(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2019-101385
(P2019-101385A)
(43)公開日　令和1年6月24日(2019.6.24)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 15/04	(2013.01)	G10L 15/04	300A	
G10L 15/00	(2013.01)	G10L 15/04	300B	
		G10L 15/00	200C	

審査請求 未請求　請求項の数 6　OL　（全 19 頁）

(21)出願番号　特願2017-235977 (P2017-235977)
(22)出願日　平成29年12月8日 (2017.12.8)

(71)出願人　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　100099759
弁理士　青木　篤
(74)代理人　100119987
弁理士　伊坪　公一
(74)代理人　100133835
弁理士　河野　努
(74)代理人　100135976
弁理士　宮本　哲夫
(72)発明者　鷲尾　信之
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

(54)【発明の名称】音声処理装置、音声処理方法及び音声処理用コンピュータプログラム

(57)【要約】
【課題】音声信号中で複数の話者の何れかが発話を開始したタイミングを誤検出しても、発話した話者に応じた処理を音声信号に適用できる音声処理装置を提供する。
【解決手段】音声処理装置は、第１の音声入力部（１１－１）により生成された第１の音声信号及び第２の音声入力部（１１－２）により生成された第２の音声信号の少なくとも一方に基づいて、何れかの話者が発話を開始したタイミングを検出する発話区間開始検出部（２４）と、検出された発話を開始したタイミングを修正するか否かを判定する開始タイミング修正部（２６）と、発話を開始したタイミングが修正されると、修正されたタイミング以降の第１及び第２の音声信号に基づいて、発話した話者を特定する話者特定部（２５）と、特定された話者に応じた処理を、修正された発話を開始したタイミング以降の第１及び第２の音声信号の少なくとも一方に対して実行する音声処理部（２８）とを有する。
【選択図】図２

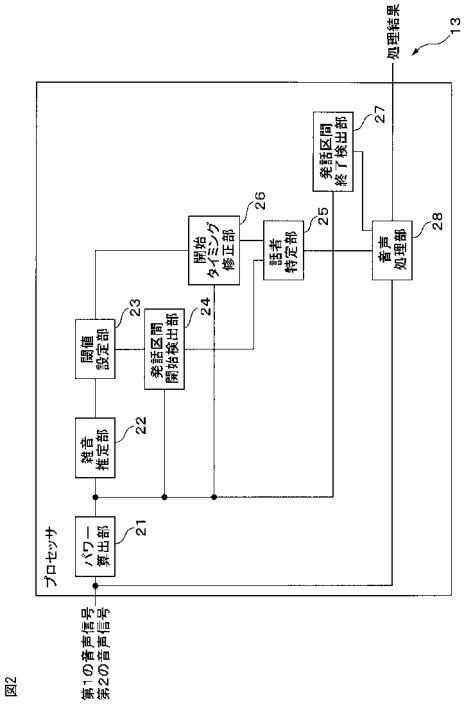


図2

【特許請求の範囲】
【請求項１】
第１の音声入力部により生成された第１の音声信号及び第２の音声入力部により生成された第２の音声信号の少なくとも一方に基づいて、複数の話者の何れかが発話を開始したタイミングを検出する発話区間開始検出部と、
検出された前記発話を開始したタイミング以降における、前記第１の音声信号及び前記第２の音声信号の少なくとも一方に基づいて、前記発話を開始したタイミングを修正するか否かを判定する開始タイミング修正部と、
前記発話を開始したタイミングが修正されると、修正された前記発話を開始したタイミング以降の前記第１の音声信号及び前記第２の音声信号に基づいて、前記複数の話者のうちの発話した話者を特定する話者特定部と、
特定された前記話者に応じた処理を、修正された前記発話を開始したタイミング以降の前記第１の音声信号及び前記第２の音声信号の少なくとも一方に対して実行する音声処理部と、
を有する音声処理装置。
【請求項２】
前記話者特定部は、前記発話を開始したタイミングが検出されると、当該タイミング以降の前記第１の音声信号及び前記第２の音声信号に基づいて、前記複数の話者のうちの発話した話者を特定し、
前記音声処理部は、前記発話を開始したタイミングが検出されたときに特定された前記話者に応じた第１の処理を、前記第１の音声信号及び前記第２の音声信号の少なくとも一方に対して実行し、
前記音声処理部は、前記発話を開始したタイミングが修正されたときに前記第１の処理を停止する、請求項１に記載の音声処理装置。
【請求項３】
前記音声処理部は、前記発話を開始したタイミングが検出されたときに特定された前記話者と、前記発話を開始したタイミングが修正されたときに特定された前記話者とが異なる場合、前記第１の処理を停止する、請求項２に記載の音声処理装置。
【請求項４】
前記発話区間開始検出部は、前記第１の音声信号及び前記第２の音声信号のそれぞれについて、当該音声信号を分割した所定長を持つフレームごとに、前記音声信号の周期性の強さを表すピッチゲインを算出し、前記第１の音声信号及び前記第２の音声信号の少なくとも一方について前記ピッチゲインが所定の閾値以上となるフレームを、前記発話を開始したタイミングとして検出し、
前記開始タイミング修正部は、前記第１の音声信号及び前記第２の音声信号の少なくとも一方について、前記発話を開始したタイミングが検出されたときの前記ピッチゲインよりも前記ピッチゲインが所定のオフセット以上大きくなるフレームを検出すると、当該フレームを前記発話を開始したタイミングとして修正する、請求項１～３の何れか一項に記載の音声処理装置。
【請求項５】
第１の音声入力部により生成された第１の音声信号及び第２の音声入力部により生成された第２の音声信号の少なくとも一方に基づいて、複数の話者の何れかが発話を開始したタイミングを検出し、
検出された前記発話を開始したタイミング以降における、前記第１の音声信号及び前記第２の音声信号の少なくとも一方に基づいて、前記発話を開始したタイミングを修正するか否かを判定し、
前記発話を開始したタイミングが修正されると、修正された前記発話を開始したタイミング以降の前記第１の音声信号及び前記第２の音声信号に基づいて、前記複数の話者のうちの発話した話者を特定し、
特定された前記話者に応じた処理を、修正された前記発話を開始したタイミング以降の

前記第１の音声信号及び前記第２の音声信号の少なくとも一方に対して実行する、
ことを含む音声処理方法。
【請求項６】
第１の音声入力部により生成された第１の音声信号及び第２の音声入力部により生成された第２の音声信号の少なくとも一方に基づいて、複数の話者の何れかが発話を開始したタイミングを検出し、
検出された前記発話を開始したタイミング以降における、前記第１の音声信号及び前記第２の音声信号の少なくとも一方に基づいて、前記発話を開始したタイミングを修正するか否かを判定し、
前記発話を開始したタイミングが修正されると、修正された前記発話を開始したタイミング以降の前記第１の音声信号及び前記第２の音声信号に基づいて、前記複数の話者のうちの発話した話者を特定し、
特定された前記話者に応じた処理を、修正された前記発話を開始したタイミング以降の前記第１の音声信号及び前記第２の音声信号の少なくとも一方に対して実行する、
ことをコンピュータに実行させるための音声処理用コンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、例えば、話者の声を表す音声信号を処理する音声処理装置、音声処理方法及び音声処理用コンピュータプログラムに関する。
【背景技術】
【０００２】
音声信号から話者が発した語句を認識し、認識した語句を他の言語に翻訳したり、認識した語句をクエリとしてネットワークまたはデータベース上で探索するといったアプリケーションが開発されている。このようなアプリケーションでは、音声信号中で話者が発話している区間が検出され、検出された区間に対してアプリケーションに応じた音声処理が実行される。
【０００３】
場合によっては、複数の話者のそれぞれの声が音声処理の対象となり、かつ、話者に応じて、実行すべき処理が異なることがある。そこで、音声入力部に入力された２以上の使用者の音声信号を、使用者ごとに分離し、分離された使用者ごとの音声信号を認識し、その認識結果を表示部の使用者ごとに対応する表示領域に表示させる技術が提案されている（例えば、特許文献１を参照）。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２０１５－１０６０１４号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
しかしながら、音声処理を行う装置の周囲の環境に応じて、音声信号に含まれる雑音成分の大きさが変動する。そのため、話者が発話していないにもかかわらず、音声信号中に含まれる雑音により、話者が発話を開始したタイミングが誤検出されることがある。このような場合、上記の技術では、音声信号中において実際には発話していない一方の話者の声として分離された区間中において、他方の話者が発話を開始すると、他方の話者が発話している区間も、発話していない方の話者と関連付けられてしまう。その結果として、発話している話者の声を含む区間に対して、発話していない方の話者に対する音声処理が行われてしまうことがある。
【０００６】
一つの側面では、本発明は、音声信号中で複数の話者の何れかが発話を開始したタイミ

ングを誤検出しても、発話した話者に応じた処理を音声信号に適用できる音声処理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】

【０００７】

一つの実施形態によれば、音声処理装置が提供される。この音声処理装置は、第１の音声入力部により生成された第１の音声信号及び第２の音声入力部により生成された第２の音声信号の少なくとも一方に基づいて、複数の話者の何れかが発話を開始したタイミングを検出する発話区間開始検出部と、検出された発話を開始したタイミング以降における、第１の音声信号及び第２の音声信号の少なくとも一方に基づいて、何れかの話者が発話を開始したタイミングを修正するか否かを判定する開始タイミング修正部と、何れかの話者が発話を開始したタイミングが修正されると、修正された発話を開始したタイミング以降の第１の音声信号及び第２の音声信号に基づいて、複数の話者のうちの発話した話者を特定する話者特定部と、特定された話者に応じた処理を、修正された発話を開始したタイミング以降の第１の音声信号及び第２の音声信号の少なくとも一方に対して実行する音声処理部とを有する。

【発明の効果】

【０００８】

音声信号中で複数の話者の何れかが発話を開始したタイミングを誤検出しても、発話した話者に応じた処理を音声信号に適用できる。

【図面の簡単な説明】

【０００９】

【図１】一つの実施形態による音声処理装置の概略構成図である。

【図２】音声処理に関する音声処理装置のプロセッサの機能ブロック図である。

【図３】本実施形態による話者特定の説明図である。

【図４】発話区間開始タイミングの修正についての説明図である。

【図５】話者と音声処理の対応関係の一例を示す図である。

【図６】発話区間の開始タイミングの修正と音声処理の関係の一例を示す図である。

【図７】音声処理の動作フローチャートである。

【図８】実施形態またはその変形例による音声処理装置が実装されたサーバクライアントシステムの概略構成図である。

【発明を実施するための形態】

【００１０】

以下、図を参照しつつ、実施形態による音声処理装置について説明する。

この音声処理装置は、音声信号中で複数の話者の何れかが発話している区間（以下、単に発話区間と呼ぶ）を検出し、検出した発話区間において発話した話者を特定する。そしてこの音声処理装置は、発話区間に対して、特定した話者に応じた処理を実行する。ここで、この音声処理装置は、雑音の大きさの変動などにより発話区間の開始タイミングを誤検出した場合に備えて、発話区間の開始検出後の音声信号に基づいて、発話区間の開始を修正すべきか否か判定する。この音声処理装置は、発話区間の開始タイミングを修正すると、修正された開始タイミングから実際の発話区間が開始されたものとして、発話した話者を再度特定する。そしてこの音声処理装置は、再特定された話者に応じた処理を、再検出された開始タイミング以降の発話区間に対して実行する。

【００１１】

この音声処理装置は、音声信号を利用するユーザインターフェースを採用する様々な装置、例えば、ナビゲーションシステム、電話会議システム、携帯電話機またはコンピュータなどに実装できる。本実施形態では、この音声処理装置は、話者ごとに異なる言語の翻訳処理を行う多言語翻訳装置に実装されるものとする。

【００１２】

図１は、一つの実施形態による音声処理装置の概略構成図である。音声処理装置１は、二つのマイクロホン１１－１、１１－２と、二つのアナログ／デジタルコンバータ１２－

１、１２－２と、プロセッサ１３と、メモリ１４と、表示装置１５とを有する。なお、音声処理装置１は、さらに、スピーカ（図示せず）及び他の機器と通信するための通信インターフェース（図示せず）を有していてもよい。

【００１３】

マイクロホン１１－１、１１－２は、それぞれ、音声入力部の一例であり、互いに対して所定の間隔を空けて設置される。例えば、マイクロホン１１－１は、マイクロホン１１－２よりも、複数の話者のうちの一人（便宜上、第１の話者と呼ぶ）の近くに設置される。また、マイクロホン１１－２は、マイクロホン１１－１よりも、複数の話者のうちの他の一人（便宜上、第２の話者と呼ぶ）の近くに設置される。そしてマイクロホン１１－１、１１－２は、それぞれ、複数の話者の何れかの声を含む、音声処理装置１の周囲の音を集音し、その音の強度に応じたアナログ音声信号を生成する。そしてマイクロホン１１－１は、そのアナログ音声信号をアナログ／デジタルコンバータ（以下、Ａ／Ｄコンバータと表記する）１２－１へ出力する。同様に、マイクロホン１１－２は、生成したアナログ音声信号をＡ／Ｄコンバータ１２－２へ出力する。

【００１４】

Ａ／Ｄコンバータ１２－１は、マイクロホン１１－１から受け取ったアナログ音声信号を所定のサンプリングレートでサンプリングすることにより、その音声信号をデジタル化する。なお、サンプリングレートは、例えば、音声信号から話者の声を解析するために必要な周波数帯域がナイキスト周波数以下となるよう、例えば、16kHz～32kHzに設定される。そしてＡ／Ｄコンバータ１２－１は、デジタル化された音声信号をプロセッサ１３へ出力する。同様に、Ａ／Ｄコンバータ１２－２は、マイクロホン１１－２から受け取ったアナログ音声信号を所定のサンプリングレートでサンプリングすることにより、その音声信号をデジタル化し、デジタル化された音声信号をプロセッサ１３へ出力する。

【００１５】

なお、以下では、Ａ／Ｄコンバータ１２－１によりデジタル化された、マイクロホン１１－１からの音声信号を第１の音声信号と呼び、Ａ／Ｄコンバータ１２－２によりデジタル化された、マイクロホン１１－２からの音声信号を第２の音声信号と呼ぶ。

【００１６】

プロセッサ１３は、例えば、Central Processing Unit(CPU)と、読み書き可能なメモリ回路と、その周辺回路とを有する。プロセッサ１３は、数値演算回路をさらに有していてもよい。そしてプロセッサ１３は、第１の音声信号及び第２の音声信号から、何れかの話者が発話している発話区間を検出するとともに、その発話区間にて発話している話者を特定する。そしてプロセッサ１３は、発話区間に対して、特定した話者に対応する言語についての音声認識処理を実行し、認識された語句を、特定した話者に対応する言語以外の言語に翻訳し、その翻訳結果を表示装置１５に表示させる。

【００１７】

さらに、プロセッサ１３は、一旦発話区間の開始タイミングを検出した後に、発話区間の開始タイミングを修正するか否か判定する。そして発話区間の開始タイミングが修正された場合には、プロセッサ１３は、修正された発話区間の開始タイミング以降における、第１及び第２の音声信号に基づいて、発話している話者を再度特定する。プロセッサ１３は、修正された開始タイミング以降の発話区間に対して、再特定された話者に対応する言語についての音声認識処理及び翻訳処理を実行する。

なお、音声処理の詳細については後述する。

【００１８】

メモリ１４は、例えば、読み書き可能な不揮発性の半導体メモリと、読み書き可能な揮発性の半導体メモリとを有する。さらに、メモリ１４は、磁気記録媒体あるいは光記録媒体及びそのアクセス装置を有していてもよい。そしてメモリ１４は、プロセッサ１３上で実行される音声処理で利用される各種のデータ及び音声処理の途中で生成される各種のデータを記憶する。

【００１９】

表示装置１５は、例えば、液晶ディスプレイまたは有機ＥＬディスプレイとすることができる。そして表示装置１５は、プロセッサ１３から受け取った表示用のデータ、例えば、何れかの話者が発話した内容、あるいは、その内容を話者が使用した言語（例えば、日本語）から他言語（例えば、英語）に翻訳して得られた文字列を表示する。

【００２０】

以下、プロセッサ１３の詳細について説明する。

【００２１】

図２は、音声処理に関するプロセッサ１３の機能ブロック図である。プロセッサ１３は、パワー算出部２１と、雑音推定部２２と、閾値設定部２３と、発話区間開始検出部２４と、話者特定部２５と、開始タイミング修正部２６と、発話区間終了検出部２７と、音声処理部２８とを有する。

プロセッサ１３が有するこれらの各部は、例えば、プロセッサ１３上で動作するコンピュータプログラムにより実現される機能モジュールである。あるいは、プロセッサ１３が有するこれらの各部は、その各部の機能を専用の回路として、プロセッサ１３に組み込まれてもよい。

【００２２】

プロセッサ１３は、第１及び第２の音声信号のそれぞれを所定長を持つフレームを処理単位として音声処理を実行する。フレーム長は、例えば、10msec～20msecに設定される。そのため、プロセッサ１３は、第１及び第２の音声信号のそれぞれをフレームごとに分割し、各フレームをパワー算出部２１及び音声処理部２８へ入力する。

【００２３】

パワー算出部２１は、第１及び第２の音声信号のそれぞれについて、フレームが入力される度に、そのフレームについてのパワーを算出する。パワー算出部２１は、例えば、フレームごとに、次式に従ってパワーを算出する。

【数１】

$$Spow(k)=\sum_{n=0}^{N-1}s_k(n)^2 \qquad (1)$$

ここで、$S_k(n)$は、最新のフレーム（現フレームとも呼ぶ）のn番目のサンプリング点の信号値を表す。kはフレーム番号である。またNは、一つのフレームに含まれるサンプリング点の総数を表す。そしてSpow(k)は、現フレームのパワーを表す。

【００２４】

なお、パワー算出部２１は、各フレームについて、複数の周波数のそれぞれごとにパワーを算出してもよい。この場合、パワー算出部２１は、フレームごとに、第１及び第２の音声信号を、時間周波数変換を用いて時間領域から周波数領域のスペクトル信号に変換する。なお、パワー算出部２１は、時間周波数変換として、例えば、高速フーリエ変換(Fast Fourier Transform,FFT)を用いることができる。そしてパワー算出部２１は、第１及び第２の音声信号のそれぞれについて、周波数ごとに、その周波数に含まれるスペクトル信号の２乗和を、その周波数のパワーとして算出できる。そしてパワー算出部２１は、フレームごとに、人の声が含まれる周波数帯域（例えば、100Hz～20kHz）に含まれる各周波数のパワーの和を、そのフレームのパワーとして算出してもよい。

【００２５】

パワー算出部２１は、第１及び第２の音声信号のそれぞれについて、フレームごとのパワーを、雑音推定部２２、発話区間開始検出部２４、話者特定部２５、開始タイミング修正部２６及び発話区間終了検出部２７へ出力する。

【００２６】

雑音推定部２２は、第１及び第２の音声信号のそれぞれについて、フレームごとに、そ

のフレームにおける音声信号中の推定雑音成分を算出する。本実施形態では、雑音推定部２２は、直前のフレームにおける推定雑音成分を、現フレームのパワーを用いて次式に従って更新することで、現フレームの推定雑音成分を算出する。

【数２】

$$Noise(k)=\beta\cdot Noise(k-1)+(1-\beta)\cdot Spow(k) \qquad (2)$$

ここで、Noise(k-1)は、直前のフレームにおける推定雑音成分を表し、Noise(k)は、現フレームにおける推定雑音成分を表す。またβは、忘却係数であり、例えば、0.9に設定される。

【００２７】

なお、パワーが周波数ごとに算出されている場合には、雑音推定部２２は、（２）式に従って、推定される雑音成分を周波数ごとに算出してもよい。この場合には、（２）式において、Noise(k-1)、Noise(k)及びSpow(k)は、それぞれ、着目する周波数についての直前のフレームの推定雑音成分、現フレームの推定雑音成分、パワーとなる。

【００２８】

雑音推定部２２は、第１及び第２の音声信号のそれぞれについて、フレームごとの推定雑音成分を閾値設定部２３へ出力する。

なお、後述する発話区間開始検出部２４により、現フレームが何れかの話者の声を含む発話区間に含まれるフレームであると判定されることがある。この場合には、雑音推定部２２は、現フレームの推定雑音成分Noise(k)を、Noise(k-1)で置換して、（２）式に従って再度現フレームの推定雑音成分を算出してもよい。これにより、雑音推定部２２は、雑音成分のみを含み、信号成分を含まないと推定されるフレームに基づいて雑音成分を推定できるので、雑音成分の推定精度を向上できる。

【００２９】

あるいは、雑音推定部２２は、現フレームのパワーが所定の閾値以下である場合に限り、（２）式に従って推定雑音成分を更新すればよい。そして現フレームのパワーが所定の閾値より大きい場合には、雑音推定部２２は、Noise(k)=Noise(k-1)とすればよい。なお、所定の閾値は、例えば、Noise(k-1)に所定のオフセット値を加算した値とすることができる。

【００３０】

閾値設定部２３は、第１及び第２の音声信号のそれぞれについて、推定雑音成分に基づいて発話区間を検出するための閾値を設定する。例えば、閾値設定部２３は、発話区間が検出されていない間、フレームごとに閾値を設定する。例えば、閾値設定部２３は、第１の音声信号についての現フレームの推定雑音成分に、所定のオフセット値を加算した値を第１の音声信号に対する閾値とする。同様に、閾値設定部２３は、第２の音声信号についての現フレームの推定雑音成分に、所定のオフセット値を加算した値を第２の音声信号に対する閾値とすればよい。

【００３１】

あるいは、閾値設定部２３は、現フレームについての、第１の音声信号についての推定雑音成分と第２の音声信号についての推定雑音成分の平均値に所定のオフセット値を加算した値を、第１の音声信号及び第２の音声信号に共通する閾値としてもよい。あるいはまた、閾値設定部２３は、現フレームについての、第１の音声信号についての推定雑音成分と第２の音声信号についての推定雑音成分のうちの大きい方に所定のオフセット値を加算した値を、第１の音声信号及び第２の音声信号に共通する閾値としてもよい。

【００３２】

閾値設定部２３は、第１及び第２の音声信号のそれぞれについて、発話区間の開始が検出されるまで、フレームごとに、設定した閾値を発話区間開始検出部２４へ通知する。

【００３３】

発話区間開始検出部２４は、フレームごとに、そのフレームの第１の音声信号のパワー及び第２の音声信号のパワーの少なくとも一方と閾値とを比較することで、発話区間が開始されたタイミングを検出する。

【００３４】

例えば、発話区間開始検出部２４は、直前のフレームまで第１及び第２の音声信号の何れについてもパワーが対応する閾値未満であり、かつ、第１及び第２の音声信号の少なくとも一方について、現フレームのパワーが対応する閾値以上となった場合、発話区間が開始されたと判定する。そして発話区間開始検出部２４は、現フレームを発話区間の開始タイミングとする。

【００３５】

あるいは、発話区間開始検出部２４は、フレームごとに、第１の音声信号及び第２の音声信号のうち、パワーの大きい方を、対応する閾値と比較してもよい。そして発話区間開始検出部２４は、直前のフレームまで、パワーの大きい方が対応する閾値未満となり、かつ、現フレームにおいて、パワーの大きい方が対応する閾値以上となる場合に、現フレームを発話区間の開始タイミングとして検出してもよい。

【００３６】

あるいはまた、発話区間開始検出部２４は、第１の音声信号及び第２の音声信号の少なくとも一方について、所定数のフレームにわたって連続してパワーが対応する閾値以上となった場合、発話区間が開始されたと判定してもよい。そして発話区間開始検出部２４は、その連続するフレームのうちの最初にパワーが閾値以上となったフレームを、発話区間の開始タイミングとして検出してもよい。

【００３７】

発話区間開始検出部２４は、発話区間が開始されたと判定すると、その旨を話者特定部２５及び開始タイミング修正部２６へ通知する。

【００３８】

話者特定部２５は、発話区間の開始が検出されると、その発話区間において発話している話者を特定する。例えば、話者特定部２５は、第１及び第２の音声信号のそれぞれについて、発話区間開始検出直後の所定数（例えば、1～5）のフレームのパワーの平均値を算出する。そして話者特定部２５は、マイクロホン１１－１、１１－２のうち、パワーの平均値が高い方の音声信号を取得したマイクロホンと対応する話者（例えば、そのマイクロホンに近い方の話者）が発話したと判定する。

【００３９】

図３は、本実施形態による話者特定の説明図である。この例では、左から、マイクロホン１１－１、マイクロホン１１－２の順に各マイクロホンは設置されている。そしてマイクロホン１１－１よりも左側に第１の話者３０１が位置し、マイクロホン１１－２よりも右側に第２の話者３０２が位置している。したがって、第１の話者３０１に対して、マイクロホン１１－２よりもマイクロホン１１－１の方が近い。そのため、第１の話者３０１が発話している場合、マイクロホン１１－１により集音された第１の音声信号のパワーの方が、マイクロホン１１－２により集音された第２の音声信号のパワーよりも大きいと推定される。したがって、発話区間開始検出直後における、第１の音声信号のパワーの平均値が第２の音声信号のパワーの平均値よりも大きい場合、第１の話者３０１が発話していると判定される。

【００４０】

同様に、第２の話者３０２に対して、マイクロホン１１－１よりもマイクロホン１１－２の方が近い。そのため、第２の話者３０２が発話している場合、マイクロホン１１－２により集音された第２の音声信号のパワーの方が、マイクロホン１１－１により集音された第１の音声信号のパワーよりも大きいと推定される。したがって、発話区間開始検出直後における、第２の音声信号のパワーの平均値が第１の音声信号のパワーの平均値よりも大きい場合、第２の話者３０２が発話していると判定される。

【００４１】

なお、話者が３名いることが想定される場合、話者特定部２５は、発話区間開始検出直後の第１の音声信号のパワーの平均値と第２の音声信号のパワーの平均値との比較結果に基づいて、３名の話者のうちの何れが発話したかを判定してもよい。例えば、話者特定部２５は、第１の音声信号のパワーの平均値と第２の音声信号のパワーの平均値との差の絶対値を所定のパワー差閾値と比較する。そして話者特定部２５は、その差の絶対値がパワー差閾値以下である場合、マイクロホン１１－１とマイクロホン１１－２の並び方向に対する法線方向に位置する話者が発話したと判定してもよい。一方、話者特定部２５は、その差の絶対値がパワー差閾値よりも大きく、かつ、第１の音声信号のパワーの平均値が第２の音声信号のパワーの平均値よりも大きい場合、マイクロホン１１－２よりもマイクロホン１１－１に近い方の話者が発話したと判定する。また、話者特定部２５は、その差の絶対値がパワー差閾値よりも大きく、かつ、第２の音声信号のパワーの平均値が第１の音声信号のパワーの平均値よりも大きい場合、マイクロホン１１－１よりもマイクロホン１１－２に近い方の話者が発話したと判定する。

【００４２】

あるいは、話者特定部２５は、発話区間開始直後の所定数のフレームにおける第１の音声信号と第２の音声信号に基づいて、音源方向を推定し、推定した音源方向の話者が発話していると判定してもよい。この場合、話者特定部２５は、例えば、発話区間開始検出直後の所定数のフレームについて、第１の音声信号と第２の音声信号間の正規化相互相関値を、互いに対する時間差をずらしながら算出する。話者特定部２５は、正規化相互相関値が最も高くなる時間差を遅延時間として特定する。そして話者特定部２５は、マイクロホン１１－１とマイクロホン１１－２間の距離と、遅延時間とに基づいて、音源方向を推定すればよい。推定された音源方向が、マイクロホン１１－１とマイクロホン１１－２の並び方向に対する法線方向よりも、マイクロホン１１－１側を向いている場合、話者特定部２５は、マイクロホン１１－２よりもマイクロホン１１－１に近い方の話者が発話したと判定する。なお、以下では、マイクロホン１１－１とマイクロホン１１－２の並び方向に対する法線方向を、マイクロホンの並び方向に対する法線方向と呼ぶ。一方、推定された音源方向が、マイクロホンの並び方向に対する法線方向よりも、マイクロホン１１－２側を向いている場合、話者特定部２５は、マイクロホン１１－１よりもマイクロホン１１－２に近い方の話者が発話したと判定する。なお、話者が３名いることが想定される場合、話者特定部２５は、推定された音源方向が、マイクロホンの並び方向に対する法線方向から±45°未満である場合、その法線方向に位置する話者が発話したと判定してもよい。また、話者特定部２５は、推定された音源方向とマイクロホンの並び方向に対する法線方向とのなす角が45°以上であり、かつ、その法線方向よりもマイクロホン１１－１側を向いている場合、マイクロホン１１－１に近い方の話者が発話したと判定する。さらに、話者特定部２５は、推定された音源方向とマイクロホンの並び方向に対する法線方向とのなす角が45°以上であり、かつ、その法線方向よりもマイクロホン１１－２側を向いている場合、マイクロホン１１－２に近い方の話者が発話したと判定する。

【００４３】

なお、話者特定部２５は、開始タイミング修正部２６により、発話区間の開始タイミングが修正されると、修正後の発話区間の開始タイミングから所定数のフレームの第１及び第２の音声信号に対して上記と同様の処理を行って、再度話者を特定する。

【００４４】

話者特定部２５は、特定した話者を音声処理部２８へ通知する。

【００４５】

開始タイミング修正部２６は、発話区間開始検出部２４により、発話区間の開始が検出されてからの第１及び第２の音声信号のそれぞれに基づいて、発話区間の開始タイミングを修正するか否か判定する。

【００４６】

雑音が急に大きくなることにより、発話区間開始検出部２４が、雑音が急に大きくなったタイミングを、発話区間の開始タイミングと誤検出することがある。発話区間の開始タ

イミングが誤検出された後に、何れかの話者が発話を開始すると、第１及び第２の音声信号のパワーは実際の発話開始後にさらに大きくなる。そのため、誤検出された発話区間の開始タイミング直後における第１及び第２の音声信号のパワーに対して、実際の発話区間における、第１及び第２の音声信号のパワーの最大値は相対的に大きくなる。

【００４７】

一方、何れかの話者が発話を継続している間、第１及び第２の音声信号には、その話者の声が含まれるので、何れかの話者が発話を継続している期間中の第１及び第２の音声信号のパワーは、そのパワーの最大値と比較してそれほど低下しない。

【００４８】

そこで、開始タイミング修正部２６は、発話区間の開始が検出された後の第１及び第２の音声信号のそれぞれについて、パワーの最大値を検出する。そして開始タイミング修正部２６は、検出したパワーの最大値に対するパワーの低下量が所定のパワー差以上となるフレームが所定数連続すると、その連続するフレームのうちの最初のフレームを、発話区間の開始タイミングに修正する。また、開始タイミング修正部２６は、第１及び第２の音声信号のそれぞれについて、発話区間検出用の閾値を、パワーの最大値から所定のパワー差を減じた値に更新する。なお、所定のパワー差は、例えば、何れかの話者が発話を継続している期間における、その話者の声による想定されるパワーの最大値とパワーの最小値の差に設定される。

【００４９】

なお、開始タイミング修正部２６は、発話区間の開始タイミングについての修正判定に利用する各フレームのパワーとして、パワー算出部２１により算出された値そのものを用いてもよい。あるいは、開始タイミング修正部２６は、その修正判定に利用する各フレームのパワーとして、パワー算出部２１により算出された値から、推定雑音成分を減じた値を用いてもよい。あるいはまた、開始タイミング修正部２６は、その修正判定に利用する各フレームのパワーとして、パワーの移動平均値を算出し、その移動平均値を用いてもよい。

【００５０】

図４は、発話区間開始タイミングの修正についての説明図である。図４において、横軸は時間を表し、縦軸はパワーを表す。波形４０１は、着目する音声信号のパワーの時間変化を表す。また、波形４０２は、推定雑音成分のパワーの時間変化を表す。さらに、波形４０３は、発話区間検出用の閾値Thの時間変化を表す。

【００５１】

この例では、時刻t0～時刻t1までは、着目する音声信号のパワーは閾値Th未満となっているので、発話区間ではないと判定される。そして時刻t1の直前より、例えば、雑音が急激に大きくなることで着目する音声信号のパワーが上昇する。このとき、雑音の増加が急激であるため、閾値Thに、雑音の増加が反映されず、その結果として、時刻t1にて、着目する音声信号のパワーが閾値Th以上となる。そこで、時刻t1にて、発話区間開始検出部２４により、発話区間が開始したと判定される。

【００５２】

時刻t1より後の時刻t2の直前にて、何れかの話者が実際に発話を開始することで、時刻t2の直前より、着目する音声信号のパワーはさらに大きくなる。その結果、時刻t2以降の各フレームにおいて、閾値Thが、発話区間中のパワーの最大値Pmaxから所定のパワー差αだけ低下した値(Pmax-α)未満となる。そのため、時刻t2に、発話区間の開始タイミングが修正される。また閾値Thは、(Pmax-α)に更新される。その後、発話区間の開始検出後において、着目する音声信号のパワーが更新後の閾値Th未満となる最初のフレームの直前のフレームの時刻t3において、発話区間が終了したと判定される。

【００５３】

このように、閾値Thが更新されることで、時刻t1～時刻t2までの雑音しか含まれない区間が発話区間から除外されるので、発話区間が正確に求められる。

【００５４】

なお、変形例によれば、開始タイミング修正部２６は、第１及び第２の音声信号のうち、発話区間の開始検出後のパワーの最大値が大きい方の音声信号についてのみ、上記の処理を行って、発話区間の開始タイミングを修正するか否かを判定してもよい。これは、発話区間の開始検出後のパワーの最大値が大きい方の音声信号の方が、他方の音声信号よりも発話している話者の声を多く含んでいると想定されるためである。このように、一方の音声信号のみに基づいて発話区間の開始タイミングを修正するか否かを判定することで、開始タイミング修正部２６は、演算量を削減できる。

【００５５】

開始タイミング修正部２６は、発話区間の開始タイミングを修正すると、その旨を話者特定部２５へ通知する。そして上記のように、話者特定部２５は、発話区間の開始タイミングを修正したことが通知されると、発話区間にて発話している話者を再度特定する。さらに、開始タイミング修正部２６は、発話区間の開始タイミングを修正すると、第１及び第２の音声信号のそれぞれについて、更新された閾値Thを発話区間終了検出部２７へ通知する。

【００５６】

発話区間終了検出部２７は、発話区間の開始が検出された以降の各フレームの第１及び第２の音声信号の少なくとも一方のパワーに基づいて発話区間が終了したか否か判定する。

【００５７】

例えば、発話区間終了検出部２７は、マイクロホン１１－１と１１－２のうち、話者特定部２５により特定された話者に近い方のマイクロホンにより集音された音声信号（以下、着目音声信号と呼ぶ）のフレームのパワーを発話区間検出用の閾値と比較する。発話区間終了検出部２７は、直前のフレームにおける着目音声信号のパワーが発話区間検出用の閾値以上であり、かつ、現フレームにおける着目音声信号のパワーが発話区間検出用の閾値未満である場合、直前のフレームにて発話区間が終了したと判定する。

【００５８】

あるいは、発話区間終了検出部２７は、着目音声信号のパワーが発話区間検出用の閾値未満となるフレームが所定数連続すると、最初に着目音声信号のパワーが発話区間検出用の閾値未満となったフレームの直前のフレームにて発話区間が終了したと判定してもよい。

【００５９】

あるいはまた、発話区間終了検出部２７は、第１の音声信号及び第２の音声信号のそれぞれについて、上記の何れかの発話区間終了検出処理を実行してもよい。そして発話区間終了検出部２７は、第１の音声信号及び第２の音声信号の何れか一方、あるいは両方について、発話区間が終了したと判定される条件を満たした場合に、発話区間が終了したと判定してもよい。

【００６０】

なお、発話区間検出用の閾値が、開始タイミング修正部２６により更新されている場合には、発話区間終了検出部２７は、更新後の閾値を利用すればよい。この場合も、一旦発話区間が終了したと判定された後に、再度発話区間の開始を検出する際には、閾値設定部２３により算出される、推定雑音成分に基づく閾値が利用されればよい。

【００６１】

発話区間終了検出部２７は、発話区間の終了を検出すると、その旨を音声処理部２８へ通知する。

【００６２】

音声処理部２８は、発話区間の開始が検出されると、発話していると特定された話者に対応する音声処理を実行する。その際、音声処理部２８は、第１及び第２の音声信号の何れに対して音声処理を実行してもよいが、例えば、マイクロホン１１－１とマイクロホン１１－２のうち、特定された話者に近い方のマイクロホンにて集音された音声信号に対して音声処理を実行する。発話中の話者に近い方のマイクロホンにより集音された音声信号

の信号対雑音比の方が、発話中の話者から遠い方のマイクロホンにより集音された音声信号の信号対雑音比よりも高いことが想定される。そのため、音声処理部２８は、発話していると特定された話者に近い方のマイクロホンにて集音された音声信号に対して音声処理を実行することで、より適切な音声処理結果が得られる。

【００６３】

図５は、話者と音声処理の対応関係の一例を示す図である。本実施形態では、マイクロホン１１－１に近い方の第１の話者５０１が日本語を話すと想定され、一方、マイクロホン１１－２に近い方の第２の話者５０２が英語を話す想定される。したがって、特定された話者が第１の話者５０１である場合、音声処理部２８は、第１の音声信号に対して日本語を対象言語とする音声認識処理を実行し、認識された発話内容に対して日本語から英語への自動翻訳処理を実行する。一方、特定された話者が第２の話者５０２である場合、音声処理部２８は、第２の音声信号に対して英語を対象言語とする音声認識処理を実行し、認識された発話内容に対して英語から日本語への自動翻訳処理を実行する。

【００６４】

例えば、音声処理部２８は、発話区間中に話者が発話した内容を認識するために、処理対象となる方の音声信号の発話区間中の各フレームから、話者の声の特徴を表す複数の特徴量を抽出する。そのような特徴量として、例えば、メル周波数ケプストラムの所定の次数の係数が用いられる。そして音声処理部２８は、例えば、各フレームの特徴量を、隠れマルコフモデルによる音響モデルに適用することで、発話区間内の音素系列を認識する。そして音声処理部２８は、単語ごとの音素系列を表す単語辞書を参照して、発話区間の音素系列と一致する単語の組み合わせを検出することで、発話区間内の発話内容を認識する。そして音声処理部２８は、発話内容に応じた単語の組み合わせに対して自動翻訳処理を行って、その発話内容を他言語に翻訳する。なお、音声処理部２８は、自動翻訳処理として、様々な自動翻訳手法の何れを適用してもよい。そして音声処理部２８は、翻訳された発話内容に応じた文字列を表示装置１５に表示する。あるいは、音声処理部２８は、その翻訳された文字列に音声合成処理を適用して、その文字列に対応した合成音声信号を生成し、その合成音声信号をスピーカ（図示せず）を介して再生してもよい。

【００６５】

なお、話者が３名いることが想定され、特定された話者が第１及び第２の話者の何れでもない場合には、音声処理部２８は、発話区間の第１及び第２の音声信号の何れか一方に対して、日本語及び英語以外の言語を対象とする音声認識処理を実行してもよい。あるいは、音声処理部２８は、特定された話者が第１及び第２の話者の何れでもない場合には、前回適用された言語の音声認識処理を実行してもよい。

【００６６】

また、音声処理部２８は、音声処理の開始後、かつ、発話区間の終了が通知される前に、話者特定部２５から特定された話者が再度通知され、かつ、前回通知された話者と再度通知された話者が異なっている場合には、既に開始している音声処理を停止する。そして音声処理部２８は、再度通知された話者に対応する音声処理を実行する。これにより、発話区間の開始タイミングが誤検出されることで、特定された話者が誤っている場合でも、誤って特定された話者に対応する音声処理が継続することが防止される。

【００６７】

図６は、発話区間の開始タイミングの修正と音声処理の関係の一例を示す図である。図６において、横軸は時間を表す。波形６０１は、第１及び第２の音声信号の一方の波形の一例である。この例では、時刻t1～時刻t2までは、音声信号には雑音成分のみが含まれ、話者の声は含まれないものとする。一方、時刻t2～時刻t3において、マイクロホン１１－２に近い方の話者が発話しているとする。

【００６８】

時刻t1において、発話区間の開始が誤検出され、マイクロホン１１－１に近い方の第１の話者が発話していると判定されたとする。この場合、誤検出された区間６０２において、音声処理部２８により、日本語を認識対象とする音声認識処理が実行されることになる

。そして発話区間の開始が修正されなければ、実際に発話が開始された時刻t2以降においても、日本語を認識対象とする音声認識処理が継続されるため、話者の発話内容が正確に認識されない。

【００６９】

一方、本実施形態では、時刻t2にて発話区間の開始タイミングが修正され、修正された発話区間の開始タイミングにて再度発話している話者が特定される。そのため、実際の発話区間６０３では、実際に発話している、マイクロホン１１－２に近い方の第２の話者に対応する、英語を認識対象とする音声認識処理が実行される。したがって、音声処理部２８は、実際に発話している話者の発話内容を正確に認識することができる。

なお、誤検出された区間に対する日本語を認識対象とする音声認識処理は、修正された発話区間の開始タイミングにて停止される。

【００７０】

図７は、本実施形態による、音声処理の動作フローチャートである。プロセッサ１３は、フレームごとに、下記の動作フローチャートに従って音声処理を実行する。

【００７１】

パワー算出部２１は、第１及び第２の音声信号のそれぞれについて、現フレームのパワーPを算出する（ステップＳ１０１）。雑音推定部２２は、第１及び第２の音声信号のそれぞれについて、現フレームのパワーPと、直前のフレームにおける推定雑音成分とに基づいて、現フレームの推定雑音成分を算出する（ステップＳ１０２）。

【００７２】

閾値設定部２３は、直前のフレームが発話区間内か否か判定する（ステップＳ１０３）。直前のフレームが発話区間外であれば（ステップＳ１０３－Ｎｏ）、第１及び第２の音声信号のそれぞれについて、閾値設定部２３は、推定雑音成分に基づいて閾値Thを設定する（ステップＳ１０４）。そして発話区間開始検出部２４は、第１及び第２の音声信号のそれぞれについて、現フレームのパワーPが閾値Th以上か否か判定する（ステップＳ１０５）。

【００７３】

第１及び第２の音声信号の両方について、現フレームのパワーPが閾値Th未満であれば（ステップＳ１０５－Ｎｏ）、発話区間開始検出部２４は、現フレームは発話区間に含まれないと判定する。そしてプロセッサ１３は、音声処理を終了する。一方、第１及び第２の音声信号の少なくとも一方について、現フレームのパワーPが閾値Th以上であれば（ステップＳ１０５－Ｙｅｓ）、発話区間開始検出部２４は、現フレームから発話区間が開始したと判定する（ステップＳ１０６）。そして発話区間開始検出部２４は、現フレームを発話区間の開始タイミングとして検出する。話者特定部２５は、開始した発話区間において発話した話者を、第１及び第２の音声信号に基づいて特定する（ステップＳ１０７）。さらに、音声処理部２８は、第１及び第２の音声信号の何れかに対して、特定された話者に応じた処理を実行する（ステップＳ１０８）。その後、プロセッサ１３は、現フレームにおける音声処理を終了する。

【００７４】

また、ステップＳ１０３において、直前のフレームが発話区間に含まれる場合（ステップＳ１０３－Ｙｅｓ）、発話区間の開始タイミングが既に検出されている。そこで開始タイミング修正部２６は、第１及び第２の音声信号のそれぞれについて、発話区間開始後のパワーの最大値Pmaxから所定のパワー差αだけ減じた値よりも閾値Thが小さいフレームが所定数以上継続したか否か判定する（ステップＳ１０９）。

【００７５】

第１及び第２の音声信号の少なくとも一方について、現フレームにおいて、$(Pmax-\alpha)>Th$となるフレームが所定数以上連続している場合（ステップＳ１０９－Ｙｅｓ）、開始タイミング修正部２６は、閾値Thを$(Pmax-\alpha)$に更新する。そして開始タイミング修正部２６は、発話区間の開始タイミングをその連続するフレームのうちの最初のフレームに修正する（ステップＳ１１０）。その後、プロセッサ１３は、ステップＳ１０７以降の処理を

実行する。なお、この場合、ステップＳ１０８において、音声処理部２８は、発話区間の開始タイミングの修正前後において、特定された話者が異なる場合、発話区間の開始タイミングの修正前に行っている音声処理を停止する。

【００７６】

一方、第１及び第２の音声信号の両方について、現フレームにおいて、(Pmax-α)>Thとなるフレームの継続数が所定数未満であれば（ステップＳ１０９－Ｎｏ）、開始タイミング修正部２６は、発話区間の開始タイミングを修正しない。一方、発話区間終了検出部２７は、第１及び第２の音声信号のうち、音声処理部２８の音声処理対象となる方の音声信号の現フレームのパワーPが閾値Th未満か否か判定する（ステップＳ１１１）。パワーPが閾値Th未満であれば（ステップＳ１１１－Ｙｅｓ）、発話区間終了検出部２７は、直前のフレームにて発話区間が終了したと判定する（ステップＳ１１２）。そしてプロセッサ１３は、音声処理部２８に対して発話区間の終了を通知する。一方、パワーPが閾値Th以上であれば（ステップＳ１１１－Ｎｏ）、発話区間終了検出部２７は、現フレームも発話区間内に含まれると判定する。そしてプロセッサ１３は、ステップＳ１０８の処理を実行する。

【００７７】

以上に説明してきたように、この音声処理装置は、発話区間の開始が検出されると、その発話区間にて発話した話者を特定し、特定した話者に応じた音声処理を、第１及び第２の音声信号の少なくとも一方に対して実行する。そしてこの音声処理装置は、一旦発話区間の開始が検出された後に、発話区間の開始タイミングが修正されると、修正された開始タイミングにて、複数の話者のうち、その発話区間にて発話した話者を再度特定する。そして音声処理装置は、再特定された話者に応じた音声処理を第１及び第２の音声信号の少なくとも一方に対して実行する。そのため、この音声処理装置は、各音声信号中で複数の話者の何れかが発話を開始したタイミングを誤検出しても、発話した話者に応じた処理を音声信号に適用できる。

【００７８】

なお、変形例によれば、音声処理部２８は、音声認識処理及び自動翻訳処理以外の処理を実行してもよい。例えば、第１の話者の周囲がエコーを生じ易く、第２の話者の周囲に雑音源が存在するとする。この場合、第１の話者が発話していると判定された場合、音声処理部２８は、発話区間における第１及び第２の音声信号の少なくとも一方に対してエコー除去処理を実行してもよい。一方、第２の話者が発話していると判定された場合、音声処理部２８は、発話区間における第１及び第２の音声信号の少なくとも一方に対して雑音除去処理を実行してもよい。

【００７９】

また、発話区間開始検出部２４及び開始タイミング修正部２６は、各フレームのパワー以外の、音声信号に含まれる、話者の声を表す特徴量に基づいて、発話区間の開始タイミングの検出、及びその開始タイミングの修正判定を行ってもよい。例えば、発話区間開始検出部２４は、第１及び第２の音声信号の各フレームから、音の周期性の強さを表すピッチゲインを算出する。そして発話区間開始検出部２４は、第１及び第２の音声信号の少なくとも一方について、直前のフレームのピッチゲインが閾値未満となり、かつ、現フレームのピッチゲインが閾値以上となる場合、発話区間の開始を検出してもよい。なお、ピッチゲインg_{pitch}は、例えば、次式に従って算出される。

【数３】

$$g_{pitch}=\frac{C(d_{\max})}{\sum_{n=0}^{N-1}s_k(n)\cdot s_k(n)}$$
$$C(d)=\sum_{n=0}^{N-1}s_k(n)\cdot s_k(n-d)\quad(d=d_{low},\ldots,d_{high}) \qquad (3)$$

ここで、C(d)は、着目する音声信号の長期自己相関である。またd∈{$d_{low},\ldots,d_{high}$}は、遅延量である。そして$S_k(n)$は、現フレームkのn番目の信号値である。またNは、フレームに含まれるサンプリング点の総数を表す。なお、(n-d)が負となる場合、直前のフレームの対応する信号値（すなわち、フレーム区間の重複がない場合、$S_{k-1}(N-(n-d))$）が$S_k(n-d)$として用いられる。そして遅延量dの範囲{$d_{low},\ldots,d_{high}$}は、人の声の基本周波数(100～300Hz)に相当する遅延量が含まれるように設定される。ピッチゲインは、基本周波数において最も高くなるためである。例えば、サンプリングレートが16kHzである場合、d_{low}=40、d_{high}=286に設定される。さらに、d_{max}は、長期自己相関C(d)の最大値C(d_{max})に対応する遅延量であり、この遅延量はピッチ周期に相当する。

【００８０】

一般に、ピッチゲインは、発話が開始された直後において最も大きく、発話が継続するにつれて小さくなる。そこで、開始タイミング修正部２６は、第１及び第２の音声信号の少なくとも一方について、発話区間の開始を検出された直後の所定数のフレームのピッチゲインの最大値と、発話区間の開始検出後の各フレームのピッチゲインを比較する。そして開始タイミング修正部２６は、そのピッチゲインの最大値よりも所定のオフセット値以上ピッチゲインが大きくなるフレームを検出すると、そのフレームに発話区間の開始タイミングを修正すればよい。

【００８１】

なお、この変形例の場合、発話区間終了検出部２７は、発話区間の開始検出後において、第１及び第２の音声信号の両方についてピッチゲインが閾値未満となる最初のフレームにおいて、発話区間が終了したと判定してもよい。あるいは、発話区間終了検出部２７は、第１及び第２の音声信号の両方についてピッチゲインが閾値未満となるフレームが所定数連続する場合、ピッチゲインが閾値未満となった最初のフレームにおいて、発話区間が終了したと判定してもよい。また、発話区間終了検出部２７は、パワーとピッチゲインが共に閾値未満となる最初のフレームにおいて、発話区間が終了したと判定してもよい。

【００８２】

上記の実施形態または変形例による音声処理装置は、サーバクライアント型のシステムに実装されてもよい。

図８は、上記の実施形態またはその変形例による音声処理装置が実装されたサーバクライアントシステムの概略構成図である。

サーバクライアントシステム１００は、端末１１０とサーバ１２０とを有し、端末１１０とサーバ１２０とは、通信ネットワーク１３０を介して互いに通信可能となっている。なお、サーバクライアントシステム１００が有する端末１１０は複数存在してもよい。同様に、サーバクライアントシステム１００が有するサーバ１２０は複数存在してもよい。

【００８３】

端末１１０は、二つのマイクロホン１１１－１、１１１－２と、メモリ１１２と、通信インターフェース１１３と、プロセッサ１１４と、表示装置１１５とを有する。マイクロホン１１１、メモリ１１２及び通信インターフェース１１３は、例えば、プロセッサ１１４とバスを介して接続されている。

【００８４】

マイクロホン１１１－１、１１１－２は、それぞれ、音声入力部の一例である。マイクロホン１１１－１は、アナログ信号である第１の音声信号を取得し、第１の音声信号をＡ／Ｄコンバータ（図示せず）へ出力する。Ａ／Ｄコンバータによりデジタル化された第１の音声信号はプロセッサ１１４へ出力される。同様に、マイクロホン１１１－２は、アナログ信号である第２の音声信号を取得し、第２の音声信号をＡ／Ｄコンバータ（図示せず）へ出力する。Ａ／Ｄコンバータによりデジタル化された第２の音声信号はプロセッサ１１４へ出力される。

【００８５】

メモリ１１２は、例えば、不揮発性の半導体メモリ及び揮発性の半導体メモリを有する。そしてメモリ１１２は、端末１１０を制御するためのコンピュータプログラム、端末１１０の識別情報、発話区間検出処理で利用される各種のデータ及びコンピュータプログラムなどを記憶する。

【００８６】

通信インターフェース１１３は、端末１１０を通信ネットワーク１３０に接続するためのインターフェース回路を有する。そして通信インターフェース１１３は、プロセッサ１１４から受け取った音声信号を、端末１１０の識別情報とともに通信ネットワーク１３０を介してサーバ１２０へ送信する。

【００８７】

プロセッサ１１４は、CPUとその周辺回路を有する。そしてプロセッサ１１４は、第１及び第２の音声信号を、端末１１０の識別情報とともに、通信インターフェース１１３及び通信ネットワーク１３０を介してサーバ１２０へ送信する。またプロセッサ１１４は、サーバ１２０から受け取った、各音声信号に対する処理結果を表示装置１１５に表示するか、あるいは、その処理結果に対応する合成音声信号をスピーカ（図示せず）を介して再生する。

【００８８】

表示装置１１５は、例えば、液晶ディスプレイまたは有機ＥＬディスプレイであり、各音声信号に対する処理結果を表示する。

【００８９】

サーバ１２０は、通信インターフェース１２１と、メモリ１２２と、プロセッサ１２３とを有する。通信インターフェース１２１及びメモリ１２２は、プロセッサ１２３とバスを介して接続されている。

【００９０】

通信インターフェース１２１は、サーバ１２０を通信ネットワーク１３０に接続するためのインターフェース回路を有する。そして通信インターフェース１２１は、第１及び第２の音声信号と端末１１０の識別情報とを端末１１０から通信ネットワーク１３０を介して受信してプロセッサ１２３に渡す。

【００９１】

メモリ１２２は、例えば、不揮発性の半導体メモリ及び揮発性の半導体メモリを有する。そしてメモリ１２２は、サーバ１２０を制御するためのコンピュータプログラムなどを記憶する。またメモリ１２２は、音声処理を実行するためのコンピュータプログラム及び各端末から受信した各音声信号を記憶してもよい。

【００９２】

プロセッサ１２３は、CPUとその周辺回路を有する。そしてプロセッサ１２３は、上記の実施形態または変形例による音声処理装置のプロセッサの各機能を実現する。そしてプロセッサ１２３は、受信した第１及び第２の音声信号に対する音声処理結果を通信インターフェース１２１及び通信ネットワーク１３０を介して端末１１０へ送信する。

【００９３】

なお、端末１１１０のプロセッサ１１４が、上記の実施形態または変形例による音声処理装置のプロセッサの各機能のうち、音声処理部２８以外の処理を実行してもよい。この場合、端末１１０は、発話区間中の第１及び第２の音声信号の少なくとも何れかと、特定

された話者を表す情報とをサーバ１２０へ送信すればよい。また、端末１１０は、発話区間の開始タイミングを修正した場合には、修正された発話区間の開始タイミング及び再特定された話者を表す情報をサーバ１２０へ送信する。そしてサーバ１２０のプロセッサ１２３は、受信した第１及び第２の音声信号の少なくとも一方に対して、音声処理部２８の処理を実行すればよい。

【００９４】

上記の実施形態または変形例による発話区間検出装置のプロセッサが有する各機能をコンピュータに実現させるコンピュータプログラムは、磁気記録媒体または光記録媒体といったコンピュータによって読み取り可能な媒体に記録された形で提供されてもよい。

【００９５】

ここに挙げられた全ての例及び特定の用語は、読者が、本発明及び当該技術の促進に対する本発明者により寄与された概念を理解することを助ける、教示的な目的において意図されたものであり、本発明の優位性及び劣等性を示すことに関する、本明細書の如何なる例の構成、そのような特定の挙げられた例及び条件に限定しないように解釈されるべきものである。本発明の実施形態は詳細に説明されているが、本発明の精神及び範囲から外れることなく、様々な変更、置換及び修正をこれに加えることが可能であることを理解されたい。

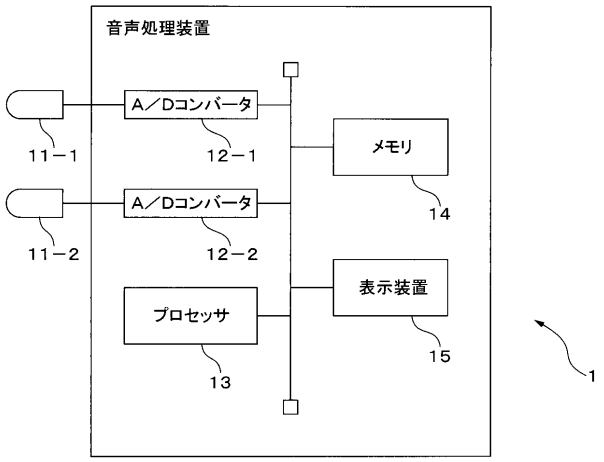
【符号の説明】

【００９６】

１　音声処理装置
１１－１、１１－２　マイクロホン
１２－１、１２－２　アナログ／デジタルコンバータ
１３　プロセッサ
１４　メモリ
１５　表示装置
２１　パワー算出部
２２　雑音推定部
２３　閾値設定部
２４　発話区間開始検出部
２５　話者特定部
２６　開始タイミング修正部
２７　発話区間終了検出部
２８　音声処理部
１００　サーバクライアントシステム
１１０　端末
１１１－１、１１１－２　マイクロホン
１１２　メモリ
１１３　通信インターフェース
１１４　プロセッサ
１１５　表示装置
１２０　サーバ
１２１　通信インターフェース
１２２　メモリ
１２３　プロセッサ
１３０　通信ネットワーク

【図１】

図1



【図２】

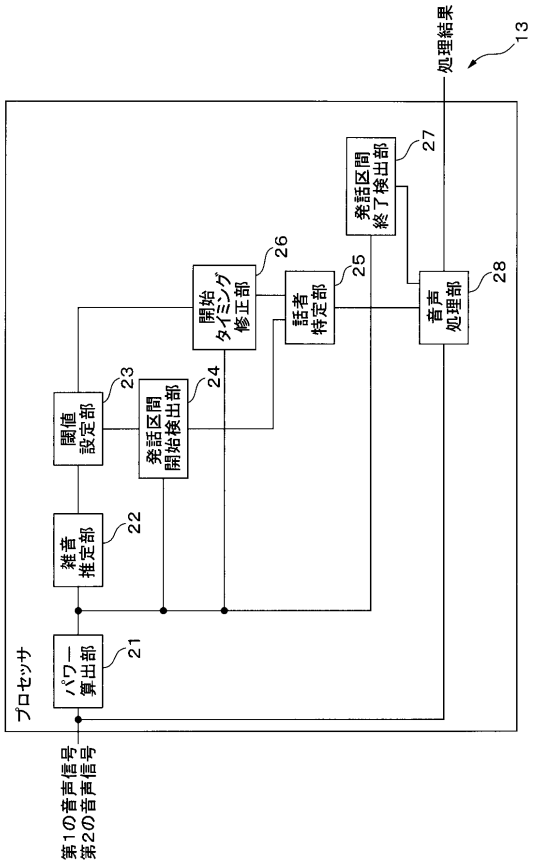
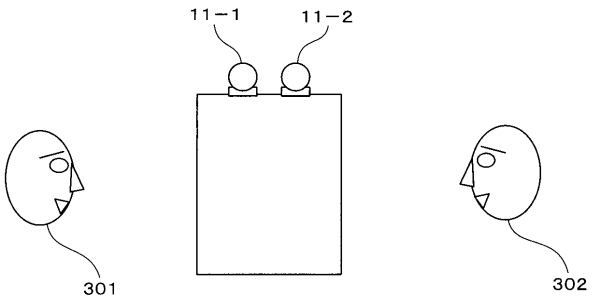


図2

【図３】

図3



【図４】

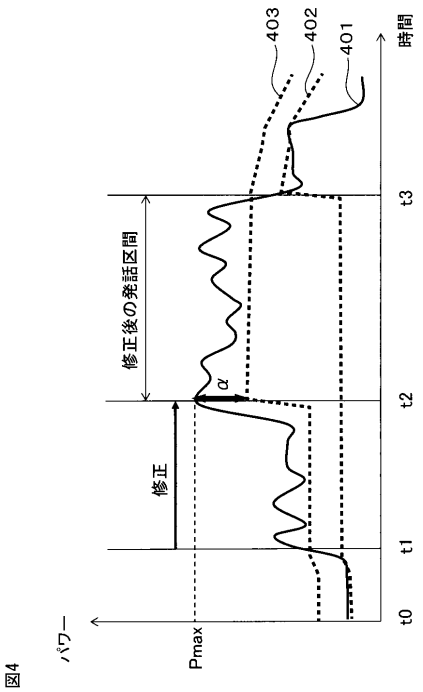
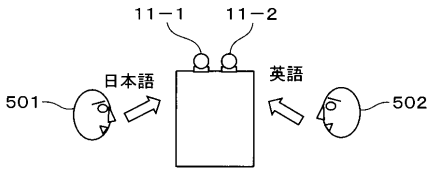


図4

【図５】

図5



話者の位置	話者特定結果	言語判定
右側：英語話者 左側：日本語話者	左側の話者	日本語
	右側の話者	英国語

【図６】

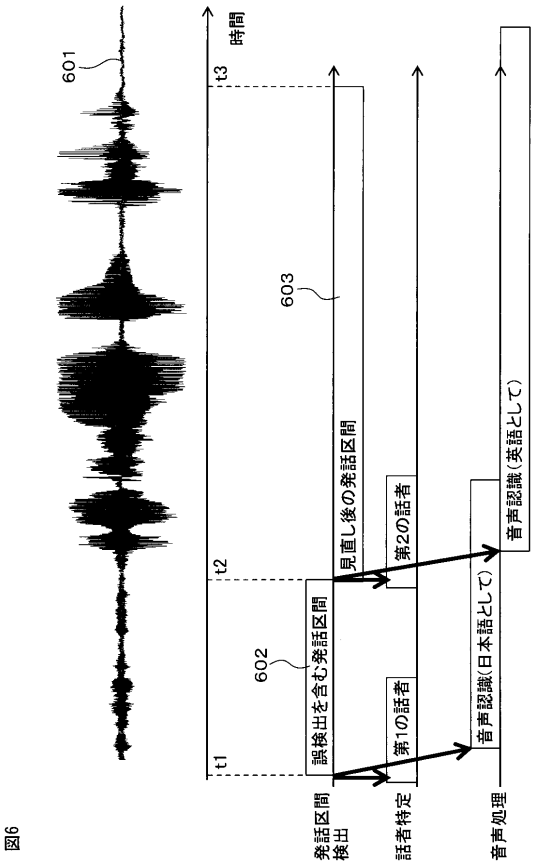
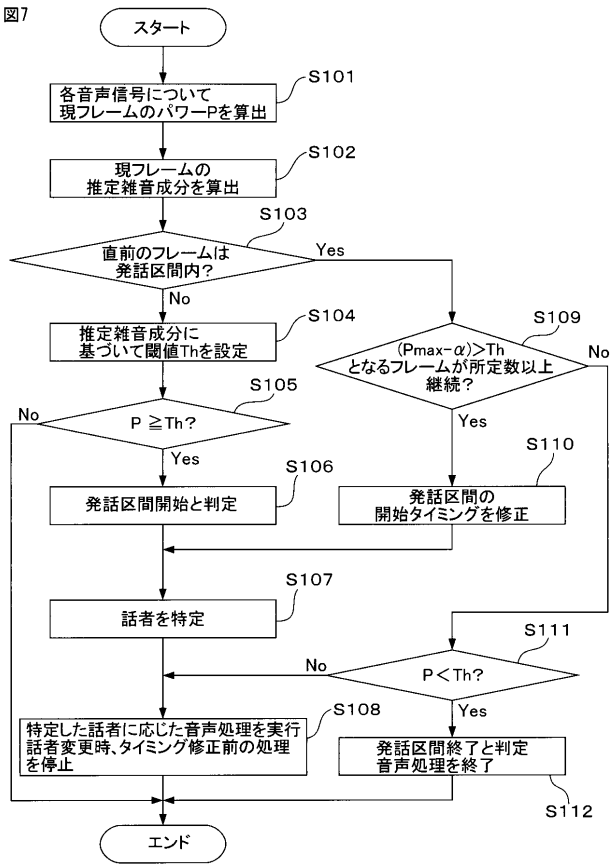


図6

【図７】

図7



【図８】

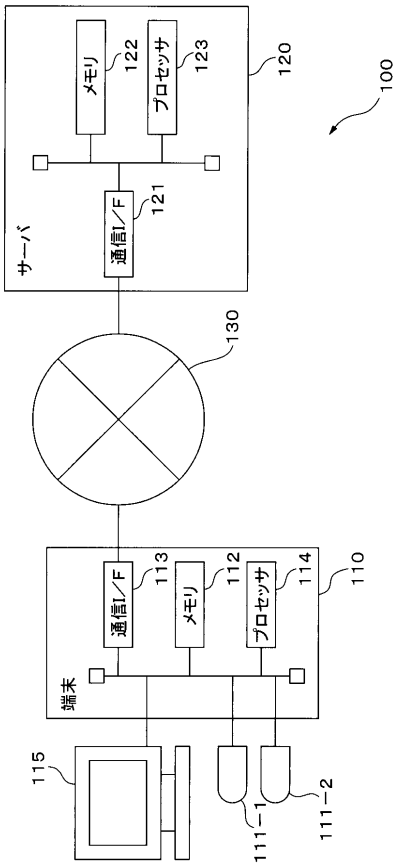


図8