

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号
特許第4497911号
(P4497911)

(45) 発行日 平成22年7月7日 (2010.7.7)

(24) 登録日 平成22年4月23日 (2010.4.23)

(51) Int.Cl.

F I

G 1 O L 11/02 (2006.01)

G 1 O L 15/04 (2006.01)

G 1 O L 11/02

G 1 O L 15/04 3 O O B

請求項の数 5 (全 18 頁)

(21) 出願番号	特願2003-418646 (P2003-418646)	(73) 特許権者	000001007
(22) 出願日	平成15年12月16日 (2003.12.16)		キヤノン株式会社
(65) 公開番号	特開2005-181458 (P2005-181458A)		東京都大田区下丸子3丁目30番2号
(43) 公開日	平成17年7月7日 (2005.7.7)	(74) 代理人	100076428
審査請求日	平成18年12月8日 (2006.12.8)		弁理士 大塚 康德
		(74) 代理人	100112508
			弁理士 高柳 司郎
		(74) 代理人	100115071
			弁理士 大塚 康弘
		(74) 代理人	100116894
			弁理士 木村 秀二
		(72) 発明者	ガーナー フィリップ
			東京都大田区下丸子3丁目30番2号 キ
			ヤノン株式会社内

最終頁に続く

(54) 【発明の名称】 信号検出装置および方法、ならびにプログラム

(57) 【特許請求の範囲】

【請求項 1】

入力したフレーム f における信号系列の特徴量として前記信号系列のスペクトルパワーベクトル s^2_f を抽出する第1の抽出手段と、

前記信号系列に含まれる雑音成分の特徴量として前記信号系列の雑音推定ベクトル μ_f を抽出する第2の抽出手段と、

前記スペクトルパワーベクトル s^2_f の k 番目の要素を s^2_k 、前記雑音推定ベクトル μ_f の k 番目の要素を μ_k 、ベクトル要素の個数を S 、予め設定した信号対雑音比を ω 、低域不完全ガンマ関数を γ とするとき、前記信号系列が音声として検出される確からしさを表す音声尤度と前記信号系列が雑音として検出される確からしさを表す雑音尤度との比を表すフレーム f における尤度比 L_f を、次式、

$$L_f = \prod_{k=1}^S \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s_k^2}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s_k^2}{\mu_k} + \omega \right) - 1 \right]$$

に従い計算する尤度比計算手段と、

前記尤度比計算手段により計算された尤度比 L_f に基づいて前記信号系列の状態を判断する判断手段と、

を有することを特徴とする信号検出装置。

【請求項 2】

入力したフレーム f における信号系列の特徴量として前記信号系列のスペクトルマグニチュードベクトル s_f を抽出する第 1 の抽出手段と、

前記信号系列に含まれる雑音成分の特徴量として前記信号系列の雑音推定ベクトル μ_f を抽出する第 2 の抽出手段と、

前記スペクトルマグニチュードベクトル s_f の k 番目の要素を s_k 、前記雑音推定ベクトル μ_f の k 番目の要素を μ_k 、ベクトル要素の個数を S 、予め設定した信号対雑音比を ω 、低域不完全ガンマ関数を γ とするとき、前記信号系列が音声として検出される確からしさを表す音声尤度と前記信号系列が雑音として検出される確からしさを表す雑音尤度との比を表すフレーム f における尤度比 L_f を、次式、

$$L_f = \prod_{k=1}^S \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s_k}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s_k}{\mu_k} + \omega \right) - 1 \right]$$

に従い計算する尤度比計算手段と、

前記尤度比計算手段により計算された尤度比 L_f に基づいて前記信号系列の状態を判断する判断手段と、

を有することを特徴とする信号検出装置。

【請求項 3】

信号検出装置によって実行される信号検出方法であって、

第 1 の抽出手段が、入力したフレーム f における信号系列の特徴量として前記信号系列のスペクトルパワーベクトル s^2_f を抽出する第 1 の抽出ステップと、

第 2 の抽出手段が、前記信号系列に含まれる雑音成分の特徴量として前記信号系列の雑音推定ベクトル μ_f を抽出する第 2 の抽出ステップと、

前記スペクトルパワーベクトル s^2_f の k 番目の要素を s^2_k 、前記雑音推定ベクトル μ_f の k 番目の要素を μ_k 、ベクトル要素の個数を S 、予め設定した信号対雑音比を ω 、低域不完全ガンマ関数を γ とするとき、尤度比計算手段が、前記信号系列が音声として検出される確からしさを表す音声尤度と前記信号系列が雑音として検出される確からしさを表す雑音尤度との比を表すフレーム f における尤度比 L_f を、次式、

$$L_f = \prod_{k=1}^S \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s^2_k}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s^2_k}{\mu_k} + \omega \right) - 1 \right]$$

に従い計算する尤度比計算ステップと、

判断手段が、前記尤度比計算ステップで計算された尤度比 L_f に基づいて前記信号系列の状態を判断する判断ステップと、

を有することを特徴とする信号検出方法。

【請求項 4】

信号検出装置によって実行される信号検出方法であって、

第 1 の抽出手段が、入力したフレーム f における信号系列の特徴量として前記信号系列のスペクトルマグニチュードベクトル s_f を抽出する第 1 の抽出ステップと、

第 2 の抽出手段が、前記信号系列に含まれる雑音成分の特徴量として前記信号系列の雑音推定ベクトル μ_f を抽出する第 2 の抽出ステップと、

前記スペクトルマグニチュードベクトル s_f の k 番目の要素を s_k 、前記雑音推定ベクトル μ_f の k 番目の要素を μ_k 、ベクトル要素の個数を S 、予め設定した信号対雑音比を ω 、低域不完全ガンマ関数を γ とするとき、尤度比計算手段が、前記信号系列が音声として検出される確からしさを表す音声尤度と前記信号系列が雑音として検出される確からしさを表す雑音尤度との比を表すフレーム f における尤度比 L_f を、次式、

10

20

30

40

50

$$L_f = \prod_{k=1}^s \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s_k}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s_k}{\mu_k} + \omega \right) - 1 \right]$$

に従い計算する尤度比計算ステップと、
 判断手段が、前記尤度比計算ステップで計算された尤度比 L_f に基づいて前記信号系列
 の状態を判断する判断ステップと、
 を有することを特徴とする信号検出方法。

【請求項 5】

コンピュータを、請求項 1 又は 2 に記載の信号検出装置が有する各手段として機能させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、音響信号等の信号を検出するための装置および方法に関する。

【背景技術】

【0002】

たとえば音声処理の分野では、音声区間を検出する技術が多くの場面で必要とされる。音声区間の検出は一般に、VAD (Voice Activity Detection) と呼ばれるが、speech activity detection と呼ばれる。

【0003】

音声区間検出 (VAD) が必要とされる典型的な場面としては、次の 2 つが挙げられる。

【0004】

第 1 は、音声通信システムである。図 1 に、音声通信システムにおける音声の送受信手続きの一例を示す。基本的には、送信側において入力された音声信号は前処理部 11 で所定の前処理が行われ、符号化器 13 で符号化された後、その符号化音声は通信回線 15 を介して受信側に送られる。そして、受信側では、復号化器 16 でその符号化音声を復号化して音声を出力することになる。このように音声信号は通信回線 15 を介して他の場所に送られるのであるが、その通信回線 15 には何らかの制限がある。例えば、利用料金が低い、伝送容量が小さい、といった事情からくる制限である。このような制限に対処するために、VAD 12 が用いられる。この VAD 12 を用いることにより、利用者が発声していない間は通信を行わないように指示することができ、その結果、利用料金を削減したり、あるいは、他の利用者がその間、その通信回線を利用することができるようにすることが可能となる。また、必ずしもそうである必要はないのだが、VAD 12 および符号化器 13 の前段に設けられる前処理部 11 を、図示の如く共通のものとすることができ。また、VAD 12 を設けることにより、符号化器 13 自体は、音声中のポーズ区間と長時間の無音区間との区別を行う必要がなくなる。

【0005】

第 2 は、音声認識システムである。図 2 に、VAD を含む音声認識システムの処理例を示す。同図において、VAD 22 は、音声認識 (ASR) 部 24 における音声認識処理で背景雑音を音声として認識してしまうことを防止する役割を果たす。すなわち、VAD 22 の機能は、雑音を単語として変換してしまう誤りを起こさないようにすることである。加えて、VAD 22 を設けることにより、一般的に多くの計算機資源を利用する音声認識システムにおいて、システム全体の処理能力をよりうまく管理することが可能になる。例えば、音声により携帯機器などの制御を可能にするのである。具体的には、VAD は、利用者が発声していない区間と、利用者がコマンドを指示している区間と区別する。この結果、装置は、音声認識を行っていないときは他の機能に専念し、他方、利用者が発声している時は音声認識 (ASR) に専念するように制御することが可能になる。ここにおいて

10

20

30

40

50

も、VAD22および音声認識部24の前段に設けられる前処理部21を、図示の如く共通のものとして行うことができる。更に、本例においては、音声端点検出部23は、VADの信号を利用し、発声の開始～終了の区間と、単語間のポーズ区間との区別を行う。これは、音声認識部24では、単語内のギャップをなくした単語全体を音声として受け取る必要があるからである。

【0006】

さて、音声区間を精度よく検出するためには、背景雑音の存在を考慮する必要がある。背景雑音は時々刻々と変動するので、その変動を追跡して、それを音声区間の検出尺度に反映させなければならない。しかし、これを高精度に実現することは容易ではない。従来より、このような観点からさまざまな提案がなされている。以下、簡単に従来例について説明する。

10

【0007】

従来の音声区間検出手法の典型例は、エネルギーや零交差数といった時間領域における分析結果を用いるものである。しかし、時間領域処理から得られるパラメータは雑音に弱いという欠点がある。これに対し、特許文献1は、時間領域ではなく周波数領域での分析に基づき音声区間を精度よく検出する手法を開示している。

【0008】

特許文献2および非特許文献1は、雑音抑圧における音声検出技術に関する。これらの文献には、信号対雑音比(SNR)が音声区間の検出に有用な尺度となることが記載されている。

20

【0009】

特許文献3および4は、雑音推定値を連続的に更新する音声区間検出の技術を開示する。雑音推定部は第2の補助的なVADにより制御される。

【0010】

特許文献5は、スペクトルバンド毎のサブディビジョン(サブ判断)を用いた音声区間検出の技術を開示している。

【0011】

また、非特許文献2は、尤度比に基づいた音声区間検出の技術を開示している。ここでは、音声と雑音パラメータのみを用いている。

【0012】

30

【特許文献1】米国特許第5,692,104号明細書

【特許文献2】米国特許第5,432,859号明細書

【特許文献3】米国特許第5,749,067号明細書

【特許文献4】米国特許第6,061,647号明細書

【特許文献5】米国特許第5,963,901号明細書

【非特許文献1】Jin Yang, "Frequency domain noise suppression approaches in mobile telephone systems", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume II, p. 363-366, 1993.

【非特許文献2】Jongseo Sohn and Wonyong Sung, "A Voice Activity Detector employing soft decision based noise spectrum adaptation.", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, p. 365-368, May 1998.

40

【発明の開示】

【発明が解決しようとする課題】

【0013】

(問題点1)

上記したような先行技術には、典型的な音声信号と背景雑音の間の信号対雑音比を指定する方法がない。このために、ある種の雑音が誤って音声として通知されることになる。この点、本発明の一つの特徴は、信号対雑音比をあらかじめ設定する手段を提供することにより、MAP法(事後確立最大化法)による定式化をしたことにある。このことにより

50

、ある種の雑音において、音声検出の感度を低くすることが可能となる。

【 0 0 1 4 】

(問題点 2)

上記した典型的な先行技術は、音声信号のスペクトル形状に関して何ら仮定をしていない。このために、単に音の大きい雑音も誤って音声として通知されることになる。この点、本発明のもう一つの特徴は、ある種の雑音（周波数形状が平坦であるもの）と音声（周波数形状が平坦ではないもの）との区別を可能にする差分スペクトル尺度を用いている点にある。

【 0 0 1 5 】

(問題点 3)

上記の先行技術では、背景雑音が現れる区間のみを用いて雑音追尾の更新が行われる。このような区間においては、低周波の変動のみを精度よく追尾できるように、最小追尾率を用いる必要がある。先行技術では、明示的な最小値が与えられていないため、特に前述のMAP法を用いた場合には、高周波の変動をも追尾してしまう可能性がある。本発明のもう一つの特徴は、最小追尾率を有する雑音追尾方法にある。

【 0 0 1 6 】

このように、本発明の目的は、さまざまな種類の背景雑音に対して頑健な信号検出の技術を提供することである。

【課題を解決するための手段】

【 0 0 1 7 】

上記した課題は本発明の信号検出装置および方法、ならびにプログラムによって解決される。本発明の一側面に係る信号検出装置は、入力したフレーム f における信号系列の特徴量として前記信号系列のスペクトルパワーベクトル s^2_f を抽出する第1の抽出手段と、前記信号系列に含まれる雑音成分の特徴量として前記信号系列の雑音推定ベクトル μ_f を抽出する第2の抽出手段と、前記スペクトルパワーベクトル s^2_f の k 番目の要素を s^2_k 、前記雑音推定ベクトル μ_f の k 番目の要素を μ_k 、ベクトル要素の個数を S 、予め設定した信号対雑音比を ω 、低域不完全ガンマ関数を γ とするとき、前記信号系列が音声として検出される確からしさを表す音声尤度と前記信号系列が雑音として検出される確からしさを表す雑音尤度との比を表すフレーム f における尤度比 L_f を、次式、

$$L_f = \prod_{k=1}^S \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s_k^2}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s_k^2}{\mu_k} + \omega \right) - 1 \right]$$

に従い計算する尤度比計算手段と、前記尤度比計算手段により計算された尤度比 L_f に基づいて前記信号系列の状態を判断する判断手段とを有することを特徴とする。

【 0 0 1 8 】

本発明の別の側面に係る信号検出装置は、入力したフレーム f における信号系列の特徴量として前記信号系列のスペクトルマグニチュードベクトル s_f を抽出する第1の抽出手段と、前記信号系列に含まれる雑音成分の特徴量として前記信号系列の雑音推定ベクトル μ_f を抽出する第2の抽出手段と、前記スペクトルマグニチュードベクトル s_f の k 番目の要素を s_k 、前記雑音推定ベクトル μ_f の k 番目の要素を μ_k 、ベクトル要素の個数を S 、予め設定した信号対雑音比を ω 、低域不完全ガンマ関数を γ とするとき、前記信号系列が音声として検出される確からしさを表す音声尤度と前記信号系列が雑音として検出される確からしさを表す雑音尤度との比を表すフレーム f における尤度比 L_f を、次式、

$$L_f = \prod_{k=1}^S \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s_k}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s_k}{\mu_k} + \omega \right) - 1 \right]$$

10

20

30

40

50

に従い計算する尤度比計算手段と、前記尤度比計算手段により計算された尤度比 L_f に基づいて前記信号系列の状態を判断する判断手段とを有することを特徴とする。

【 0 0 1 9 】

本発明のさらに別の側面に係る信号検出方法は、信号検出装置によって実行される信号検出方法であって、第 1 の抽出手段が、入力したフレーム f における信号系列の特徴量として前記信号系列のスペクトルパワーベクトル s^2_f を抽出する第 1 の抽出ステップと、第 2 の抽出手段が、前記信号系列に含まれる雑音成分の特徴量として前記信号系列の雑音推定ベクトル μ_f を抽出する第 2 の抽出ステップと、前記スペクトルパワーベクトル s^2_f の k 番目の要素を s^2_k 、前記雑音推定ベクトル μ_f の k 番目の要素を μ_k 、ベクトル要素の個数を S 、予め設定した信号対雑音比を ω 、低域不完全ガンマ関数を γ とするとき、尤度比計算手段が、前記信号系列が音声として検出される確からしさを表す音声尤度と前記信号系列が雑音として検出される確からしさを表す雑音尤度との比を表すフレーム f における尤度比 L_f を、次式、

$$L_f = \prod_{k=1}^S \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s_k^2}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s_k^2}{\mu_k} + \omega \right) - 1 \right]$$

に従い計算する尤度比計算ステップと、判断手段が、前記尤度比計算ステップで計算された尤度比 L_f に基づいて前記信号系列の状態を判断する判断ステップとを有することを特徴とする。

【 0 0 2 0 】

本発明のさらに別の側面に係る信号検出方法は、信号検出装置によって実行される信号検出方法であって、第 1 の抽出手段が、入力したフレーム f における信号系列の特徴量として前記信号系列のスペクトルマグニチュードベクトル s_f を抽出する第 1 の抽出ステップと、第 2 の抽出手段が、前記信号系列に含まれる雑音成分の特徴量として前記信号系列の雑音推定ベクトル μ_f を抽出する第 2 の抽出ステップと、前記スペクトルマグニチュードベクトル s_f の k 番目の要素を s_k 、前記雑音推定ベクトル μ_f の k 番目の要素を μ_k 、ベクトル要素の個数を S 、予め設定した信号対雑音比を ω 、低域不完全ガンマ関数を γ とするとき、尤度比計算手段が、前記信号系列が音声として検出される確からしさを表す音声尤度と前記信号系列が雑音として検出される確からしさを表す雑音尤度との比を表すフレーム f における尤度比 L_f を、次式、

$$L_f = \prod_{k=1}^S \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s_k}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s_k}{\mu_k} + \omega \right) - 1 \right]$$

に従い計算する尤度比計算ステップと、判断手段が、前記尤度比計算ステップで計算された尤度比 L_f に基づいて前記信号系列の状態を判断する判断ステップとを有することを特徴とする。

【発明の効果】

【 0 0 2 3 】

本発明によれば、さまざまな種類の背景雑音に対して頑健な信号検出を行うことができる。

【発明を実施するための最良の形態】

【 0 0 2 4 】

以下、図面を参照して本発明の好適な実施形態について詳細に説明する。

【 0 0 2 5 】

まず、本明細書で使用する用語の意味について注釈しておく。「無音」の用語は「非音声」と同じ意味で用いる。よって、この「無音」は、“silence”のみならず“noise (雑

10

20

30

40

50

音) " を含みうることに留意されたい。

【 0 0 2 6 】

また、以下ではいくつかの数式を示して本発明の信号検出処理を説明するが、一般に信号のベクトル表現についてはスカラ値と区別するためにボールド体で表示するところ、以下の説明ではそのように区別する表示態様とはなっていないことに留意されたい。したがって、信号がベクトルを意味するときは「ベクトル」の文字を表示するが、当業者にとってスカラ値との区別が容易であるときは省略する場合もある。

【 0 0 2 7 】

以下では一実施形態として、図 2 に示したような音声認識システムにおいて本発明の音声区間検出を適用することを考える。もっとも、本発明はこの他に、例えば図 1 に示したような音声通信システムにも適用可能であることはいうまでもない。

【 0 0 2 8 】

さて、本発明は、一般的なコンピュータシステムで実現することができるものである。もちろん、専用のハードウェアロジックにより実現することもできるが、ここではコンピュータシステムで実現するものとする。

【 0 0 2 9 】

図 3 は、実施形態におけるコンピュータシステムの構成を示す図である。図示のように、本コンピュータシステムは、システム全体の制御をつかさどる CPU 1、ブートプログラム等を記憶している ROM 2、主記憶装置として機能する RAM 3をはじめ、以下の構成を備える。

【 0 0 3 0 】

HDD 4 はハードディスク装置であって、ここに OS のほか、音声認識プログラム、ならびにその音声認識プログラムに呼び出されて動作する VAD プログラムが格納されている。もっとも、このコンピュータシステムが別の機器に組み込まれるなどの場合、これらのプログラムは HDD ではなく ROM 2 に格納しておくことも可能である。VRAM 5 は表示しようとするイメージデータを展開するメモリであり、ここにイメージデータ等を展開することで CRT 6 に表示させることができる。7 および 8 は、入力デバイスとしてのキーボードおよびマウスである。また、9 は音声を入力するためのマイクロフォン、10 はマイクロフォン 9 からの信号をディジタル信号に変換する A/D コンバータである。

【 0 0 3 1 】

図 4 は、実施形態における信号検出処理を実現する機能ブロック図である。以下では、この図を参照しながら VAD の各処理を説明する。

【 0 0 3 2 】

(特徴抽出)

マイクロフォン 9 より入力された音響信号 (音声および背景雑音が含まれうる) は、A/D コンバータ 10 によって例えば 11.025 kHz でサンプリングされ、フレーム処理モジュール 32 により、256 サンプルからなるフレームに分割される。このフレームは例えば 110 サンプル毎に生成される。これにより、これらのフレームは、隣接するフレームとオーバーラップすることになる。この構成によれば、100 フレームでおよそ 1 秒となる。各フレームは、ハミング窓処理の後、ハートレー変換処理が行われ、同じ周波数に対応するハートレー変換の 2 つの出力結果は 2 乗和がとられ、ペリオドグラム (periodogram) が形成される。ペリオドグラムは一般に、パワースペクトル密度 (PSD : Power Spectral Density) として知られている。256 サンプルのフレームでは、PSD は 128 個の値 (bin) からなる。

【 0 0 3 3 】

PSD は、メル化処理モジュール 34 によって、メル帯域の値 (bin) を用いてそのサイズが縮小される (例えば 32 点)。メル化処理モジュール 34 は、等間隔に線形変換された周波数特性を聴覚特性尺度 (メル尺度) の空間に変換する。メル化された各点の値は、PSD が窓をオーバーラップさせて形成されているため、大きな相関ができる。本実施形態においては、このように生成された 32 個のメル尺度信号を音声区間検出の特徴量

10

20

30

40

50

として使用する。音声認識の分野では、メルによる表現は一般的に用いられている。メルスペクトルを対数化処理を行った後、コサイン変換を施すことによりメルケプストラムに変換する、という処理を行うのがその典型的な使われ方である。しかし、この音声区間検出処理では、直接メル化した値を用いる。このように本実施形態では、メル尺度信号を特徴量として用いるが、その他の尺度による特徴量を用いてもよい。

【 0 0 3 4 】

(雑音追跡)

メル尺度信号は、雑音追跡モジュール 3 6 および音声区間検出尺度計算モジュール 3 8 に入力される。雑音追跡モジュール 3 6 は、入力されたメル尺度信号において、緩やかに変動する背景雑音を追跡する。この追跡には、過去に音声区間検出尺度計算モジュール 3 8 で計算された音声区間検出尺度の平均が利用される。

10

【 0 0 3 5 】

音声区間検出尺度については後ほど詳細に説明するが、本発明は音声区間検出尺度として尤度比を用いる。フレーム f における尤度比 L_f は例えば次式で定義される。

【 0 0 3 6 】

【 数 1 】

$$L_f = \frac{p(s_f^2 | \text{speech})}{p(s_f^2 | \text{noise})}$$

20

【 0 0 3 7 】

上式において、 s_f^2 はフレーム f において観測された 3 2 次元の特徴 $\{s_1^2, s_2^2, \dots, s_S^2\}$ で構成されるベクトルである。分子はフレーム f が音声として検出される確からしさを示す尤度、分母はフレーム f が雑音として検出される確からしさを示す尤度を表している。なお、本明細書で記述される全ての式において、直接スペクトルマグニチュードのベクトル $s_f = \{s_1, s_2, \dots, s_S\}$ をスペクトル尺度として用いることも可能ではあるが、ここでは、特に断らない限り、P S D から計算した特徴ベクトルである 2 乗として表す。

【 0 0 3 8 】

30

雑音追跡モジュール 3 6 による雑音の追跡は、典型的には、次の単一極フィルタ型の式で表される。

【 0 0 3 9 】

【 数 2 】

$$\mu_f = (1 - \rho_\mu) s_f^2 + \rho_\mu \mu_{f-1}$$

【 0 0 4 0 】

ただし、 μ_f は、フレーム f における 3 2 次元の雑音推定ベクトルであり、 ρ_μ は雑音更新フィルタ成分の特徴量の極であって最低更新値の意味を持つ。

40

【 0 0 4 1 】

本実施形態の雑音追跡は、次式で定義される。

【 0 0 4 2 】

【 数 3 】

$$\mu_f = \frac{1 - \rho_\mu}{1 + L_f} s_f^2 + \frac{\rho_\mu + L_f}{1 + L_f} \mu_{f-1}$$

【 0 0 4 3 】

50

なお、スペクトルパワー s^2 のかわりにスペクトルマグニチュード s を用いる場合には、尤度比は次式で表される。

【 0 0 4 4 】

【 数 4 】

$$\mu_f = \frac{1-\rho_\mu}{1+L_f} s_f + \frac{\rho_\mu + L_f}{1+L_f} \mu_{f-1}$$

【 0 0 4 5 】

10

上記したとおり、 L_f はフレーム f における尤度比である。ここで、 L_f を 0 に近づけると、この雑音追跡は上記した数 2 の単一極フィルタ型の式となることが分かる。この場合、極は最小追跡率として機能する。一方、 L_f の値を大きく（1 よりも更に大きく）すると、次式に近づくことが理解されよう。

【 0 0 4 6 】

【 数 5 】

$$\mu_f = \mu_{f-1}$$

【 0 0 4 7 】

20

このように、本実施形態における雑音成分の抽出には、過去のフレームにおける雑音成分の特徴量と、その過去のフレームにおける尤度比とに基づいて雑音を追跡する処理が含まれる。

【 0 0 4 8 】

（音声区間検出尺度の計算）

本発明は音声区間検出尺度として数 1 に示した尤度比を用いることは先述したとおりである。以下では、尤度比の算出する 3 つの手法を示す。

【 0 0 4 9 】

（1）最尤法（ML：Maximum Likelihood method）

最尤法（ML：Maximum Likelihood method）は、例えば次式により表される。これは、非特許文献 2 として掲げた、Jongseo Sohn らによる “A Voice Activity Detector employing soft decision based noise spectrum adaptation.”（Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, p. 365-368, May 1998）にも開示されている。

30

【 0 0 5 0 】

【 数 6 】

$$p(s_f^2 | \text{speech}) = \prod_{k=1}^s \frac{1}{\Pi(\lambda_k + \mu_k)} \exp\left(-\frac{s_k^2}{\lambda_k + \mu_k}\right)$$

40

【 0 0 5 1 】

【 数 7 】

$$p(s_f^2 | \text{noise}) = \prod_{k=1}^s \frac{1}{\Pi \mu_k} \exp\left(-\frac{s_k^2}{\mu_k}\right)$$

【 0 0 5 2 】

よって、

【数 8】

$$L_f = \prod_{k=1}^S \frac{\mu_k}{\lambda_k + \mu_k} \exp\left(-\frac{\lambda_k}{\lambda_k + \mu_k} \cdot \frac{s_k^2}{\mu_k}\right)$$

【0053】

ここで、 k は特徴ベクトルのインデックスであり、 S は特徴ベクトルの特徴（ベクトル要素）の個数（本実施形態では 32）である。また、 μ_k は、フレーム f における雑音推定ベクトル μ_f の k 番目の要素を表し、 λ_k はベクトル λ_f （後述する）の k 番目の要素を表し、また、 s_k^2 は、ベクトル s_f^2 の k 番目の要素を表している。なお、この算出手順を図示すると、図 5 のようになる。

10

【0054】

この最尤法を用いた音声区間検出尺度の計算では、ベクトル λ_f の k 番目の要素の値 λ_k を求める必要がある。この λ_f は、フレーム f における音声の分散（スペクトルパワー s^2 のかわりにスペクトルマグニチュード s を用いる場合は、標準偏差）の推定値で、同図中、音声分散推定 50 により求められる。本実施形態では、次式で表されるスペクトルサブトラクション法により λ_f を求める。

【0055】

【数 9】

$$\lambda_f = \max(s_f^2 - \alpha \mu_f, \beta s_f^2)$$

20

【0056】

ただし、 α および β は適当な固定値であり、本実施形態ではそれぞれ、例えば、 $\alpha = 1$ 、 $\beta = 0.3$ である。

【0057】

（2）最大事後確率法（MAP：Maximum a-posteriori method）

ところで、上記（1）の最尤法を用いた計算方法では、ベクトル λ_f の計算が必要とされる。この計算には、スペクトルサブトラクション法や、判定結果により導出する推定（"decision directed" estimation）のような処理を必要とする。そこで、最尤法を用いるかわりに最大事後確率法（MAP：Maximum a-posteriori method）を用いることもできる。MAP による方法では、ベクトル λ_f の計算を避けることができるという利点がある。この計算手順を図示すると、図 6 のようになる。この場合では、61 で示される雑音尤度計算は、上記の最尤法の場合（図 5 の 52 で示される雑音尤度計算）と同様であるが、音声尤度計算は上記の最尤法と異なり、次式により計算する。

30

【0058】

【数 10】

$$p(s_f^2 | \text{speech}) = \prod_{k=1}^S \frac{1}{\Gamma(0, \omega) \mu_k \left(\frac{s_k^2}{\mu_k} + \omega \right)} \left(1 - \exp\left(-\frac{s_k^2}{\mu_k} - \omega\right) \right)$$

40

【0059】

ただし、 ω は、実験的に決まる事前の信号対雑音比（SNR）であり、 $\gamma(*, *)$ は、低域不完全（lower incomplete）ガンマ関数である。この結果、尤度比は次式で表されることになる。

【0060】

【数 1 1】

$$L_f = \prod_{k=1}^s \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s_k^2}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s_k^2}{\mu_k} + \omega \right) - 1 \right]$$

【0 0 6 1】

本実施形態では、 $\omega = 100$ とした。なお、スペクトルパワー s^2 のかわりにスペクトルマグニチュード s を用いる場合には、尤度比は次式で表される。

【0 0 6 2】

10

【数 1 2】

$$L_f = \prod_{k=1}^s \frac{1}{e^{\omega} \gamma(0, \omega) \left(\frac{s_k}{\mu_k} + \omega \right)} \left[\exp \left(\frac{s_k}{\mu_k} + \omega \right) - 1 \right]$$

【0 0 6 3】

(3) 差分特徴最尤法 (Differential feature ML method)

ところで、上記した2つの計算法は、特徴量を直接用いる方法に基づいている。これと異なる方法として、(時間ドメインではなく)特徴量ドメインにおいて、音声区間検出計算を行う前にローパスフィルタを施す方法がある。特徴量がスペクトルの場合には、下記の2つの利点が生まれる。

20

【0 0 6 4】

(a) オフセット (DC) が除去される。つまり、広帯域にわたる雑音成分が除かれることを意味する。これは、手を叩いた音や堅いもの同士が衝突した音などの、短時間の広帯域雑音 (インパルス) に対して実質的に有効である。これらの音は、雑音追跡の速度より非常に速い速度である。

(b) また、メル化する処理の導入により生じる、相関を除くことができる。

典型的なローパスフィルタは、次の再帰式により表される。

【0 0 6 5】

30

$$x'_k = x_k - x_{k+1}$$

【0 0 6 6】

ここで、スペクトルの場合には、 $x_k = s^2_k$ である。

【0 0 6 7】

本実施形態では例えば、次のように打ち切りを行っている。通常のフィルタでは、ベクトル x' を生成することになる。

【0 0 6 8】

$$x'_1 = x_1 - x_2,$$

$$x'_2 = x_2 - x_3,$$

...

$$x'_{s-1} = x_{s-1} - x_s$$

40

【0 0 6 9】

その結果、ベクトルは $s - 1$ の要素からなることになる。本実施形態で用いた打ち切りフィルタでは、一つおきの値を用い、その結果、ベクトルは $s / 2$ の要素からなっている。

【0 0 7 0】

$$x'_1 = x_1 - x_2,$$

$$x'_2 = x_3 - x_4,$$

...

$$x'_{s/2} = x_{s-1} - x_s$$

50

【 0 0 7 1 】

図 7 は、この計算手順を示す図である。この場合において、7 2 で計算される音声尤度と 7 3 で計算される雑音尤度との比（尤度比）は、両スペクトルの要素のどちらが大きいのかに依存する。すなわち、 $s_{2k-1}^2 > s_{2k}^2$ の場合、音声尤度 $P(s_f^2 | \text{speech})$ および雑音尤度 $P(s_f^2 | \text{noise})$ はそれぞれ次式で表される。

【 0 0 7 2 】

【 数 1 3 】

$$p(s_f^2 | \text{speech}) = \prod_{k=1}^{S/2} \frac{1}{\lambda_{2k} + \mu_{2k} + \lambda_{2k-1} + \mu_{2k-1}} \exp\left(-\frac{s_{2k-1}^2 - s_{2k}^2}{\lambda_{2k-1} + \mu_{2k-1}}\right), \quad 10$$

【 0 0 7 3 】

【 数 1 4 】

$$p(s_f^2 | \text{noise}) = \prod_{k=1}^{S/2} \frac{1}{\mu_{2k} + \mu_{2k-1}} \exp\left(-\frac{s_{2k-1}^2 - s_{2k}^2}{\mu_{2k-1}}\right).$$

【 0 0 7 4 】

しかし、 $s_{2k}^2 > s_{2k-1}^2$ の場合には、音声尤度 $P(s_f^2 | \text{speech})$ および雑音尤度 $P(s_f^2 | \text{noise})$ はそれぞれ次式で表される。

20

【 0 0 7 5 】

【 数 1 5 】

$$p(s_f^2 | \text{speech}) = \prod_{k=1}^{S/2} \frac{1}{\lambda_{2k} + \mu_{2k} + \lambda_{2k-1} + \mu_{2k-1}} \exp\left(-\frac{s_{2k}^2 - s_{2k-1}^2}{\lambda_{2k} + \mu_{2k}}\right),$$

【 0 0 7 6 】

【 数 1 6 】

$$p(s_f^2 | \text{noise}) = \prod_{k=1}^{S/2} \frac{1}{\mu_{2k} + \mu_{2k-1}} \exp\left(-\frac{s_{2k}^2 - s_{2k-1}^2}{\mu_{2k}}\right). \quad 30$$

【 0 0 7 7 】

したがって、尤度比は次のように表される。

【 0 0 7 8 】

【 数 1 7 】

$$L_f = \prod_{k=1}^{S/2} \frac{\mu_{2k} + \mu_{2k-1}}{\lambda_{2k} + \mu_{2k} + \lambda_{2k-1} + \mu_{2k-1}} \exp\left(\frac{\lambda_{2k-1}}{\lambda_{2k-1} + \mu_{2k-1}} \cdot \frac{s_{2k-1}^2 - s_{2k}^2}{\mu_{2k-1}}\right), \text{ if } s_{2k-1}^2 > s_{2k}^2,$$

40

$$L_f = \prod_{k=1}^{S/2} \frac{\mu_{2k} + \mu_{2k-1}}{\lambda_{2k} + \mu_{2k} + \lambda_{2k-1} + \mu_{2k-1}} \exp\left(\frac{\lambda_{2k}}{\lambda_{2k} + \mu_{2k}} \cdot \frac{s_{2k}^2 - s_{2k-1}^2}{\mu_{2k}}\right), \text{ if } s_{2k-1}^2 < s_{2k}^2.$$

【 0 0 7 9 】

なお、スペクトルパワー s^2 のかわりにスペクトルマグニチュード s を用いる場合には、尤度比は次式で表される。

【 0 0 8 0 】

【数 18】

$$L_f = \prod_{k=1}^{S/2} \frac{\mu_{2k} + \mu_{2k-1}}{\lambda_{2k} + \mu_{2k} + \lambda_{2k-1} + \mu_{2k-1}} \exp\left(\frac{\lambda_{2k-1}}{\lambda_{2k-1} + \mu_{2k-1}} \cdot \frac{S_{2k-1} - S_{2k}}{\mu_{2k-1}}\right), \text{ if } S_{2k-1} > S_{2k},$$

$$L_f = \prod_{k=1}^{S/2} \frac{\mu_{2k} + \mu_{2k-1}}{\lambda_{2k} + \mu_{2k} + \lambda_{2k-1} + \mu_{2k-1}} \exp\left(\frac{\lambda_{2k}}{\lambda_{2k} + \mu_{2k}} \cdot \frac{S_{2k} - S_{2k-1}}{\mu_{2k}}\right), \text{ if } S_{2k-1} < S_{2k}.$$

10

【0081】

(類似度計算)

上述の全ての L_f の計算は、次式のとおりに定式化される。

【0082】

【数 19】

$$L_f = \prod_{k=1}^S L_k$$

【0083】

20

しかし、一般的には、 L_f は様々な相関があるので、これらを掛け合わせると非常に大きな値になる。そこで、次式のように、 L_k を $1 / (\kappa S)$ 乗することにより、これを抑える。

【0084】

【数 20】

$$L_f = \prod_{k=1}^S L_k^{\frac{1}{\kappa S}}$$

【0085】

30

これを対数尤度で表現すると、次式のようになる。

【0086】

【数 21】

$$\log L_f = \sum_{k=1}^S \frac{1}{\kappa S} \log L_k$$

【0087】

この場合では、 $\kappa S = 1$ のとき、各々の要素の尤度の幾何学的な平均を求めていることに相当する。本実施形態においては、対数の形式を用いており、 κS は場合に応じて最適化を行う。ここでは、概ね 0.5 ~ 2 の間の値をとる。

40

【0088】

(信号検出アルゴリズムの詳細)

図 8 は、本実施形態における信号検出処理を示すフローチャートである。このフローチャートに対応するプログラムは HDD 4 に記憶されている VAD プログラムに含まれ、RAM 3 にロードされた後、CPU 1 によって実行されるものである。

【0089】

この処理は S1 を初期ステップとして開始し、ステップ S2 でフレームインデックス (frame index) を 0 に設定し、ステップ S3 で、現在のフレームインデックスに対応するフレームを読み込む。

【0090】

50

次に、ステップ S 4 で、フレームインデックスが 0 (初期フレーム) か否かを判断する。ここで、フレームインデックスが 0 のときはステップ S 10 に進み、音声区間検出尺度としての尤度比を 0 に設定し、その後、ステップ S 11 で、その初期フレームの値を雑音推定値に設定してステップ S 12 に進む。

【0091】

一方、ステップ S 4 で、フレームインデックスが 0 でないと判断された場合には、ステップ S 5 に進み、上述した方法で音声の分散の推定を行う。次に、ステップ S 6 で、フレームインデックスが所定値 (例えば 10) 未満であるか否かを判断する。ここで、フレームインデックスが 10 未満の場合にはステップ S 8 に進み、尤度比 = 0 を維持する。他方、フレームインデックスが所定値以上の場合にはステップ S 7 に進み、音声区間検出尺度としての尤度比を計算する。次に、ステップ S 9 で、ステップ S 7 または S 8 で決定された尤度比を用いて雑音推定の更新を行う。このような処理により、雑音推定を信頼性のある値と仮定することが可能になる。

【0092】

次に、ステップ S 12 で、尤度比を所定のしきい値と比較して、2 値 (音声または雑音を示す値) を生成する。MAP を用いる場合には、しきい値は例えば 0 であり、他の場合には例えば 2.5 である。

【0093】

次に、ステップ S 13 で、ステップ S 12 での尤度比のしきい値との比較結果に基づいて、後述するような音声の端点検出が行われる。

【0094】

そして、ステップ S 14 でフレームインデックスをインクリメントしてステップ S 3 に戻り、次のフレームについて処理を繰り返す。

【0095】

以上説明した実施形態によれば、音声区間検出尺度として尤度比を用いることで、さまざまな種類の背景雑音に対して耐性のある音声区間検出を行うことができる。

【0096】

とりわけ、尤度比の算出に最大事後確率法 (MAP) を導入することで、推定される SNR に対する VAD の調整が容易になる。これにより、レベルの低い音声に高レベルの雑音が混入しても音声を精度よく検出することができる。

【0097】

また、差分特徴最尤法を用いれば、周波数全域にわたりパワーがほぼ等しい雑音 (例えば、足音のような振動音や風や息のような聞き分けにくい音を含む。) に対してもロバストな性能が得られる。

【0098】

(他の実施形態)

上述した実施形態は、音声認識などの音声に関わる内容で説明したが、本発明は例えば、機械などの音や動物の音などの音声以外の音響信号にも適用することが可能である。また、ソナーや動物の音などのように、人には聞くことができないような領域の音響情報にも利用できるものである。さらに、レーダーや無線信号などの電磁波などの信号などにも利用できる。

【0099】

以上、本発明の実施形態を詳述したが、本発明は、複数の機器から構成されるシステムに適用してもよいし、また、一つの機器からなる装置に適用してもよい。

【0100】

なお、本発明は、前述した実施形態の機能を実現するソフトウェアのプログラムを、システムあるいは装置に直接あるいは遠隔から供給し、そのシステムあるいは装置のコンピュータがその供給されたプログラムコードを読み出して実行することによっても達成される。その場合、プログラムの機能を有していれば、その形態はプログラムである必要はない。

10

20

30

40

50

【 0 1 0 1 】

従って、本発明の機能処理をコンピュータで実現するために、そのコンピュータにインストールされるプログラムコード自体も本発明を実現するものである。つまり、本発明の特許請求の範囲には、本発明の機能処理を実現するためのコンピュータプログラム自体も含まれる。

【 0 1 0 2 】

その場合、プログラムの機能を有していれば、オブジェクトコード、インタプリタにより実行されるプログラム、OSに供給するスクリプトデータ等、プログラムの形態を問わない。

【 0 1 0 3 】

プログラムを供給するための記録媒体としては、例えば、フレキシブルディスク、ハードディスク、光ディスク、光磁気ディスク、MO、CD-ROM、CD-R、CD-RW、磁気テープ、不揮発性のメモリカード、ROM、DVD(DVD-ROM、DVD-R)などがある。

【 0 1 0 4 】

その他、プログラムの供給方法としては、クライアントコンピュータのブラウザを用いてインターネットのホームページに接続し、そのホームページから本発明のコンピュータプログラムそのもの、もしくは圧縮され自動インストール機能を含むファイルをハードディスク等の記録媒体にダウンロードすることによっても供給できる。また、本発明のプログラムを構成するプログラムコードを複数のファイルに分割し、それぞれのファイルを異なるホームページからダウンロードすることによっても実現可能である。つまり、本発明の機能処理をコンピュータで実現するためのプログラムファイルを複数のユーザに対してダウンロードさせるWWWサーバも、本発明のクレームに含まれるものである。

【 0 1 0 5 】

また、本発明のプログラムを暗号化してCD-ROM等の記憶媒体に格納してユーザに配布し、所定の条件をクリアしたユーザに対し、インターネットを介してホームページから暗号化を解く鍵情報をダウンロードさせ、その鍵情報を使用することにより暗号化されたプログラムを実行してコンピュータにインストールさせて実現することも可能である。

【 0 1 0 6 】

また、コンピュータが、読み出したプログラムを実行することによって、前述した実施形態の機能が実現される他、そのプログラムの指示に基づき、コンピュータ上で稼動しているOSなどが、実際の処理の一部または全部を行い、その処理によっても前述した実施形態の機能が実現され得る。

【 0 1 0 7 】

さらに、記録媒体から読み出されたプログラムが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれた後、そのプログラムの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるCPUなどが実際の処理の一部または全部を行い、その処理によっても前述した実施形態の機能が実現される。

【図面の簡単な説明】

【 0 1 0 8 】

【図1】音声通信システムにおける音声の送受信手続きの一例を示す図である。

【図2】VADを含む音声認識システムの処理例を示す図である。

【図3】実施形態におけるコンピュータシステムの構成を示す図である。

【図4】実施形態における信号検出処理を実現する機能ブロック図である。

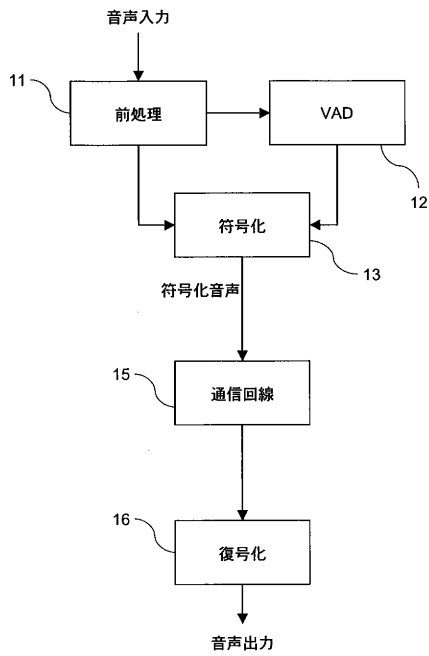
【図5】最尤法を用いた音声区間検出尺度の計算手順を示す図である。

【図6】最大事後確率法を用いた音声区間検出尺度の計算手順を示す図である。

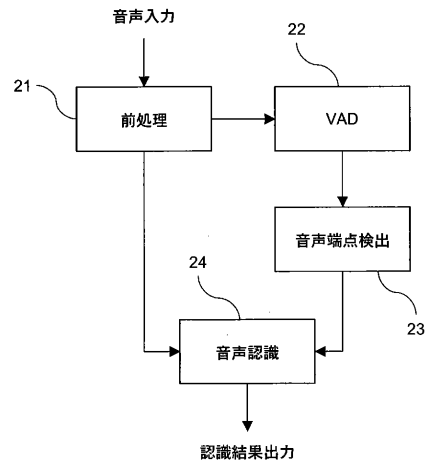
【図7】差分特徴最尤法を用いた音声区間検出尺度の計算手順を示す図である。

【図8】実施形態における信号検出処理を示すフローチャートである。

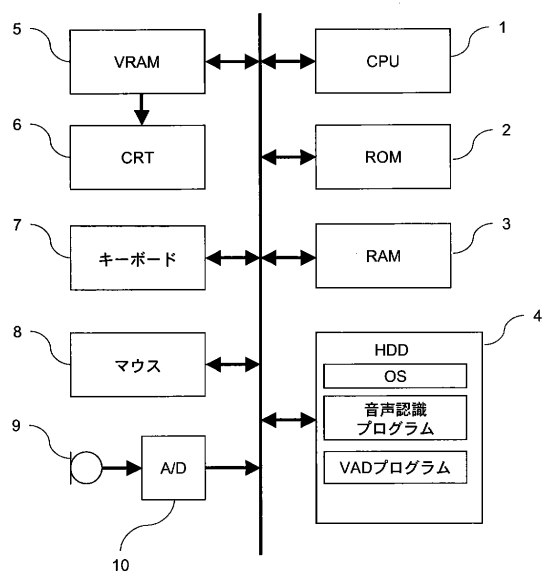
【図 1】



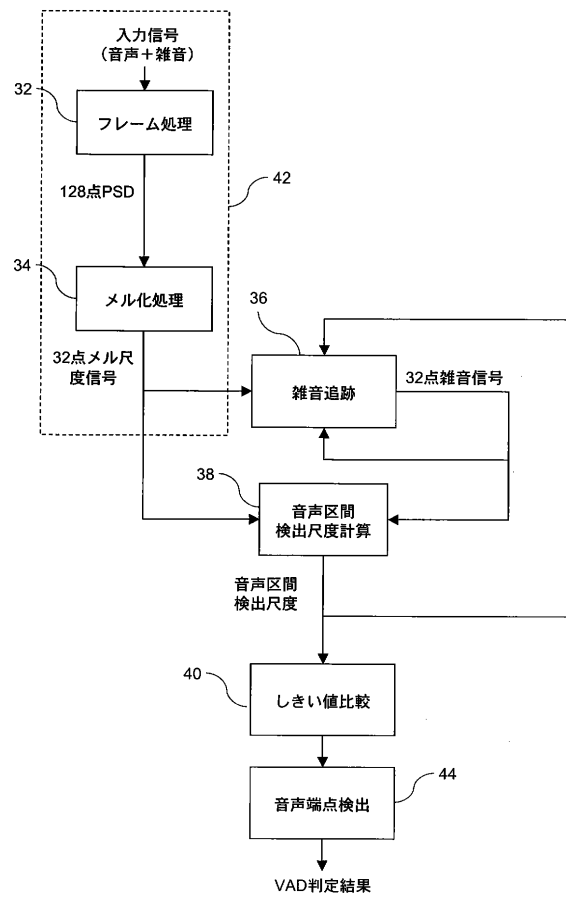
【図 2】



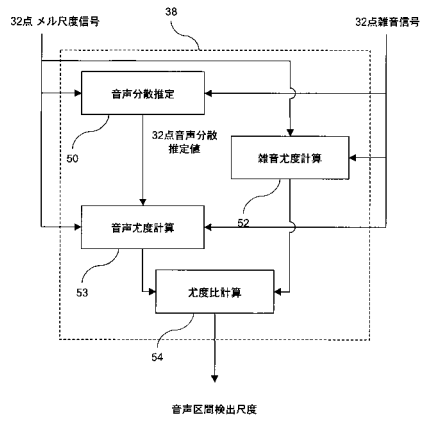
【図 3】



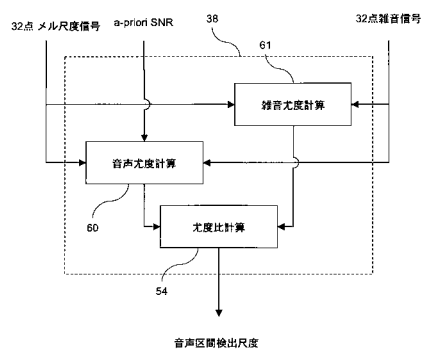
【図 4】



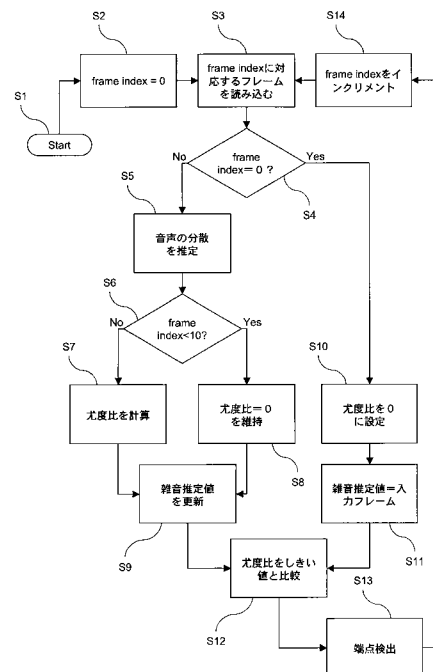
【図 5】



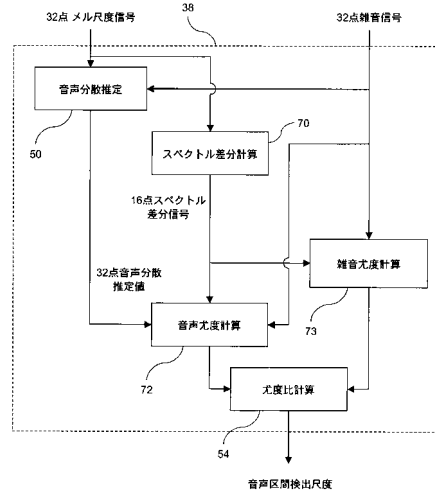
【図 6】



【図 8】



【図 7】



フロントページの続き

- (72)発明者 深田 俊明
東京都大田区下丸子3丁目30番2号 キヤノン株式会社内
- (72)発明者 小森 康弘
東京都大田区下丸子3丁目30番2号 キヤノン株式会社内

審査官 菊池 智紀

- (56)参考文献 特開平04-058297(JP,A)
特開平05-273992(JP,A)
特開2000-330597(JP,A)
特開2002-055691(JP,A)

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