会話分析装置１０は、図１に示した機能部以外にも既知のコンピュータが有する各種の機能部、例えば各種の入力デバイスや音声出力デバイスなどを始め、他の装置との通信を制御する通信インターフェースなどの機能部を有するものとする。
【００１９】
音声記憶部１１は、音声データを記憶する記憶部である。この音声記憶部１１は、第１の音声データ１２Ａと、第２の音声データ１２Ｂと、第３の音声データ１２Ｃとを記憶する。
【００２０】

これら第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃは、話者Ａ～話者Ｃが装着する接話マイク３０によって集音された音声信号がＡ／Ｄ変換されたデジタルデータである。このうち、第１の音声データ１２Ａには、話者Ａの音声だけでなく、話者Ｂおよび話者Ｃの音声も含み得るが、話者Ａから接話マイク３０Ａまでの距離が話者Ｂや話者Ｃに比べて接近している。よって、第１の音声データ１２Ａに含まれる音声は、話者Ａと話者Ｂや話者Ｃとの間で同時に発話がなされていた場合でも、話者Ａによって発話された音声のエネルギーが最も高くなる。同様に、第２の音声データ１２Ｂに含まれる音声は、話者Ｂによって発話された音声のエネルギーが最も高くなり、第３の音声データ１２Ｃに含まれる音声は、話者Ｃによって発話された音声のエネルギーが最も高くなる。

【００２１】

なお、上記の音声記憶部１１などの記憶部には、半導体メモリ素子や記憶装置を採用できる。例えば、半導体メモリ素子としては、ＶＲＡＭ（Video Random Access Memory）、ＲＡＭ（Random Access Memory)、ＲＯＭ（Read Only Memory）やフラッシュメモリ（flash memory）などが挙げられる。また、記憶装置としては、ハードディスク、光ディスクなどの記憶装置が挙げられる。

【００２２】

ここで、話者によって発話される有声音および無声音について説明する。図２は、有声音および無声音の一例を示す図である。図２の例では、サンプリング周波数が１６ｋＨｚである接話マイクを用いて取得した音声データが示されている。図２の例では、横軸は時間を示し、縦軸は周波数を示し、図中の濃淡はスペクトルエントロピーの大小を示す。

【００２３】

図２に示すように、有声音Ｖ（Voiced）は、スペクトルエントロピーの変化が大きく、無声音Ｕ（Unvoiced）よりも低い周波数の音である。有声音の一例としては、母音「ａ」、「ｉ」、「ｕ」、「ｅ」、「ｏ」などが挙げられる。また、無声音Ｕは、有声音Ｖよりも高い周波数の音である。無声音の一例としては、母音以外の音、例えば「ｓ」、「ｐ」、「ｈ」などが挙げられる。これら有声音および無声音の特徴は、話者によって発話される言語に依存せず、日本語、英語や中国語などの任意の言語において共通する。

【００２４】

次に、有声音および無声音と発話領域および沈黙領域との関係について説明する。図３は、発話領域および沈黙領域の一例を示す図である。発話領域は、話者によって発話がなされている領域を指し、無声音領域および有声音領域を含む。なお、図３の例では、話者によって「ＷａＴａＳｈｉＷａ　Ｃｈｏｕ　ＤｅＳｕ」と発話された場合を示す。

【００２５】

図３に示す発話の例では、「ＷａＴａＳｈｉＷａ」の発話領域４０と、「Ｃｈｏｕ」の発話領域４１と、「ＤｅＳｕ」の発話領域４２との間に、沈黙領域４３および沈黙領域４４が存在することを示す。このうち、発話領域４０には、無声音「Ｗ」、有声音「ａ」、無声音「Ｔ」、有声音「ａ」、無声音「Ｓｈ」、有声音「ｉ」、無声音「Ｗ」、有声音「ａ」が含まれる。また、発話領域４１には、無声音「Ｃｈ」、有声音「ｏｕ」が含まれる。さらに、発話領域４２には、無声音「Ｄ」、有声音「ｅ」、無声音「Ｓ」、有声音「ｕ」が含まれる。

【００２６】

図１の説明に戻り、会話分析装置１０は、複数の話者によってなされる会話の各場面において各話者のうち誰が発話しているのかを判別する話者判別装置５０を有する。

【００２７】

ここで、本実施例に係る話者判別装置５０は、接話マイク３０Ａ～３０Ｃから第１の音声データ、第２の音声データ及び第３の音声データを取得する。さらに、本実施例に係る話者判別装置５０は、第１の音声データ、第２の音声データ及び第３の音声データを所定の区間のフレームにフレーム化する。さらに、本実施例に係る話者判別装置５０は、第１の確率モデルに基づいて、フレームが有声音領域または無声音領域のいずれであるかを識

別する。さらに、本実施例に係る話者判別装置５０は、各音声データにおける同一区間のフレームで有声音領域が重複して識別された場合に、当該同一区間で有声音領域と識別されたフレームのうち最大のエネルギーを持つフレームの識別結果を有効化する。その上で、本実施例に係る話者判別装置５０は、第２の確率モデルに基づいて、有効化された後のフレームの識別結果から各々の音声データにおける発話領域および沈黙領域を識別する。

【００２８】

図４を用いて、上記の話者判別方法について説明する。図４は、話者判別方法を説明するための図である。図４の上段には、各フレームの有声音領域または無声音領域の識別結果が図示されている。図４の中段には、最大のエネルギーを持つフレームの有声音領域の識別結果が有効化された後の各フレームの識別結果が図示されている。図４の下段には、各々の音声データにおける発話領域および沈黙領域の識別結果が図示されている。

【００２９】

図４の上段に示すように、話者判別装置５０は、第１の確率モデルに基づいて、第１の音声データ、第２の音声データ及び第３の音声データからフレーム化した各フレームが有声音領域または無声音領域のいずれであるかを識別する。ここで、図４の例では、記号「●」、記号「○」、記号「△」がそれぞれ音声データのフレームを表し、記号「●」及び記号「○」が有声音領域であることを示し、記号「△」が無声音領域であることを示す。図４に示す記号「●」のフレームは、図４に示す記号「○」のフレームよりも高いエネルギーを有することを示す。これら第１の音声データ、第２の音声データおよび第３の音声データの識別結果からは、話者Ａ～話者Ｃのうち話者Ｂと話者Ｃが会話しており、話者Ｂが話者Ｃよりも大声で発話していることが推定できる。なお、以下では、第１の音声データから得られた各フレームのことを観測順に第１フレーム（１）・・・第１フレーム（ｎ）と記載する場合がある。また、第２の音声データから得られた各フレームのことを観測順に第２フレーム（１）・・・第２フレーム（ｍ）と記載する場合がある。さらに、第３の音声データから得られた各フレームのことを観測順に第３フレーム（１）・・・第３フレーム（ｍ）と記載する場合がある。

【００３０】

また、図４の中段に示すように、話者判別装置５０は、各音声データの同一の区間のフレームで有声音領域が重複する場合に、最大のエネルギーを持つフレームの識別結果を有効化する。この例では、第２の音声データ及び第３の音声データを構成するフレームのうち、下記のように、同一区間のフレームで互いに識別結果が有声音領域と識別されている。すなわち、第２フレーム（１）と第３フレーム（１）、第２フレーム（６）と第３フレーム（６）、第２フレーム（１０）と第３フレーム（１０）において互いの識別結果が有声音領域と識別されている。さらに、第２フレーム（１３）と第３フレーム（１３）、第２フレーム（１８）と第３フレーム（１８）において互いの識別結果が有声音領域と識別されている。この場合には、いずれのフレームについても第２の音声データのエネルギーの方が高いので、第３フレーム（１）、第３フレーム（６）、第３フレーム（１０）、第３フレーム（１３）及び第３フレーム（１８）の識別結果が有声音から無声音に置き換えられる。

【００３１】

さらに、図４の下段に示すように、話者判別装置５０は、第２の確率モデルに基づいて、有効化後のフレームの識別結果から各々の音声データにおける発話領域および沈黙領域を識別する。この例では、第２の音声データのフレームのうち下線が引かれた領域が話者Ｂの発話領域として識別されている。さらに、第３の音声データのフレームのうち下線が引かれた領域が話者Ｃの発話領域として識別されている。この場合には、話者Ｂの発話領域と話者Ｃの発話領域が重複するフレーム、すなわち第２フレーム（７）～第２フレーム（１３）の区間が同時発話として判別される。

【００３２】

このように、本実施例に係る話者判別装置５０は、各音声データにおける同一区間のフレームで有声音領域が重複して識別された場合に、最大のエネルギーを持つフレームの識

別結果だけを有効化して各々の音声データの発話領域および沈黙領域を識別する。このため、本実施例に係る話者判別装置５０は、各音声データを構成する同一区間のフレーム間で閾値を用いて判定せずとも、話者を判別することができる。さらに、本実施例に係る話者判別装置５０では、話者の判別に複雑なアルゴリズムを用いる必要はなく、事前に学習を行う必要もない。したがって、本実施例に係る話者判別装置５０によれば、話者の判別を簡易かつ正確に行うことができる。

【００３３】

また、本実施例に係る話者判別装置５０は、各音声データにおける同一区間のフレームで有声音領域が単独で識別された場合には、エネルギーの大小に関係なく、有声音領域と識別された識別結果を維持する。一般に、発話は、有声音と無声音が混在して構成されるので、複数の話者によって同時に発話された場合でも、同時発話で有声音領域が完全に重複する可能性は低く、有声音領域が単独で識別される機会が残る可能性は高い。例えば、図４の下段の例で言えば、話者Ｃの発話の音量が話者Ｂの発話の音量よりも低くても、第３フレーム（７）、第３フレーム（９）及び第３フレーム（１２）の識別結果は有声音のまま維持される。それゆえ、本実施例に係る話者判別装置５０では、話者が発話する音量に開きがある場合でも、同時発話を判別することもできる。

【００３４】

さらに、話者判別装置５０を詳細に説明する。図１に示すように、話者判別装置５０は、取得部５１と、フレーム化部５２と、第１の識別部５３と、有効化部５４と、第２の識別部５５とを有する。

【００３５】

取得部５１は、第１の音声データ、第２の音声データおよび第３の音声データを取得する処理部である。一態様としては、取得部５１は、音声記憶部１１に記憶された第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃを読み出す。他の一態様としては、取得部５１は、登録部３１によってＡ／Ｄ変換された第１の音声データ、第２の音声データおよび第３の音声データをストリームデータとして取得することもできる。更なる一態様としては、取得部５１は、ネットワークを介して図示しない外部装置から第１の音声データ、第２の音声データおよび第３の音声データを取得することもできる。

【００３６】

フレーム化部５２は、取得部５１によって取得された第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃを所定の区間のフレームにフレーム化する処理部である。一態様としては、フレーム化部５２は、第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃそれぞれの長さを比較する。そして、フレーム化部５２は、第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃの長さの差が許容誤差範囲内でない場合には、図示しない表示部等にエラーメッセージを出力し、以降の処理を行わない。一方、フレーム化部５２は、第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データの長さが同一であるか、あるいは許容誤差範囲内である場合には、下記のような処理を実行する。すなわち、フレーム化部５２は、第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃをフレーム化する。一例を挙げれば、フレーム化部５２は、下記の式（１）、式（２）を用いて、各々の音声データを、長さを２５６ｍｓとするフレーム化を行う。このとき、フレーム化部５２は、前後のフレームの重複部分の長さが１２８ｍｓとなるようにする。なお、上記のフレームの長さ、前後のフレームの重複部分の長さは、あくまでも一例であり、任意の値を採用できる。

Ｓ＝ｆｌｏｏｒ（Ｙ／Ｘ）・・・・・・・・・・・・・・・・・式（１）

ｍ＝ｆｌｏｏｒ（（Ｓ－２５６）／１２８）＋１・・・・・・・・式（２）

なお、「ｆｌｏｏｒ（ｘ）」は、ｘ以下の最大の整数を算出するための関数であり、Ｙは、第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃそれぞれのデータ量（byte）であり、Ｘは、１（byte）のデータに対応する長さ（ms）である。

【００３７】
第１の識別部５３は、第１の確率モデルに基づいて、フレーム化部５２によってフレーム化されたフレームが有声音領域または無声音領域のいずれであるかを識別する処理部である。一態様としては、第１の識別部５３は、第１フレーム（１）～第１フレーム（ｍ）、第２フレーム（１）～第２フレーム（ｍ）、第３フレーム（１）～第３フレーム（ｍ）の各々の音声データごとに、下記の処理を実行する。すなわち、第１の識別部５３は、自己相関係数のピークの数、自己相関係数のピークの最大値及びスペクトルエントロピーの３つの特徴量を抽出する。さらに、第１の識別部５３は、先に抽出した３つの特徴量それぞれの平均値および標準偏差を各々の音声データごとに算出する。その上で、第１の識別部５３は、確率モデルである隠れマルコフモデル（Hidden Markov Model；HMM）を用いて、有声音領域および無声音領域を各々の音声データごとに識別する。

【００３８】
ここで、有声音領域および無声音領域の識別方法について説明する。図５は、隠れマルコフモデルにおける状態遷移図の一例を示す図である。図５に示すように、第１の識別部５３は、上記の３つの特徴量、並びに、各特徴量の平均値および標準偏差を観測結果（observation）とし、ＥＭ法（Expectation-Maximization algorithm）を用いて、状態遷移確率（transition possibility）P_tを算出する。

【００３９】
かかる状態遷移確率P_tは、例えば、有声音の状態のままでいる確率、有声音の状態から無声音の状態に遷移する確率、無声音の状態のままでいる確率、無声音の状態から有声音の状態に遷移する確率を指す。図５に示す例で言えば、発話は、有声音および無声音の両方とも同一の確率で開始すると仮定して、発話の開始における有声音および無声音の状態の確率がいずれも「０．５」と設定されている。さらに、初期の状態遷移確率P_tとして、有声音の状態のままでいる確率が「０．９５」に設定されるとともに、有声音の状態から無声音の状態に遷移する確率が「０．０５」に設定されている。さらに、初期の状態遷移確率P_tとして、無声音の状態のままでいる確率が「０．９５」に設定されるとともに、無声音の状態から有声音の状態に遷移する確率が「０．０５」に設定されている。このような設定の下、第１の識別部５３は、状態遷移確率P_tを算出することを所定回数繰り返す。これによって、精度の高い状態遷移確率P_tを算出することができる。

【００４０】
さらに、第１の識別部５３は、上記の３つの特徴量、並びに、各特徴量の平均値および標準偏差を観測結果とし、ビタビアルゴリズム（Viterbi algorithm）により、観測確率（observation possibility）P_oを各々の音声データごとに算出する。ここで、観測確率P_oは、例えば、有声音の状態から観測（observed）を出力する確率、有声音の状態から非観測（not observed）を出力する確率、無声音の状態から観測を出力する確率および無声音の状態から非観測を出力する確率である。なお、観測確率は、出力確率（emission possibility）とも称される。

【００４１】
これら状態遷移確率P_tおよび観測確率P_oを算出した後に、第１の識別部５３は、上記の３つの特徴量に基づいて、ビタビアルゴリズムを用いて、次のような処理を実行する。すなわち、第１の識別部５３は、発話が行われている各フレームにおいて発話されている音が有声音であるか、あるいは無声音であるかを識別する。その上で、第１の識別部５３は、有声音と識別された領域を有声音領域とし、無声音と識別された領域を無声音領域とする。

【００４２】
このように、第１の識別部５３は、自己相関係数のピークの数、自己相関係数のピークの最大値及びスペクトルエントロピーなどの特徴量を用いて、有声音領域および無声音領域を識別する。したがって、第１の識別部５３では、周囲のノイズの影響によって有声音領域および無声音領域を識別する精度が低下することを抑制できる。また、第１の識別部５３は、周囲のノイズに強い特徴量を用いるため、第１の音声データ１２Ａ、第２の音声

データ１２Ｂ及び第３の音声データ１２Ｃをフレーム化する場合に、フレームの個数をより少なくすることができる。それゆえ、第１の識別部５３では、より簡易な処理で有声音領域および無声音領域を識別できる。

【００４３】

有効化部５４は、各音声データにおける同一区間のフレームで有声音領域が重複する場合に、当該同一区間で有声音領域と識別されたフレームのうち最大のエネルギーを持つフレームの識別結果を有効化する。

【００４４】

一態様としては、有効化部５４は、各音声データにおける同一区間のフレームで第１の識別部５３による識別結果を比較する。このとき、有効化部５４は、同一区間のフレームで有声音領域が重複する場合に、当該有声音領域と識別されたフレームのエネルギーを演算する。そして、有効化部５４は、当該同一区間で有声音領域と識別されたフレームのうち最大エネルギーを持つフレームを特定する。その上で、有効化部５４は、最大エネルギーを持つフレーム以外の識別結果を有声音領域から無声音領域に置き換える。その後、有効化部５４は、各音声データ間で同一区間のフレームを全て処理するまで、識別結果の比較、フレームの特定、識別結果の置き換えを繰り返し実行する。なお、上記のエネルギーは、各々の音声データのフレームに高速フーリエ変換、いわゆるＦＦＴ（Fast Fourier Transform）を実行して周波数解析を行った上で周波数成分ごとの振幅値を平均化することにより算出される。

【００４５】

ここで、有効化部５４による識別結果の置換要領について説明する。図６は、有声音領域および無声音領域の識別結果の一例を示す図である。図７は、図６に示した識別結果の置換結果の一例を示す図である。図６に示すように、「１２時００分００．０００秒」から「１２時００分００．０１０秒」までの区間では、第１フレーム、第２フレーム及び第３フレームの全ての識別結果が有声音領域と識別されている。この場合には、有効化部５４は、図７に示すように、第１フレーム、第２フレーム及び第３フレームのうちエネルギーが最大である第１フレームを除き、第２フレーム及び第３フレームの識別結果を有声音領域「Ｖ」から無声音領域「Ｕ」へ置き換える。また、図６に示す「１２時００分００．０１０秒」から「１２時００分００．０２０秒」までの区間では、第１フレーム及び第２フレームの識別結果が有声音領域と識別されている。この場合には、有効化部５４は、図７に示すように、第１フレーム及び第２フレームのうちエネルギーが最大である第１フレームの識別結果を維持する一方で、最大でない第２フレームの識別結果を有声音領域「Ｖ」から無声音領域「Ｕ」へ置き換える。さらに、図６に示すように、「１２時００分００．０２０秒」から「１２時００分００．０３０秒」までの区間では、第２フレームの識別結果だけが有声音領域と識別されている。この場合には、有効化部５４は、同一区間のフレームで有声音領域が重複しないので、図７に示すように、第２フレームの識別結果を維持する。

【００４６】

第２の識別部５５は、第２の確率モデルに基づいて、有効化部５４による有効化がなされた後のフレームの識別結果から各々の音声データにおける発話領域および沈黙領域を識別する処理部である。

【００４７】

ここで、発話領域および沈黙領域の識別方法について説明する。図８は、隠れマルコフモデルにおける状態遷移図の一例を示す図である。図８に示す状態遷移確率P_tおよび観測確率P_oは、予め定められた値である。かかる状態遷移確率P_tは、例えば、沈黙の状態である沈黙状態のままでいる確率、沈黙状態から発話の状態である発話状態に遷移する確率、発話状態のままでいる確率および発話状態から沈黙状態に遷移する確率を示す。図８に示す例で言えば、発話は、有声音および無声音の両方とも同一の確率で開始すると仮定して、発話の開始における沈黙状態および発話状態の確率がいずれも「０．５」に設定されている。また、状態遷移確率P_tとして、沈黙状態のままでいる確率が「０．９９９

」に設定されるとともに、沈黙状態から発話状態に遷移する確率が「０．００１」に設定されている。さらに、状態遷移確率P_tとして、発話状態のままでいる確率が「０．９９９」設定されるとともに、発話状態から沈黙状態に遷移する確率が「０．００１」に設定されている。

【００４８】

また、観測確率P_oは、例えば、沈黙状態において無声音が検出される確率、沈黙状態において有声音が検出される確率、発話状態において無声音が検出される確率、および発話状態において有声音が検出される確率を指す。図８の例で言えば、観測確率P_oとして、沈黙状態において無声音が検出される確率が「０．９９」に設定されるとともに、沈黙状態において有声音が検出される確率が「０．０１」に設定されている。また、観測確率P_oとして、発話状態において無声音が検出される確率が「０．５」に設定されるとともに、発話状態において有声音が検出される確率が「０．５」に設定されている。

【００４９】

なお、図８の例では、発話状態において無声音が検出される確率および発話状態において有声音が検出される確率をともに「０．５」に設定する場合を例示したが、同時発話の場合には他の話者よりも音量が小さい発話を行う話者の無声音が増加することも想定される。よって、発話状態において無声音が検出される確率を「０．５」よりも大きく設定することにより、他の話者よりも音量が小さい発話を行う話者の無声音の増加を抑制することもできる。

【００５０】

このような設定の下、第２の識別部５５は、ビタビアルゴリズムを用いて、有効化部５４による有効化がなされた後の有声音および無声音から、各々の音声データにおける沈黙領域および発話領域であるかを識別する。これによって、第１の音声データにおける話者Ａの発話領域および沈黙領域、第２の音声データにおける話者Ｂの発話領域および沈黙領域、さらには、第３の音声データにおける話者Ｃの発話領域および沈黙領域が識別される。

【００５１】

会話分析装置１０の説明に戻り、抽出部１３は、各々の音声データから会話特性を抽出する処理部である。一態様としては、抽出部１３は、第２の識別部５５によって識別された第１の音声データにおける話者Ａの発話領域をもとに有声音領域の数、有声音領域の長さの平均値および有声音領域の長さの標準偏差を算出する。また、抽出部１３は、第２の識別部５５によって識別された第１の音声データにおける話者Ａの発話領域をもとに発話領域の数、発話領域の長さの平均値および発話領域の長さの標準偏差を算出する。さらに、抽出部１３は、第２の識別部５５によって識別された第１の音声データにおける話者Ａの沈黙領域をもとに、沈黙領域の数、沈黙領域の長さの平均値および沈黙領域の長さの標準偏差を算出する。

【００５２】

また、抽出部１３は、会話全体の時間の長さに対する話者Ａの発話時間の長さの割合を算出する。このとき、抽出部１３は、話者Ａの発話領域の長さの合計を、話者Ａの発話時間の長さとして、上記の割合を算出する。また、抽出部１３は、話者Ｂの発話時間に対する話者Ａの発話時間の割合を算出する。さらに、抽出部１３は、話者Ｃの発話時間に対する話者Ａの発話時間の割合も算出する。また、抽出部１３は、話者Ａの発話領域をもとに、音量の標準偏差およびスペクトルエントロピーの標準偏差を算出する。さらに、抽出部１３は、話者Ａの発話領域をもとに算出した音量の標準偏差と、スペクトルエントロピーの標準偏差との和を、変化の度合いとして算出する。なお、ここでは、話者Ａの会話特性を抽出する場合を例示したが、話者Ｂおよび話者Ｃについても、上記の話者Ａと同様にして、会話特性を抽出する。

【００５３】

このようにして算出された有声音領域の数、有声音領域の長さの平均値および有声音領域の長さの標準偏差の各会話特性は、有声音の長さがどの位長いのかを示す指標となる。

また、発話領域の数、発話領域の長さの平均値、および発話領域の長さの標準偏差の各会話特性は、対応する人物が、常に会話において長く続けて話すのか、あるいは少ししか話さないのかを示す指標となる。また、沈黙領域の数、沈黙領域の長さの平均値および沈黙領域の長さの標準偏差の各会話特性は、話者の話し方が、長く続けて話すのか、あるいは中断（沈黙）を多くはさみながら話すのかを示す指標となる。また、会話全体の時間の長さに対するある人物の発話時間の長さの割合および他の人物の発話時間に対するある人物の発話時間の割合R_tの各会話特性は、会話の参加状態を示す指標となる。また、音量の標準偏差、スペクトルエントロピーの標準偏差および変化の度合いの各会話特性は、感情の変化が激しい情熱的な話者であるのか、あるいは感情の変化が小さい静かな話者であるのかを示す指標となる。

【００５４】

分析部１４は、抽出部１３によって抽出された会話特性に基づいて、会話スタイルを分析する処理部である。一態様としては、分析部１４は、他の人物の発話時間に対するある人物の発話時間の割合R_tが、所定値、例えば１．５以上である場合には、この「ある人物」は、会話においてよく話す人物であると分析する。また、分析部１４は、割合R_tが所定値、例えば０．６６以下である場合には、この「ある人物」は、会話においてあまり話さない、いわゆる聞き役の人物であると分析する。なお、分析部１４は、割合R_tが、所定値、例えば０．６６より大きく、１．５未満である場合には、会話に対する参加状況において両者は対等であると分析する。

【００５５】

他の一態様としては、分析部１４は、ある人物の発話領域の数に対する有声音領域の数の割合および発話領域の長さの平均値が、他の人物の発話領域の数に対する有声音領域の数の割合および発話領域の長さの平均値よりも大きい場合には、次のように分析する。すなわち、分析部１４は、「ある人物」は会話において長く続けて話しがちな人物であると分析する。また、分析部１４は、ある人物の沈黙領域の長さの平均値が他の人物の沈黙領域の長さの平均値よりも大きく、かつある人物の沈黙領域の長さの標準偏差が所定値、例えば、６．０以上である場合には、次のように分析する。すなわち、分析部１４は、「ある人物」は、相手の話を聞いて、相手の内容に合わせて自分の発話を中断するため、発話の長さが一定しない人物であると分析する。

【００５６】

更なる一態様としては、分析部１４は、ある人物の音量の標準偏差、スペクトルエントロピーの標準偏差または変化の度合いが、それぞれに対応する基準値以上である場合には、「ある人物」は感情の変化が激しい情熱的な話者であると分析する。また、分析部１４は、ある人物の音量の標準偏差、スペクトルエントロピーの標準偏差または変化の度合いが、それぞれに対応する基準値未満である場合には、「ある人物」は感情の変化が小さい静かな話者であると分析する。

【００５７】

他の一態様としては、分析部１４は、ある人物と他の人物との関係を分析することもできる。例えば、分析部１４は、他の人物の発話時間に対するある人物の発話時間の割合R_tが所定値、例えば１．０以上である場合には、「ある人物」は「他の人物」に対してよく話しかけているため、ある人物と他の人物との関係が友達や家族であると分析できる。一方、割合R_tが所定値、例えば１．０未満である場合には、この「ある人物」は「他の人物」の話を聞こうとしているため、ある人物と他の人物との関係が会社の同僚やビジネスパートナーであると分析できる。

【００５８】

更なる一態様としては、分析部１４は、ある人物と他の人物との会話においてある人物の発話領域の長さの平均値が所定値、例えば、１．８５（ｓ）以上である場合には、ある人物と他の人物との関係が友達や家族であると分析できる。これは、「ある人物」が「他の人物」に対してよく話しかけているためである。一方、分析部１４は、ある人物と他の人物との会話においてある人物の発話領域の長さの平均値が所定値、例えば、１．８５（

ｓ）未満である場合には、ある人物と他の人物との関係が会社の同僚やビジネスパートナーであると分析できる。

【００５９】

他の一態様としては、分析部１４は、ある人物と他の人物との会話においてある人物の沈黙領域の長さの平均値が所定値、例えば、３．００（ｓ）以下である場合には、同様の理由で、ある人物と他の人物との関係が友達や家族であると分析できる。一方、分析部１４は、ある人物の沈黙領域の長さの平均値が所定値、例えば、３．００（ｓ）より大きい場合には、ある人物と他の人物との関係が会社の同僚やビジネスパートナーであると分析できる。

【００６０】

更なる一態様としては、分析部１４は、ある人物と他の人物との会話においてある人物の変化の度合いが所定値、例えば、０．３３以上である場合には、同様の理由で、ある人物と他の人物との関係が友達や家族であると分析できる。一方、分析部１４は、ある人物の変化の度合いが所定値、例えば、０．３３未満である場合には、ある人物と他の人物との関係が会社の同僚やビジネスパートナーであると分析できる。

【００６１】

これらの分析を行った後に、分析部１４は、分析結果を所定の出力先の装置、例えば会話分析装置１０が有する表示部や話者Ａ～話者Ｃが利用する情報処理装置などに出力することができる。

【００６２】

なお、話者判別装置５０、抽出部１３及び分析部１４には、各種の集積回路や電子回路を採用できる。また、話者判別装置５０に含まれる機能部の一部を別の集積回路や電子回路とすることもできる。例えば、集積回路としては、ＡＳＩＣ（Application Specific Integrated Circuit）が挙げられる。また、電子回路としては、ＣＰＵ（Central Processing Unit）やＭＰＵ（Micro Processing Unit）などが挙げられる。

【００６３】

続いて、本実施例に係る会話分析装置の処理の流れについて説明する。なお、ここでは、会話分析装置１０によって実行される（１）会話分析処理を説明した後に、話者判別装置５０によって実行される（２）有効化処理を説明する。

【００６４】

（１）会話分析処理

図９及び図１０は、実施例１に係る会話分析処理の手順を示すフローチャートである。この会話分析処理は、一例として、図示しない入力部から会話分析処理を実行する指示を受け付けた場合に処理が起動する。

【００６５】

図９に示すように、取得部５１は、第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃを取得する（ステップＳ１０１）。そして、フレーム化部５２は、第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃそれぞれの長さが同一であるか否かを判定する（ステップＳ１０２）。なお、ここで言う「同一」は、長さの差が許容誤差範囲内である場合も含む。

【００６６】

このとき、各々の音声データの長さが同一でない場合（ステップＳ１０２否定）には、フレーム化部５２は、エラーメッセージを図示しない表示部に出力し（ステップＳ１０３）、処理を終了する。

【００６７】

一方、各々の音声データの長さが同一である場合（ステップＳ１０２肯定）には、フレーム化部５２は、第１の音声データ１２Ａ、第２の音声データ１２Ｂ及び第３の音声データ１２Ｃをフレーム化する（ステップＳ１０４）。

【００６８】

その後、第１の識別部５３は、自己相関係数のピークの数、自己相関係数のピークの最

大値およびスペクトルエントロピーの３つの特徴量を各々の音声データごとに抽出する（ステップＳ１０５）。そして、第１の識別部５３は、各々の音声データごとに抽出した３つの特徴量それぞれの平均値および標準偏差を算出する（ステップＳ１０６）。

【００６９】

続いて、第１の識別部５３は、変数Ｎに０を設定し（ステップＳ１０７）、隠れマルコフモデルにおける有声音および無声音の状態遷移について初期の状態遷移確率$Ｐ_{t}$を設定する（ステップＳ１０８）。

【００７０】

そして、第１の識別部５３は、変数Ｎの値を１つインクリメントする（ステップＳ１０９）。このとき、変数Ｎの値が５以上でない場合（ステップＳ１１０否定）には、第１の識別部５３は、各々の音声データごとに抽出した上記の３つの特徴量、並びに、各特徴量の平均値および標準偏差を観測結果とし、ＥＭ法を用いて、状態遷移確率$Ｐ_{t}$を算出し（ステップＳ１１１）、ステップＳ１０９へ移行する。

【００７１】

一方、変数Ｎの値が５以上である場合（ステップＳ１１０肯定）には、第１の識別部５３は、各々の音声データごとに抽出した上記の３つの特徴量、並びに、各特徴量の平均値および標準偏差を観測結果とし、ＥＭ法を用いて、状態遷移確率$Ｐ_{t}$を算出する（ステップＳ１１２）。

【００７２】

そして、第１の識別部５３は、各々の音声データごとに抽出した上記の３つの特徴量、並びに、各特徴量の平均値および標準偏差を観測結果とし、ビタビアルゴリズムを用いて、観測確率$Ｐ_{o}$を算出する（ステップＳ１１３）。

【００７３】

その後、第１の識別部５３は、各々の音声データごとに抽出した上記の３つの特徴量に基づいて、ビタビアルゴリズムを用いて、次のような処理を行う。すなわち、第１の識別部５３は、発話が行われている各フレームにおいて、発話されている音が有声音であるか、あるいは無声音であるかを識別する。そして、第１の識別部５３は、有声音が検出された領域を有声音領域とし、無声音が検出された領域を無声音領域とする（ステップＳ１１４）。

【００７４】

ここで、有効化部５４は、各音声データにおける同一区間のフレームで有声音領域が重複する場合に、当該同一区間で有声音領域と識別されたフレームのうち最大のエネルギーを持つフレームの識別結果を有効化する「有効化処理」を実行する（ステップＳ１１５）。

【００７５】

その後、第２の識別部５５は、有効化部５４による有効化後の有声音および無声音に基づいて、ビタビアルゴリズムを用いて、沈黙状態であるか、あるいは発話状態であるかを検出することで、沈黙領域および発話領域を識別する（ステップＳ１１６）。

【００７６】

続いて、抽出部１３は、図１０に示すように、ある話者が発話したと特定されたフレームから、有声音領域の数、有声音領域の長さの平均値および有声音領域の長さの標準偏差を算出する（ステップＳ１１７）。

【００７７】

さらに、抽出部１３は、ある話者が発話したと特定されたフレームから、発話領域の数、発話領域の長さの平均値および発話領域の長さの標準偏差を算出する（ステップＳ１１８）。その後、抽出部１３は、ある話者の沈黙領域のフレームから、沈黙領域の数、沈黙領域の長さの平均値および沈黙領域の長さの標準偏差を算出する（ステップＳ１１９）。

【００７８】

そして、抽出部１３は、会話全体の時間の長さに対するある話者の発話時間の長さの割合を算出する（ステップＳ１２０）。さらに、抽出部１３は、他の話者の発話時間に対す

るある話者の発話時間の割合を算出する（ステップＳ１２１）。
【００７９】
続いて、抽出部１３は、ある話者が発話したと特定されたフレームから、音量の標準偏差およびスペクトルエントロピーの標準偏差を算出する（ステップＳ１２２）。抽出部１３は、ある話者が発話したと特定されたフレームから算出した音量の標準偏差と、スペクトルエントロピーの標準偏差との和を、変化の度合いとして算出する（ステップＳ１２３）。
【００８０】
そして、全ての話者の会話特性を抽出するまで（ステップＳ１２４否定）、上記のステップＳ１１７～ステップＳ１２３までの処理を繰り返し実行する。その後、全ての話者の会話特性を抽出すると（ステップＳ１２４肯定）、分析部１４は、抽出部１３によって抽出された会話特性に基づいて、会話スタイルを分析する（ステップＳ１２５）。最後に、分析部１４は、分析結果を所定の出力先の装置へ出力し（ステップＳ１２６）、処理を終了する。
【００８１】
（２）有効化処理
図１１は、実施例１に係る有効化処理の手順を示すフローチャートである。この有効化処理は、図９に示したステップＳ１１５に対応する処理であり、有声音領域および無声音領域が識別された後に処理が起動する。
【００８２】
図１１に示すように、有効化部５４は、各音声データにおける同一区間のフレームで第１の識別部５３による識別結果を比較する（ステップＳ３０１）。このとき、同一区間のフレームで有声音領域が重複する場合（ステップＳ３０２肯定）には、有効化部５４は、当該有声音領域と識別されたフレームのエネルギーを演算する（ステップＳ３０３）。なお、同一区間のフレームで有声音領域が重複しない場合（ステップＳ３０２否定）には、ステップＳ３０６へ移行する。
【００８３】
そして、有効化部５４は、当該同一区間で有声音領域と識別されたフレームのうち最大エネルギーを持つフレームを特定する（ステップＳ３０４）。その上で、有効化部５４は、最大エネルギーを持つフレーム以外の識別結果を有声音領域から無声音領域に置き換える（ステップＳ３０５）。
【００８４】
その後、各音声データ間で同一区間のフレームを全て処理するまで（ステップＳ３０６否定）、上記のステップＳ３０１～ステップＳ３０５までの処理を繰り返し実行する。そして、各音声データ間で同一区間のフレームを全て処理すると（ステップＳ３０６肯定）、処理を終了する。
【００８５】
［実施例１の効果］
上述してきたように、本実施例に係る話者判別装置５０は、各音声データにおける同一区間のフレームで有声音領域が重複して識別された場合に、最大のエネルギーを持つフレームの識別結果だけを有効化して各々の音声データの発話領域および沈黙領域を識別する。このため、本実施例に係る話者判別装置５０は、各音声データを構成する同一区間のフレーム間で閾値を用いて判定せずとも、話者を判別することができる。さらに、本実施例に係る話者判別装置５０では、話者の判別に複雑なアルゴリズムを用いる必要はなく、事前に学習を行う必要もない。したがって、本実施例に係る話者判別装置５０によれば、話者の判別を簡易かつ正確に行うことができる。
【００８６】
また、本実施例に係る話者判別装置５０は、各音声データにおける同一区間のフレームで有声音領域が単独で識別された場合には、エネルギーの大小に関係なく、有声音領域と識別された識別結果を維持する。一般に、発話は、有声音と無声音が混在して構成される

ので、複数の話者によって同時に発話された場合でも、同時発話で有声音領域が完全に重複する可能性は低く、有声音領域が単独で識別される機会が残る可能性は高い。それゆえ、本実施例に係る話者判別装置５０では、話者が発話する音量に開きがある場合でも、同時発話を判別することもできる。

【００８７】

さらに、本実施例に係る話者判別装置５０は、当該同一区間で有声音領域と識別されたフレームのうち最大のエネルギーを持つフレーム以外の識別結果を無声音領域に置き換える。このため、本実施例に係る話者判別装置５０では、識別情報の置換という簡易な処理によって最大のエネルギーを持つフレームの識別結果だけを有効化できる結果、話者の判別を簡易に実現できる。

【実施例２】

【００８８】

さて、これまで開示の装置に関する実施例について説明したが、本発明は上述した実施例以外にも、種々の異なる形態にて実施されてよいものである。そこで、以下では、本発明に含まれる他の実施例を説明する。

【００８９】

［エネルギー］

例えば、上記の実施例１では、最大エネルギーを持つフレームの識別結果だけを有効化する場合を例示したが、エネルギーに関連する他の指標が最大となるフレームの識別結果だけを有効化することもできる。一例としては、開示の装置は、フレームで観測される振幅の最大値および最小値の差が最大であるフレームの識別結果だけを有効化することもできる。この場合には、エネルギーの演算処理よりも簡易な演算により、識別結果の置換を実現できる。

【００９０】

［マイク］

また、上記の実施例１では、接話型マイクロホンを適用する場合を例示したが、開示の装置はこれに限定されず、必ずしもマイクを装着する話者以外の他の話者をマイクから遠ざける必要はない。例えば、指向性を持つマイクを適用することができる。この場合には、話者Ａが発話する方向の感度が他の方向の感度よりも強くなるように話者Ａまたは指向性マイクを配置し、また、話者Ｂおよび話者Ｃについても同様にして指向性マイクを用いればよい。なお、指向性マイクを用いる場合についても、話者は複数であればよく、２人であっても４人以上であっても開示の装置を適用できる。

【００９１】

［分散および統合］

また、図示した各装置の各構成要素は、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各装置の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷や使用状況などに応じて、任意の単位で機能的または物理的に分散・統合して構成することができる。例えば、話者判別装置５０、抽出部１３または分析部１４を会話分析装置の外部装置としてネットワーク経由で接続するようにしてもよい。また、話者判別装置５０、抽出部１３または分析部１４を別の装置がそれぞれ有し、ネットワーク接続されて協働することで、上記の話者判別装置の機能を実現するようにしてもよい。

【００９２】

［話者判別プログラム］

また、上記の実施例で説明した各種の処理は、予め用意されたプログラムをパーソナルコンピュータやワークステーションなどのコンピュータで実行することによって実現することができる。そこで、以下では、図１２を用いて、上記の実施例と同様の機能を有する話者判別プログラムを実行するコンピュータの一例について説明する。

【００９３】

図１２は、実施例１及び実施例２に係る話者判別プログラムを実行するコンピュータの

一例について説明するための図である。図１２に示すように、コンピュータ１００は、操作部１１０ａと、スピーカ１１０ｂと、マイク１１０ｃと、ディスプレイ１２０と、通信部１３０とを有する。さらに、このコンピュータ１００は、ＣＰＵ１５０と、ＲＯＭ１６０と、ＨＤＤ１７０と、ＲＡＭ１８０と有する。これら１１０～１８０の各部はバス１４０を介して接続される。

【００９４】

ＨＤＤ１７０には、図１２に示すように、上記の実施例１で示した取得部５１と、フレーム化部５２と、第１の識別部５３と、有効化部５４と、第２の識別部５５と同様の機能を発揮する話者判別プログラム１７０ａが予め記憶される。この話者判別プログラム１７０ａについては、図１に示した各々の取得部５１、フレーム化部５２、第１の識別部５３、有効化部５４及び第２の識別部５５の各構成要素と同様、適宜統合又は分離しても良い。すなわち、ＨＤＤ１７０に格納される各データは、常に全てのデータがＨＤＤ１７０に格納される必要はなく、処理に必要なデータのみがＨＤＤ１７０に格納されれば良い。

【００９５】

そして、ＣＰＵ１５０が、話者判別プログラム１７０ａをＨＤＤ１７０から読み出してＲＡＭ１８０に展開する。これによって、図１２に示すように、話者判別プログラム１７０ａは、話者判別プロセス１８０ａとして機能する。この話者判別プロセス１８０ａは、ＨＤＤ１７０から読み出した各種データを適宜ＲＡＭ１８０上の自身に割り当てられた領域に展開し、この展開した各種データに基づいて各種処理を実行する。なお、話者判別プロセス１８０ａは、図１に示した取得部５１、フレーム化部５２、第１の識別部５３、有効化部５４及び第２の識別部５５にて実行される処理、例えば図９～図１１に示す処理を含む。また、ＣＰＵ１５０上で仮想的に実現される各処理部は、常に全ての処理部がＣＰＵ１５０上で動作する必要はなく、処理に必要な処理部のみが仮想的に実現されれば良い。

【００９６】

なお、上記の話者判別プログラム１７０ａについては、必ずしも最初からＨＤＤ１７０やＲＯＭ１６０に記憶させておく必要はない。例えば、コンピュータ１００に挿入されるフレキシブルディスク、いわゆるＦＤ、ＣＤ－ＲＯＭ、ＤＶＤディスク、光磁気ディスク、ＩＣカードなどの「可搬用の物理媒体」に各プログラムを記憶させる。そして、コンピュータ１００がこれらの可搬用の物理媒体から各プログラムを取得して実行するようにしてもよい。また、公衆回線、インターネット、ＬＡＮ、ＷＡＮなどを介してコンピュータ１００に接続される他のコンピュータまたはサーバ装置などに各プログラムを記憶させておき、コンピュータ１００がこれらから各プログラムを取得して実行するようにしてもよい。

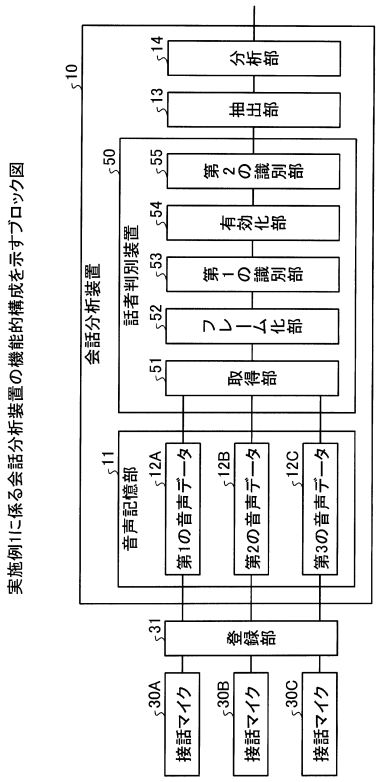
【符号の説明】

【００９７】

１０　会話分析装置
１１　音声記憶部
１２Ａ　第１の音声データ
１２Ｂ　第２の音声データ
１２Ｃ　第３の音声データ
３０Ａ，３０Ｂ，３０Ｃ　接話マイク
３１　登録部
５０　話者判別装置
５１　取得部
５２　フレーム化部
５３　第１の識別部
５４　有効化部
５５　第２の識別部

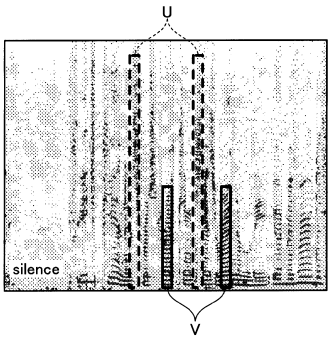
【図１】

実施例1に係る会話分析装置の機能的構成を示すブロック図



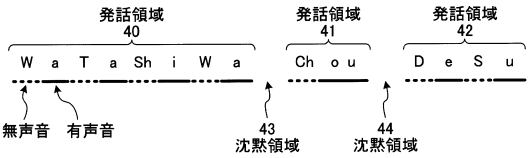
【図２】

有声音および無声音の一例を示す図



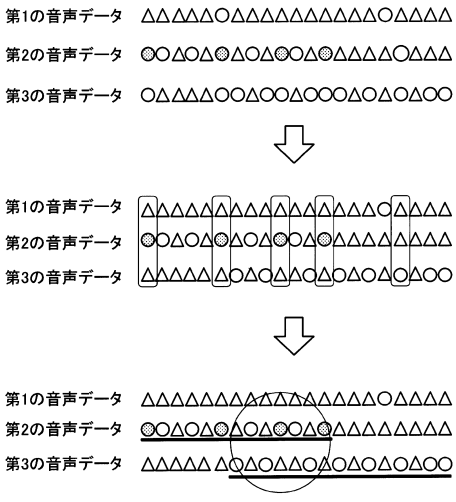
【図３】

発話領域および沈黙領域の一例を示す図



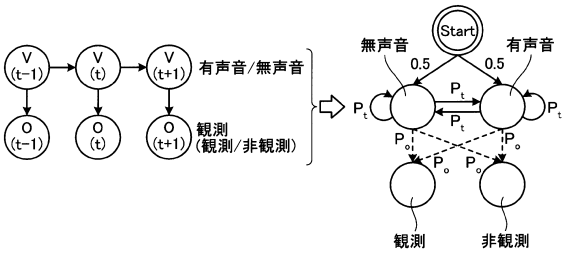
【図４】

話者判別方法を説明するための図



【図５】

隠れマルコフモデルにおける状態遷移図の一例を示す図



【図６】

有声音領域および無声音領域の識別結果の一例を示す図

時刻	第1音声データ		第2音声データ		第3音声データ	
	V or U	energy	V or U	energy	V or U	energy
12:00:00.000	V	200	V	20	V	10
12:00:00.010	V	200	V	100	U	
12:00:00.020	U		V	100	U	
12:00:00.030	U		U		U	
⋮	⋮	⋮	⋮	⋮	⋮	⋮

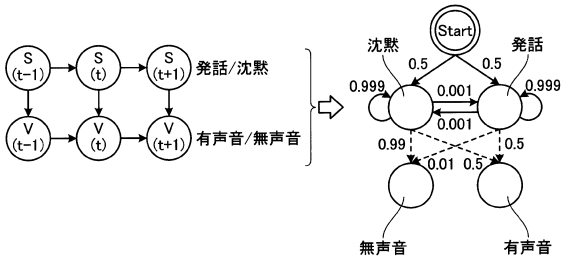
【図７】

図6に示した識別結果の置換結果の一例を示す図

時刻	第1音声データ	第2音声データ	第3音声データ
	V or U	V or U	V or U
12:00:00.000	V	U	U
12:00:00.010	V	U	U
12:00:00.020	U	V	U
12:00:00.030	U	U	U
⋮	⋮	⋮	⋮

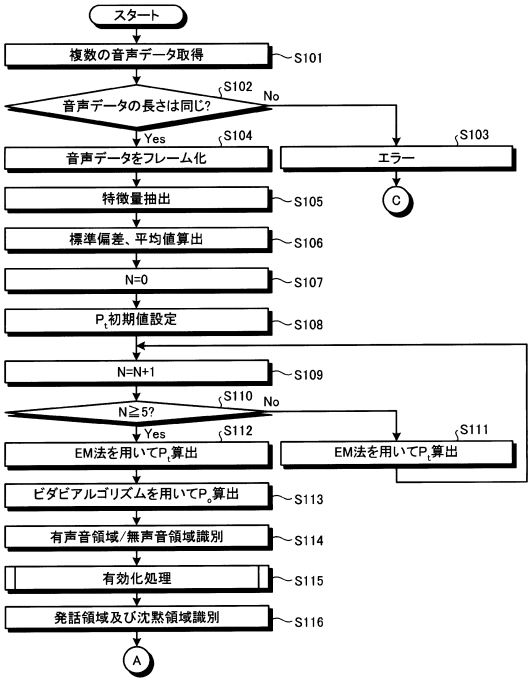
【図８】

隠れマルコフモデルにおける状態遷移図の一例を示す図



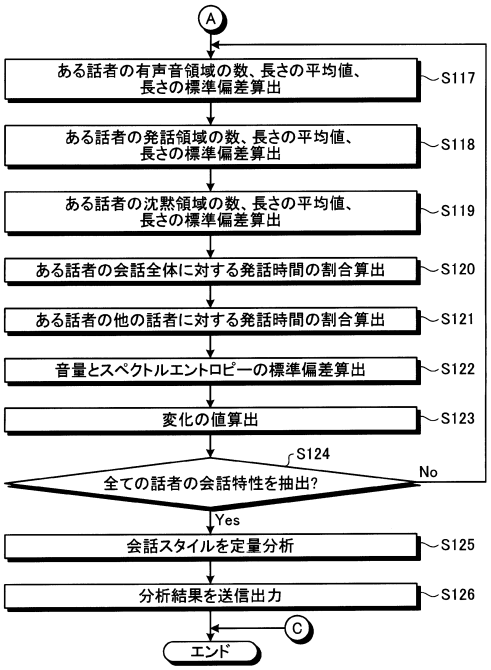
【図９】

実施例1に係る会話分析処理の手順を示すフローチャート



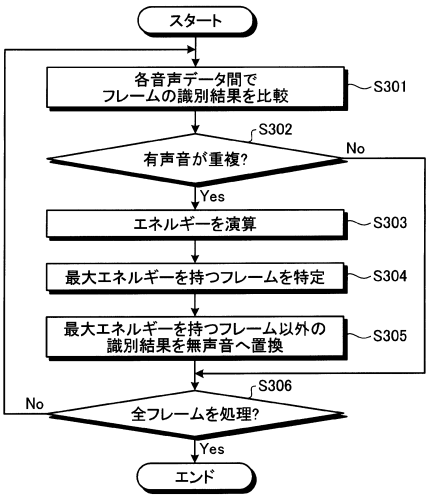
【図１０】

実施例1に係る会話分析処理の手順を示すフローチャート



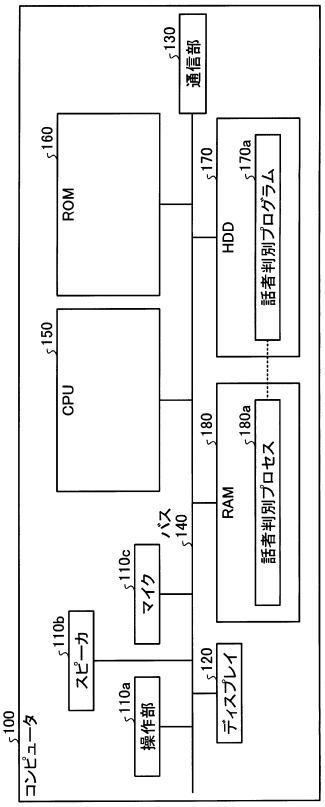
【図１１】

実施例1に係る有効化処理の手順を示すフローチャート



【図１２】

実施例1及び実施例2に係る話者判別プログラムを実行するコンピュータの一例について説明するための図



フロントページの続き

(56)参考文献 特開２００６－２０８４８２（ＪＰ，Ａ）
特開２００１－３４３９８３（ＪＰ，Ａ）
特開平０７－０５６５９８（ＪＰ，Ａ）
特開２００１－３４３９８５（ＪＰ，Ａ）
特開昭５８－０９５３９９（ＪＰ，Ａ）
特開２００６－０８６８７７（ＪＰ，Ａ）
米国特許第０５１９７１１３（ＵＳ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ １５／００－１５／３４
Ｇ１０Ｌ ２５／００－２５／９３