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

特許第6233625号
(P6233625)

(45)発行日　平成29年11月22日(2017.11.22)　(24)登録日　平成29年11月2日(2017.11.2)

(51)Int.Cl.		F I		
G10L 25/54	(2013.01)	G10L	25/54	
G10L 15/10	(2006.01)	G10L	15/10	500Z
G06F 17/30	(2006.01)	G06F	17/30	350C
		G06F	17/30	170E

請求項の数 7　(全 20 頁)

(21)出願番号　特願2013-37542(P2013-37542)
(22)出願日　平成25年2月27日(2013.2.27)
(65)公開番号　特開2014-115605(P2014-115605A)
(43)公開日　平成26年6月26日(2014.6.26)
審査請求日　平成27年2月9日(2015.2.9)
(31)優先権主張番号　特願2012-251809(P2012-251809)
(32)優先日　平成24年11月16日(2012.11.16)
(33)優先権主張国　日本国(JP)

前置審査

(73)特許権者　000002185
ソニー株式会社
東京都港区港南１丁目７番１号
(74)代理人　100121131
弁理士　西川　孝
(74)代理人　100082131
弁理士　稲本　義雄
(72)発明者　澁谷　崇
東京都港区港南１丁目７番１号　ソニー株式会社内
(72)発明者　安部　素嗣
東京都港区港南１丁目７番１号　ソニー株式会社内

最終頁に続く

(54)【発明の名称】音声処理装置および方法、並びにプログラム

(57)【特許請求の範囲】
【請求項１】
同定対象となるコンテンツの入力信号のスペクトログラムを二次関数で近似した近似結果、および前記スペクトログラムのピーク周辺のスペクトルを二次関数で近似した近似結果に基づいて、距離関数を用いて各時間周波数領域における信号を平滑化フィルタによりフィルタリングすることで、正弦波らしさの時間平均量に基づく第１の音響特徴量を算出するとともに、前記各時間周波数領域における信号を一次微分フィルタによりフィルタリングすることで、正弦波らしさの時間変化量に基づく、前記第１の音響特徴量とは異なる第２の音響特徴量を算出する入力信号処理部と、
予め用意したコンテンツの参照信号に基づいて、前記第１の音響特徴量と前記第２の音響特徴量とを算出する参照信号処理部と、
前記入力信号の前記第１の音響特徴量および前記第２の音響特徴量と、前記参照信号の前記第１の音響特徴量および前記第２の音響特徴量とに基づいて、前記入力信号と前記参照信号の類似度を計算するマッチング処理部と
を備える音声処理装置。
【請求項２】
前記マッチング処理部は、前記入力信号と前記参照信号の前記第１の音響特徴量に基づいて、各時間周波数領域におけるコンテンツの信号らしさを示すマスクパターンを生成し、前記マスクパターン、前記第１の音響特徴量、および前記第２の音響特徴量に基づいて前記類似度を計算する

請求項１に記載の音声処理装置。
【請求項３】
前記マッチング処理部は、前記入力信号の前記第１の音響特徴量と、前記参照信号の前記第１の音響特徴量との類似度をさらに算出し、前記マスクパターン、前記第１の音響特徴量の前記類似度、および前記第２の音響特徴量に基づいて、前記入力信号と前記参照信号の前記類似度を計算する
請求項２に記載の音声処理装置。
【請求項４】
前記マッチング処理部は、前記第１の音響特徴量の前記類似度に対する前記入力信号の寄与率よりも、前記第１の音響特徴量の前記類似度に対する前記参照信号の寄与率をより大きくして、前記第１の音響特徴量の前記類似度を算出する
請求項３に記載の音声処理装置。
【請求項５】
前記第２の音響特徴量は、前記入力信号または前記参照信号のスペクトログラムに基づいて算出され、時間軸および周波数軸において前記第１の音響特徴量と同じ粒度を有する
請求項１乃至請求項４の何れか一項に記載の音声処理装置。
【請求項６】
同定対象となるコンテンツの入力信号のスペクトログラムを二次関数で近似した近似結果、および前記スペクトログラムのピーク周辺のスペクトルを二次関数で近似した近似結果に基づいて、距離関数を用いて各時間周波数領域における信号を平滑化フィルタによりフィルタリングすることで、正弦波らしさの時間平均量に基づく第１の音響特徴量を算出するとともに、前記各時間周波数領域における信号を一次微分フィルタによりフィルタリングすることで、正弦波らしさの時間変化量に基づく、前記第１の音響特徴量とは異なる第２の音響特徴量を算出し、
予め用意したコンテンツの参照信号に基づいて、前記第１の音響特徴量と前記第２の音響特徴量とを算出し、
前記入力信号の前記第１の音響特徴量および前記第２の音響特徴量と、前記参照信号の前記第１の音響特徴量および前記第２の音響特徴量とに基づいて、前記入力信号と前記参照信号の類似度を計算する
ステップを含む音声処理方法。
【請求項７】
同定対象となるコンテンツの入力信号のスペクトログラムを二次関数で近似した近似結果、および前記スペクトログラムのピーク周辺のスペクトルを二次関数で近似した近似結果に基づいて、距離関数を用いて各時間周波数領域における信号を平滑化フィルタによりフィルタリングすることで、正弦波らしさの時間平均量に基づく第１の音響特徴量を算出するとともに、前記各時間周波数領域における信号を一次微分フィルタによりフィルタリングすることで、正弦波らしさの時間変化量に基づく、前記第１の音響特徴量とは異なる第２の音響特徴量を算出し、
予め用意したコンテンツの参照信号に基づいて、前記第１の音響特徴量と前記第２の音響特徴量とを算出し、
前記入力信号の前記第１の音響特徴量および前記第２の音響特徴量と、前記参照信号の前記第１の音響特徴量および前記第２の音響特徴量とに基づいて、前記入力信号と前記参照信号の類似度を計算する
ステップを含む処理をコンピュータに実行させるプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本技術は音声処理装置および方法、並びにプログラムに関し、特に、より高精度に任意のコンテンツの同定を行なうことができるようにした音声処理装置および方法、並びにプログラムに関する。

【背景技術】
【０００２】
例えば、コンテンツを構成する音声の信号を参照信号として、任意の機器において参照信号に基づいて再生された音声を収音して得られた入力信号と、参照信号とに基づいて一致検索を行なえば、コンテンツを同定することができる。この場合、元の音源から出力された音声は、残響や雑音が混入されて収音されるので、入力信号に基づく音声は、コンテンツの音声に残響や雑音が重畳されたものとなる。
【０００３】
このようなコンテンツ同定技術として、例えばＣＤ（Compact Disc）等のクリーンな音楽の信号を参照信号として、音楽以外の音が混入してしまった入力信号から、その背景音楽を同定する楽曲同定技術がある。
【０００４】
楽曲同定技術では、クリーンな音楽の参照信号から抽出された音響特徴量と、入力信号から抽出された音響特徴量とのマッチング処理により、楽曲の同定が行なわれる。ここで、入力信号には雑音が混入していることが前提となっており、入力信号から得られる音響特徴量は雑音の影響を受けてしまう。
【０００５】
そこで、例えばマッチング処理においては、マスクパターンが用いられる。このマスクパターンは音響特徴量を構成する要素のうち、信頼のおける要素を表現する情報である。マスクパターンが用いられるマッチング処理では、マスクパターンに基づいて、多次元の音響特徴量を構成する各要素が、信頼のおける要素と信頼のおけない要素に分けられて、信頼のおける要素のみが用いられてマッチングが行なわれる。
【０００６】
このようにマスクパターンを利用する楽曲同定技術として、例えば時間周波数成分を持つ特徴行列に対して、所定の時間周波数領域をマスクする複数のマスクパターンを予め用意しておき、楽曲同定を行うものが提案されている（例えば、特許文献１参照）。
【０００７】
この技術では、入力信号の特徴行列と、データベース内の楽曲の特徴行列、つまり参照信号の特徴行列に対して、予め用意した全てのマスクパターンを用いて算出した類似度のうちの最大値を入力信号と楽曲の類似度として楽曲同定が行なわれる。なお、この楽曲同定では、入力信号によらない固定の複数マスクパターンを記憶しておき、それらのマスクパターンが用いられてマッチング処理が行なわれる。
【先行技術文献】
【特許文献】
【０００８】
【特許文献１】特開２００９－２７６７７６号公報
【発明の概要】
【発明が解決しようとする課題】
【０００９】
しかしながら、上述した技術では、コンテンツの同定は音楽の一致検索に特化されており、一般の任意のコンテンツ、例えば放送番組等のコンテンツについてコンテンツを同定することはできなかった。例えば、放送番組のコンテンツでは、音楽の流れていないシーンの音声信号を入力信号として検索したい場合もあるが、そのような場合には上述した技術によりコンテンツの同定を行なうことができない。
【００１０】
また、上述した技術では音声の残響の影響が考慮されていないため、高精度なコンテンツ同定を実現することができない場合があった。すなわち、実使用環境では入力信号は残響の影響を受け、この残響は検索に悪影響を及ぼす。そのため、残響の強い環境では、コンテンツの一致検索の精度が低下してしまう。
【００１１】

さらに、特許文献１に記載の技術では、固定のマスクパターンが使用されている。ところが、入力信号に乗っている混入雑音に関しては、どの時間にどのような特性の雑音が乗るかは予測不能であるから、入力信号に最適なマスクパターンを予め用意するのは困難である。したがって、予め用意したマスクパターンでは高精度にコンテンツを同定することができないことがある。

【００１２】

本技術は、このような状況に鑑みてなされたものであり、より高精度に任意のコンテンツの同定を行なうことができるようにするものである。

【課題を解決するための手段】

【００１３】

本技術の一側面の音声処理装置は、同定対象となるコンテンツの入力信号のスペクトログラムを二次関数で近似した近似結果、および前記スペクトログラムのピーク周辺のスペクトルを二次関数で近似した近似結果に基づいて、距離関数を用いて各時間周波数領域における信号を平滑化フィルタによりフィルタリングすることで、正弦波らしさの時間平均量に基づく第１の音響特徴量を算出するとともに、前記各時間周波数領域における信号を一次微分フィルタによりフィルタリングすることで、正弦波らしさの時間変化量に基づく、前記第１の音響特徴量とは異なる第２の音響特徴量を算出する入力信号処理部と、予め用意したコンテンツの参照信号に基づいて、前記第１の音響特徴量と前記第２の音響特徴量とを算出する参照信号処理部と、前記入力信号の前記第１の音響特徴量および前記第２の音響特徴量と、前記参照信号の前記第１の音響特徴量および前記第２の音響特徴量とに基づいて、前記入力信号と前記参照信号の類似度を計算するマッチング処理部とを備える。

【００１４】

前記マッチング処理部には、前記入力信号と前記参照信号の前記第１の音響特徴量に基づいて、各時間周波数領域におけるコンテンツの信号らしさを示すマスクパターンを生成させ、前記マスクパターン、前記第１の音響特徴量、および前記第２の音響特徴量に基づいて前記類似度を計算させることができる。

【００１５】

前記マッチング処理部には、前記入力信号の前記第１の音響特徴量と、前記参照信号の前記第１の音響特徴量との類似度をさらに算出させ、前記マスクパターン、前記第１の音響特徴量の前記類似度、および前記第２の音響特徴量に基づいて、前記入力信号と前記参照信号の前記類似度を計算させることができる。

【００１６】

前記マッチング処理部には、前記第１の音響特徴量の前記類似度に対する前記入力信号の寄与率よりも、前記第１の音響特徴量の前記類似度に対する前記参照信号の寄与率をより大きくして、前記第１の音響特徴量の前記類似度を算出させることができる。

【００１７】

前記第２の音響特徴量を、前記入力信号または前記参照信号のスペクトログラムに基づいて算出し、時間軸および周波数軸において前記第１の音響特徴量と同じ粒度を有するようにすることができる。

【００１８】

本技術の一側面の音声処理方法またはプログラムは、同定対象となるコンテンツの入力信号のスペクトログラムを二次関数で近似した近似結果、および前記スペクトログラムのピーク周辺のスペクトルを二次関数で近似した近似結果に基づいて、距離関数を用いて各時間周波数領域における信号を平滑化フィルタによりフィルタリングすることで、正弦波らしさの時間平均量に基づく第１の音響特徴量を算出するとともに、前記各時間周波数領域における信号を一次微分フィルタによりフィルタリングすることで、正弦波らしさの時間変化量に基づく、前記第１の音響特徴量とは異なる第２の音響特徴量を算出し、予め用意したコンテンツの参照信号に基づいて、前記第１の音響特徴量と前記第２の音響特徴量とを算出し、前記入力信号の前記第１の音響特徴量および前記第２の音響特徴量と、前記

参照信号の前記第１の音響特徴量および前記第２の音響特徴量とに基づいて、前記入力信号と前記参照信号の類似度を計算するステップを含む。

【００１９】

本技術の一側面においては、同定対象となるコンテンツの入力信号のスペクトログラムを二次関数で近似した近似結果、および前記スペクトログラムのピーク周辺のスペクトルを二次関数で近似した近似結果に基づいて、距離関数が用いられて各時間周波数領域における信号を平滑化フィルタによりフィルタリングすることで、正弦波らしさの時間平均量に基づく第１の音響特徴量が算出されるとともに、前記各時間周波数領域における信号を一次微分フィルタによりフィルタリングすることで、正弦波らしさの時間変化量に基づく、前記第１の音響特徴量とは異なる第２の音響特徴量が算出され、予め用意したコンテンツの参照信号に基づいて、前記第１の音響特徴量と前記第２の音響特徴量とが算出され、前記入力信号の前記第１の音響特徴量および前記第２の音響特徴量と、前記参照信号の前記第１の音響特徴量および前記第２の音響特徴量とに基づいて、前記入力信号と前記参照信号の類似度が計算される。

【発明の効果】

【００２０】

本技術の一側面によれば、より高精度に任意のコンテンツの同定を行なうことができる。

【図面の簡単な説明】

【００２１】

【図１】マスクパターンについて説明する図である。

【図２】音声処理装置の構成例を示す図である。

【図３】入力信号処理部の構成例を示す図である。

【図４】参照信号処理部の構成例を示す図である。

【図５】マッチング処理部の構成例を示す図である。

【図６】音響特徴量について説明する図である。

【図７】一致検索処理を説明するフローチャートである。

【図８】音響特徴量ＩＡ１の抽出処理を説明するフローチャートである。

【図９】音響特徴量ＩＡ２の抽出処理を説明するフローチャートである。

【図１０】コンピュータの構成例を示す図である。

【発明を実施するための形態】

【００２２】

以下、図面を参照して、本技術を適用した実施の形態について説明する。

【００２３】

〈第１の実施の形態〉

〈本技術の特徴〉

本技術は、多機能携帯型電話機やタブレット型端末装置等の携帯端末装置の録音機能を使用して、ユーザが別の機器で視聴しているテレビ番組やラジオ番組、ストリーミング配信コンテンツ等の任意のコンテンツを同定するためのものである。

【００２４】

処理対象となる音声をテレビジョン受像機やラジオ、パーソナルコンピュータ等の機器のスピーカから出力させ、その音声を携帯端末装置で録音させると、録音により得られる音声には、機器のスピーカから出力された音声が携帯端末装置までの空間を通ることによって、その空間での音声の残響も含まれてしまう。また、録音により得られる音声には、機器のスピーカから出力された音声以外の音声が混入する（以下、これを混入雑音と呼ぶこととする）。

【００２５】

本技術では、これらの残響や混入雑音に頑健なコンテンツの一致検索を行うことが目的となる。より一般的には、所定の音源が空間を通過することによって残響や混入雑音が重畳された音源と、元の音源（ドライソース）との一致検索を行うことが本技術の目的とな

る。
【００２６】
まず、本技術の特徴について説明する。本技術は、以下の５つの特徴を有している。
【００２７】
特徴（１）
マスクパターンの生成には入力信号と参照信号について算出された、切り出された各々の時間周波数領域内での正弦波らしさの指標が用いられる
特徴（２）
正弦波らしさの指標は、微小時間内でのスペクトル形状の安定性によって定量化される
特徴（３）
上記の正弦波らしさは残響に頑健な指標とされる
特徴（４）
マスクパターンは入力信号の情報だけでなく、参照信号の情報も用いて生成される
特徴（５）
入力信号と参照信号との類似度算出の際に、それらを等価に扱うのではなく入力信号よりも参照信号に重きをおいた類似度計算が行なわれる
【００２８】
例えば、本技術では図１に示す入力信号のスペクトログラムと参照信号のスペクトログラムが得られる。なお、図１において、縦軸および横軸は、それぞれ周波数および時間を示している。
【００２９】
図１では、図中、右側には参照信号のスペクトログラムが示されており、図中、左側には入力信号のスペクトログラムが示されている。
【００３０】
入力信号のスペクトログラム、つまり時間周波数領域において、実線で表される成分は参照信号にも含まれている音声成分を示しており、点線で表される成分は参照信号には含まれていない、混入雑音の成分を示している。
【００３１】
本技術では、マスクパターンを生成することで、図中、斜線部分の領域である、信頼のおける時間周波数領域が特定され、その領域のみが用いられて入力信号と参照信号のマッチング処理が行なわれる。
【００３２】
このような本技術によれば、以下の効果を得ることができる。
【００３３】
効果（１）
音楽の流れているシーンだけでなく、音楽の流れていないシーンを使って、そのコンテンツの同定を行うことができる
効果（２）
残響のある空間においても、視聴番組等のコンテンツの同定が可能である
効果（３）
入力信号に元の参照信号に含まれていた音声以外の音声（混入雑音）が含まれていても、視聴番組等のコンテンツの同定が可能である
【００３４】
〈音声処理装置の構成例〉
次に、本技術を適用した具体的な実施の形態について説明する。
【００３５】
図２は、本技術を適用した音声処理装置の一実施の形態の構成例を示す図である。
【００３６】
音声処理装置１１は、入力信号処理部２１、参照信号処理部２２、およびマッチング処理部２３から構成される。

【００３７】
音声処理装置１１には、予め用意されたコンテンツの音声の参照信号と、所定の機器で再生された参照信号に基づく音声を、他の機器で録音（収音）することで得られた、同定対象のコンテンツの音声の入力信号とが入力される。例えば、入力信号は音声処理装置１１による録音で得られた音声信号であってもよい。
【００３８】
なお、参照信号として、例えば複数のコンテンツの音声信号が入力される。また、音声処理装置１１には、参照信号のコンテンツ属性データも入力される。ここで、コンテンツ属性データとは、コンテンツ名（番組名）や放送日時、出演者など、コンテンツに関するデータである。
【００３９】
入力信号処理部２１は、供給された入力信号を解析して２種類の音響特徴量ＩＡ１および音響特徴量ＩＡ２を生成し、マッチング処理部２３に供給する。
【００４０】
参照信号処理部２２は、供給された、コンテンツの元音源である参照信号を解析して、音響特徴量ＩＡ１および音響特徴量ＩＡ２に対応する２種類の音響特徴量ＲＡ１および音響特徴量ＲＡ２を生成し、マッチング処理部２３に供給する。
【００４１】
ここで、音響特徴量ＩＡ１と音響特徴量ＲＡ１は同じ特徴量（同じ種別の特徴量）であり、音響特徴量ＩＡ２と音響特徴量ＲＡ２も同じ特徴量である。以下、音響特徴量ＩＡ１と音響特徴量ＲＡ１を特に区別する必要のない場合、単に音響特徴量Ａ１とも称し、音響特徴量ＩＡ２と音響特徴量ＲＡ２を特に区別する必要のない場合、単に音響特徴量Ａ２とも称することとする。
【００４２】
マッチング処理部２３は、入力信号処理部２１からの音響特徴量ＩＡ１および音響特徴量ＩＡ２と、参照信号処理部２２からの音響特徴量ＲＡ１および音響特徴量ＲＡ２とに基づいて、入力信号と参照信号とのマッチング処理を行い、コンテンツを同定する。また、マッチング処理部２３は、供給されたコンテンツ属性データのうち、マッチング処理により同定されたコンテンツのコンテンツ属性データと、マッチング処理の結果とを出力する。
【００４３】
〈入力信号処理部の構成例〉
また、図２の入力信号処理部２１は、より詳細には図３に示すように構成される。図３に示す入力信号処理部２１は、入力信号切り出し部５１、時間周波数変換部５２、音響特徴量抽出部５３、および音響特徴量抽出部５４から構成される。
【００４４】
入力信号切り出し部５１は、供給された入力信号の所定時間長の区間を切り出して、時間周波数変換部５２に供給する。時間周波数変換部５２は、入力信号切り出し部５１から供給された入力信号に対する時間周波数変換を行なって、入力信号を対数振幅スペクトログラムに変換し、音響特徴量抽出部５３および音響特徴量抽出部５４に供給する。
【００４５】
音響特徴量抽出部５３は、時間周波数変換部５２から供給された対数振幅スペクトログラムに基づいて音響特徴量ＩＡ１を算出し、マッチング処理部２３に供給する。音響特徴量抽出部５４は、時間周波数変換部５２から供給された対数振幅スペクトログラムに基づいて音響特徴量ＩＡ２を算出し、マッチング処理部２３に供給する。
【００４６】
ここで、音響特徴量ＩＡ１と音響特徴量ＩＡ２について説明する。
【００４７】
例えば、音響特徴量ＩＡ１と音響特徴量ＩＡ２は全て行列で表現され、２つの軸は時間成分と周波数成分である。それぞれの行列は、以下のような特徴を有している。

【００４８】
すなわち、音響特徴量ＩＡ１は、各々の時間周波数領域において、入力信号のその領域内での正弦波らしさを表現する特徴行列である。
【００４９】
また、音響特徴量ＩＡ２は入力信号と参照信号のマッチングに用いる特徴量であり、信号の個性を表現する特徴行列である。但し、音響特徴量ＩＡ２の時間軸および周波数軸の粒度は音響特徴量ＩＡ１の時間軸および周波数軸の粒度と同じとされる。
【００５０】
さらに、入力信号処理部２１が音響特徴量ＩＡ１と音響特徴量ＩＡ２を算出する具体的な処理について説明する。
【００５１】
まず、入力信号切り出し部５１は、常に入力され続けている入力信号からある一定時間（例えば、５秒間）の信号を切り出して、時間周波数変換部５２に供給する。時間周波数変換部５２は、切り出された入力信号を対数振幅スペクトログラム（以下、単にスペクトログラムとも称する）に変換する。
【００５２】
また、音響特徴量抽出部５３は、スペクトログラムを、分割された時間周波数領域内のスペクトログラムの正弦波らしさを数値化した中間特徴量に変換する。
【００５３】
すなわち、正弦波らしさの数値化にはスペクトログラムの微小時間内の安定性が利用される。楽器音や人の音声は雑音と違い、微小時間（例えば、0.020秒）内では周波数がほぼ一定の正弦波と見なすことができ、スペクトログラムの形状がほぼ一定となる。
【００５４】
音響特徴量抽出部５３は、この性質を利用して、微小時間内でのスペクトログラムの安定性を周波数帯ごとに数値化し、これを正弦波らしさの指標とする。具体的には、音響特徴量抽出部５３は、スペクトログラムの時間フレームごとにピーク検出処理を行ない、ピーク周辺の時間周波数領域について対数振幅スペクトログラムを次式（１）に示す双二次関数ｇ（ｋ，ｎ）で近似する。
【００５５】
【数１】

$$g(k,n)=\bar{a}k^2+\bar{b}k+\bar{c} \qquad \cdots(1)$$

【００５６】
なお、式（１）において、ｋはスペクトログラムの周波数bin番号を示しており、ｎはスペクトログラムの時間フレーム番号を示している。また、対数振幅スペクトログラムの近似は、最小二乗法等の最適化手法によって行なわれる。
【００５７】
次に、音響特徴量抽出部５３は検出されたピーク周辺の時間周波数領域の各時間フレームの対数振幅スペクトルを次式（２）に示す二次関数f_n（ｋ）で近似する。
【００５８】
【数２】

$$f_n(k)=a_nk^2+b_nk+c_n \qquad \cdots(2)$$

【００５９】
この近似も最小二乗法等の最適化手法によって行なわれる。
【００６０】
さらに、音響特徴量抽出部５３は、双二次関数ｇ（ｋ，ｎ）および二次関数f_n（ｋ）の２種類の関数への近似によって得られた係数を用いて、正弦波らしさを次式（３）により算出する。
【００６１】

【数３】

$$\eta(n,k)=1-\alpha\sqrt{\sum\{D_1(a_n,\bar{a})+D_2(b_n,\bar{b})\}} \quad \cdots(3)$$

【００６２】

なお、式（３）において、αは正の値を持つパラメータであり、Ｄ（ｘ，ｙ）、つまりＤ$_1$（ｘ，ｙ）およびＤ$_2$（ｘ，ｙ）は距離関数を示している。

【００６３】

また、正弦波を時間周波数変換した場合には、二次関数の二次の係数の理論値ａが存在する。この理論値と算出された二次関数の二次の係数の近さを加味して、次式（４）により正弦波らしさが算出されるようにしてもよい。

【００６４】

【数４】

$$\eta(n,k)=1-\alpha\sqrt{\sum\{D_1(a_n,\bar{a})+D_2(b_n,\bar{b})+D_3(a_n,a)\}} \quad \cdots(4)$$

【００６５】

なお、η（ｎ，ｋ）は各ピークにおける正弦波らしさを意味するので、η（ｎ，ｋ）＜０となった場合はη（ｎ，ｋ）＝０とされる。これによって、η（ｎ，ｋ）は０から１の値を取ることとなる。

【００６６】

また、ピークに該当しない周波数binについてはη（ｎ，ｋ）＝０とされ、当該時間フレームについて、各周波数binの正弦波らしさの情報を持ったベクトルが得られる。この正弦波らしさは残響にロバストな特徴量であるため、最終的に残響に頑健な検索を行うことが可能である。

【００６７】

以上のようにして得られたベクトルを、時間フレームをずらしながら算出していき、得られたベクトルを時系列に並べて、時間軸方向にダウンサンプリングを行ったものが音響特徴量ＩＡ１である。ダウンサプリングには平滑化フィルタ（ローパスフィルタ）が用いられる。フィルタリングされた値は各周波数における正弦波らしさの時間平均値を意味する。

【００６８】

なお、得られた音響特徴量ＩＡ１の各要素について、量子化や対数関数、指数関数、シグモイド関数等の非線形処理が施されるようにしてもよい。

【００６９】

また、音響特徴量抽出部５４では、スペクトログラムが音響特徴量ＩＡ２に変換される。

【００７０】

例えば、音響特徴量ＩＡ１と同様に算出された正弦波らしさの行列に対して時間軸方向に一次微分フィルタを施し、ダウンサンプリングしたものが音響特徴量ＩＡ２とされる。一次微分フィルタによってフィルタリングされた値は各周波数における正弦波らしさの時間変化量を意味する。

【００７１】

なお、得られた音響特徴量ＩＡ２の各要素についても、量子化や対数関数、指数関数、シグモイド関数等の非線形処理が施されるようにしてもよい。さらに、音響特徴量ＩＡ２は信号の個性を表現するものであればよく、例えば、一定の時間区間内のスペクトルの時間平均を正規化したものなどを用いるようにしてもよい。

【００７２】

〈参照信号処理部の構成例〉

また、図２の参照信号処理部２２は、より詳細には図４に示すように構成される。図４に示す参照信号処理部２２は、参照信号切り出し部８１、時間周波数変換部８２、音響特徴量抽出部８３、および音響特徴量抽出部８４から構成される。

【００７３】
参照信号切り出し部８１は、供給された参照信号の所定時間長の区間を切り出して、時間周波数変換部８２に供給する。時間周波数変換部８２は、参照信号切り出し部８１から供給された参照信号に対する時間周波数変換を行なって、参照信号を対数振幅スペクトログラムに変換し、音響特徴量抽出部８３および音響特徴量抽出部８４に供給する。
【００７４】
音響特徴量抽出部８３は、時間周波数変換部８２から供給された対数振幅スペクトログラムに基づいて音響特徴量ＲＡ１を算出し、マッチング処理部２３に供給する。音響特徴量抽出部８４は、時間周波数変換部８２から供給された対数振幅スペクトログラムに基づいて音響特徴量ＲＡ２を算出し、マッチング処理部２３に供給する。
【００７５】
なお、音響特徴量抽出部８３と音響特徴量抽出部８４は、音響特徴量抽出部５３と音響特徴量抽出部５４に対応しており、音響特徴量ＩＡ１および音響特徴量ＩＡ２と同じ時間軸・周波数軸粒度の音響特徴量ＲＡ１および音響特徴量ＲＡ２を出力する。
【００７６】
また、参照信号から抽出された音響特徴量ＲＡ１および音響特徴量ＲＡ２は、直接マッチング処理部２３に供給されてもよいし、記憶装置に供給されてデータベースとして保存されてもよい。但し、音響特徴量ＲＡ１および音響特徴量ＲＡ２が記憶装置に供給される場合には、音響特徴量ＲＡ１および音響特徴量ＲＡ２は、参照信号のメタデータ（番組名、放送日時、出演者など）、つまりコンテンツ属性データとセットで保存される必要がある。
【００７７】
〈マッチング処理部の構成例〉
さらに、図２のマッチング処理部２３は、より詳細には図５に示すように構成される。図５に示すマッチング処理部２３は、マスクパターン生成部１１１、類似度計算部１１２、および比較統合部１１３から構成される。
【００７８】
マスクパターン生成部１１１は、音響特徴量抽出部５３からの音響特徴量ＩＡ１と、音響特徴量抽出部８３からの音響特徴量ＲＡ１とに基づいてマスクパターンを生成し、マスクパターンおよび音響特徴量Ａ１間の類似度を類似度計算部１１２に供給する。マスクパターンは、各時間周波数領域におけるコンテンツの信号らしさの信頼度、つまり信頼のおける時間周波数領域を示している。
【００７９】
類似度計算部１１２は、音響特徴量抽出部５４からの音響特徴量ＩＡ２、音響特徴量抽出部８４からの音響特徴量ＲＡ２、およびマスクパターン生成部１１１からのマスクパターンと類似度に基づいて、入力信号の参照信号との類似度を算出する。また、類似度計算部１１２は、算出した類似度とともに、供給されたコンテンツ属性データを比較統合部１１３に供給する。
【００８０】
比較統合部１１３は、類似度計算部１１２から供給された類似度に基づいて、参照信号のコンテンツと、入力信号に含まれるコンテンツが同一であるかを判定し、その判定結果とともにコンテンツ属性データを出力する。
【００８１】
マッチング処理部２３では、参照信号と入力信号の類似度が計算される。例えば図６に示すように、一定時間（例えば、５秒間）の入力信号に参照信号の断片が含まれている場合、一般に入力信号の音響特徴量ＩＡ１および音響特徴量ＩＡ２の行列は、参照信号の音響特徴量ＲＡ１および音響特徴量ＲＡ２に比べ時間方向の成分数が少ない。
【００８２】
そのため、参照信号の音響特徴量ＲＡ１および音響特徴量ＲＡ２の行列から入力信号の音響特徴量ＩＡ１および音響特徴量ＩＡ２の時間方向の長さと同じ長さの部分行列が切り

出されて類似度計算が行なわれる。この部分行列の切り出しについては、切り取り可能な部分行列すべてについて切り出しが行なわれる。切り出し処理はマスクパターン生成部１１１および類似度計算部１１２で行われる。

【００８３】

図６では、縦方向および横方向は、それぞれ周波数および時間を示している。また、矢印Ｑ１１乃至矢印Ｑ１４により示される長方形は、それぞれ参照信号の音響特徴量ＲＡ１、参照信号の音響特徴量ＲＡ２、入力信号の音響特徴量ＩＡ１、および入力信号の音響特徴量ＩＡ２を表している。

【００８４】

この例では、参照信号から抽出された音響特徴量ＲＡ１および音響特徴量ＲＡ２は、入力信号から抽出された音響特徴量ＩＡ１および音響特徴量ＩＡ２よりも、図中、横方向、つまり時間方向に長くなっており、時間方向の成分数が多いことが分かる。

【００８５】

そのため、音響特徴量ＲＡ１や音響特徴量ＲＡ２の一部分が切り出されて部分行列とされ、類似度計算に用いられる。

【００８６】

次に、マッチング処理部２３で行なわれる具体的な処理について説明する。

【００８７】

マスクパターン生成部１１１は、入力信号の音響特徴量ＩＡ１と参照信号の音響特徴量ＲＡ１からマスクパターンを生成し、さらに音響特徴量Ａ１同士の類似度を算出する。マスクパターンは、音響特徴量Ａ１と同様に時間軸と周波数軸を持った二次元の行列で表現される。

【００８８】

例えば、入力信号の音響特徴量ＩＡ１と参照信号の音響特徴量ＲＡ１から、正弦波が存在しない時間周波数領域をマスクする行列がマスクパターンとして生成される。より具体的には、例えば次式（５）を計算することによりマスクパターンが生成される。

【００８９】

【数５】

$$W_{f(t+u)} = S^{(1)}_{fu} A^{(1)}_{f(t+u)} \qquad \cdots (5)$$

【００９０】

なお、式（５）において、$W_{f(t+u)}$はマスクパターンの行列要素を示しており、$S^{(1)}_{fu}$は入力信号の音響特徴量ＩＡ１の行列要素を示しており、$A^{(1)}_{f(t+u)}$は参照信号の音響特徴量ＲＡ１の部分行列の要素を示している。

【００９１】

また、ｆは各行列の周波数成分を示しており、ｕは各行列の時間成分を示しており、ｔは部分行列の時間オフセットを示している。

【００９２】

このようにして算出されたマスクパターンは、後段の類似度計算部１１２において、時間周波数領域ごとの重みとして用いられる。つまり、マスクパターンの行列要素$W_{f(t+u)}$の値の大きい時間周波数領域に重きを置いた類似度が算出される。

【００９３】

音響特徴量Ａ１同士の類似度は、２つの特徴量の近さを定量化した非負の指標で、例えば次式（６）によって算出される。

【００９４】

【数６】

$$R^{(1)}(t) = \frac{\sum S^{(1)}_{fu} A^{(1)}_{f(t+u)}}{\left(\sum S^{(1)p}_{fu}\right)^{1/p} \cdot \left(\sum A^{(1)q}_{f(t+u)}\right)^{1/q}} \qquad \cdots (6)$$

【００９５】
なお、式（６）において、Ｒ（１）（ｔ）はＳ（１）ｆｕとＡ（１）ｆ（ｔ＋ｕ）の類似度を示している。また、ｐおよびｑは、入力信号の音響特徴量ＩＡ１と参照信号の音響特徴量ＲＡ１の類似度に対する寄与率を調整するパラメータである。すなわち、ｐおよびｑは１／ｐ＋１／ｑ＝１を満たす１以上の値を持つ重み係数である。
【００９６】
例えば、ｐをｑより大きくすることで参照信号に含まれている音声に重きを置いた類似度計算が行なわれ、入力信号に参照信号とは関係のない混入雑音が含まれていても、その影響を軽減したマッチングを行うことができる。なお、音響特徴量同士の類似度として、上記の類似度の他に、二乗誤差や絶対誤差など２つの行列の相違に基づいて算出される値が用いられてもよい。
【００９７】
また、類似度計算部１１２は、入力信号の音響特徴量ＩＡ２と参照信号の音響特徴量ＲＡ２、マスクパターン、および音響特徴量Ａ１同士の類似度を用いて、最終的な類似度計算を行なう。
【００９８】
類似度計算部１１２により算出される類似度は、時間周波数領域における正弦波らしさの情報を持つマスクパターンを各時間周波数領域における信頼度と見なして重み付けして定量化した、入力信号の音響特徴量ＩＡ２と参照信号の音響特徴量ＲＡ２の近さの指標である。さらに、音響特徴量Ａ１同士の類似度も加味されて、例えば次式（７）の計算により類似度Ｒ（ｔ）が算出される。
【００９９】
【数７】

$$R(t)=\frac{\sum W_{f(t+u)}\exp\left(-\beta\left(S^{(2)}_{fu}-A^{(2)}_{f(t+u)}\right)^2\right)}{\sum W_{f(t+u)}}R^{(1)}(t) \quad \cdots(7)$$

【０１００】
なお、式（７）において、Ａ（２）ｆ（ｔ＋ｕ）は参照信号の音響特徴量ＲＡ２の部分行列を示しており、Ｓ（２）ｆｕは入力信号の音響特徴量ＩＡ２の行列を示している。また、βは正の値を持つパラメータである。
【０１０１】
さらに、類似度計算には、上記の式（７）の計算以外に、二乗誤差や絶対誤差などの２つの行列（入力信号の音響特徴量ＩＡ２と参照信号の音響特徴量ＲＡ２）の相違に基づいて計算される値が用いられてもよい。
【０１０２】
比較統合部１１３では、類似度計算部１１２で算出された類似度から参照信号のコンテンツと、入力信号に含まれるコンテンツが同一のものであるかが判定される。
【０１０３】
同一コンテンツであるかの判定方法は、複数の参照信号について得られた類似度のうち、所定の閾値を超えたもので最大の類似度を持つ参照信号が入力信号に含まれているコンテンツであると判定する方法とされる。なお、全ての参照信号の類似度が閾値を超えなかった場合は、参照信号の中に該当するコンテンツはないという判定がなされる。
【０１０４】
また、ここで用いられる閾値は、常に固定の値とされてもよいし、入力信号と複数の参照信号から得られた複数の類似度から統計的に設定されるようにしてもよい。
【０１０５】
〈一致検索処理の説明〉
ところで、音声処理装置１１に入力信号と参照信号が供給され、コンテンツ同定が指示

されると、音声処理装置１１は一致検索処理を行なってコンテンツの同定を行なう。以下、図７のフローチャートを参照して、音声処理装置１１による一致検索処理について説明する。

【０１０６】

ステップＳ１１において、入力信号切り出し部５１は、供給された入力信号を切り出して時間周波数変換部５２に供給する。例えば、一定時間分の入力信号が切り出される。

【０１０７】

ステップＳ１２において、時間周波数変換部５２は、入力信号切り出し部５１から供給された入力信号に対して時間周波数変換を行い、入力信号を対数振幅スペクトログラムに変換し、音響特徴量抽出部５３および音響特徴量抽出部５４に供給する。

【０１０８】

ステップＳ１３において、音響特徴量抽出部５３は音響特徴量ＩＡ１の抽出処理を行なって、入力信号の音響特徴量ＩＡ１を算出し、マッチング処理部２３のマスクパターン生成部１１１に供給する。

【０１０９】

ここで、図８のフローチャートを参照して、ステップＳ１３の処理に対応する、音響特徴量抽出部５３による音響特徴量ＩＡ１の抽出処理について説明する。

【０１１０】

ステップＳ５１において、音響特徴量抽出部５３は、時間周波数変換部５２から供給された対数振幅スペクトログラムについて、時間フレームを選択する。

【０１１１】

ステップＳ５２において、音響特徴量抽出部５３は、対数振幅スペクトログラムの選択した時間フレームについてピーク検出を行なう。

【０１１２】

ステップＳ５３において、音響特徴量抽出部５３は、検出されたピーク周辺の時間周波数領域の対数振幅スペクトルを２種類の二次関数で近似する。例えば、対数振幅スペクトルが式（１）および式（２）に示した関数で近似される。

【０１１３】

ステップＳ５４において、音響特徴量抽出部５３は、近似された二次関数の係数から正弦波らしさの指標へ変換し、保存する。例えば式（３）のη（ｎ，ｋ）が正弦波らしさの指標として算出される。

【０１１４】

ステップＳ５５において、音響特徴量抽出部５３は入力信号の全時間フレームを処理したか否かを判定する。ステップＳ５５においてまだ全ての時間フレームを処理していないと判定された場合、処理はステップＳ５１に戻り、上述した処理が繰り返される。

【０１１５】

これに対して、ステップＳ５５において、全時間フレームを処理したと判定された場合、ステップＳ５６において、音響特徴量抽出部５３は、保存した正弦波らしさの指標のベクトルを時系列順に並べて行列化する。

【０１１６】

ステップＳ５７において、音響特徴量抽出部５３は、行列化された正弦波らしさの指標、つまり正弦波らしさの行列に対して時間軸方向にフィルタリングを施し、正弦波らしさの時間平均量を算出する。例えば、平滑化フィルタによりフィルタリングが行なわれる。

【０１１７】

ステップＳ５８において、音響特徴量抽出部５３は、フィルタリングにより得られた正弦波らしさの時間平均量に対して時間軸方向にリサンプリングを行い、音響特徴量ＩＡ１とする。音響特徴量抽出部５３が、このようにして入力信号から抽出した音響特徴量ＩＡ１をマスクパターン生成部１１１に供給すると、音響特徴量ＩＡ１の抽出処理は終了し、その後、処理は図７のステップＳ１４へと進む。

【０１１８】

ステップＳ１４において、音響特徴量抽出部５４は音響特徴量ＩＡ２の抽出処理を行なって、入力信号の音響特徴量ＩＡ２を算出し、マッチング処理部２３の類似度計算部１１２に供給する。

【０１１９】

ここで、図９のフローチャートを参照して、ステップＳ１４の処理に対応する、音響特徴量抽出部５４による音響特徴量ＩＡ２の抽出処理について説明する。なお、ステップＳ９１乃至ステップＳ９６の処理は、図８のステップＳ５１乃至ステップＳ５６の処理と同様であるので、その説明は省略する。

【０１２０】

ステップＳ９６の処理が行なわれると、正弦波らしさの行列が得られるので、ステップＳ９７において、音響特徴量抽出部５４は、正弦波らしさの行列に対して時間軸方向にフィルタリングを施し、正弦波らしさの時間変化量を算出する。例えば、一次微分フィルタによりフィルタリングが行なわれる。

【０１２１】

ステップＳ９８において、音響特徴量抽出部５４は、フィルタリングにより得られた正弦波らしさの時間変化量に対して時間軸方向にリサンプリングを行い、音響特徴量ＩＡ２とする。音響特徴量抽出部５４が、このようにして入力信号から抽出した音響特徴量ＩＡ２を類似度計算部１１２に供給すると、音響特徴量ＩＡ２の抽出処理は終了し、その後、処理は図７のステップＳ１５へと進む。

【０１２２】

図７のフローチャートの説明に戻り、ステップＳ１５において、参照信号切り出し部８１は、供給された参照信号を切り出して時間周波数変換部８２に供給する。

【０１２３】

ステップＳ１６において、時間周波数変換部８２は、参照信号切り出し部８１から供給された参照信号に対して時間周波数変換を行い、参照信号を対数振幅スペクトログラムに変換し、音響特徴量抽出部８３および音響特徴量抽出部８４に供給する。

【０１２４】

ステップＳ１７において、音響特徴量抽出部８３は音響特徴量ＲＡ１の抽出処理を行なって、参照信号の音響特徴量ＲＡ１を算出し、マッチング処理部２３のマスクパターン生成部１１１に供給する。

【０１２５】

また、ステップＳ１８において、音響特徴量抽出部８４は音響特徴量ＲＡ２の抽出処理を行なって、参照信号の音響特徴量ＲＡ２を算出し、マッチング処理部２３の類似度計算部１１２に供給する。

【０１２６】

なお、これらのステップＳ１７およびステップＳ１８の処理は、ステップＳ１３およびステップＳ１４の処理と同様であるので、その説明は省略する。但し、ステップＳ１７およびステップＳ１８の処理では、処理対象の信号が入力信号ではなく参照信号とされる。

【０１２７】

ステップＳ１９において、マスクパターン生成部１１１は、音響特徴量抽出部５３からの音響特徴量ＩＡ１と、音響特徴量抽出部８３からの音響特徴量ＲＡ１とに基づいてマスクパターンを生成する。例えば、マスクパターン生成部１１１は式（５）の演算を行なうことでマスクパターンを生成する。

【０１２８】

ステップＳ２０において、マスクパターン生成部１１１は、音響特徴量Ａ１の類似度を算出する。例えば、マスクパターン生成部１１１は式（６）により音響特徴量Ａ１の類似度を算出する。そして、マスクパターン生成部１１１は、算出されたマスクパターンと、音響特徴量Ａ１の類似度とを類似度計算部１１２に供給する。

【０１２９】

ステップＳ２１において類似度計算部１１２は、音響特徴量抽出部５４からの音響特徴

量ＩＡ２、音響特徴量抽出部８４からの音響特徴量ＲＡ２、およびマスクパターン生成部１１１からのマスクパターンと類似度に基づいて、入力信号と参照信号の最終的な類似度を算出する。

【０１３０】

例えば、類似度計算部１１２は、式（７）を計算することで、入力信号と参照信号、つまり入力信号のコンテンツと参照信号のコンテンツとの類似度を算出し、コンテンツ属性データとともに比較統合部１１３に供給する。

【０１３１】

ステップＳ２２において、比較統合部１１３は、類似度計算部１１２から供給された類似度に基づいて、参照信号のコンテンツと、入力信号に含まれるコンテンツが同一コンテンツであるかの判定を行なう。

【０１３２】

例えば、比較統合部１１３は、複数の参照信号について得られた類似度のうち、所定の閾値を超える最大のものを特定し、特定された類似度の参照信号のコンテンツが、入力信号のコンテンツであるとする。比較統合部１１３は、このようにして特定された入力信号のコンテンツのコンテンツ属性データと、コンテンツ同定の判定結果とを出力し、一致検索処理は終了する。

【０１３３】

以上のようにして、音声処理装置１１は、入力信号と参照信号から正弦波らしさを示す音響特徴量Ａ１を算出するとともに、音響特徴量Ａ１からマスクパターンを生成し、そのマスクパターンと、信号の個性を示す音響特徴量Ａ２とから類似度を算出する。

【０１３４】

このように、入力信号から得られた音響特徴量ＩＡ１と、参照信号から得られた音響特徴量ＲＡ１とからマスクパターンを生成すれば、残響や混入雑音に頑健なものを得ることができる。これにより、より高精度に任意のコンテンツの同定を行なうことができる。

【０１３５】

ところで、上述した一連の処理は、ハードウェアにより実行することもできるし、ソフトウェアにより実行することもできる。一連の処理をソフトウェアにより実行する場合には、そのソフトウェアを構成するプログラムが、コンピュータにインストールされる。ここで、コンピュータには、専用のハードウェアに組み込まれているコンピュータや、各種のプログラムをインストールすることで、各種の機能を実行することが可能な、例えば汎用のパーソナルコンピュータなどが含まれる。

【０１３６】

図１０は、上述した一連の処理をプログラムにより実行するコンピュータのハードウェアの構成例を示すブロック図である。

【０１３７】

コンピュータにおいて、CPU（Central Processing Unit）７０１，ROM（Read Only Memory）７０２，RAM（Random Access Memory）７０３は、バス７０４により相互に接続されている。

【０１３８】

バス７０４には、さらに、入出力インターフェース７０５が接続されている。入出力インターフェース７０５には、入力部７０６、出力部７０７、記録部７０８、通信部７０９、及びドライブ７１０が接続されている。

【０１３９】

入力部７０６は、キーボード、マウス、マイクロホン、撮像素子などよりなる。出力部７０７は、ディスプレイ、スピーカなどよりなる。記録部７０８は、ハードディスクや不揮発性のメモリなどよりなる。通信部７０９は、ネットワークインターフェースなどよりなる。ドライブ７１０は、磁気ディスク、光ディスク、光磁気ディスク、又は半導体メモリなどのリムーバブルメディア７１１を駆動する。

【０１４０】

以上のように構成されるコンピュータでは、CPU７０１が、例えば、記録部７０８に記録されているプログラムを、入出力インターフェース７０５及びバス７０４を介して、RAM７０３にロードして実行することにより、上述した一連の処理が行われる。

【０１４１】

コンピュータ（CPU７０１）が実行するプログラムは、例えば、パッケージメディア等としてのリムーバブルメディア７１１に記録して提供することができる。また、プログラムは、ローカルエリアネットワーク、インターネット、デジタル衛星放送といった、有線または無線の伝送媒体を介して提供することができる。

【０１４２】

コンピュータでは、プログラムは、リムーバブルメディア７１１をドライブ７１０に装着することにより、入出力インターフェース７０５を介して、記録部７０８にインストールすることができる。また、プログラムは、有線または無線の伝送媒体を介して、通信部７０９で受信し、記録部７０８にインストールすることができる。その他、プログラムは、ROM７０２や記録部７０８に、あらかじめインストールしておくことができる。

【０１４３】

なお、コンピュータが実行するプログラムは、本明細書で説明する順序に沿って時系列に処理が行われるプログラムであっても良いし、並列に、あるいは呼び出しが行われたとき等の必要なタイミングで処理が行われるプログラムであっても良い。

【０１４４】

また、本技術の実施の形態は、上述した実施の形態に限定されるものではなく、本技術の要旨を逸脱しない範囲において種々の変更が可能である。

【０１４５】

例えば、本技術は、１つの機能をネットワークを介して複数の装置で分担、共同して処理するクラウドコンピューティングの構成をとることができる。

【０１４６】

また、上述のフローチャートで説明した各ステップは、１つの装置で実行する他、複数の装置で分担して実行することができる。

【０１４７】

さらに、１つのステップに複数の処理が含まれる場合には、その１つのステップに含まれる複数の処理は、１つの装置で実行する他、複数の装置で分担して実行することができる。

【０１４８】

さらに、本技術は、以下の構成とすることも可能である。

【０１４９】

［１］

同定対象となるコンテンツの入力信号に基づいて、各時間周波数領域における信号の正弦波らしさを示す第１の音響特徴量と、前記第１の音響特徴量とは異なる第２の音響特徴量とを算出する入力信号処理部と、

予め用意したコンテンツの参照信号に基づいて、前記第１の音響特徴量と前記第２の音響特徴量とを算出する参照信号処理部と、

前記入力信号の前記第１の音響特徴量および前記第２の音響特徴量と、前記参照信号の前記第１の音響特徴量および前記第２の音響特徴量とに基づいて、前記入力信号と前記参照信号の類似度を計算するマッチング処理部と

を備える音声処理装置。

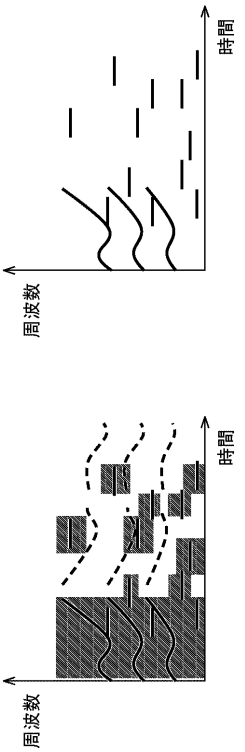
［２］

前記マッチング処理部は、前記入力信号と前記参照信号の前記第１の音響特徴量に基づいて、各時間周波数領域におけるコンテンツの信号らしさを示すマスクパターンを生成し、前記マスクパターン、前記第１の音響特徴量、および前記第２の音響特徴量に基づいて前記類似度を計算する

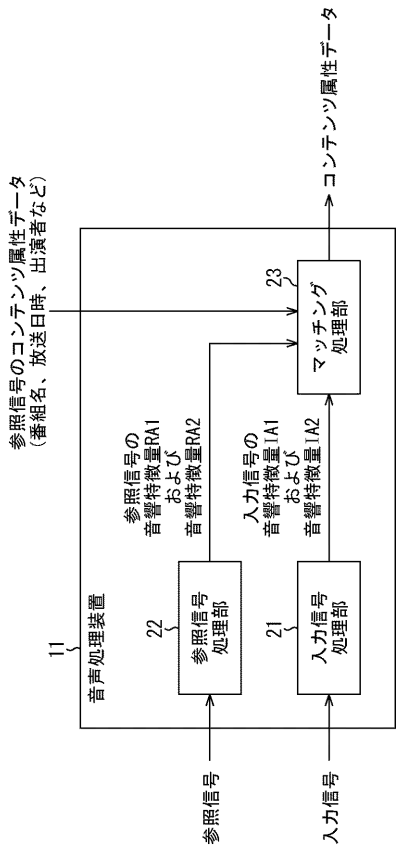
［１］に記載の音声処理装置。

［３］
前記マッチング処理部は、前記入力信号の前記第１の音響特徴量と、前記参照信号の前記第１の音響特徴量との類似度をさらに算出し、前記マスクパターン、前記第１の音響特徴量の前記類似度、および前記第２の音響特徴量に基づいて、前記入力信号と前記参照信号の前記類似度を計算する
［２］に記載の音声処理装置。
［４］
前記マッチング処理部は、前記第１の音響特徴量の前記類似度に対する前記入力信号の寄与率よりも、前記第１の音響特徴量の前記類似度に対する前記参照信号の寄与率をより大きくして、前記第１の音響特徴量の前記類似度を算出する
［３］に記載の音声処理装置。
［５］
前記第２の音響特徴量は、前記入力信号または前記参照信号のスペクトログラムに基づいて算出され、時間軸および周波数軸において前記第１の音響特徴量と同じ粒度を有する
［１］乃至［４］の何れかに記載の音声処理装置。
【符号の説明】
【０１５０】
１１　音声処理装置，　２１　入力信号処理部，　２２　参照信号処理部，　２３　マッチング処理部，　５３　音響特徴量抽出部，　５４　音響特徴量抽出部，　８３　音響特徴量抽出部，　８４　音響特徴量抽出部，　１１１　マスクパターン生成部，　１１２　類似度計算部，　１１３　比較統合部

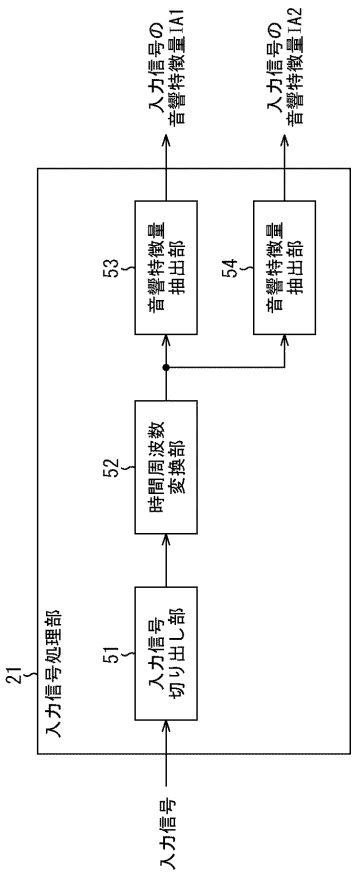
【図１】
図1



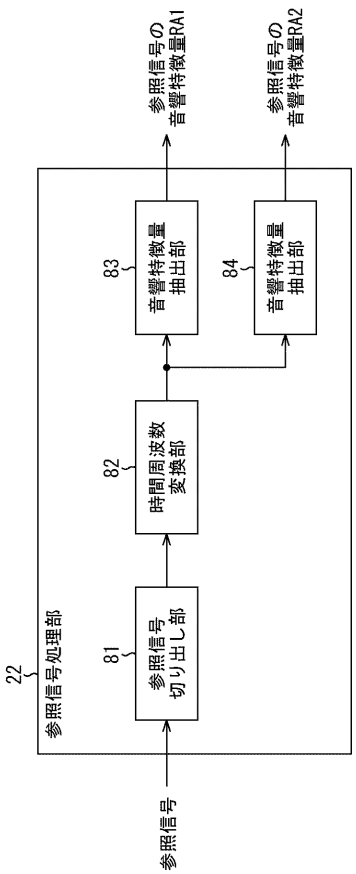
【図２】
図2



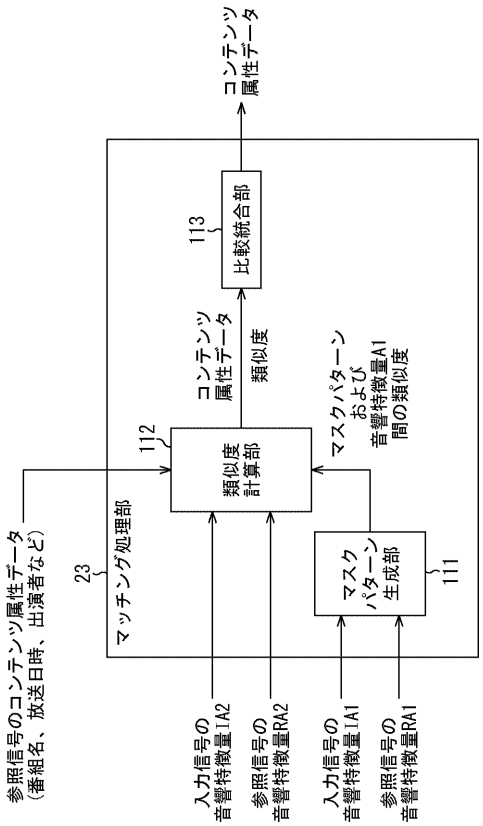
【図３】
図3



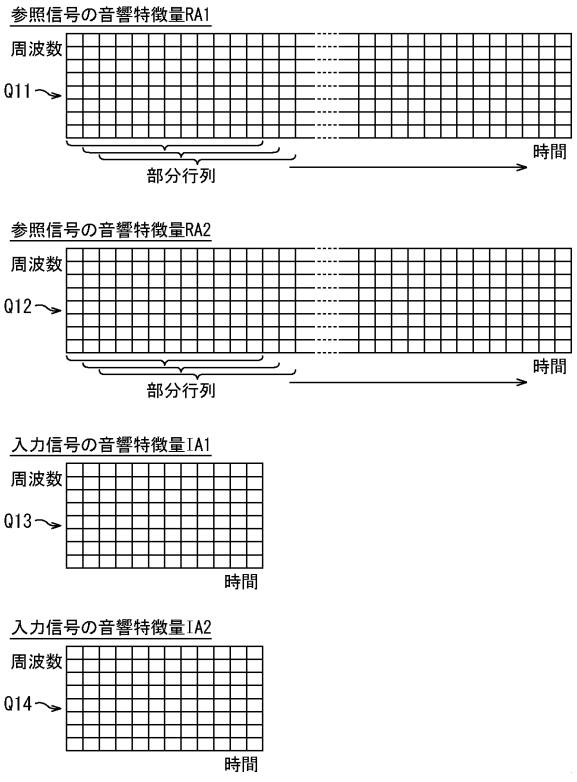
【図４】
図4



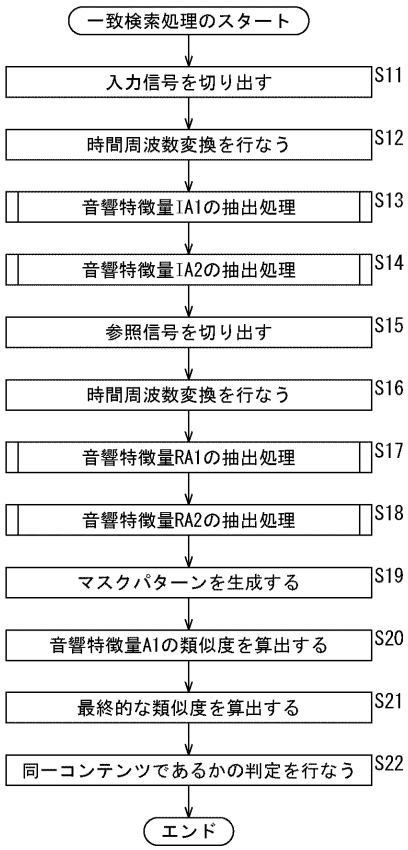
【図５】
図5



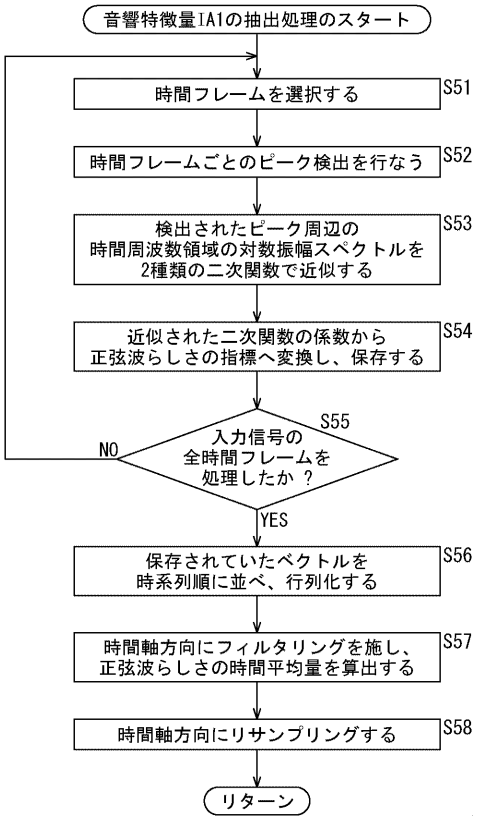
【図６】
図6



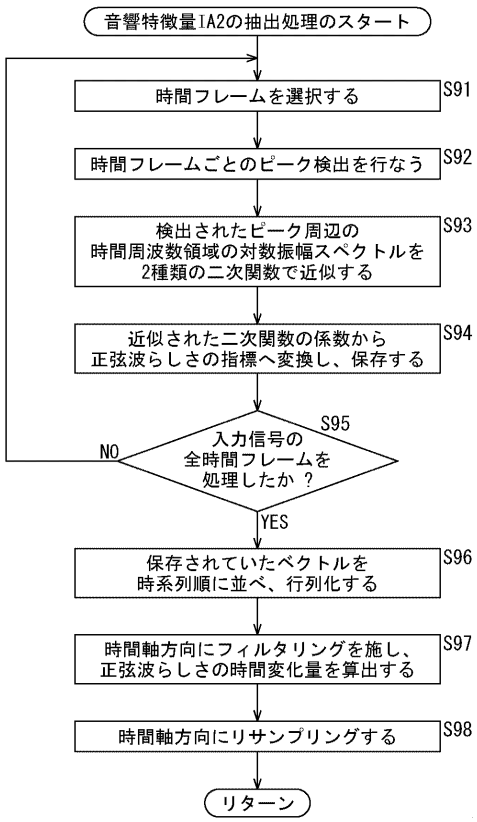
【図７】
図7



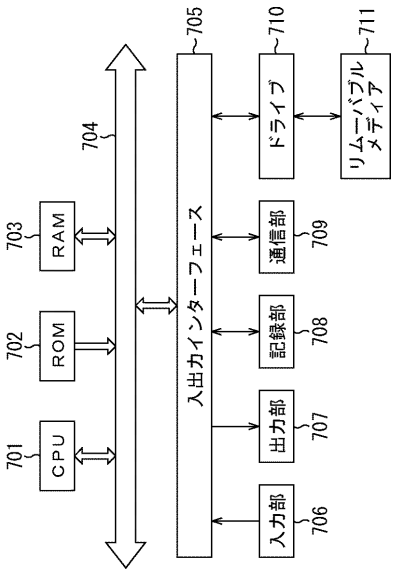
【図８】
図8



【図９】
図9



【図１０】
図10



フロントページの続き

(72)発明者　西口　正之
東京都港区港南１丁目７番１号　ソニー株式会社内

審査官　菊池　智紀

(56)参考文献　特開２０１２－２２６０８０（ＪＰ，Ａ）
特開２０１２－０９８３６０（ＪＰ，Ａ）
澁谷崇 他，"トーン構造記述子を用いた高速背景音楽検索"，情報処理学会研究報告，２０１１年　7月２９日，Vol.2011-MUS-91，No.17，pp.1-8

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　２５／００－２５／９３，
１５／００－１５／３４
Ｇ０６Ｆ　１７／３０