(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6759898号
(P6759898)
(45)発行日　令和2年9月23日(2020.9.23)　(24)登録日　令和2年9月7日(2020.9.7)

(51)Int.Cl.　F I
G10L 15/04 (2013.01)　G10L 15/04 300B
G10L 25/84 (2013.01)　G10L 25/84

請求項の数 6　(全 21 頁)

(21)出願番号　特願2016-175765(P2016-175765)
(22)出願日　平成28年9月8日(2016.9.8)
(65)公開番号　特開2018-40982(P2018-40982A)
(43)公開日　平成30年3月15日(2018.3.15)
審査請求日　令和1年6月11日(2019.6.11)

(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　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　鈴木　政直
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

最終頁に続く

(54)【発明の名称】発話区間検出装置、発話区間検出方法及び発話区間検出用コンピュータプログラム

(57)【特許請求の範囲】
【請求項１】
話者の声が表された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するピッチゲイン算出部と、
前記フレームごとに、前記音声信号の信号対雑音成分比を算出する信号対雑音成分比算出部と、
前記フレームごとの前記信号対雑音成分比に基づいて前記音声信号に信号成分が含まれる有音区間を検出する有音区間検出部と、
前記有音区間内において直前のフレームが発話区間でなく、かつ、現フレームの前記ピッチゲインが第１の閾値以上となる場合に前記現フレームから前記発話区間が開始されたと判定し、かつ、前記発話区間が継続している場合において前記ピッチゲインが前記第１の閾値よりも小さい第２の閾値未満となると前記発話区間が終了すると判定する発話区間検出部と、
を有する発話区間検出装置。
【請求項２】
前記発話区間検出部は、前記発話区間が開始されたと判定されたフレームにおける前記ピッチゲインが大きいほど、前記第２の閾値を高くする、請求項１に記載の発話区間検出装置。
【請求項３】
前記発話区間検出部は、前記信号対雑音成分比が大きいフレームほど、当該フレームに

おける前記第１の閾値及び前記第２の閾値を高くする、請求項１に記載の発話区間検出装置。
【請求項４】
前記発話区間検出部は、前記発話区間が継続している場合において前記ピッチゲインが前記第２の閾値未満となる期間が一定期間継続すると前記発話区間が終了したと判定する、請求項１～３の何れか一項に記載の発話区間検出装置。
【請求項５】
話者の声が表された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、
前記フレームごとに、前記音声信号の信号対雑音成分比を算出し、
前記フレームごとの前記信号対雑音成分比に基づいて前記音声信号に信号成分が含まれる有音区間を検出し、
前記有音区間内において直前のフレームが発話区間でなく、かつ、現フレームの前記ピッチゲインが第１の閾値以上となる場合に前記現フレームから前記発話区間が開始されたと判定し、かつ、前記発話区間が継続している場合において前記ピッチゲインが前記第１の閾値よりも小さい第２の閾値未満となると前記発話区間が終了すると判定する、
ことを含む発話区間検出方法。
【請求項６】
話者の声が表された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、
前記話者が発話していない非発話区間が継続している場合において前記ピッチゲインが第１の閾値以上となると前記話者が発話している発話区間が開始されたと判定し、かつ、前記発話区間が継続している場合において前記ピッチゲインが前記第１の閾値よりも小さい第２の閾値未満となると前記発話区間が終了すると判定し、
前記発話区間が開始されたと判定されたフレームにおける前記ピッチゲインが大きいほど、前記第２の閾値を高くする、
ことをコンピュータに実行させるための発話区間検出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、音声信号から話者が発話している区間を検出する発話区間検出装置、発話区間検出方法及び発話区間検出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、話者が発話している区間を特定するために、例えば、話者自身がそのようなアプリケーションが実装された装置を操作して、音声信号の録音開始及び録音終了を指示することが求められる。しかし、そのようなアプリケーションが利用される環境によっては、話者がそのような操作を行えないことがある。例えば、話者が両手を使用する何らかの作業を行っている場合には、話者は音声信号の録音開始及び録音終了を指示する操作を行えない。
【０００３】
一方、音声信号において有音か無音かを判定する技術が提案されている（例えば、特許文献１を参照）。この技術では、入力音声信号の一定区間ごとに音声信号のパワー及びピッチパラメータなどが算出される。そして有音の第１の一定区間に続く次の第２の一定区間に対応するピッチパラメータが所定の閾値より低い場合に、その第２の一定区間が無音の一定区間として判定される。
【先行技術文献】
【特許文献】

【０００４】
【特許文献１】特開平１１－１３３９９７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、話者が発話している区間において、ピッチパラメータは一定ではない。そのため、ピッチパラメータと固定された閾値との比較に基づいて有音か否かを判定する上記の技術では、音声信号中で話者が発話している区間を正確に検出できないことがある。
【０００６】
一つの側面では、本発明は、音声信号中で話者が発話している区間を検出できる発話区間検出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、発話区間検出装置が提供される。この発話区間検出装置は、話者の声が表された音声信号を分割した所定長を持つフレームごとに、音声信号の周期性の強さを表すピッチゲインを算出するピッチゲイン算出部と、話者が発話していない非発話区間が継続している場合においてピッチゲインが第１の閾値以上となると話者が発話している発話区間が開始されたと判定し、かつ、発話区間が継続している場合においてピッチゲインが第１の閾値よりも小さい第２の閾値未満となると発話区間が終了すると判定する発話区間検出部とを有する。
【発明の効果】
【０００８】
音声信号中で話者が発話している区間を検出できる。
【図面の簡単な説明】
【０００９】
【図１】一つの実施形態による発話区間検出装置の概略構成図である。
【図２】発話区間検出処理に関する処理部の機能ブロック図である。
【図３】発話区間検出処理の概要を説明する図である。
【図４】発話区間検出処理の動作フローチャートである。
【図５】変形例による、発話区間と閾値の関係を示す図である。
【図６】SN比と第１の閾値の関係の一例を示す図である。
【図７】変形例による、発話区間の判定に関する状態遷移図である。
【図８】変形例による、発話区間検出処理の動作フローチャートである。
【図９】（ａ）は、変形例による、発話区間開始からの経過時間と第２の閾値との関係の一例を表す図である。（ｂ）は、変形例による、発話区間終了からの経過時間と第１の閾値との関係の一例を表す図である。
【図１０】実施形態またはその変形例による発話区間検出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、実施形態による発話区間検出装置について説明する。
この発話区間検出装置は、音声信号中で話者が発話している区間（以下、単に発話区間と呼ぶ）では、人の声の特性上、ある程度の周期性が認められることから、音の周期性の強さを表すピッチゲインに基づいて発話区間を検出する。これにより、この発話区間検出装置は、話者の声以外でも大きな値を取り得るパワーまたは信号対雑音比を利用するよりも、発話区間をより正確に検出できる。
【００１１】
ここで、話者が連続して発声していると、語尾にかけて話者の呼気圧が低下し、声門閉鎖の周期性が弱くなることが知られている（例えば、上村幸雄、「呼気流・呼気圧と調音

音声学」、人文 6、 pp.247-291、2007年を参照）。このことから、発明者は、発話区間では、語尾にかけてピッチゲインが減衰することに着目した。そこで、この発話区間検出装置は、発話区間の開始を検出する際に用いられる、ピッチゲインに対する第１の閾値よりも、発話区間の終了を検出する際に用いられる、ピッチゲインに対する第２の閾値を低く設定する。

【００１２】

なお、この発話区間検出装置は、音声認識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、携帯電話機またはコンピュータなどに実装できる。

【００１３】

図１は、一つの実施形態による発話区間検出装置の概略構成図である。発話区間検出装置１は、マイクロホン１１と、アナログ／デジタルコンバータ１２と、処理部１３と、記憶部１４とを有する。なお、発話区間検出装置１は、さらに、スピーカ（図示せず）、表示装置（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。

【００１４】

マイクロホン１１は、音声入力部の一例であり、話者の声を含む、発話区間検出装置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１２（以下、Ａ／Ｄコンバータと表記する）へ出力する。Ａ／Ｄコンバータ１２は、アナログの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は、デジタル化された音声信号を処理部１３へ出力する。なお、以下では、デジタル化された音声信号を、単に音声信号と呼ぶ。

【００１５】

処理部１３は、例えば、一つまたは複数のプロセッサと、読み書き可能なメモリ回路と、その周辺回路とを有する。そして処理部１３は、音声信号に対して発話区間検出処理を実行することで発話区間を検出する。処理部１３は、発話区間中の音声信号に対して、例えば、音声認識処理を実行して、話者が発声した語句を認識する。さらに、処理部１３は、認識した語句に応じた処理、例えば、認識した語句を、予め設定された言語の語句に翻訳したり、あるいは、認識した語句をクエリとしてネットワークなどの探索処理を行う。なお、発話区間検出処理の詳細については後述する。

【００１６】

記憶部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、記憶部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。そして記憶部１４は、処理部１３上で実行される発話区間検出処理で利用される各種のデータ及び発話区間検出処理の途中で生成される各種のデータを記憶する。例えば、記憶部１４は、ピッチゲインに対する閾値を記憶する。さらに、記憶部１４は、発話区間中の音声信号に対して行われる処理に関するプログラム及びそのプログラムで利用される各種のデータを記憶してもよい。

【００１７】

以下、処理部１３の詳細について説明する。

【００１８】

図２は、発話区間検出処理に関する処理部１３の機能ブロック図である。処理部１３は、パワー算出部２１と、雑音推定部２２と、信号対雑音比算出部２３と、有音判定部２４と、ピッチゲイン算出部２５と、発話区間検出部２６とを有する。

処理部１３が有するこれらの各部は、例えば、処理部１３が有するプロセッサ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、処理部１３が有するこれらの各部は、その各部の機能を実現する一つまたは複数の集積回路であ

ってもよい。また処理部１３は、音声信号を所定長を持つフレームを処理単位として発話区間検出処理を実行する。フレーム長は、例えば、10msec～20msecに設定される。そのため、処理部１３は、音声信号をフレームごとに分割し、各フレームをパワー算出部２１及びピッチゲイン算出部２５へ入力する。

【００１９】

パワー算出部２１は、フレームが入力される度に、そのフレームについての音声信号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。

【００２０】

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、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、周波数帯域ごとに、その周波数帯域に含まれるスペクトル信号の２乗和を、その周波数帯域のパワーとして算出できる。

【００２１】

パワー算出部２１は、フレームごとのパワーを雑音推定部２２及び信号対雑音比算出部２３へ出力する。

【００２２】

雑音推定部２２は、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いて推定雑音成分を、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。

【００２３】

なお、パワーが周波数帯域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帯域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、着目する周波数帯域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。

【００２４】

雑音推定部２２は、フレームごとの推定雑音成分を信号対雑音比算出部２３へ出力する。

なお、後述する有音判定部２４により、現フレームが何らかの信号成分を含む有音フレームであると判定された場合には、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。

【００２５】

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。

【００２６】

信号対雑音比算出部２３は、フレームごとに、信号対雑音比（以下では、単にSN比と表記する）を算出する。例えば、信号対雑音比算出部２３は、次式に従ってSN比を算出する。

【数３】

$$SNR(k)=10\cdot\log_{10}\frac{Spow(k)}{Noise(k)} \qquad (3)$$

ここで、SNR(k)は、現フレームのSN比を表す。なお、パワー及び推定雑音成分が周波数帯域ごとに算出されている場合には、信号対雑音比算出部２３は、（３）式に従って、SN比を周波数帯域ごとに算出してもよい。この場合には、（３）式において、Noise(k)、Spow(k)及びSNR(k)は、それぞれ、着目する周波数帯域についての現フレームの推定雑音成分、パワー、SN比となる。

【００２７】

信号対雑音比算出部２３は、フレームごとのSN比を有音判定部２４へ出力する。

【００２８】

有音判定部２４は、フレームごとに、そのフレームのSN比に基づいてそのフレームが有音区間に含まれるか否かを判定する。なお、有音区間は、その区間中の音声信号中に何らかの信号成分が含まれると推定される区間である。そのため、発話区間は有音区間に含まれると想定される。そこで、発話区間の検出対象となる区間として有音区間を特定することで、発話区間検出装置１は、発話区間の検出精度を向上できる。

【００２９】

本実施形態では、有音判定部２４は、フレームごとに、そのフレームのSN比を有音判定閾値Thsnrと比較する。なお、有音判定閾値Thsnrは、例えば、音声信号中に推定雑音成分以外の信号成分が含まれることに相当する値、例えば、2～3に設定される。そして有音判定部２４は、SN比が有音判定閾値Thsnr以上であれば、そのフレームは有音区間に含まれると判定する。一方、有音判定部２４は、SN比が有音判定閾値Thsnr未満であれば、そのフレームは有音区間に含まれない、すなわち、無音区間に含まれると判定する。なお、有音判定部２４は、SN比が有音判定閾値Thsnr以上となるフレームが一定期間（例えば、1秒間）連続した時点で、有音区間に入ったと判定してもよい。また、有音判定部２４は、それ以前のフレームが有音区間に含まれると判定されている状態で、SN比が有音判定閾値Thsnr未満となるフレームが一定期間連続した時点で、有音区間が終了したと判定してもよい。

【００３０】

さらに、周波数帯域ごとにSN比が算出されている場合には、有音判定部２４は、SN比が有音判定閾値Thsnr以上となる周波数帯域の数が所定数以上となる場合に、そのフレームは有音区間に含まれると判定してもよい。なお、所定数は、例えば、SN比が算出される周

波数帯域の総数の1/2とすることができる。あるいは、有音判定部２４は、解析対象となる周波数が含まれる周波数帯域についてSN比が有音判定閾値Thsnr以上となる場合に、そのフレームは有音区間に含まれると判定してもよい。

【００３１】

あるいは、有音判定部２４は、フレームごとのパワーそのものに基づいて、フレームごとに有音区間に含まれるか否かを判定してもよい。この場合には、有音判定部２４は、現フレームのパワーが所定の閾値以上であれば、現フレームは有音区間に含まれ、現フレームのパワーが所定の閾値未満であれば、現フレームは無音区間に含まれると判定してもよい。この場合、所定の閾値は、現フレームの推定雑音成分が大きくなるほど、高くなるように設定されてもよい。

【００３２】

有音判定部２４は、フレームごとに、有音区間に含まれるか否かの判定結果を表す情報を雑音推定部２２及びピッチゲイン算出部２５に通知する。なお、有音区間に含まれるか否かの判定結果を表す情報は、例えば、有音区間である場合に"1"となり、無音区間である場合に"0"となるフラグとすることができる。

【００３３】

なお、発話区間検出部２６が発話区間の開始を検出した後において、発話区間の終了を検知するよりも前に、有音判定部２４が現フレームについて無音区間に属すると判定した場合、有音判定部２４は、直前のフレームまでで発話区間が終了したと判定してもよい。

【００３４】

ピッチゲイン算出部２５は、有音区間に含まれる各フレームについて、音の周期性の強さを表すピッチゲインを算出する。なお、ピッチゲインは、ピッチ予測利得とも呼ばれる。ピッチゲイン算出部２５は、有音区間に含まれる各フレームについて同一の処理を実行するので、以下では、一つのフレームに対する処理について説明する。

【００３５】

ピッチゲインを算出するために、ピッチゲイン算出部２５は、先ず、音声信号の長期自己相関C(d)を、遅延量$d \in \{d_{low},\ldots,d_{high}\}$について算出する。

【数４】

$$C(d)=\sum_{n=0}^{N-1} s_k(n)\cdot s_k(n-d) \quad (d=d_{low},\ldots,d_{high}) \qquad (4)$$

上記のように、$S_k(n)$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、$S_{k-1}(N-(n-d))$）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。

【００３６】

ピッチゲイン算出部２５は、遅延量の範囲に含まれる遅延量dごとに長期自己相関C(d)を算出すると、長期自己相関C(d)のうちの最大値$C(d_{max})$を求める。なお、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。そしてピッチゲイン算出部２５は、次式に従ってピッチゲインg_{pitch}を算出する。

【数５】

$$g_{pitch}=\frac{C(d_{max})}{\sum_{n=0}^{N-1} s_k(n)\cdot s_k(n)} \qquad (5)$$

【００３７】
ピッチゲイン算出部２５は、有音区間内のフレームについてピッチゲインg_{pitch}を算出する度に、ピッチゲインg_{pitch}を発話区間検出部２６へ出力する。
【００３８】
発話区間検出部２６は、有音区間内の各フレームについて、ピッチゲインg_{pitch}を発話区間検出用の閾値と比較することで、発話区間を検出する。すなわち、発話区間検出部２６は、話者が発話していない非発話区間が継続している場合においてピッチゲインg_{pitch}が第１の閾値以上となると話者が発話している発話区間が開始されたと判定する。一方、発話区間検出部２６は、発話区間が継続している場合においてピッチゲインが第１の閾値よりも小さい第２の閾値未満となると発話区間が終了すると判定する。
【００３９】
本実施形態では、発話区間検出部２６は、現フレームの直前のフレームが発話区間でない場合、相対的に高い、発話区間開始検出用の第１の閾値とピッチゲインとを比較する。なお、直前のフレームが発話区間に含まれるか否かは、例えば、記憶部１４に保存されている、発話区間か否かを表すフラグを参照することで判定される。そして発話区間検出部２６は、ピッチゲインが第１の閾値以上である場合、現フレームから発話区間が開始されたと判定する。そして発話区間検出部２６は、発話区間か否かを表すフラグを、発話区間であることを表す値（例えば、'1'）に更新定する。
【００４０】
一方、現フレームの直前のフレームが発話区間に含まれている場合、相対的に低い、発話区間終了検出用の第２の閾値とピッチゲインとを比較する。そして発話区間検出部２６は、ピッチゲインが第２の閾値未満である場合、直前のフレームまでで発話区間は終了したと判定する。そして発話区間検出部２６は、発話区間か否かを表すフラグを、非発話区間であることを表す値（例えば、'0'）に更新する。
【００４１】
図３は、本実施形態による、発話区間検出処理の概要を説明する図である。図３の各グラフにおいて、横軸は時間を表す。１番上のグラフでは、縦軸はSN比を表す。上から２番目のグラフでは、縦軸は有音区間か無音区間かの判定結果を表す。また、上から３番目のグラフでは、縦軸はピッチゲインを表す。そして一番下のグラフでは、縦軸は発話区間か否かの判定結果を表す。
【００４２】
一番上のグラフにおいて、折れ線３０１は、SN比の時間変化を表す。上から２番目のグラフにおいて、折れ線３０２は、時刻ごとの有音区間か無音区間かの判定結果を表す。折れ線３０１に示されるように、時刻t1にてSN比が有音判定閾値Thsnr以上となり、その後、時刻t4まで継続してSN比は有音判定閾値Thsnr以上となる。時刻t4以降、SN比は、有音判定閾値Thsnr未満となる。その結果、折れ線３０２に示されるように、時刻t1から時刻t4までの区間が有音区間と判定され、その前後は、無音区間と判定される。
【００４３】
上から３番目のグラフにおいて、折れ線３０３は、ピッチゲインの時間変化を表す。また一番下のグラフにおいて、折れ線３０４は、時刻ごとの発話区間か否かの判定結果を表す。折れ線３０３に示されるように、ピッチゲインは、時刻t1から上昇を開始し、時刻t2にて第１の閾値Th1以上となる。その後しばらくしてからピッチゲインはピークとなり、以降徐々に減衰する。そして時刻t3にて、ピッチゲインは第１の閾値Th1よりも低い、第２の閾値Th2未満となる。その結果、折れ線３０４に示されるように、時刻t2から時刻t3までの区間が発話区間と判定される。なお、仮に、発話区間の終了の判定にも閾値Th1が用いられると、時刻t2'にてピッチゲインは閾値Th1未満となるので、本来の発話区間よりも短い区間しか発話区間として検出されないことになる。しかし上記のように、発話区間の終了の判定に利用される閾値Th2を、発話区間の開始の判定に利用される閾値Th1よりも小さくすることで、発話区間検出部２６は、発話区間を適切に検出できる。
【００４４】

発話区間検出部２６は、発話区間が開始されたタイミングと発話区間が終了したタイミングとを処理部１３に出力する。

【００４５】

処理部１３は、発話区間が検出されると、例えば、発話区間中に話者が発話した内容を認識するために、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。そして処理部１３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り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。そして処理部１３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。さらに処理部１３は、その発話内容と、処理部１３にて実行されるアプリケーションとに応じた処理を実行してもよい。例えば、処理部１３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してもよい。そして処理部１３は、他言語に翻訳された発話内容に応じた文字列を表示装置（図示せず）に表示してもよい。あるいは、処理部１３は、その翻訳された文字列に音声合成処理を適用して、その文字列を表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。あるいは、処理部１３は、発話内容に応じた単語の組み合わせをクエリとして、発話区間検出装置１と接続されたネットワーク上で探索処理を実行してもよい。あるいはまた、処理部１３は、発話内容を表す文字列と、発話区間検出装置１が実装された装置の操作コマンドとを比較し、発話内容を表す文字列が何れかの操作コマンドと一致する場合に、その操作コマンドに応じた処理を実行してもよい。

【００４６】

図４は、本実施形態による、発話区間検出処理の動作フローチャートである。処理部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発話区間検出処理を実行する。

【００４７】

パワー算出部２１は、音声信号の現フレームのパワーを算出する（ステップＳ１０１）。雑音推定部２２は、現フレームのパワーと、直前のフレームにおける推定雑音成分に基づいて、現フレームの推定雑音成分を算出する（ステップＳ１０２）。そして信号対雑音比算出部２３は、現フレームのパワーと推定雑音成分に基づいて、現フレームのSN比SNR(k)を算出する（ステップＳ１０３）。

【００４８】

有音判定部２４は、現フレームのSN比SNR(k)が有音判定閾値Thsnr以上か否か判定する（ステップＳ１０４）。現フレームのSN比SNR(k)が有音判定閾値Thsnr未満であれば（ステップＳ１０４－Ｎｏ）、有音判定部２４は、現フレームは有音区間には含まれないと判定する。そして処理部１３は、発話区間検出処理を終了する。

【００４９】

一方、現フレームのSN比が有音判定閾値Thsnr以上であれば（ステップＳ１０４－Ｙｅｓ）、有音判定部２４は、現フレームは有音区間に含まれると判定する。そしてピッチゲイン算出部２５は、現フレームのピッチゲインg_{pitch}を算出する（ステップＳ１０５）。

【００５０】

発話区間検出部２６は、直前のフレームが発話区間に含まれるか否か判定する（ステップＳ１０６）。直前のフレームが発話区間に含まれない場合（ステップＳ１０６－Ｎｏ）、発話区間検出部２６は、現フレームのピッチゲインg_{pitch}が相対的に高い第１の閾値Th1以上か否か判定する（ステップＳ１０７）。現フレームのピッチゲインg_{pitch}が第１の閾値Th1以上であれば（ステップＳ１０７－Ｙｅｓ）、発話区間検出部２６は、現フレームから発話区間が開始したと判定し、発話区間が開始したことを表す情報を出力する（ステップＳ１０８）。また、発話区間検出部２６は、発話区間か否かを表すフラグを、発話区間であることを表す値に更新する。

【００５１】

一方、現フレームのピッチゲインg_{pitch}が第１の閾値Th1未満であれば（ステップＳ１０７－Ｎｏ）、発話区間検出部２６は、現フレームは発話区間に含まれないと判定する。そして処理部１３は、発話区間検出処理を終了する。

【００５２】

また、ステップＳ１０６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Ｓ１０６－Ｙｅｓ）、発話区間検出部２６は、現フレームのピッチゲインg_{pitch}が相対的に低い第２の閾値Th2未満か否か判定する（ステップＳ１０９）。現フレームのピッチゲインg_{pitch}が第２の閾値Th2未満であれば（ステップＳ１０９－Ｙｅｓ）、発話区間検出部２６は、直前のフレームまでで発話区間が終了したと判定し、発話区間が終了したことを表す情報を出力する（ステップＳ１１０）。また、発話区間検出部２６は、発話区間か否かを表すフラグを、非発話区間であることを表す値に更新する。

【００５３】

一方、現フレームのピッチゲインg_{pitch}が第２の閾値Th2以上であれば（ステップＳ１０９－Ｎｏ）、発話区間検出部２６は、現フレームにおいても発話区間は継続していると判定する。そして処理部１３は、発話区間検出処理を終了する。

【００５４】

以上に説明してきたように、この発話区間検出装置は、発話区間の開始を検出する際のピッチゲインに対する閾値よりも、発話区間の終了を検出する際のピッチゲインに対する閾値を低く設定する。そのため、この発話区間検出装置は、発話の継続に応じてピッチゲインが小さくなっても、発話区間を適切に検出できる。

【００５５】

なお、変形例によれば、発話区間検出部２６は、発話区間開始時におけるピッチゲインに基づいて、第２の閾値Th2を調整してもよい。例えば、発話区間検出部２６は、次式に示されるように、第１の閾値Th1に対する発話区間開始時におけるピッチゲイン$g_{pitch}(t_{start})$の比を第２の閾値Th2に乗じて得られる値を、調整後の第２の閾値Th2'としてもよい。すなわち、発話区間開始時におけるピッチゲインが大きいほど、調整後の第２の閾値Th2'も大きくなる。

【数６】

$$Th2' = \frac{g_{pitch}(t_{start})}{Th1} \cdot Th2 \qquad (6)$$

【００５６】

この場合、発話区間検出部２６は、ピッチゲインが調整後の第２の閾値Th2'未満となったときに発話区間が終了したと判定すればよい。

【００５７】

図５は、この変形例による、発話区間と閾値の関係を示す図である。図５において、横軸は時間を表し、縦軸はピッチゲインを表す。折れ線５０１は、ピッチゲインの時間変化を表す。また折れ線５０２は、発話区間の検出に利用される閾値の時間変化を表す。この例では、時刻t1において最初にピッチゲイン$g_{pitch}(t1)$が第１の閾値Th1以上となるので、時刻t1にて発話区間が開始したと判定される。そして、比$(g_{pitch}(t1)/Th1)$に基づいて調整された第２の閾値Th2'が算出される。その後、時刻t2において、ピッチゲイン$g_{pitch}(t2)$が調整された第２の閾値Th2'未満となるので、時刻t2において発話区間が終了したと判定される。

【００５８】

この変形例によれば、発話区間開始時のピッチゲインに基づいて第２の閾値が調整されるので、発話区間検出部２６は、話者の声の特徴に応じて適切に第２の閾値を調整できる。その結果として、発話区間検出部２６は、発話区間をより適切に検出できる。

【００５９】

また他の変形例によれば、発話区間検出部２６は、音声信号のSN比に基づいて、第１の閾値Th1及び第２の閾値Th2を調整してもよい。

【００６０】

一般に、SN比が低いほど、音声信号に含まれる雑音成分の比率が高いので、音声信号の周期性も低下する。そこで、この変形例によれば、発話区間検出部２６は、現フレームのSN比が低いほど、第１の閾値Th1及び第２の閾値Th2を低く設定する。

【００６１】

図６は、SN比と第１の閾値の関係の一例を示す図である。図６において、横軸はSN比を表し、縦軸は、第１の閾値を表す。そして折れ線６００は、SN比と第１の閾値の関係を表す。折れ線６００に示されるように、SN比がSNRlow以下のときは、第１の閾値は、Thlowに設定される。そしてSN比がSNRlowより大きく、かつ、SNRhigh未満のときは、SN比が大きくなるにつれて第１の閾値も線形に増加する。そしてSN比がSNRhigh以上となると、第１の閾値はThhighに設定される。なお、SNRlow及びSNRhighは、例えば、18dB及び30dBに設定される。また、Thlow及びThhighは、例えば、0.5及び0.7に設定される。第２の閾値Th2についても同様に、SN比がSNRlowより大きく、かつ、SNRhigh未満のときにSN比が大きくなるにつれて線形に増加するように設定されればよい。また、SN比がSNRlow以下のときの第２の閾値は、例えば、0.4に設定され、SN比がSNRhigh以上の時の第２の閾値は、例えば、0.6に設定される。なお、図６に示されるように、SN比と第１及び第２の閾値との関係を表す参照テーブルが予め記憶部１４に保存され、発話区間検出部２６は、その参照テーブルを参照して、SN比に対応する第１及び第２の閾値の値を設定すればよい。

【００６２】

この変形例によれば、発話区間検出部２６は、音声信号のSN比に応じて、発話区間の検出に利用されるピッチゲインに対する第１及び第２の閾値を適切に決定できる。なお、ピッチゲインに対する第１及び第２の閾値がフレームごとに急激に変動することを抑制するために、発話区間検出部２６は、図６に示される関係に従って、有音区間開始時のフレームのSN比に応じて第１及び第２の閾値を決定してもよい。

【００６３】

また、SN比による第１及び第２の閾値の調整と、ピッチゲインによる第２の閾値の調整は組み合わせられてもよい。この場合には、発話区間検出部２６は、例えば、SN比に基づいて決定された第１及び第２の閾値を（６）式における閾値Th1及び閾値Th2とすることで、調整後の閾値Th2'を算出すればよい。

【００６４】

また、雑音が大きい環境では、雑音の影響により、ピッチゲインの算出値に含まれる誤差が相対的に大きくなることがある。そのため、発話区間が終了していなくても、瞬間的にピッチゲインが第２の閾値未満となることがある。

【００６５】

そこでさらに他の変形例によれば、発話区間検出部２６は、発話区間の開始後において、ピッチゲインが第２の閾値未満となることが一定の監視区間にわたって継続した場合に、発話区間が終了したと判定してもよい。なお、発明者による実験によれば、雑音が比較的小さい環境（例えば、SN比が30dB）では、ピッチゲインの値は、発話区間中、継続して0.6以上となった。一方、発話区間以外では、ピッチゲインが1秒以上継続することはなかった。このことから、上記の監視区間は、例えば、1秒間に設定される。

【００６６】

図７は、この変形例による、発話区間の判定に関する状態遷移図である。状態遷移図７００において、状態１～状態３は、それぞれ、互いに異なる発話区間の検出状態を表す。具体的に、状態１は、直前のフレームが発話区間及び監視区間中でないこと、すなわち、非発話区間中であることを表す。また状態２は、直前のフレームが発話区間中であることを表す。そして状態３は、直前のフレームが監視区間中であることを表す。

【００６７】

状態１において、現フレームのピッチゲインが第１の閾値Th1未満であれば、発話区間

の検出状態は変化しない。すなわち、現フレームは、非発話区間内に含まれる。一方、状態１において、現フレームのピッチゲインが第１の閾値Th1以上であれば、発話区間の検出状態は状態１から状態２に遷移する。すなわち、現フレームから発話区間が開始となる。

【００６８】

状態２において、現フレームのピッチゲインが第１の閾値Th1よりも低い第２の閾値Th2以上であれば、発話区間の検出状態は変化しない。すなわち、現フレームは、発話区間内に含まれる。一方、状態２において、現フレームのピッチゲインが第２の閾値Th2未満であれば、発話区間の検出状態は状態２から状態３に遷移する。すなわち、現フレームから監視区間が開始となる。

【００６９】

状態３において、現フレームのピッチゲインが第２の閾値Th2以上となれば、発話区間の検出状態は状態３から状態２に遷移する。すなわち、現フレームまで発話区間は継続していると判定され、監視区間は一旦終了する。一方、現フレームのピッチゲインが第２の閾値Th2未満であり、かつ、監視区間開始からの継続時間（その継続時間に相当するフレーム数をNframeと表記する）が一定期間（閾値ThN）に達していなければ、発話区間の検出状態は変化しない。すなわち、現フレームは、監視区間内に含まれる。そして、現フレームのピッチゲインが第２の閾値Th2未満であり、かつ、監視区間開始からの継続時間が一定期間に達していれば、発話区間の検出状態は状態３から状態１に遷移する。すなわち、現フレームにて、発話区間が終了したと判定される。

【００７０】

図８は、この変形例による、発話区間検出処理の動作フローチャートである。なお、図４に示される、上記の実施形態による発話区間検出処理と比較して、ステップＳ１０５までは同じであるため、図８では、ステップＳ１０５以降の処理について説明する。

【００７１】

発話区間検出部２６は、直前のフレームが非発話区間に含まれるか否か判定する（ステップＳ２０１）。すなわち、発話区間検出部２６は、直前のフレームにおける発話区間の検出状態が状態１か否か判定する。直前のフレームが非発話区間に含まれる場合（ステップＳ２０１－Ｙｅｓ）、発話区間検出部２６は、現フレームのピッチゲインg_{pitch}が相対的に高い第１の閾値Th1以上か否か判定する（ステップＳ２０２）。現フレームのピッチゲインg_{pitch}が第１の閾値Th1以上であれば（ステップＳ２０２－Ｙｅｓ）、発話区間検出部２６は、現フレームから発話区間が開始したと判定し、発話区間が開始したことを表す情報を出力する（ステップＳ２０３）。すなわち、発話区間の検出状態が状態１から状態２へ遷移する。

【００７２】

一方、現フレームのピッチゲインg_{pitch}が第１の閾値Th1未満であれば（ステップＳ２０２－Ｎｏ）、発話区間検出部２６は、現フレームは発話区間に含まれないと判定する。すなわち、発話区間の検出状態は状態１のまま維持される。そして処理部１３は、発話区間検出処理を終了する。

【００７３】

また、ステップＳ２０１において、直前のフレームが非発話区間に含まれない場合（ステップＳ２０１－Ｎｏ）、発話区間検出部２６は、直前のフレームが発話区間に含まれるか否か判定する（ステップＳ２０４）。すなわち、発話区間検出部２６は、直前のフレームにおける発話区間の検出状態が状態２か否か判定する。直前のフレームが発話区間に含まれる場合（ステップＳ２０４－Ｙｅｓ）、現フレームのピッチゲインg_{pitch}が相対的に低い第２の閾値Th2未満か否か判定する（ステップＳ２０５）。現フレームのピッチゲインg_{pitch}が第２の閾値Th2未満であれば（ステップＳ２０５－Ｙｅｓ）、発話区間検出部２６は、監視区間を開始する（ステップＳ２０６）。すなわち、発話区間の検出状態が状態２から状態３へ遷移する。そして発話区間検出部２６は、監視区間が継続する時間を表す、監視区間開始からのフレーム数Nframeを1に設定する。一方、現フレームのピッチゲ

インg$_{pitch}$が第２の閾値Th2以上であれば（ステップＳ２０５－Ｎｏ）、発話区間検出部２６は、現フレームにおいても発話区間は継続していると判定する。すなわち、発話区間の検出状態は状態２のまま維持される。そして処理部１３は、発話区間検出処理を終了する。

【００７４】

また、ステップＳ２０４において、直前のフレームが発話区間に含まれない場合（ステップＳ２０４－Ｎｏ）、監視区間が継続中（状態３）である。この場合、発話区間検出部２６は、現フレームのピッチゲインg$_{pitch}$が第２の閾値Th2以上か否か判定する（ステップＳ２０７）。現フレームのピッチゲインg$_{pitch}$が第２の閾値Th2以上であれば（ステップＳ２０７－Ｙｅｓ）、発話区間検出部２６は、監視区間を終了する（ステップＳ２０８）。すなわち、発話区間の検出状態が状態３から状態２へ遷移する。そして発話区間検出部２６は、Nframeを0にリセットする。

【００７５】

一方、現フレームのピッチゲインg$_{pitch}$が第２の閾値Th2未満であれば（ステップＳ２０７－Ｎｏ）、発話区間検出部２６は、Nframeを１インクリメントする（ステップＳ２０９）。そして発話区間検出部２６は、Nframeが監視区間の長さの閾値を表すフレーム数ThN以上となったか否か判定する（ステップＳ２０１）。なお、ThNは、例えば、1秒間に相当するフレーム数に設定される。NframeがThN以上であれば（ステップＳ２１０－Ｙｅｓ）、発話区間検出部２６は、現フレームにおいて発話区間が終了したと判定し、発話区間が終了したことを表す情報を出力する（ステップＳ２１１）。すなわち、発話区間の検出状態が状態３から状態１へ遷移する。なお、この場合において、発話区間検出部２６は、監視区間が開始した時点で発話区間が終了したと遡って判定してもよい。

【００７６】

一方、NframeがThN未満であれば（ステップＳ２１０－Ｎｏ）、発話区間検出部２６は、現フレームにおいても監視区間は継続していると判定する。すなわち、発話区間の検出状態は状態３のまま維持される。そして処理部１３は、発話区間検出処理を終了する。

【００７７】

この変形例によれば、発話区間検出部２６は、音声信号中の雑音成分により、ピッチゲインの誤差が大きくなる場合でも、発話区間が終了するタイミングを適切に検出できる。

【００７８】

さらに他の変形例によれば、発話区間検出部２６は、発話区間が開始してからの経過時間に応じて第２の閾値を調整してもよい。同様に、発話区間検出部２６は、発話区間が終了してからの経過時間に応じて第１の閾値を調整してもよい。

【００７９】

図９（ａ）は、この変形例による、発話区間開始からの経過時間と第２の閾値Th2との関係の一例を表す。また図９（ｂ）は、この変形例による、発話区間終了からの経過時間と第１の閾値Th1との関係の一例を表す。図９（ａ）及び図９（ｂ）において、横軸は時間を表し、縦軸は閾値を表す。そして図９（ａ）に示される折れ線９０１は、発話区間開始からの経過時間と第２の閾値Th2との関係を表す。また図９（ｂ）に示される折れ線９０２は、発話区間終了からの経過時間と第１の閾値Th1との関係を表す。

【００８０】

図９（ａ）に示される例では、時刻t1にて発話区間が開始したとする。折れ線９０１に示されるように、時刻t1から時刻t2にかけて、経過時間に応じて第２の閾値Th2は、第１の閾値Th1と同じ値であるTh2highから線形に減少する。そして時刻t2以降、一定値Th2lowとなる。同様に、図９（ｂ）に示される例では、時刻t1にて発話区間が終了したとする。折れ線９０２に示されるように、時刻t1から時刻t2にかけて、経過時間に応じて第１の閾値Th1は、第２の閾値Th2と同じ値Th1lowから線形に増加する。そして時刻t2以降、一定値Th1highとなる。なお、時刻t1から時刻t2までの間隔は、例えば、1秒未満、より具体的には、0.2秒～0.4秒に設定されることが好ましい。

【００８１】

このように、第１の閾値及び第２の閾値を時間経過に応じて滑らかに変化させることで、発話区間検出部２６は、ピッチゲインの時間変動が大きい場合でも、発話区間をより適切に検出することができる。

【００８２】

さらに他の変形例によれば、ピッチゲイン算出部２５は、音声信号の線形予測成分に対する残差信号の長期自己相関に基づいてピッチゲインを算出してもよい。なお、残差信号の長期自己相関は、音声信号から短期相関成分を取り除いた残りの自己相関を表す。この場合、ピッチゲイン算出部２５は、音声信号の線形予測係数を算出する。その際、ピッチゲイン算出部２５は、例えば、TTC標準JT-G722.2規格の5.2.2章で規定されている方法に従って線形予測係数を算出すればよい。そしてピッチゲイン算出部２５は、次式に従って残差信号res(n)を算出する。

【数７】

$$res(n)=s_k(n)+\sum_{i=1}^{p}a(i)\cdot s_k(n-i)\quad (n=0,1,...,N-1) \tag{7}$$

ここでa(i)は、線形予測係数であり、pは、線形予測係数の次数（例えば、16）である。

【００８３】

ピッチゲイン算出部２５は、残差信号の長期自己相関$C_{res}(d)$を次式に従って算出する。

【数８】

$$C_{res}(d)=\sum_{n=0}^{N-1}res(n)\cdot res(n-d)\quad (d=d_{low},...,d_{high}) \tag{8}$$

なお、遅延量dの最小値d_{low}及び最大値d_{high}は、上記の実施形態における（４）式と同様に、人の声の基本周波数に相当する遅延量が含まれるように設定される。

【００８４】

ピッチゲイン算出部２５は、遅延量の範囲に含まれる遅延量dごとに残差信号の長期自己相関$C_{res}(d)$を算出すると、その長期自己相関$C_{res}(d)$のうちの最大値$C_{res}(d_{max})$を求める。なお、d_{max}は、長期自己相関$C_{res}(d)$の最大値$C_{res}(d_{max})$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。そしてピッチゲイン算出部２５は、次式に従ってピッチゲインg_{pitch}を算出すればよい。

【数９】

$$g_{pitch}=\frac{C_{res}(d_{max})}{\sum_{n=0}^{N-1}res(n)\cdot res(n)} \tag{9}$$

【００８５】

また、上記の実施形態または変形例において、発話区間検出装置１は、有音区間を検出せずに、音声信号から発話区間を直接検出してもよい。すなわち、ピッチゲイン算出部２５は、全てのフレームについてピッチゲインを算出し、発話区間検出部２６は、有音区間か否かにかかわらず、ピッチゲインと第１の閾値Th1または第２の閾値Th2との比較結果により、発話区間を検出すればよい。

【００８６】

これにより、発話区間の検出精度が若干低下する可能性があるものの、発話区間の検出に要する演算量が削減される。この場合、処理部１３が有する各部のうち、有音判定部２

４が省略されてもよい。また、第１の閾値Th1及び第２の閾値Th2の調整にSN比が利用されない場合には、パワー算出部２１、雑音推定部２２及び信号対雑音比算出部２３も省略されてもよい。

【００８７】

また上記の実施形態または変形例による発話区間検出装置は、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１０は、上記の何れかの実施形態またはその変形例による発話区間検出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。

【００８８】

端末１１０は、音声入力部１１１と、記憶部１１２と、通信部１１３と、制御部１１４とを有する。音声入力部１１１、記憶部１１２及び通信部１１３は、例えば、制御部１１４とバスを介して接続されている。

【００８９】

音声入力部１１１は、例えば、オーディオインターフェースとＡ／Ｄコンバータを有する。そして音声入力部１１１は、例えば、マイクロホンからアナログ信号である音声信号を取得し、その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。そして音声入力部１１１は、デジタル化された音声信号を制御部１１４へ出力する。

【００９０】

記憶部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。そして記憶部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。

【００９１】

通信部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。そして通信部１１３は、制御部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。

【００９２】

制御部１１４は、一つまたは複数のプロセッサとその周辺回路を有する。そして制御部１１４は、音声信号を、端末１１０の識別情報とともに、通信部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。また制御部１１４は、サーバ１２０から受け取った、音声信号に対する処理結果をディスプレイ（図示せず）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。

【００９３】

サーバ１２０は、通信部１２１と、記憶部１２２と、処理部１２３とを有する。通信部１２１及び記憶部１２２は、処理部１２３とバスを介して接続されている。

【００９４】

通信部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。そして通信部１２１は、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を介して受信して処理部１２３に渡す。

【００９５】

記憶部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。そして記憶部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。また記憶部１２２は、発話区間検出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音声信号を記憶していてもよい。

【００９６】

処理部１２３は、一つまたは複数のプロセッサとその周辺回路を有する。そして処理部１２３は、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処理部の各機能を実現する。さらに処理部１２３は、検出された発話区間に対して音声認識などの所定の処理を実行してその処理結果を求める。そして処理部１２３は、その処理結果を通信部１２１及び通信ネットワーク１３０を介して端末１１０へ送信する。

【００９７】

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処理部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。

【００９８】

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。

【００９９】

以上説明した実施形態及びその変形例に関し、更に以下の付記を開示する。

（付記１）

話者の声が表された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するピッチゲイン算出部と、

前記話者が発話していない非発話区間が継続している場合において前記ピッチゲインが第１の閾値以上となると前記話者が発話している発話区間が開始されたと判定し、かつ、前記発話区間が継続している場合において前記ピッチゲインが前記第１の閾値よりも小さい第２の閾値未満となると前記発話区間が終了すると判定する発話区間検出部と、

を有する発話区間検出装置。

（付記２）

前記フレームごとに、前記音声信号の信号対雑音成分比を算出する信号対雑音成分比算出部と、

前記フレームごとの前記信号対雑音成分比に基づいて前記音声信号に信号成分が含まれる有音区間を検出する有音区間検出部とをさらに有し、

前記発話区間検出部は、前記有音区間内において直前のフレームが前記発話区間でなく、かつ、現フレームの前記ピッチゲインが前記第１の閾値以上となる場合に前記現フレームから前記発話区間が開始されたと判定する、付記１に記載の発話区間検出装置。

（付記３）

前記発話区間検出部は、前記発話区間が開始されたと判定されたフレームにおける前記ピッチゲインが大きいほど、前記第２の閾値を高くする、付記１または２に記載の発話区間検出装置。

（付記４）

前記フレームごとに、前記音声信号の信号対雑音成分比を算出する信号対雑音成分比算出部をさらに有し、

前記発話区間検出部は、前記信号対雑音成分比が大きいフレームほど、当該フレームにおける前記第１の閾値及び前記第２の閾値を高くする、付記１に記載の発話区間検出装置。

（付記５）

前記フレームごとに、前記音声信号の信号対雑音成分比を算出する信号対雑音成分比算出部と、

前記フレームごとの前記信号対雑音成分比に基づいて前記音声信号に信号成分が含まれる有音区間を検出する有音区間検出部とをさらに有し、

前記発話区間検出部は、前記有音区間が開始されたと判定したフレームにおける前記信号対雑音成分比が大きいほど、前記第１の閾値及び前記第２の閾値を高くする、付記１に記載の発話区間検出装置。

（付記６）

前記発話区間検出部は、前記発話区間が継続している場合において前記ピッチゲインが前記第２の閾値未満となる期間が一定期間継続すると前記発話区間が終了したと判定する、付記１～５の何れかに記載の発話区間検出装置。

（付記７）

話者の声が表された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、

前記話者が発話していない非発話区間が継続している場合において前記ピッチゲインが第１の閾値以上となると前記話者が発話している発話区間が開始されたと判定し、かつ、前記発話区間が継続している場合において前記ピッチゲインが前記第１の閾値よりも小さい第２の閾値未満となると前記発話区間が終了すると判定する、

ことを含む発話区間検出方法。

（付記８）

話者の声が表された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、

前記話者が発話していない非発話区間が継続している場合において前記ピッチゲインが第１の閾値以上となると前記話者が発話している発話区間が開始されたと判定し、かつ、前記発話区間が継続している場合において前記ピッチゲインが前記第１の閾値よりも小さい第２の閾値未満となると前記発話区間が終了すると判定する、

ことをコンピュータに実行させるための発話区間検出用コンピュータプログラム。

（付記９）

話者の声が表された音声信号を取得するマイクロホンと、

前記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、

前記話者が発話していない非発話区間が継続している場合において前記ピッチゲインが第１の閾値以上となると前記話者が発話している発話区間が開始されたと判定し、かつ、前記発話区間が継続している場合において前記ピッチゲインが前記第１の閾値よりも小さい第２の閾値未満となると前記発話区間が終了すると判定するように構成されたプロセッサと、

を有する発話区間検出装置。

【符号の説明】

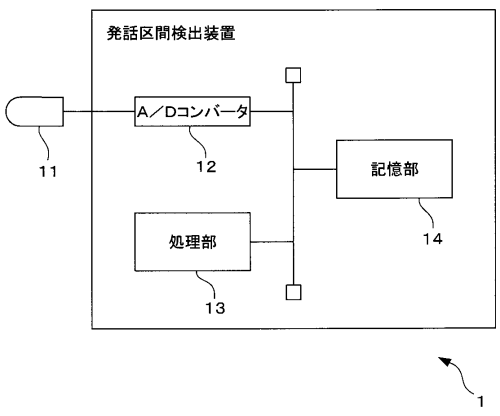
【０１００】

１　　発話区間検出装置
１１　　マイクロホン
１２　　アナログ／デジタルコンバータ
１３　　処理部
１４　　記憶部
２１　　パワー算出部
２２　　雑音推定部
２３　　信号対雑音比算出部
２４　　有音判定部
２５　　ピッチゲイン算出部
２６　　発話区間検出部
１００　　サーバクライアントシステム
１１０　　端末
１１１　　音声入力部
１１２　　記憶部

１１３　　通信部
１１４　　制御部
１２０　　サーバ
１２１　　通信部
１２２　　記憶部
１２３　　処理部
１３０　　通信ネットワーク

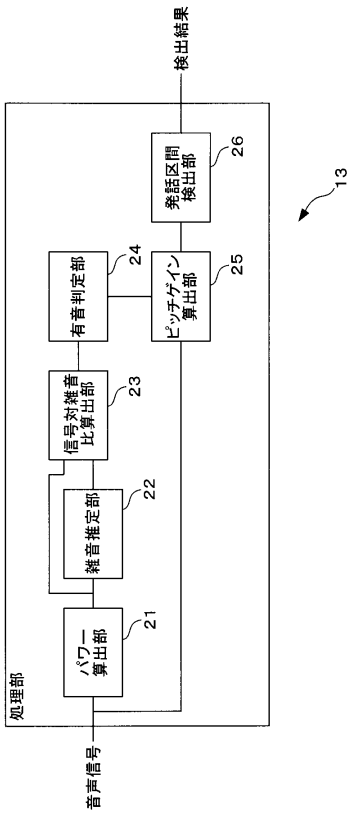
【図１】

図1



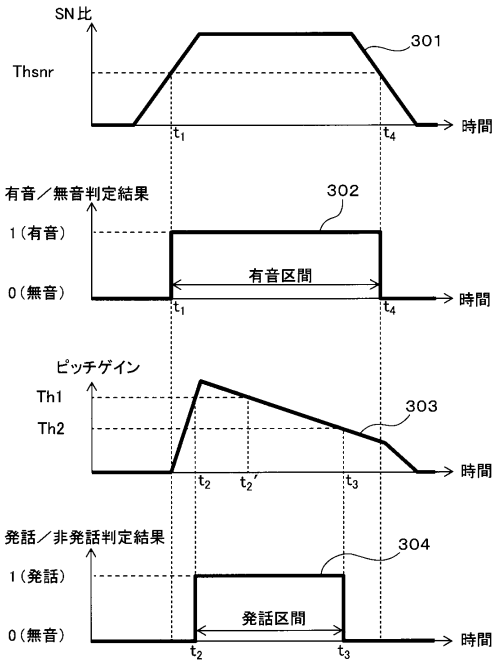
【図２】

図2



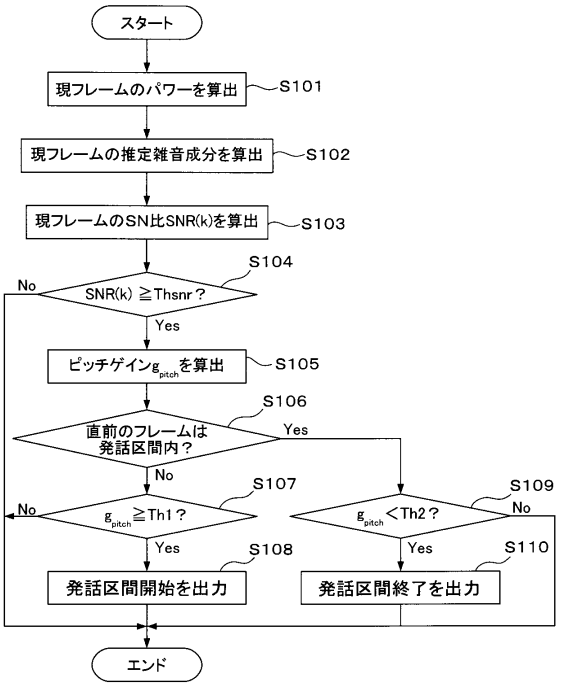
【図３】

図3



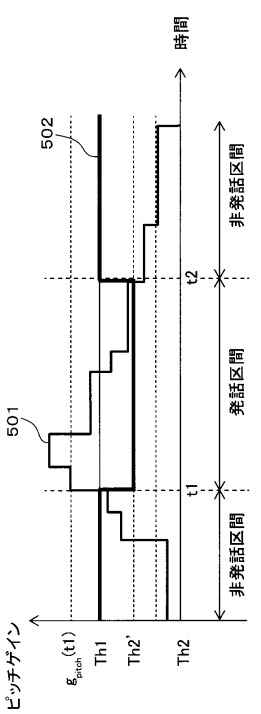
【図４】

図4



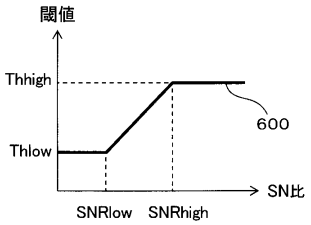
【図５】

図5



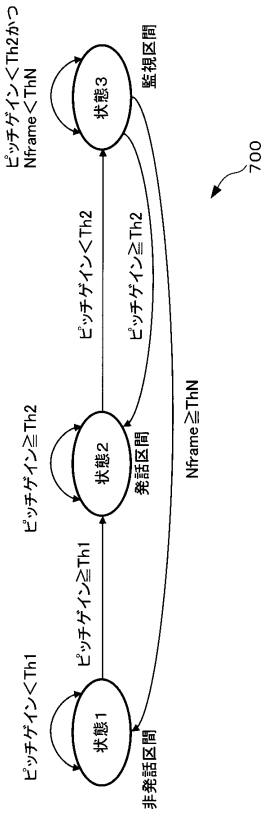
【図６】

図6



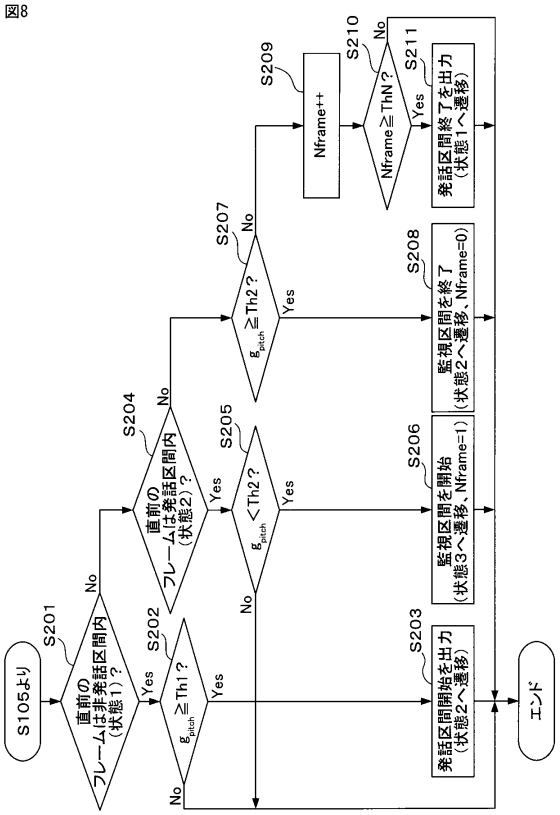
【図７】

図7



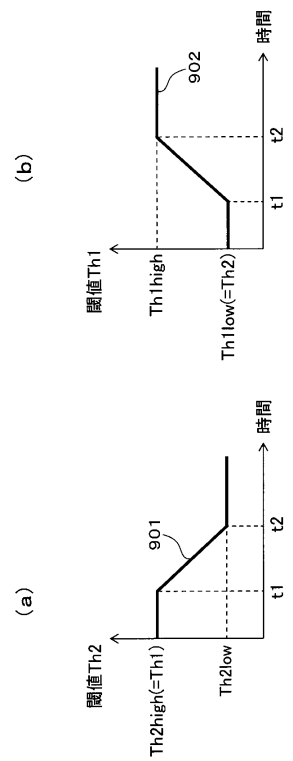
【図８】

図8



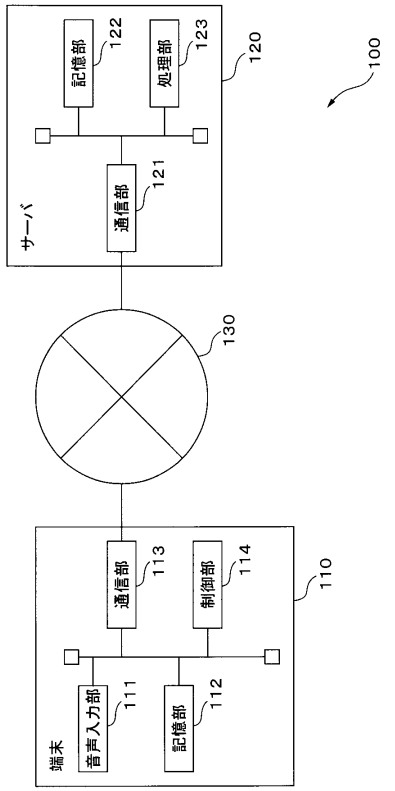
【図９】

図9



【図１０】

図10



フロントページの続き

(72)発明者　塩田　千里
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
(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

審査官　渡部　幸和

(56)参考文献　特開平０７－１５２３９５（ＪＰ，Ａ）
特開平０６－０８３３９１（ＪＰ，Ａ）
特開２０１５－００４７０３（ＪＰ，Ａ）

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