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

特許第4521673号
(P4521673)

(45)発行日　平成22年8月11日(2010.8.11)　(24)登録日　平成22年6月4日(2010.6.4)

(51)Int.Cl.　F I
G10L 15/04 (2006.01)　G10L 15/04　300A
G10L 11/02 (2006.01)　G10L 11/02

請求項の数 13　(全 23 頁)

(21)出願番号　特願2004-101094(P2004-101094)
(22)出願日　平成16年3月30日(2004.3.30)
(65)公開番号　特開2005-31632(P2005-31632A)
(43)公開日　平成17年2月3日(2005.2.3)
審査請求日　平成19年3月28日(2007.3.28)
(31)優先権主張番号　特願2003-174416(P2003-174416)
(32)優先日　平成15年6月19日(2003.6.19)
(33)優先権主張国　日本国(JP)

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　スーン　フランク　ガーピン
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　中村　哲
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　葦苅　豊
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

最終頁に続く

(54)【発明の名称】発話区間検出装置、コンピュータプログラム及びコンピュータ

(57)【特許請求の範囲】
【請求項１】
音声データを逐次フレーム化するためのフレーム化手段と、
前記フレーム化手段によりフレーム化された音声のエネルギ値をフレームごとに算出し、ＦＩＦＯ（Ｆｉｒｓｔ－Ｉｎ　Ｆｉｒｓｔ－Ｏｕｔ）形式で第１の個数のフレームのエネルギ値を記憶するためのフレームエネルギ算出及び記憶手段と、
前記フレームエネルギ算出及び記憶手段に、第２の個数のフレームのエネルギ値が格納されたことに応答して、前記第２の個数のフレームのエネルギ値を所定の統計的手法に従って処理することにより、前記音声データに含まれる環境雑音の推定値の初期値を算出するための初期値算出手段と、
前記推定値の初期値と、フレームエネルギ算出及び記憶手段に逐次記憶される音声のエネルギ値とに基づいて、前記音声データに含まれる環境雑音の変化に追従して変化する様に、前記発話区間を検出するためのエネルギ値のしきい値をフレームごとに逐次算出するための手段と、
前記しきい値に基づいて、前記第２の個数のフレーム以降のフレームの中で、前記音声データの発話区間の開始位置又は終了位置に対応するフレームを推定するための発話区間推定手段とを含む、発話区間検出装置であって、
前記初期値算出手段は、
前記第２の個数のフレームを、各フレームのエネルギ値の大きさによって、第１のエネルギ値を中心とする第１のクラスタと、前記第１のエネルギよりも大きな第２のエネルギ

値を中心とする第２のクラスタとにクラスタ化するための手段と、
前記第１のエネルギ値を前記環境雑音の推定値の初期値として出力するための手段とを含む、発話区間検出装置。
【請求項２】
前記クラスタ化するための手段は、
前記第２の個数のフレームを前記第１及び第２のクラスタにクラスタ化するための境界値を決定するための手段と、
前記境界値よりも小さなエネルギ値を持つフレームを前記第１のクラスタに、それ以外のフレームを前記第２のクラスタに、それぞれ分類するための手段とを含む、請求項１に記載の発話区間検出装置。
【請求項３】
前記境界値を決定するための手段は、
前記第２の個数のフレームのうち、エネルギ値をキーとしてソートしたときに予め定める第１のソート順位及び第２のソート順位となる二つのフレームを選択するための手段と、
前記選択された二つのフレームのエネルギ値の平均値を算出するための第１の平均値算出手段と、
前記第１の平均値算出手段により算出された平均値より小さいエネルギ値を持つか否かを基準として、前記第２の個数のフレームを第１及び第２のグループに分類するための手段と、
前記第１及び第２のグループに属するフレームのエネルギ値の平均値をそれぞれ算出するための第２の平均値算出手段と、
前記第２の平均値算出手段により算出された二つの平均値の平均値をさらに算出し、前記境界値として出力するための第３の平均値算出手段とを含む、請求項２に記載の発話区間検出装置。
【請求項４】
前記しきい値をフレームごとに逐次算出するための手段は、
前記フレームエネルギ算出及び記憶手段に格納されているフレームのエネルギ値と、前記環境雑音の推定値の初期値とに基づいて、前記フレームエネルギ算出及び記憶手段に格納されているフレームの環境雑音のエネルギ値をフレームごとに推定するための手段と、
前記フレームエネルギ算出及び記憶手段に格納されているフレームのエネルギ値のうち、定常的な背景雑音及び発話音声の合計のエネルギ値の最大値をフレームごとに逐次推定するための手段と、
前記推定された環境雑音のエネルギ値と、前記推定された背景雑音及び発話音声の合計のエネルギ値とに基づいて、前記発話区間を検出するためのエネルギのしきい値をフレームごとに算出するための手段とを含む、請求項１に記載の発話区間検出装置。
【請求項５】
前記発話区間推定手段は、前記しきい値に基づいて、前記第２の個数のフレーム以降のフレームの状態を判定するための手段を含み、
前記状態は、非発話状態を含み、
前記環境雑音のエネルギ値をフレームごとに逐次推定するための手段は、
１フレーム前の時点において推定された前記環境雑音のエネルギ値を記憶するための手段と、
前記環境雑音の推定値の初期値が算出された時点で前記記憶するための手段に前記環境雑音の推定値の初期値を記憶させるための手段と、
前記記憶するための手段に記憶された値、前記フレームエネルギ算出及び記憶手段に含まれるフレームのエネルギ値、及び前記フレームの状態を判定する手段による判定結果に基づいて、以下の式
$b(t)=b(t-1)\times\alpha+E(t)\times(1-\alpha)$　（状態が非発話状態の場合）
$b(t)=b(t-1)$　（状態が非発話状態以外の場合）

ただしαは所定の忘却係数、Ｅ（ｔ）は時刻ｔにおけるフレームのエネルギ値、に従って時刻ｔにおける背景雑音ｂ（ｔ）を算出するための手段とを含み、

前記記憶するための手段は、算出された前記背景雑音ｂ（ｔ）を記憶する、請求項４に記載の発話区間検出装置。

【請求項６】

前記合計のエネルギ値の最大値をフレームごとに推定するための手段は、

前記フレームエネルギ算出及び記憶手段に格納されているフレームを、エネルギ値をキーとしてソートするための手段と、

前記ソートするための手段によりソートされた結果所定の順位となるフレームのエネルギ値を前記合計のエネルギ値の最大値Ｅmax(t)として選択するための手段を含む、請求項５に記載の発話区間検出装置。

【請求項７】

前記しきい値をフレームごとに逐次算出するための手段は、

時刻ｔにおける発話開始位置検出のためのしきい値Ｅｔｈ1（ｔ）を、

Eth1(t)＝b(t)＋max(β，Emax(t)－b(t))×第１の定数、

ただしβは音声データ信号の最低ダイナミックレンジとして予め定められた定数、

に従って算出するための手段を含む、請求項６に記載の発話区間検出装置。

【請求項８】

前記しきい値をフレームごとに逐次算出するための手段は、さらに、

時刻ｔにおける発話終了位置検出のためのしきい値Ｅｔｈ2（ｔ）を、

Eth2(t)＝b(t)＋max(β，Emax(t)－b(t))×第２の定数、

ただし第２の定数＜第１の定数、

βは音声データ信号の最低ダイナミックレンジとして予め定められた前記定数、

に従って算出するための手段を含む、請求項７に記載の発話区間検出装置。

【請求項９】

さらに、発話の先頭からの各フレームの音声データの最大エネルギ値又は所定のデフォルト基準値のいずれか大きい方を用いて各フレームの音声データを正規化し、各フレームの音声特徴パラメータとして出力するための音声エネルギ正規化手段を含む、請求項１～請求項８のいずれかに記載の発話区間検出装置。

【請求項１０】

前記音声エネルギ正規化手段は、

正規化の基準値を記憶するための基準値記憶手段と、

前記フレームエネルギ算出及び記憶手段により算出された音声エネルギが、前記基準値記憶手段に記憶された基準値を超えていることを検出し、検出信号を出力するための検出手段と、

前記検出手段により出力される前記検出信号に応答して、前記基準値記憶手段に記憶された基準値を、前記フレームエネルギ算出及び記憶手段により算出された値で置換するための手段と、

前記フレームエネルギ算出及び記憶手段により算出された音声エネルギ値を、前記基準値記憶手段に記憶された基準値で除算することにより、当該フレームの音声エネルギを正規化するための除算手段とを含む、請求項９に記載の発話区間検出装置。

【請求項１１】

前記発話区間推定手段により、発話区間の終了位置に対応するフレームが推定されたことに応答して、前記基準値記憶手段の記憶内容を、所定のデフォルト値で置換するための手段をさらに含む、請求項１０に記載の発話区間検出装置。

【請求項１２】

前記所定のデフォルト値を、前記発話区間検出装置の起動時に与えられたオプション値に基づいて設定するための手段をさらに含む、請求項１０又は請求項１１に記載の発話区間検出装置。

【請求項１３】

コンピュータにより実行されると、当該コンピュータを請求項１から請求項１２のいずれかに記載の発話区間検出装置として動作させる、発話区間検出のためのコンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

この発明は音声認識処理などの前処理として発話区間を検出するための装置に関し、特に、実時間での音声認識処理において、環境雑音による発話区間の誤検出を避けることができる発話区間検出装置、ならびにフレームごとの特徴量として正規化した音声エネルギを算出するための音声エネルギ正規化装置に関する。

【背景技術】

【０００２】

音声認識などの処理においては、音声認識に先立って入力信号中の発話区間とそれ以外の区間（無音区間と呼ぶ。）との区別をすることが必要である。さもなければ、発話のない区間を音声認識することにより無意味な結果がもたらされるためである。

【０００３】

従来、このような発話区間（又は無音区間）の検出は、入力される音声信号のパワー（エネルギ）を算出し、その値が予め定められたしきい値以上になれば発話区間、しきい値未満であれば無音区間とする、という手法により行なわれている。このとき、そうした条件の成立が持続した時間をも考慮にいれて発話区間又は無音区間の検出がされるのが通常である。

【０００４】

そのような技術が特許文献１に開示されている。特許文献１は、音声付の映像情報から要約を自動的に作成するために、要約の対象となる個所を抽出するための技術を開示している。音声付の映像では、その内容（ジャンル）により、環境雑音の大きさが異なることが知られている。例えばニュース番組では環境雑音が小さく、スポーツ中継等の番組では環境雑音が大きい、などである。そのため、同じしきい値を用いて発話区間を検出しようとすると、映像情報のジャンルによりその結果が異なってしまうという問題がある。そのために特許文献１に開示の技術では、映像情報に、そのジャンルを示す付帯情報をもたせておき、付帯情報に従って各ジャンルに予め割当てられたしきい値を選択している。

【０００５】

【特許文献１】特開２００３－１０１９３９（段落２０９、２１０、図１及び図７）

【発明の開示】

【発明が解決しようとする課題】

【０００６】

しかし、上記した特許文献１に記載の技術では、一つの映像情報には一種類のしきい値しか使用できない。そのため、番組の中で環境雑音が変化した場合には、発話区間の検出に問題が生じるという問題がある。

【０００７】

特に、実時間の音声認識を行なう場合には、上記したような付帯情報が利用可能となるとは考えられない。また、電話による自動応答などに音声認識を用いる場合、音声信号の背景に存在する環境雑音がどのようなものになるかは予想できない。たとえば突発的な環境雑音が生じた場合、発話区間の検出を誤る可能性が高い。

【０００８】

また、音声認識においては発話中の音声エネルギの最大値で各フレームの音声エネルギを正規化した特徴量を用いると有効であることが知られている。しかしそのためには、発話の終了まで待って発話中での最大パワーを算出した後、算出された最大パワーを用いて当該発話中の各フレームの音声エネルギを正規化する必要がある。しかし、発話の終了まで待っていると実時間の音声認識を行なうことができないという問題がある。

【０００９】

従って、本発明の目的は、環境雑音にかかわらず発話区間の検出を適切に行なうことができる発話区間検出装置を提供することである。

【００１０】

本発明の他の目的は、環境雑音が変化しても発話区間の検出を適切に行なうことができる発話区間検出装置を提供することである。

【００１１】

本発明のさらに他の目的は、環境雑音が変化しても発話区間の検出を実時間で適切に行なうことができる発話区間検出装置を提供することである。

【００１２】

本発明のさらに他の目的は、突発的な環境雑音の変化があっても発話区間の検出を実時間で適切に行なうことができる発話区間検出装置を提供することである。

【００１３】

本発明の他の目的は、実時間で各フレームの音声エネルギを正規化することができる音声エネルギ正規化装置を提供することである。

【課題を解決するための手段】

【００１４】

本発明の第１の局面に係る発話区間検出装置は、音声データを逐次フレーム化するためのフレーム化手段と、フレーム化手段によりフレーム化された音声のエネルギ値をフレームごとに算出し、ＦＩＦＯ（Ｆｉｒｓｔ－Ｉｎ　Ｆｉｒｓｔ－Ｏｕｔ）形式で第１の個数のフレームのエネルギ値を記憶するフレームエネルギ算出及び記憶手段と、フレームエネルギ算出及び記憶手段に、第２の個数のフレームのエネルギ値が格納されたことに応答して、第２の個数のフレームのエネルギ値を所定の統計的手法に従って処理することにより、音声データに含まれる環境雑音の推定値の初期値を算出するための初期値算出手段と、推定値の初期値と、フレームエネルギ算出及び記憶手段に逐次記憶される音声のエネルギ値とに基づいて、音声データに含まれる環境雑音の変化に追従して変化する様に、発話区間を検出するためのエネルギ値のしきい値をフレームごとに逐次算出するための手段と、しきい値に基づいて、第２の個数のフレーム以降のフレームの中で、音声データの発話区間の開始位置又は終了位置に対応するフレームを推定するための発話区間推定手段とを含む。

【００１５】

環境雑音の推定値の初期値が、第２の個数のフレームのエネルギ値を統計的に処理することにより算出される。以後は、この推定値の初期値と、フレームエネルギ算出及び記憶手段に逐次記憶される音声のエネルギ値とに基づいて、音声データに含まれる環境雑音の変化に追従して変化する様に、発話区間を検出するためのエネルギ値のしきい値をフレームごとに逐次算出する。そのしきい値を用いて音声データの発話区間の開始位置又は終了位置に対応するフレームを推定する。しきい値が、環境雑音の変化に追従して変化するので、正確に発話区間の開始位置又は終了位置を推定できる。

【００１６】

好ましくは、初期値算出手段は、第２の個数のフレームを、各フレームのエネルギ値の大きさによって、第１のエネルギ値を中心とする第１のクラスタと、第１のエネルギよりも大きな第２のエネルギ値を中心とする第２のクラスタとにクラスタ化するための手段と、第１のエネルギ値を環境雑音の推定値の初期値として出力するための手段とを含む。

【００１７】

音声信号には、環境雑音と発話音声とが含まれる。各フレームをクラスタ化すると、環境雑音のみのフレームと、環境雑音と発話音声とを含むフレームとの二つのグループに分類されると思われる。フレームをエネルギの大きさに従って二つのクラスタにクラスタ化すると、エネルギの小さな第１のフレームからなるクラスタにおいて、環境雑音のみからなるフレームの占める割合が高くなる。そこで、この第１のクラスタのフレームのエネルギ値の平均を環境雑音の推定値の初期値とすれば、環境雑音の初期値を信頼性高く推定することができる。

【００１８】
より好ましくは、クラスタ化するための手段は、第２の個数のフレームを第１及び第２のクラスタにクラスタ化するための境界値を決定するための手段と、境界値よりも小さなエネルギ値を持つフレームを第１のクラスタに、それ以外のフレームを第２のクラスタに、それぞれ分類するための手段とを含む。
【００１９】
境界値を決定するための手段は、第２の個数のフレームのうち、エネルギ値をキーとしてソートしたときに予め定める第１のソート順位及び第２のソート順位となる二つのフレームを選択するための手段と、選択された二つのフレームのエネルギ値の平均値を算出するための第１の平均値算出手段と、第１の平均値算出手段により算出された平均値より小さいエネルギ値を持つか否かを基準として、第２の個数のフレームを第１及び第２のグループに分類するための手段と、第１及び第２のグループに属するフレームのエネルギ値の平均値をそれぞれ算出するための第２の平均値算出手段と、第２の平均値算出手段により算出された二つの平均値の平均値をさらに算出し、境界値として出力するための第３の平均値算出手段とを含んでもよい。
【００２０】
好ましくは、しきい値をフレームごとに逐次算出するための手段は、フレームエネルギ算出及び記憶手段に格納されているフレームのエネルギ値と、環境雑音の推定値の初期値とに基づいて、フレームエネルギ算出及び記憶手段に格納されているフレームの環境雑音のエネルギ値をフレームごとに推定するための手段と、フレームエネルギ算出及び記憶手段に格納されているフレームのエネルギ値のうち、定常的な背景雑音及び発話音声の合計のエネルギ値の最大値をフレームごとに逐次推定するための手段と、推定された環境雑音のエネルギ値と、推定された背景雑音及び発話音声の合計のエネルギ値とに基づいて、発話区間を検出するためのエネルギのしきい値をフレームごとに算出するための手段とを含む。
【００２１】
より好ましくは、発話区間推定手段は、しきい値に基づいて、第２の個数のフレーム以降のフレームの状態を判定するための手段を含み、状態は、非発話状態を含み、環境雑音のエネルギ値をフレームごとに逐次推定するための手段は、１フレーム前の時点において推定された環境雑音のエネルギ値を記憶するための手段と、環境雑音の推定値の初期値が算出された時点で記憶するための手段に環境雑音の推定値の初期値を記憶させるための手段と、記憶するための手段に記憶された値、フレームエネルギ算出及び記憶手段に含まれるフレームのエネルギ値、及びフレームの状態を判定する手段による判定結果に基づいて、以下の式

$b(t)=b(t-1)\times\alpha+E(t)\times(1-\alpha)$　（状態が非発話状態の場合）

$b(t)=b(t-1)$　（状態が非発話状態以外の場合）

ただしαは所定の忘却係数、Ｅ（ｔ）は時刻ｔにおけるフレームのエネルギ値、に従って時刻ｔにおける背景雑音ｂ（ｔ）を算出するための手段とを含み、記憶するための手段は、算出された背景雑音ｂ（ｔ）を記憶する。
【００２２】
合計のエネルギ値の最大値をフレームごとに推定するための手段は、フレームエネルギ算出及び記憶手段に格納されているフレームを、エネルギ値をキーとしてソートするための手段と、ソートするための手段によりソートされた結果所定の順位となるフレームのエネルギ値を合計のエネルギ値の最大値$Emax(t)$として選択するための手段を含んでもよい。
【００２３】
好ましくは、しきい値をフレームごとに逐次算出するための手段は、時刻ｔにおける発話開始位置検出のためのしきい値$Eth_1(t)$を、

$Eth_1(t)=b(t)+\max(\beta, Emax(t)-b(t))\times$第１の定数

に従って算出するための手段を含む。

【００２４】
さらに好ましくは、しきい値をフレームごとに逐次算出するための手段は、さらに、時刻ｔにおける発話終了位置検出のためのしきい値Ｅｔｈ$_2$（ｔ）を、
$Eth_2(t)=b(t)+\max(\beta,\ Emax(t)-b(t))\times$第２の定数
ただし第２の定数＜第１の定数、に従って算出するための手段を含む。
【００２５】
発話区間検出装置はさらに、発話の先頭からの各フレームの音声データの最大エネルギ値又は所定のデフォルト基準値のいずれか大きい方を用いて各フレームの音声データを正規化し、各フレームの音声特徴パラメータとして出力するための音声エネルギ正規化手段を含んでもよい。
【００２６】
発話の先頭からの各フレームの音声データの最大エネルギ値又は所定のデフォルト基準値のいずれか大きい方を用いて正規化するので、発話の終了まで待たずに、擬似的にではあるが実時間で正規化することが可能になる。したがって、音声特徴パラメータの一つとして音声エネルギを実時間で得ることができる。
【００２７】
好ましくは、音声エネルギ正規化手段は、正規化の基準値を記憶するための基準値記憶手段と、フレームエネルギ算出及び記憶手段により算出された音声エネルギが、基準値記憶手段に記憶された基準値を超えていることを検出し、検出信号を出力するための検出手段と、検出手段により出力される検出信号に応答して、基準値記憶手段に記憶された基準値を、フレームエネルギ算出及び記憶手段により算出された値で置換するための手段と、フレームエネルギ算出及び記憶手段により算出された音声エネルギ値を、基準値記憶手段に記憶された基準値で除算することにより、当該フレームの音声エネルギを正規化するための除算手段とを含む。
【００２８】
さらに好ましくは、発話区間検出装置は、発話区間推定手段により、発話区間の終了位置に対応するフレームが推定されたことに応答して、基準値記憶手段の記憶内容を、所定のデフォルト値で置換するための手段をさらに含む。
【００２９】
発話区間検出装置は、所定のデフォルト値を、発話区間検出装置の起動時に与えられたオプション値に基づいて設定するための手段をさらに含んでもよい。
【００３０】
本発明の第２の局面に係るコンピュータプログラムは、上記したいずれかの発話区間検出装置としてコンピュータを動作させるためのものである。
【００３１】
本発明の第３の局面にかかる音声エネルギ正規化装置は、フレーム化された音声データの正規化音声エネルギを実時間で算出するための音声エネルギ正規化装置であって、正規化の基準値を記憶するための基準値記憶手段と、フレームごとの音声データの音声エネルギを算出するための手段と、音声エネルギ算出手段により算出された音声エネルギが、基準値記憶手段に記憶された基準値を超えていることを検出し、検出信号を出力するための手段と、検出手段により出力される検出信号に応答して、基準値記憶手段に記憶された基準値を、音声エネルギ算出手段により算出された値で置換するための手段と、音声エネルギ算出手段により算出された音声エネルギを、基準値記憶手段に記憶された基準値で除算することにより、当該フレームの音声エネルギを正規化するための除算手段とを含む。
【００３２】
発話区間の最初においては、デフォルトの値を基準値として音声エネルギを正規化する。発話区間の途中でフレームの音声エネルギが基準値を超えると、フレームの音声エネルギを新たな基準値として音声エネルギを正規化する。発話区間の終了まで到達しなくても擬似的にではあるが音声エネルギの実時間での正規化が可能になる。発話区間の最初では誤差が生ずるが、実際に音声エネルギが発話区間中での最大値まで到達すると、後は正確

な正規化が行なえる。またデフォルトの値を適切に選ぶことにより、発話区間の最初に生ずる誤差も小さく抑えることができる。

【００３３】

好ましくは、音声エネルギ正規化装置は、発話区間の終了を検出して発話終了検出信号を出力するための手段と、発話終了検出信号に応答して、基準値記憶手段の記憶内容を、所定のデフォルト値で置換するための手段とをさらに含む。

【００３４】

発話区間が終了すると、基準値を再びデフォルトの値に再設定できる。音声エネルギを、フレームごとに適切な基準値を使用して正規化できる。

【００３５】

さらに好ましくは、音声エネルギ正規化装置は、所定のデフォルト値を、音声エネルギ正規化装置の起動時に与えられたオプション値に基づいて設定するための手段をさらに含む。

【００３６】

起動時のオプション値によってデフォルト値を設定できるので、様々なオプション値をデフォルト値として音声エネルギ正規化装置を動作させることができる。その結果、音声エネルギの正規化処理をより適切に実現することが容易になる。

【００３７】

本発明の第４の局面に係るコンピュータプログラムは、上記したいずれかの音声エネルギ正規化装置としてコンピュータを動作させるためのものである。

【００３８】

本発明の第５の局面に係るコンピュータは、上記した第２の局面に係るコンピュータプログラム、又は第４の局面に係るコンピュータプログラムによりプログラムされ、発話区間検出装置又は音声エネルギ正規化装置として動作する。

【発明を実施するための最良の形態】

【００３９】

本実施の形態に係る発話区間検出装置は、フレーム化して入力される音声信号に基づき、統計的手法によって発話区間検出の際のしきい値を変化させる。その際、装置の立上がり時の遅延をできるだけ少なくするとともに、突発的な雑音があっても安定して発話区間の検出を行なうことができるよう、統計的手法を工夫している。また、音声認識のための特徴量パラメータとしてフレームの正規化した音声エネルギを算出する際、実時間処理によって、擬似的な正規化ができるような工夫をしている。

【００４０】

［発話区間の検出原理］

図１に、音声信号と、本実施の形態において発話区間の検出に使用される手法で使用される様々なパラメータとを示す。図１を参照して、音声信号２０に対し、発話開始しきい値２２と発話終了しきい値２４という二つのしきい値を用いて発話の開始位置２６及び終了位置２８を判定する。これら発話開始しきい値２２及び発話終了しきい値２４は、入力波形データからフレーム単位で算出されるエネルギから統計的手法により定められる。これらを定める手法については後述する。

【００４１】

図１において、発話区間の判定の際に使用される時間的パラメータＴ１からＴ６は以下の意味を持つ。

【００４２】

Ｔ１：プリロール時間　あるフレームが発話の開始位置であると判定されたとき、そのフレームからさらにこのプリロール時間だけさかのぼった位置（図１の参照符号２６）のフレームに、発話開始フレームとしてのマークが付される。

【００４３】

Ｔ２：発話開始判定時間　発話が開始したと判定されるための第１の条件として、フレーム単位のエネルギ値が連続して発話開始しきい値を超えなければならない時間。

【００４４】
Ｔ３：最短発話時間　発話開始と判定されるために、フレーム単位のエネルギ値が連続して超えなければならない最小時間。エネルギ値が発話開始しきい値をＴ２時間連続して超え、かつＴ３時間連続して超えてはじめて発話開始と判定される。
【００４５】
Ｔ４：最長無音時間　発話状態でフレーム単位のエネルギ値が発話終了しきい値を下回っても、発話終了と判定されない最長の時間。
【００４６】
Ｔ５：発話終了判定時間　発話が終了したと判定されるための第１の条件として、フレーム単位のエネルギ値が連続して発話終了しきい値を下回らなければならない時間。エネルギ値が発話終了しきい値をＴ５時間連続して下回り、かつＴ４時間連続して下回った場合、発話終了と判定される。
【００４７】
Ｔ６：アフタロール時間　あるフレームで発話終了と判定されたとき、そのフレームからさらにこのアフタロール時間だけ下った位置のフレーム（図１の参照符号２８）に、発話終了フレームとしてのマークが付される。
【００４８】
図１の水平軸付近に記載されているＳ１からＳ４の符号は、後述する手法により決定される、各フレームの状態を示す。図２に、フレームの状態の遷移を示す。
【００４９】
図２を参照して、フレームは４つの状態（非発話状態（Ｓ１）３０、発話開始状態（Ｓ２）３２、発話状態（Ｓ３）３４，及び発話終了状態（Ｓ４）３６）の間を遷移する。状態間の遷移は以下の様にして行なわれる。
【００５０】
（１）　非発話状態（Ｓ１）３０で、フレームのエネルギ値が発話開始しきい値２２を上回ると状態は発話開始状態（Ｓ２）３２に遷移する（アーク４２）。
【００５１】
（２）　発話開始状態（Ｓ２）３２が、一定時間Ｔ３だけ継続すると状態は発話状態（Ｓ３）３４となる（アーク４８）。
【００５２】
（３）　発話開始状態（Ｓ２）３２で、フレームのエネルギ値が発話開始しきい値２２を下回ると状態は非発話状態（Ｓ１）３０に遷移する（アーク４６）。
【００５３】
（４）　発話状態（Ｓ３）３４で、フレームのエネルギ値が発話終了しきい値２４を下回ると状態は発話終了状態（Ｓ４）３６に遷移する（アーク５２）。
【００５４】
（５）　発話終了状態（Ｓ４）３６が、一定時間Ｔ４だけ継続すると状態は非発話状態（Ｓ１）３０に遷移する（アーク５８）。
【００５５】
（６）　発話終了状態（Ｓ４）３６で、フレームのエネルギ値が発話終了しきい値２４を上回ると状態は発話状態（Ｓ３）３４に戻る（アーク５４）。
【００５６】
（７）　それ以外の場合、状態は現在の状態を維持する（アーク４０、４４、５０及び５６）。
【００５７】
上記した種々のパラメータは、本実施の形態の装置では、装置の起動時に手操作により設定される。設定のないものはデフォルト値が用いられる。パラメータ設定の部分は本発明と直接関係をもたないため、以下の説明では詳細には説明しない。
【００５８】
［フレームの構成］

後述する様に、本実施の形態に係る装置は、音声入力信号をフレーム単位で処理する。図３にフレーム及びフレームシフトの概念を説明するための模式図を示す。

【００５９】

図３を参照して、各フレーム７０、７２、７４、…はフレーム長Ｔｗ＝３０ミリ秒の長さの音声信号である。本実施の形態では、このフレームを１０ミリ秒単位で時間軸上を移動させながら順次音声信号をフレーム化する。この移動量をフレームシフト量と呼ぶ。従って、本実施の形態の装置の処理対象となる音声データは、フレーム長３０ミリ秒，フレームシフト量１０ミリ秒である。

【００６０】

また、各フレームのエネルギは、当該フレーム中のデータに窓関数８０（ハミング窓）で示される値を乗算して総和を計算することにより得られる。フレームごとのエネルギの算出方法については後述する。

【００６１】

本実施の形態の装置では、通常は１００フレームのデータを統計的に処理することにより発話開始しきい値２２及び発話終了しきい値２４を動的に計算する。この様に動的な処理を行なう場合、ある程度のデータが集積されないと処理を開始することができない。他方で、あまり多くのデータを使用して統計的処理を行なおうとすると、装置が適切に動作するまでの時間的遅延が長くなり、発話の最初を正しく検出できなくなるおそれがある。

【００６２】

そこで、本実施の形態の装置では、処理の開始後、最初の４００ミリ秒までは無音状態であると仮定し、この間に４０フレーム分のデータをフレームバッファに収集する。この４０フレーム分のデータを用いて環境雑音の初期値を求め、その値を用いてさらにしきい値の初期値を決める。以後、１００フレーム分のデータが収集されるまで、フレームデータをフレームバッファに蓄積しながら、収集したデータを用いてしきい値を動的に計算する。１００フレームに達したら、以後、ＦＩＦＯ（Ｆｉｒｓｔ－Ｉｎ　Ｆｉｒｓｔ－Ｏｕｔ）形式でフレームデータを１００個に維持しながらしきい値の計算を行なう。なお、この最大のフレーム数（フレームバッファ内に記憶され使用される最大のフレーム数）をフレームバッファサイズと呼ぶことにする。また、環境雑音の初期値を求めるために使用するフレームの数を初期バッファサイズと呼ぶ。すなわち、本実施の形態の装置ではフレームバッファサイズは１００、初期バッファサイズは４０である。

【００６３】

なお、これらのフレームバッファサイズ及び初期バッファサイズは一例であって、これ以外の値を用いることも考えられる。

【００６４】

以下の説明では、入力されるフレームの番号をｔ（０≦ｔ）で表す。フレームは１０ミリ秒ごとに入力されるので、ｔはまた時刻も表す。従って、以下の説明では単に「ｔ番目のフレーム」を「時刻ｔにおけるフレーム」という表現で表すこともある。

【００６５】

こうした処理を行なうことで、処理開始時の遅延は４００ミリ秒となり、実用上の問題は見られない。通常は１００個のフレームデータを用いてしきい値を計算するので、信頼性高く発話区間の検出を行なうことができる。

【００６６】

［装置の構成］

図４は、本実施の形態に係る発話区間検出装置の構成を示す機能的ブロック図である。図４を参照して、この発話区間検出装置１００は、マイク１０２から与えられる音声信号の中で発話区間を検出するためのものである。発話区間検出装置１００は、マイク１０２から与えられる音声信号を標本化し、量子化することによりデジタル化し、さらに上記した形式のフレームデータとして１０ミリ秒ごとに出力するとともに、フレームデータを出力したことを示すフレーム出力信号１２４を出力するための音声入力部１０４と、音声入力部１０４から与えられる複数個のフレームデータを記憶するための入力バッファ１０６

とを含む。

【００６７】

発話区間検出装置１００はさらに、入力バッファ１０６からフレームデータを読出してエネルギ値などのフレーム情報を算出するためのフレーム情報算出部１０８と、フレーム情報算出部１０８の出力するフレーム情報を記憶するためのフレームバッファ１１０とを含む。フレームバッファ１１０のバッファサイズは、前述した通り１００フレーム分である。フレームバッファ１１０は、入力されたフレーム情報をＦＩＦＯ形式で１００個保持することができる。

【００６８】

本実施の形態では、フレーム情報算出部１０８は、次の式に従って時刻ｔにおけるフレームの音声エネルギＥ（ｔ）を算出する。

【００６９】

【数１】

$$E(t) = \log \frac{\sum_{i=1}^{N} |S_i| \times H_i}{N} \times 20$$

ただし、Ｎは１フレーム中のデータサンプル数、S_i（ｉ＝１～Ｎ）はデータの値、H_i（ｉ＝１～Ｎ）はハミング窓関数の値を、それぞれ示す。

【００７０】

発話区間検出装置１００はさらに、フレーム情報算出部１０８が算出したフレームの音声エネルギを、発話中の最大パワーを基準として正規化し、フレームの特徴ベクトルの一要素として入力バッファ１０６中に書込むためのフレーム音声エネルギ正規化処理部１２６を含む。フレームの音声エネルギの大きさを一発話の中の最大エネルギで正規化し、特徴量の一つとして音声認識に利用すると効果があることが認められている。しかし、そのためには発話の終了まで待ってフレームエネルギの最大値を算出する必要がある。しかしそれでは実時間処理を行なうことができない。

【００７１】

そこでフレーム音声エネルギ正規化処理部１２６は、音声エネルギのダイナミックレンジを実時間に更新することにより、擬似的にではあるが音声エネルギを実時間で正規化する機能を持つ。フレーム音声エネルギ正規化処理部１２６はそのため、図５に示すような構成を持つ。

【００７２】

図５を参照して、フレーム音声エネルギ正規化処理部１２６は、発話の先頭部分でまた音声エネルギの十分大きなフレームがないときに、最大音声エネルギのデフォルト値として使用されるデフォルト最大値を記憶するためのデフォルト最大値記憶部１３２と、発話の最初の部分ではデフォルト最大値記憶部１３２から与えられたデフォルト最大値を記憶し、発話途中でデフォルト最大値より大きな音声エネルギを持つフレームが検出された場合に、当該音声エネルギの値を記憶するための最大値記憶部１３４と、フレーム情報算出部１０８からの音声エネルギ１２８を最大値記憶部１３４に記憶されている最大値で除算し、結果を入力バッファ１０６の該当フレームの特徴量の一つとして書込むための除算部１３６と、最大値記憶部１３４の出力とフレーム情報算出部１０８からの音声エネルギ１２８とを受けて両者の値を比較し、比較結果信号１３９を最大値記憶部１３４に与えるための比較部１３８とを含む。比較結果信号１３９は、音声エネルギ１２８により示される値が最大値記憶部１３４に記憶された最大値を上回るとＨ（ハイ）レベルとなり，それ以外の場合はＬ（ロー）レベルとなる。なお、デフォルトの値は、オプションとしてこの装置（プログラム）起動時に与えられた値があれば、その値で書換えられる。

【００７３】

最大値記憶部１３４は、状態判定部１１８から与えられる信号２００によって発話が終了したことが示されると、デフォルト最大値記憶部１３２の値を新たな最大値として記憶

し、比較部１３８からの比較結果信号１３９がＨレベルとなると、音声エネルギ１２８により示される値を新たな最大値として記憶する。したがって、最大値記憶部１３４に記憶される値は、発話開始時にはデフォルト最大値記憶部１３２に記憶されていたデフォルト値となり、発話の進行とともに音声エネルギがデフォルト値を上回るものが出現するとその音声エネルギとなる。以下、発話の進行中には同様の処理が繰返される。この値を発話中の音声エネルギの最大値として使用して各フレームの音声エネルギを正規化することにより、擬似的にではあるが、実時間で音声エネルギの正規化を行なうことができる。

【００７４】

なお、デフォルトの値は予め実験により適切な値を決めておくことが望ましい。

【００７５】

発話区間検出装置１００はさらに、音声入力部１０４からのフレーム出力信号１２４を受け、入力バッファ１０６、フレーム情報算出部１０８及びフレームバッファ１１０の読出ポイント及び書込ポイント、並びにそれらへの書込み・読出しのタイミングを管理するための入出力・アドレス管理部１１４と、発話区間検出装置１００の処理開始後４００ミリ秒までの間にフレームバッファ１１０に格納されるフレームデータ１６０を読出し、初期環境雑音を算出するための初期環境雑音算出部１１２と、フレームバッファ１１０からのフレームデータ１９２、初期環境雑音算出部１１２からの初期環境雑音の推定値１９４、及び現在の状態が非発話状態（Ｓ１）３０（図２参照）か否かを示す信号１９０を受け、それらから発話開始しきい値２２及び発話終了しきい値２４を動的に算出し、しきい値の値を示す信号１９８として出力するための動的しきい値算出部１１６とを含む。

【００７６】

入力バッファ１０６、フレームバッファ１１０などは半導体記憶装置などにより実現される。入出力・アドレス管理部１１４はタイマを装備しており、音声入力部１０４による音声データのデジタル化に同期して、入力バッファ１０６、フレームバッファ１１０などへの書込みのポインタ、それらからの読出しポインタを管理する。入出力・アドレス管理部１１４はまた、起動後４００ミリ秒までのフレームを処理する際にはＨレベル、それ以後はＬレベルの値をとる初回フラグ１９６を動的しきい値算出部１１６に与える機能も持つ。動的しきい値算出部１１６の処理は、初回フラグ１９６及び信号１９０の値によって制御される。

【００７７】

発話区間検出装置１００はさらに、動的しきい値算出部１１６から出力されたしきい値の値を示す信号１９８及びフレームバッファ１１０からのフレームデータ１９２とから、後述する方法に従ってフレームの状態を判定し、状態を表す信号２００を出力するための状態判定部１１８と、状態判定部１１８の出力する状態を表す信号２００を受け、入力バッファ１０６から状態の確定したフレームに対応する入力データを読出して予め定められた算出方法によってこのフレームの音声の特徴ベクトルを算出し、さらに発話区間の開始又は終了フレームの場合には、それらを示すマークを当該特徴ベクトル１２２に付して出力するための特徴ベクトル出力部１２０とを含む。状態判定部１１８はまた、現在の状態が非発話状態（Ｓ１）３０か否かを示す信号１９０を生成し、動的しきい値算出部１１６に与える機能も持つ。

【００７８】

図６は初期環境雑音算出部１１２のブロック図であって、初期環境雑音算出部１１２は、フレームバッファ１１０から与えられるフレーム情報のうち、フレームごとのエネルギ値を昇順にソートしてソート後フレームエネルギ記憶部１４２に格納させるためのソート処理部１４０と、ソート処理部１４０に格納されたフレームごとのエネルギ値のうち、下位から２５％及び７５％の大きさにあたる位置のフレームのエネルギを算出し、それぞれ後述するクラスタリング処理のシードとなる値ｅｍ１及びｅｍ２として出力するためのシーズ算出部１４４と、この値ｅｍ１及びｅｍ２を記憶するための記憶部１４６とを含む。

【００７９】

初期環境雑音算出部１１２はさらに、記憶部１４６から値ｅｍ１及びｅｍ２を読出し、

その平均値$e_{average}$を算出するための第１の平均値算出部１４８と、第１の平均値算出部１４８が出力する平均値を境界値としてそれより大きいエネルギ値を持つか否かを基準として、ソート後フレームエネルギ記憶部１４２中の各フレームを二つのクラスタＣ１及びＣ２に分類するためのフレーム分類部１５０と、フレーム分類部１５０により得られた二つのクラスタＣ１及びＣ２の各々に属するフレームのエネルギ値の平均値Ｅｍ１及びＥｍ２を次の式に従って算出するための第２の平均値算出部１５２とを含む。

【００８０】

【数２】

$$Em1=\frac{1}{I1}\sum_{i_1=1}^{I1}E(i_1)\quad Em2=\frac{1}{I2}\sum_{i_2=I1+1}^{N}E(i_2)$$

ただし、Ｎはフレームバッファ１１０内のフレーム数、Ｉ１は$e_{average}$より小さいエネルギ値を持ち、クラスタＣ１に属するフレームの数、Ｉ２は$e_{average}$より大きいエネルギ値を持ち、クラスタＣ２に属するフレームの数を、それぞれ表す。

【００８１】

初期環境雑音算出部１１２はさらに、第２の平均値算出部１５２によって算出された二つの平均値Ｅｍ１及びＥｍ２をそれぞれ新たな値ｅｍ１及びｅｍ２として記憶部１４６に記憶させ、さらに第１の平均値算出部１４８、フレーム分類部１５０、及び第２の平均値算出部１５２に先ほどの処理を繰返し実行させ、その結果得られた平均値Ｅｍ１を初期環境雑音の推定値（ｅｍ１）１９４として図４に示す動的しきい値算出部１１６に与えるための判定部１５４とを含む。

【００８２】

以下に、第１の平均値算出部１４８、フレーム分類部１５０及び第２の平均値算出部１５２により行なわれる処理について、図４、及び図６から図９を参照して説明する。一般に、図４に示すフレームバッファ１１０に記憶されている各フレームのエネルギ値は、図７に示される様に、入力音声信号のエネルギの大きさに従って変動する。これをエネルギの大きさに従って昇順にソートすると図８の様になると想定される。ソート処理部１４０が行なうソート処理はこうした処理であり、ソート後フレームエネルギ記憶部１４２に記憶されているフレーム情報は図８に示すものに対応している。

【００８３】

図８の様にソートすることで、エネルギ値のヒストグラムを容易にとることができる。図９にその例を示す。音声信号に環境雑音と発話成分とが含まれているとすれば、環境雑音のみのフレームのエネルギ値と、発話成分を含むフレームのエネルギ値とは、それぞれ別々の値を中心として分布することになると思われる。そして、それらは図９に示されるようなヒストグラムにおいて、エネルギの比較的低い部分のピークと、エネルギの比較的高い部分のピークとの二つのピークを形成することになるであろう。

【００８４】

図６に示す第１の平均値算出部１４８、フレーム分類部１５０、及び第２の平均値算出部１５２が行なっているのは、最初にエネルギ値の２５％と７５％の部分とをピークの初期位置として、上記した二つのピークをその後の演算により求め、ソート後フレームエネルギ記憶部１４２に格納されている各フレームを、環境雑音側のピークに近いフレームと、発話部分側のピークに近いフレームとの二つのクラスタにクラスタ化する処理である。

【００８５】

図１０は、図４に示す動的しきい値算出部１１６の機能的ブロック図である。図１０を参照して、動的しきい値算出部１１６は、フレームデータ１９２を受け、フレームバッファ１１０に格納されているソート後のフレーム情報のうち、小さい方から９０％の位置にあるフレームのエネルギを、ｔ番目までのフレームバッファサイズ分の数のフレームにおける最大エネルギe_{max}（ｔ）（最大エネルギ信号１８２）として出力するための最大エネルギ算出部１７６と、フレームデータ１９２を受け、後述する式に従って環境雑音の推定値を算出するための環境雑音算出部１７０と、１フレーム分だけ前の処理で算出された

環境雑音の推定値ｂ（ｔ－１）を記憶するための記憶部１７４とを含む。

【００８６】

動的しきい値算出部１１６はさらに、記憶部１７４に記憶されている１フレーム分だけ前の推定値ｂ（ｔ－１）と、環境雑音算出部１７０から与えられる環境雑音の推定値と、初期環境雑音の推定値（ｅｍ１）１９４とを受けて、初回フラグ１９６がＨレベルであれば初期環境雑音の推定値（ｅｍ１）１９４を、初回フラグ１９６がＬレベルでかつ状態を示す信号１９０が非発話状態を示す値であれば環境雑音算出部１７０の出力を、初回フラグ１９６がＬレベルでかつ状態を示す信号１９０が非発話状態を示す値でなければ記憶部１７４の出力を、それぞれ選択してｔ番目のフレームに対する環境雑音ｂ（ｔ）として出力するための選択部１７２とを含む。選択部１７２の出力は記憶部１７４に与えられ記憶される。

【００８７】

動的しきい値算出部１１６はさらに、最大エネルギ算出部１７６からの最大エネルギ値と、選択部１７２からのｔ番目のフレームにおける環境雑音ｂ（ｔ）とに基づいて発話開始しきい値２２及び発話終了しきい値２４を動的に算出するためのしきい値算出部１７８を含む。しきい値算出部１７８の出力する、しきい値を表す信号１９８は状態判定部１１８に与えられ、状態判定に用いられる。

【００８８】

環境雑音算出部１７０は、フレームバッファ１１０に記憶されたフレームデータ１９２の中でｔ番目のフレームのエネルギＥ（ｔ）、及び記憶部１７４に記憶されたｔ－１番目のフレームに対する環境雑音ｂ（ｔ－１）とから次の式１に従って環境雑音の推定値ｂ’（ｔ）を算出する。

［式１］

$$b'(t) = b(t-1) \times \alpha + E(t) \times (1-\alpha)$$

ここで、αは予め定められた忘却係数、Ｅ（ｔ）はｔ番目のフレームのエネルギを表す。忘却係数は０以上１以下の値であるが、本実施の形態では０．８を用いる。

【００８９】

選択部１７２は、状態が非発話状態以外であれば記憶部１７４から出力されるｔ－１番目のフレームに対する環境雑音ｂ（ｔ－１）を選択する。従ってこの場合には環境雑音は変化しない。状態が非発話状態であれば、選択部１７２は、環境雑音算出部１７０から出力される環境雑音の推定値ｂ’（ｔ）を選択する。

【００９０】

従って、環境雑音算出部１７０から出力される時刻ｔにおける環境雑音ｂ（ｔ）は以下の通りの式で表される。ただしＥ（ｔ）は時刻ｔにおけるフレームのエネルギ値、αは前述の忘却係数である。

［式２］

$b(t) = b(t-1) \times \alpha + E(t) \times (1-\alpha)$　（状態が非発話状態の場合）

$b(t) = b(t-1)$　（状態が非発話状態以外の場合）

しきい値算出部１７８は以下の式に従って発話開始しきい値Eth_1及び発話終了しきい値cを動的に算出する。

［式３］

０≦ｔ＜４００ミリ秒では

$Eth1(t) = b(t) + \beta \times \gamma_1$

$Eth2(t) = b(t) + \beta \times \gamma_2$、

４００ミリ秒≦ｔでは

$Eth_1(t) = b(t) + \max(\beta, Emax(t) - b(t)) \times \gamma_1$

$Eth_2(t) = b(t) + \max(\beta, Emax(t) - b(t)) \times \gamma_2$

ただし、βは発話の最低ダイナミックレンジで、本実施の形態では２０ｄＢである。またγ_1及びγ_2はそれぞれ発話開始しきい値比率及び発話終了しきい値比率であり、それぞれ実験的に定められる、０以上で１以下の定数である。本実施の形態ではγ_1＝０．２５、

γ_2＝０．２０を用いる。

【００９１】

こうして算出された発話開始しきい値Ｅｔｈ$_1$及び発話終了しきい値Ｅｔｈ$_2$が、図１を参照して説明した発話区間の検出時の発話開始しきい値２２及び発話終了しきい値２４として用いられる。

【００９２】

［装置の動作］

以上構成を述べた装置は以下のように動作する。

【００９３】

-起動時-

起動時には、処理に必要となるバッファ及びオプションの値を格納するためのエリアを記憶装置に確保する。起動時に与えられるオプションの値を調べ、オプションの値に誤りがなければ当該オプションに、与えられた値を設定する。オプションの値が与えられなかったものにはデフォルト値を設定する。与えられたオプションの値に誤りがあれば、その旨のメッセージを表示して処理を終了する。図５に示すフレーム音声エネルギ正規化処理部１２６のデフォルト最大値記憶部１３２については、起動時にオプションの値が与えられれば、その値をデフォルトの値として記憶し、さらに最大値記憶部１３４に記憶する。オプションの値が与えられなければ、プログラム上のデフォルト値をデフォルト最大値記憶部１３２に記憶し、さらに最大値記憶部１３４に記憶する。

【００９４】

各バッファの書込みポイント及び読出しポイントをそれぞれ初期値に設定する。

【００９５】

なお、起動後、実際の処理を開始する時刻（フレーム番号）をｔ＝０とする。このときのフレームの状態は非発話状態に設定される。以後、図４に示す音声入力部１０４は、マイク１０２からの電気信号を１０ミリ秒ごとに、３０ミリ秒のフレーム長でデジタル化する。

【００９６】

-０ミリ秒から４００ミリ秒まで-

入出力・アドレス管理部１１４からの初回フラグ１９６はＨレベルである。音声入力部１０４は、発話判定に必要なデータ数が集まると、１回の処理で引き渡す数として予め定められた数のデータを入力バッファ１０６の、バッファ書込みポインタにより指定されるアドレスに書込む。

【００９７】

フレーム情報算出部１０８は、入力バッファ１０６の、読出しポインタにより指定されるアドレスから１フレーム分のデータを読出し、フレームエネルギを算出してフレームバッファ１１０の当該フレームに対応するエリアに書込む。フレーム情報算出部１０８はまた、算出されたフレームエネルギをこのフレームの音声エネルギ１２８として図５に示す除算部１３６、比較部１３８及び最大値記憶部１３４に与える。比較部１３８は、最大値記憶部１３４に記憶された値と音声エネルギ１２８により示される値とを比較し、比較結果信号１３９を最大値記憶部１３４に与える。音声エネルギ１２８により示される値が最大値記憶部１３４に記憶されている値を上回ったことが検出されると、比較結果信号１３９はＨレベルとなり、最大値記憶部１３４は比較結果信号１３９がＨレベルとなったことに応答して、これまで記憶していた値に代えて音声エネルギ１２８により表される値を記憶する。

【００９８】

除算部１３６は、音声エネルギ１２８により表される値を最大値記憶部１３４に記憶された値で除算して正規化された音声エネルギを算出する。正規化された音声エネルギ１３０は、入力バッファ１０６中で該当するフレームの、正規化音声エネルギのフィールドに書込まれる。以後、フレーム情報算出部１０８とフレーム音声エネルギ正規化処理部１２６とは、これと同様の動作を各フレームに対して繰返す。

【００９９】
初期環境雑音算出部１１２は、フレーム情報算出部１０８によりフレームバッファ１１０に書込まれたフレームエネルギを読出し、初期環境雑音の算出を行なう。時刻０ミリ秒から４００ミリ秒の間は、状態の判定は行なわない。
【０１００】
次に、図６を参照して、初期環境雑音算出部１１２の動作について説明する。ソート処理部１４０は、フレームバッファ１１０から読出したフレームエネルギの値１６０をソートし、ソート後フレームエネルギ記憶部１４２に格納する。ｔ＝０では読出されるフレームエネルギの値は一つ（Ｅ（０））だけなので、その値をソート後フレームエネルギ記憶部１４２の１番目の領域に書込む。２回目以後は、ソート後フレームエネルギ記憶部１４２に前のソートの結果が既に書込まれており、そこに新たに一つのフレームエネルギをその大きさに従った位置に追加するだけでよい（ヒープソート）。従って、ソート処理は少ない計算量で実行できる。
【０１０１】
起動後、０ミリ秒から４００ミリ秒の間は、シーズ算出部１４４以後の処理部は動作しない。
【０１０２】
-４００ミリ秒経過時-
起動後４００ミリ秒が経過すると、フレームバッファ１１０には４０個のフレームデータ（Ｅ（０）～Ｅ（３９））のエネルギ値が格納されている。この状態が図７に対応する。ソート後フレームエネルギ記憶部１４２には、これら４０個のフレームのエネルギ値を昇順にソートしたものが格納されている。この状態が図８に対応する。
【０１０３】
フレーム情報算出部１０８及びフレーム音声エネルギ正規化処理部１２６は、４００ミリ秒経過までと同様に動作する。
【０１０４】
除算部１３６は、音声エネルギ１２８により表される値を最大値記憶部１３４に記憶された値で除算して正規化された音声エネルギを算出する。正規化された音声エネルギ１３０は、入力バッファ１０６中で該当するフレームの、正規化音声エネルギのフィールドに書込まれる。
【０１０５】
シーズ算出部１４４は、ソート後フレームエネルギ記憶部１４２に格納されている４０個のフレームエネルギのうち、小さい方から２５％及び７５％に相当する値を算出する。この値は記憶部１４６に記憶され、第１の平均値算出部１４８、フレーム分類部１５０及び第２の平均値算出部１５２により行なわれるクラスタリングのシードとなる。
【０１０６】
第１の平均値算出部１４８は、記憶部１４６からこのシードｅｍ１及びｅｍ２の平均値を算出しフレーム分類部１５０に与える。フレーム分類部１５０は、全てのフレームについて、そのエネルギ値がシードｅｍ１及びｅｍ２のいずれに近いかを基準として、４０個のフレームを二つのクラスタに分類し、分類した結果を第２の平均値算出部１５２に与える。
【０１０７】
第２の平均値算出部１５２は、それら二つのクラスタの各々について、そのクラスタに属するフレームのエネルギ値の平均値Ｅｍ１及びＥｍ２を算出し判定部１５４に与える。
【０１０８】
判定部１５４は、第２の平均値算出部１５２から与えられたＥｍ１及びＥｍ２を記憶部１４６に新たなｅｍ１及びｅｍ２として記憶させ、先ほどと同じ処理を第１の平均値算出部１４８、フレーム分類部１５０，及び第２の平均値算出部１５２に実行させる。こうして再び得られたＥｍ１及びＥｍ２のうち、Ｅｍ１を初期環境雑音の推定値１９４（ｅｍ１）として動的しきい値算出部１１６に与える。

【０１０９】
図１０を参照して、動的しきい値算出部１１６の動作について説明する。動的しきい値算出部１１６の選択部１７２は、ｂ（ｔ）の初期値として初期環境雑音の推定値１９４であるｅｍ１を選択し、記憶部１７４及びしきい値算出部１７８に与える。記憶部１７４はこの値を記憶する。
【０１１０】
一方、最大エネルギ算出部１７６は、ソート後フレームエネルギ記憶部１４２に記憶されているソートされているフレームエネルギの値のうち、小さい方から９０％に相当するエネルギ値を計算し、最大エネルギ値（Ｅmax）１８２としてしきい値算出部１７８に与える。
【０１１１】
しきい値算出部１７８は、選択部１７２から与えられる環境雑音の推定値ｅｍ１と、最大エネルギ算出部１７６からの最大エネルギ値（Ｅmax）１８２とに基づき、前述の式３によって発話開始しきい値２２及び発話終了しきい値２４を算出し（１９８）、図４に示す状態判定部１１８に与える。
【０１１２】
状態判定部１１８は、動的しきい値算出部１１６から与えられる発話開始しきい値２２及び発話終了しきい値２４に基づき、図１及び図２に示す判定方法に従ってフレームの状態を判定し、その結果を表す信号２００を特徴ベクトル出力部１２０及びフレーム音声エネルギ正規化処理部１２６に与える。状態判定部１１８はまた、フレームの状態が非発話状態か否かを示す信号１９０を動的しきい値算出部１１６に与える。
【０１１３】
フレーム音声エネルギ正規化処理部１２６の最大値記憶部１３４（図５参照）は、状態を表す信号２００により発話区間が終了したことが示されると、それまで記憶していた値に代えてデフォルト最大値記憶部１３２の値を記憶する。この処理により、次の発話に対する音声エネルギの正規化処理の開始時には、最大パワーとしてデフォルトの値（又はオプションとして与えられた値）が再び使用される。
【０１１４】
特徴ベクトル出力部１２０は、状態判定部１１８の処理によって状態が確定したフレームのデータを入力バッファ１０６から読出し、そのフレームの特徴ベクトルを算出し、出力（１２２）する。特徴ベクトル出力部１２０はこのとき、当該フレームが発話開始フレーム又は発話終了フレームであれば、それを示すマークを当該特徴ベクトルに付して出力する。
【０１１５】
-４００ミリ秒から１秒まで-
入出力・アドレス管理部１１４からの初回フラグ１９６はオフとなる。４０番目のフレームの後、１００番目までのフレームについては、４０番目のフレームに対する処理とほぼ同様である。この間の処理では、フレームバッファ１１０には１０ミリ秒ごとに１フレーム分のデータが追加されていく。そして、その結果フレームバッファ１１０に格納されている全てのフレーム情報を用いて状態判定が実行される。
【０１１６】
また、図１０に示す動的しきい値算出部１１６においては、既に記憶部１７４に一つ前のフレームに対する処理で計算された環境雑音の推定値ｂ（ｔ－１）が記憶されている。環境雑音算出部１７０は、記憶部１７４に記憶された環境雑音の推定値ｂ（ｔ－１）と、フレームデータ１９２から得られるｔ番目のフレームのエネルギＥ（ｔ）とから、式１に従って環境雑音の推定値ｂ’（ｔ）を算出し選択部１７２に与える。
【０１１７】
選択部１７２は、初回フラグ１９６の値がオフなので、記憶部１７４の出力と、環境雑音算出部１７０の出力とのいずれかを状態を示す信号１９０の値に従って選択する。すなわち、信号１９０の表す状態が非発話状態であれば選択部１７２は環境雑音算出部１７０

の出力を選択し、それ以外であれば記憶部１７４の出力を選択する。選択部１７２は、選択した値を示す信号を、記憶部１７４及びしきい値算出部１７８に与える。

【０１１８】

他の点では、動的しきい値算出部１１６は、４０番目のフレームに対する処理と同様の処理を実行する。状態判定部１１８、特徴ベクトル出力部１２０、及びフレーム音声エネルギ正規化処理部１２６の動作も同様である。

【０１１９】

-１秒以後-

１０１番目のフレーム以降の処理も、４００ミリ秒から１秒までの処理とほぼ同様である。ただしこの処理では、フレームバッファ１１０に記憶されているフレーム情報に新たなフレーム情報を追加する際には、最も古いフレーム情報が削除される。すなわちフレームバッファ１１０はＦＩＦＯ形式でデータを格納する。その結果、フレームバッファ１１０には常に１００フレーム分のフレーム情報が維持される。ソート処理部１４０によるソート処理も同様である。ソート後フレームエネルギ記憶部１４２のうち、最も古いフレームのエネルギ値が削除され、新たなフレームのエネルギ値が、その大きさに従って決まる位置に書込まれる。

【０１２０】

初期環境雑音算出部１１２、動的しきい値算出部１１６、状態判定部１１８及び特徴ベクトル出力部１２０は、いずれもフレームバッファ１１０に記憶された１００フレーム分のデータに基づいて、背景雑音の推定、しきい値の算出、状態の判定、及び特徴ベクトルの作成を繰返し実行する。

【０１２１】

こうして、特徴ベクトル出力部１２０から出力されるフレームごとの特徴ベクトル１２２には、そのフレームが発話開始位置であれば発話開始マーカが、発話終了位置であれば発話終了マーカが、それぞれ付されている。このマーカにより、最初の音声データの発話区間（発話開始位置から発話終了位置まで）を検出する事ができる。

【０１２２】

また、特徴ベクトル１２２にはフレームごとの音声エネルギを正規化した値が含まれ、これを特徴量として音声認識で利用することができる。またこの音声エネルギは、発話全体にわたって調べることで算出された最大値ではなく、発話の最初からの最大値によって実時間に更新される最大値で擬似的に正規化されたものである。この様子を図１１に示す。

【０１２３】

図１１を参照して、この正規化処理により定められる音声エネルギの最大値の推移について説明する。図１１を参照して、従前の方法によれば、発話の終了まで完了した時点で発話の音声エネルギの最大値を調べ、その値によって音声エネルギを正規化する。図１１において、この音声エネルギの最大値は点線２１２とその後に続く太い実線２１８により表される。

【０１２４】

これに対し上記した実施の形態では、発話の開始時点では一定のデフォルト値（又はオプション値）２１４で、点線２１２で示される音声エネルギの最大値を近似する。さらに音声エネルギの値がこのデフォルト値より大きくなると（図１１における太い実線の曲線２１６の部分）、その値で音声エネルギの最大値の近似値を置換する。発話中で実際の音声エネルギの最大値位置に到達した後は、この近似値は実際の最大値と等しくなる（太い実線２１８の部分）。

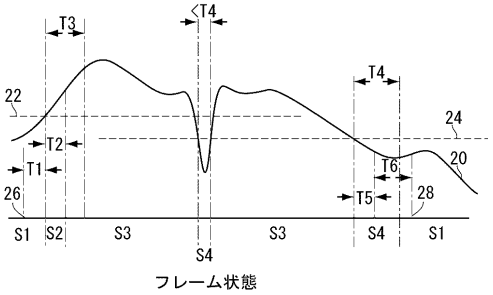
【０１２５】

この正規化処理によって、実時間で音声エネルギの正規化を行なうことができる。各発話の先頭部分ではデフォルトの値が最大値として使用されるため、多少の誤差は生じるが、デフォルトの値を適当な大きさに定めておくことにより、擬似的な正規化ではあっても十分な効果を得ることができる。

【０１２６】
-実施の形態の効果-
以上説明した本実施の形態の装置によれば、発話の開始及び終了のための発話開始しきい値及び発話終了しきい値が、実際の音声データを統計的に処理する事により、実際の音声データに従って動的に変化される。環境雑音の変化に追従して変化するしきい値を用いて発話区間の検出ができる。その結果、環境雑音の影響を最小限に抑えて、正しく発話区間を検出する事ができる。
【０１２７】
上記した実施の形態の装置では、しきい値を算出する際に用いられるフレームの最大エネルギ値として、実際の最大値の９０％のものを用いている。そのため、環境雑音の突発的な変化によるしきい値の大きな変化を抑制する事ができる。また、フレームバッファサイズだけの量のフレームに対する統計的処理によりしきい値を算出しているので、一部のフレームで突出したエネルギ値の変化があっても、しきい値にその変化が与える影響は比較的少なくて済む。その結果、安定してしきい値を算出できる。
【０１２８】
本実施の形態の装置ではさらに、フレームデータが４０個となった時点で状態の判定を開始している。統計処理にはある程度の数が必要なので、あまり少ない数のフレームデータを用いたしきい値算出では、状態判定結果の信頼性が低くなる。従って、最低で３００ミリ秒程度、望ましくは本実施の形態の装置のように４００ミリ秒程度の音声データに基づいてしきい値算出を開始する事がよい。また、処理対象のフレーム数が４０個となった時点で状態判定を開始するので、起動後、状態判定の開始までの遅延時間はほぼ４００ミリ秒程度となる。この程度の遅延の大きさは実用上で問題とならない程度である。あまり大きな遅延をとるようにすると、発話区間の開始の検出に失敗するおそれがある。また上記実施の形態では、遅延は４００ミリ秒であるが、しきい値判定には１０００ミリ秒分のデータが使用されるので、少ない遅延で信頼性の高いしきい値算出を行なう事ができる。
【０１２９】
［変形例］
上記した実施の形態では、フレームのエネルギ算出の際の窓関数としてハミング窓を用いた。しかし使用可能な窓関数はこれに限らない。ハニング窓、ブラックマン、カイザー、ブラックマン-ハリスなど種々の窓関数のうち、適切と思われるものを用いればよい。
【０１３０】
上記実施の形態では、フレームバッファサイズを１００、初期バッファサイズを４０とした。これらの値は一例であって、これ以外の組合せをとる事もできる。ただし、フレームバッファサイズをあまり大きくとると環境雑音の変化に追従してしきい値を変化させる事が困難になる。またフレームバッファサイズをあまり小さくとると、環境雑音のちょっとした変化に対応してしきい値が変化し、発話区間の検出が安定してできなくなる。また、初期バッファサイズをあまり大きくとると環境雑音の推定までの遅延が大きくなり、発話区間の先頭の検出に失敗する可能性が高くなる。また、当然の事ながら初期バッファサイズはフレームバッファサイズ以下でなければならない。従って、フレームバッファサイズとしては３００～２０００ミリ秒程度、初期バッファサイズとしては２００から５００ミリ秒程度がよい。特に、フレームバッファサイズが６００～１０００ミリ秒程度、初期バッファサイズとして３００から４５０ミリ秒程度が適している。
【０１３１】
また、上記した実施の形態では、音声エネルギの正規化のため、予め算出された固定的な値をデフォルト値として使用している。しかし本発明はそのような実施の形態には限定されない。例えば、このデフォルト値を発話の終了時に直前の発話での最大パワーによって更新することもできる。このとき、最大エネルギに対して所定の係数ａ（０＜ａ≦１、好ましくは０．７＜ａ＜０．９、さらに好ましくはａ＝０．８程度）を乗算しておくとよい。また、直前の発話だけでなく、過去の複数個の発話での最大エネルギの関数としてこのデフォルトの値を更新するようにしてもよい。

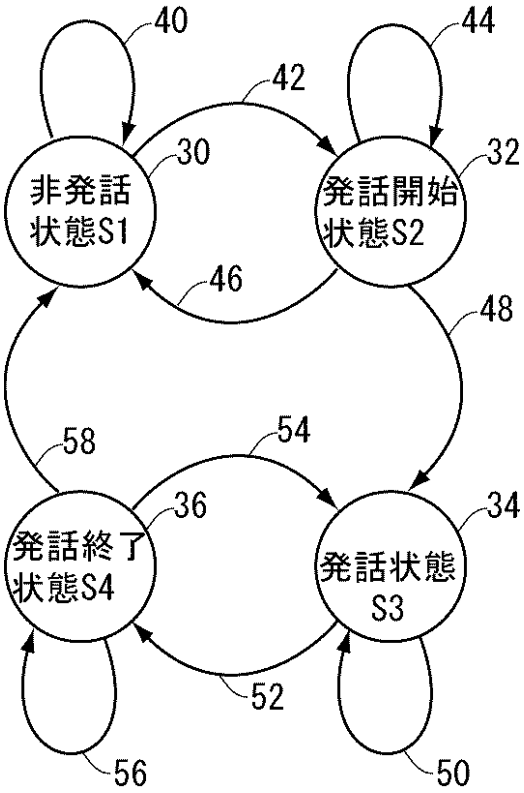
【０１３２】
また、上記した実施の形態では、フレーム内の各音声データの絶対値に窓関数の値を乗じた値の平均値の対数をとり、さらに係数２０を掛けることにより求めた対数音声エネルギを正規化したものを音声エネルギの特徴パラメータとしている。しかし本発明はそのような実施の形態には限定されず、例えば各音声データの二乗に窓関数の値を乗じた値の平均値の対数をとり、さらに係数１０を掛けることで対数音声エネルギを算出するようにした場合にも本発明は同様に適用できる。
【０１３３】
上記した実施の形態の装置は、ＤＳＰ（Ｄｉｇｉｔａｌ　Ｓｉｇｎａｌ　Ｐｒｏｃｅｓｓｏｒ）などのプロセッサと、そうしたプロセッサ上で実行されるプログラム（マイクロプログラムを含む。）とにより実現される事が想定される。上記した説明により、そうしたプログラムを作成する事は、当業者には容易であろう。
【０１３４】
今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内でのすべての変更を含む。
【図面の簡単な説明】
【０１３５】
【図１】本発明での発話区間判定の方式及びそのためのパラメータとを説明するための図である。
【図２】本発明での発話区間処理における状態遷移図である。
【図３】フレーム長及びフレームシフト量を説明するための図である。
【図４】本発明の一実施の形態に係る発話区間検出装置の機能的ブロック図である。
【図５】図４に示す装置の、音声エネルギ正規化処理部のブロック図である。
【図６】図４に示す装置の、初期環境雑音算出部の機能的ブロック図である。
【図７】フレームエネルギの変化の例を示す図である。
【図８】フレームエネルギを昇順にソートした結果を示す図である。
【図９】フレームエネルギのヒストグラムである。
【図１０】図４に示す装置の動的しきい値算出部の機能的ブロック図である。
【図１１】本発明の一実施の形態における音声エネルギ正規化処理を説明するための図である。
【符号の説明】
【０１３６】
２０　音声信号、２２　発話開始しきい値、２４　発話終了しきい値、３０　非発話状態（Ｓ１）、３２　発話開始状態（Ｓ２）、３４　発話状態（Ｓ３）、３６　発話終了状態（Ｓ４）、１００　発話区間検出装置、１０２　マイク、１０４　音声入力部、１０６　入力バッファ、１０８　フレーム情報算出部、１１０　フレームバッファ、１１２　初期環境雑音算出部、１１４　入出力・アドレス管理部、１１６　動的しきい値算出部、１１８　状態判定部、１２０　特徴ベクトル出力部、１２２　特徴ベクトル、１２４　フレーム出力信号、１２６　フレーム音声エネルギ正規化処理部、１４０　ソート処理部、１４２　ソート後フレームエネルギ記憶部、１４４　シーズ算出部、１４６、１７４　記憶部、１４８　第１の平均値算出部、１５０　フレーム分類部、１５２　第２の平均値算出部、１５４　判定部、１６０　フレームデータ、１７０　環境雑音算出部、１７２　選択部、１７６　最大エネルギ算出部、１７８　しきい値算出部

【図１】

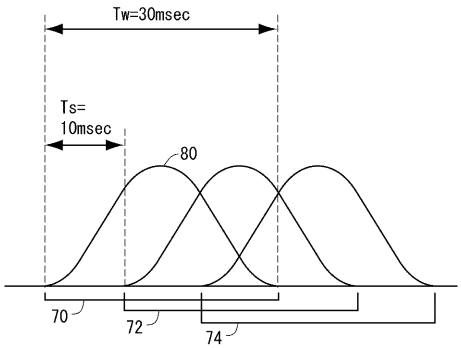


フレーム状態

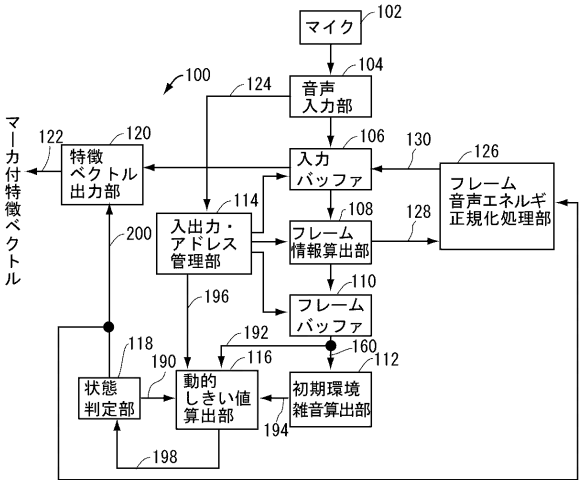
【図２】



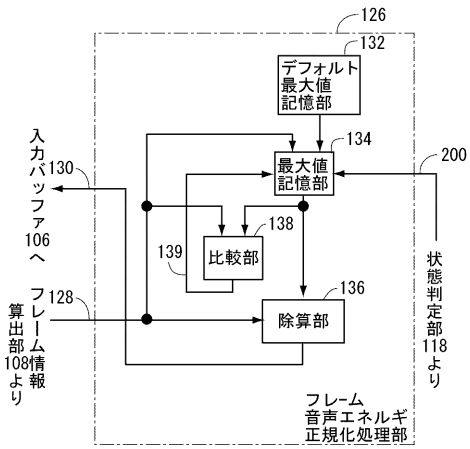
【図３】



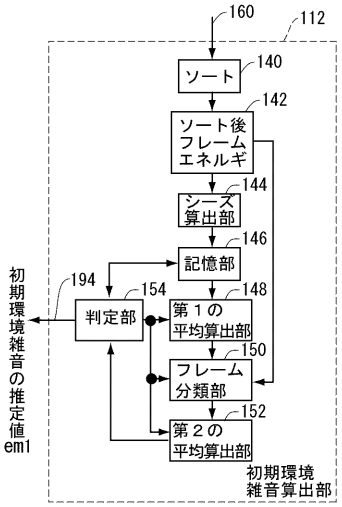
【図４】



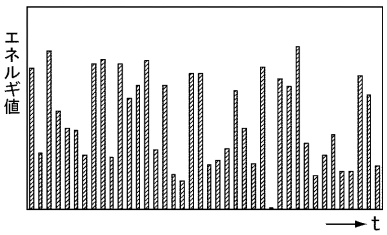
【図５】



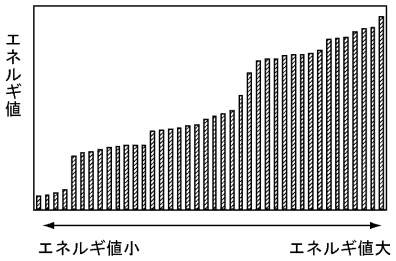
【図６】



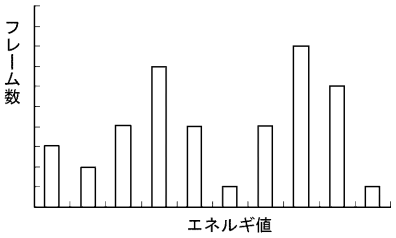
【図７】



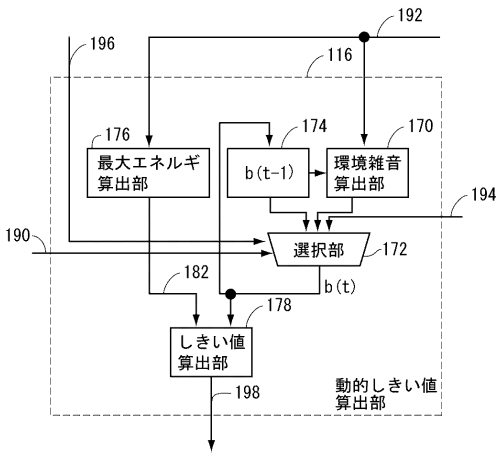
【図８】



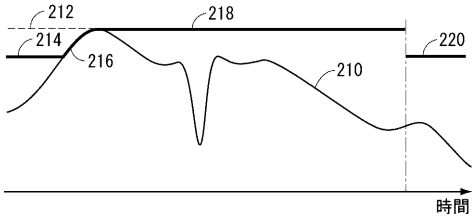
【図９】



【図１０】



【図１１】



フロントページの続き

(72)発明者　伊藤　玄
京都府相楽郡精華町光台二丁目２番地２　株式会社国際電気通信基礎技術研究所内

審査官　涌井　智則

(56)参考文献　特公平０８－０２３７５６（ＪＰ，Ｂ２）
特開２００２－２５８８８２（ＪＰ，Ａ）
特開平０８－１８７３６８（ＪＰ，Ａ）
特開平０８－０３２５２６（ＪＰ，Ａ）
特開平０８－３１４５００（ＪＰ，Ａ）
特開平１０－３０１５９３（ＪＰ，Ａ）
特開昭６１－２７３５９６（ＪＰ，Ａ）
特開昭５８－０７６８９９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１１／００－２１／０６
ＣｉＮｉｉ