(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6588212号
(P6588212)
(45)発行日　令和1年10月9日(2019.10.9)　(24)登録日　令和1年9月20日(2019.9.20)

(51)Int.Cl.　F I
G06F 16/632 (2019.01)　G06F 16/632
G10L 25/54 (2013.01)　G10L 25/54
G10L 25/18 (2013.01)　G10L 25/18

請求項の数 3　(全 17 頁)

(21)出願番号　特願2015-43586(P2015-43586)
(22)出願日　平成27年3月5日(2015.3.5)
(65)公開番号　特開2016-162411(P2016-162411A)
(43)公開日　平成28年9月5日(2016.9.5)
審査請求日　平成30年2月5日(2018.2.5)

(73)特許権者　591141784
学校法人大阪産業大学
大阪府大東市中垣内３丁目１番１号
(74)代理人　100104433
弁理士　宮園　博一
(72)発明者　高橋　徹
大阪府大東市中垣内３丁目１番１号　大阪産業大学　デザイン工学部情報システム学科内

審査官　鹿野　博嗣

最終頁に続く

(54)【発明の名称】音源検索装置および音源検索方法

(57)【特許請求の範囲】
【請求項１】
楽曲信号と音声信号とを含む混合音から特徴量を抽出する検索装置側特徴量抽出手段と、
前記抽出した特徴量を２値化する検索装置側２値化手段と、
前記検索装置側２値化手段により２値化された前記混合音の前記特徴量を検索キーとして、音源データベースから音源を検索する検索手段とを備え、
前記混合音の前記特徴量は、前記混合音のフーリエスペクトルに基づいて算出される各帯域窓の出力エネルギであるクロマスペクトルであり、
前記混合音の前記特徴量は、所定の時間長さを有する１分析フレーム毎または複数の分析フレーム毎に抽出されており、
前記検索装置側２値化手段は、前記混合音の前記特徴量が、前記１分析フレーム毎または前記複数の分析フレーム毎における特徴量のクロマスペクトルに基づいた所定の基準値以上の場合に特徴量を１とし、前記所定の基準値未満の場合に特徴量を０とするように構成されている、音源検索装置。
【請求項２】
前記音源データベースは、
データベース用楽曲信号から特徴量を抽出するデータベース側特徴量抽出手段と、
前記抽出した特徴量を２値化するデータベース側２値化手段と、
前記データベース側２値化手段により２値化された前記データベース用楽曲信号の前記

特徴量から前記音源データベースを構築する構築手段とを含み、
前記データベース用楽曲信号の前記特徴量は、前記データベース用楽曲信号のフーリエスペクトルに基づいて算出される各帯域窓の出力エネルギであるクロマスペクトルであり、
前記データベース用楽曲信号の前記特徴量は、前記所定の時間長さを有する前記１分析フレーム毎または前記複数の分析フレーム毎に抽出されており、
前記データベース側２値化手段は、前記データベース用楽曲信号の前記特徴量が、前記１分析フレーム毎または前記複数の分析フレーム毎における特徴量の前記所定の基準値以上の場合に特徴量を１とし、前記所定の基準値未満の場合に特徴量を０とするように構成されている、請求項１に記載の音源検索装置。
【請求項３】
楽曲信号と音声信号とを含む混合音から特徴量を抽出する工程と、
前記抽出した特徴量を２値化する工程と、
２値化された前記混合音の前記特徴量を検索キーとして、音源データベースから音源を検索する工程とを備え、
前記混合音の前記特徴量は、所定の時間長さを有する１分析フレーム毎または複数の分析フレーム毎に抽出されており、
前記楽曲信号と音声信号とを含む混合音から特徴量を抽出する工程は、前記混合音のフーリエスペクトルに基づいて算出される各帯域窓の出力エネルギであるクロマスペクトルを特徴量として抽出する工程を含み、
前記抽出した特徴量を２値化する工程は、前記混合音の前記特徴量が、前記１分析フレーム毎または前記複数の分析フレーム毎における特徴量のクロマスペクトルに基づいた所定の基準値以上の場合に特徴量を１とし、前記所定の基準値未満の場合に特徴量を０とする工程を含む、音源検索方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音源検索装置および音源検索方法に関する。
【背景技術】
【０００２】
従来、音源検索装置が知られている（たとえば、非特許文献１参照）。
【０００３】
上記非特許文献１には、混合音を検索キーとした音源検索装置が開示されている。この音源検索装置では、混合音から特徴量を抽出するとともに、抽出された特徴量を検索キーとして、音源データベースから音源を検索するように構成されている。ここで、特徴量としては、クロマスペクトルが用いられている。クロマスペクトルは、所定の時間長さを有する混合音（信号）の各分析フレームにおけるフーリエスペクトルを算出した後、各帯域窓の出力エネルギを算出することにより求められる。なお、クロマスペクトルの要素は、スカラー量（たとえば、単精度浮動小数点数、３２ｂｉｔ）である。そして、混合音の特徴量ベクトルと、音源データベースの音源の特徴量ベクトルとのパターンマッチング（特徴量ベクトル間のユークリッド距離）により、音源を検索するように構成されている。
【先行技術文献】
【非特許文献】
【０００４】
【非特許文献１】“特徴量間の累積距離を用いた混合音からの音源検索システムの評価”、信学技報、ｖｏｌ．１１４、ｎｏ．１９１、ｐｐ．１９－２４．
【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、上記非特許文献１に記載の音源検索装置では、クロマスペクトルが特徴

量として用いられている一方、楽曲信号の音圧の変化などに起因して検索精度が低下するという問題点がある。また、スカラー量の要素を有するクロマスペクトルを特徴量として用いているため、混合音の特徴量ベクトルと、音源データベースの音源の特徴量ベクトルとのパターンマッチング（検索）に時間がかかるという問題点がある。

【０００６】

この発明は、上記のような課題を解決するためになされたものであり、この発明の１つの目的は、混合音を検索キーとする音源検索装置および音源検索方法において、検索速度を高速化させ、かつ、検索精度を向上させることが可能な音源検索装置および音源検索方法を提供することである。

【課題を解決するための手段】

【０００７】

上記目的を達成するために、この発明の第１の局面における音源検索装置は、楽曲信号と音声信号とを含む混合音から特徴量を抽出する検索装置側特徴量抽出手段と、抽出した特徴量を２値化する検索装置側２値化手段と、検索装置側２値化手段により２値化された混合音の特徴量を検索キーとして、音源データベースから音源を検索する検索手段とを備え、混合音の特徴量は、混合音のフーリエスペクトルに基づいて算出される各帯域窓の出力エネルギであるクロマスペクトルであり、混合音の特徴量は、所定の時間長さを有する１分析フレーム毎または複数の分析フレーム毎に抽出されており、検索装置側２値化手段は、混合音の特徴量が、１分析フレーム毎または複数の分析フレーム毎における特徴量のクロマスペクトルに基づいた所定の基準値以上の場合に特徴量を１とし、所定の基準値未満の場合に特徴量を０とするように構成されている。

【０００８】

この発明の第１の局面による音源検索装置では、上記のように、抽出した特徴量を２値化する検索装置側２値化手段を備えることによって、スカラー量（たとえば、単精度浮動小数点数、３２ｂｉｔ）の要素を有する特徴量を検索キーとして音源データベースから音源を検索する場合と比べて、特徴量が２値化される分、次元が小さくなる（１ｂｉｔ）ので、検索速度を高速化させることができる。

【０００９】

また、楽曲信号に音声信号を混合した場合、音声信号が混合される分、楽曲信号の包絡（形状）が変化する。そこで、本発明では、抽出した特徴量を２値化する検索装置側２値化手段を備えることによって、２値化後の特徴量のうち、「１」の部分は、音声信号が加法的に作用している限り、「１」のままである。一方、２値化後の特徴量のうち、「０」の部分に音声信号が加法的に作用しても、２値化するための基準値を超えない限り「０」のままである。なお、２値化するためのしきい値近傍では、音声信号が混合されることにより、２値化後の特徴量の「０」または「１」が反転する場合がある一方、音声信号の出力エネルギが大きい周波数（反転する可能性がある周波数）は基本周波数の整数倍の周波数近傍のみの比較的小さい範囲であるため、反転による影響は小さいと考えられる。その結果、２値化された特徴量は、楽曲信号の包絡（形状）を表しながら、混合される音声信号に対して頑強な特徴量となる。また、楽曲信号の音圧の変化に対しても、混合音の音量の変化に伴って所定の基準値も変化させることが可能であるので、混合音の特徴量の変化（特徴量が「０」であるか、または、「１」であるかの判断の変化）が防止される。この点は、発明者の実験によって確認済みである。これらによって、検索速度を高速化させ、かつ、検索精度を向上させることができる。

【００１０】

上記第１の局面による音源検索装置において、好ましくは、音源データベースは、データベース用楽曲信号から特徴量を抽出するデータベース側特徴量抽出手段と、抽出した特徴量を２値化するデータベース側２値化手段と、データベース側２値化手段により２値化されたデータベース用楽曲信号の特徴量から音源データベースを構築する構築手段とを含み、データベース用楽曲信号の特徴量は、データベース用楽曲信号のフーリエスペクトルに基づいて算出される各帯域窓の出力エネルギであるクロマスペクトルであり、データベ

ース用楽曲信号の特徴量は、所定の時間長さを有する１分析フレーム毎または複数の分析フレーム毎に抽出されており、データベース側２値化手段は、データベース用楽曲信号の特徴量が、１分析フレーム毎または複数の分析フレーム毎における特徴量の所定の基準値以上の場合に特徴量を１とし、所定の基準値未満の場合に特徴量を０とするように構成されている。このように構成すれば、音源データベースの特徴量が２値化されるので、スカラー量（たとえば、単精度浮動小数点数、３２ｂｉｔ）の要素を有する特徴量から音源データベースが構築される場合と比べて、特徴量が２値化される分、次元が小さくなる（１ｂｉｔ）ので、音源データベースのデータベースサイズを小さくすることができる。その結果、検索速度を高速化させることができる。

【００１１】

この発明の第２の局面における音源検索方法は、楽曲信号と音声信号とを含む混合音から特徴量を抽出する工程と、抽出した特徴量を２値化する工程と、２値化された混合音の特徴量を検索キーとして、音源データベースから音源を検索する工程とを備え、混合音の特徴量は、所定の時間長さを有する１分析フレーム毎または複数の分析フレーム毎に抽出されており、楽曲信号と音声信号とを含む混合音から特徴量を抽出する工程は、混合音のフーリエスペクトルに基づいて算出される各帯域窓の出力エネルギであるクロマスペクトルを特徴量として抽出する工程を含み、抽出した特徴量を２値化する工程は、混合音の特徴量が、１分析フレーム毎または複数の分析フレーム毎における特徴量のクロマスペクトルに基づいた所定の基準値以上の場合に特徴量を１とし、所定の基準値未満の場合に特徴量を０とする工程を含む。

【００１２】

この発明の第２の局面による音源検索方法では、上記のように、抽出した特徴量を２値化する工程を備えることによって、スカラー量（たとえば、単精度浮動小数点数、３２ｂｉｔ）の要素を有する特徴量を検索キーとして音源データベースから音源を検索する場合と比べて、特徴量が２値化される分、次元が小さくなる（１ｂｉｔ）とともに、２値化された特徴量は、楽曲信号の包絡（形状）を表しながら、混合される音声信号に対して頑強でかつ楽曲信号の音圧の変化に対して不変となるので、検索速度を高速化させ、かつ、検索精度を向上させることが可能な音源検索方法を提供することができる。

【発明の効果】

【００１３】

本発明によれば、上記のように、混合音を検索キーとする音源検索装置および音源検索方法において、検索速度を高速化させ、かつ、検索精度を向上させることができる。

【図面の簡単な説明】

【００１４】

【図１】本発明の一実施形態による音源検索装置のブロック図である。

【図２】混合音の波形の模式図である。

【図３】図２の混合音のフーリエスペクトルを示す模式図である。

【図４】図３の混合音のフーリエスペクトルから求められたクロマスペクトルを示す模式図である。

【図５】本発明の一実施形態による音源データベースのブロック図である。

【図６】本発明の一実施形態による音源データベースの構築方法のフロー図である。

【図７】本発明の一実施形態による音源検索方法のフロー図である。

【図８】混合音の特徴量と音源データベースに記憶された楽曲の特徴量との間の距離の頻度を示す図である。

【図９】１楽曲分の２値化されていないクロマスペクトルを示す図である。

【図１０】図９よりも１０ｄＢ大きい１楽曲分の２値化されていないクロマスペクトルを示す図である。

【図１１】１楽曲分の２値化されたクロマスペクトルを示す図である。

【図１２】図１１よりも１０ｄＢ大きい１楽曲分の２値化されたクロマスペクトルを示す図である。
【図１３】比較例による音源検索装置の検索結果（Ｆ値）を示す図である。
【図１４】本発明の一実施形態による音源検索装置の検索結果（Ｆ値）を示す図である。
【発明を実施するための形態】
【００１５】
以下、本発明の実施形態を図面に基づいて説明する。
【００１６】
[音源検索装置の構成]
図１～図４を参照して、本実施形態による音源検索装置１００の構成について説明する。音源検索装置１００は、混合音を構成する音源を、後述する音源データベース５０から検索するように構成されている。
【００１７】
図１に示すように、本実施形態による音源検索装置１００は、特徴量抽出手段１０と、２値化手段２０と、検索手段３０とを備えている。また、本実施形態では、音源検索装置１００の検索性能を評価するために、混合音は、混合手段４０により生成されるように構成されている。なお、特徴量抽出手段１０と、２値化手段２０と、検索手段３０と、混合手段４０とは、たとえば、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）などの制御部により構成されている。なお、特徴量抽出手段１０および２値化手段２０は、本発明の「検索装置側特徴量抽出手段」および「検索装置側２値化手段」の一例である。以下、具体的に説明する。
【００１８】
（混合手段）
混合手段４０は、楽曲信号と音声信号とを混合（編集）することにより、混合音を生成するように構成されている。なお、楽曲とは、楽器による演奏のみの場合と、楽器による演奏および歌声とを含む場合とを意味する。また、音声とは、雑音（ノイズ）ではない音声を意味する。たとえば、混合音とは、テレビの番組中におけるナレーションの音声と、その背景で流されるＢＧＭとにより構成される音である。
【００１９】
混合音は、複数の音源が任意の割合で重み付け加算された音である。混合音の時間波形をｋ（ｔ）とし、Ｊ個の音源s_j（ｔ）がw_jで重み付けされたとすると、混合音は、下記の式（１）により表される。
【数１】

$$k(t)=\sum_{j=1}^{J} w_j \cdot s_j(t) \quad \cdots (1)$$

【００２０】
ここで、ｊ＝１，．．．，Ｊで、ｔは、時間を表す。音源検索の一般形は、ｋ（ｔ）を検索キーとして、Ｊ個の音源s_1（ｔ），．．．，s_J（ｔ）を音源データベース５０内から検索するものである。本実施形態において、Ｊ＝２で、重み（ｗ）は、任意であるとすると、上記の式（１）は、下記の式（２）となる。
【数２】

$$k(t)=w_1 s_1(t)+w_2 s_2(t) \quad \cdots (2)$$

【００２１】
また、２つの音源s_1（ｔ）およびs_2（ｔ）は、楽曲信号s_1（ｔ）と、音声信号s_2（ｔ）とする。このように、本実施形態の音源検索装置１００は、混合音であるｋ（ｔ）の特徴量を検索キーとして、音源データベース５０からs_1（ｔ）の特徴量を検索するように構成されている。
【００２２】

（特徴量抽出手段）
図１に示すように、特徴量抽出手段１０には、混合手段４０によって生成された楽曲信号と音声信号とを含む混合音が入力されるように構成されている。そして、特徴量抽出手段１０は、楽曲信号と音声信号とを含む混合音から特徴量を抽出するように構成されている。具体的には、混合音の特徴量は、混合音のフーリエスペクトルに基づいて算出される各帯域窓の出力エネルギであるクロマスペクトルである。以下、混合音の特徴量の抽出について、具体的に説明する。
【００２３】
混合音の特徴量ベクトルをｋ（ｎ）、楽曲の特徴量ベクトルをs_1（ｎ）、音声の特徴量ベクトルをs_2（ｎ）とする。ただし、ｎは、分析フレーム番号である。なお、分析フレームの説明は、後述する。また、各ベクトルは、Ｄ次元であり、混合音の特徴量ベクトルを、ｋ（ｎ）＝［ｋ（ｎ，１）、ｋ（ｎ，２），．．．，ｋ（ｎ，Ｄ）］^{T}とする。ここで、^{T}は、ベクトルの転置を表す。s_1（ｎ）およびs_2（ｎ）も同様に表される。
【００２４】
（分析フレーム）
次に、図２を参照して、分析フレームについて説明する。図２では、横軸は、時間（ｔ）を表し、縦軸は、混合音の振幅を表す。そして、図２の混合音の波形を、所定の時間長さＴ（たとえば、１ｓ）毎に取り出す。具体的には、ある時刻ｔ１を先頭に、所定の時間長さＴの分析フレームに窓（たとえば、ハミング窓）をかけて取り出す。また、時刻ｔ２から一定の時間（フレームシフト長）経過後の時刻ｔ２を先頭に、所定の時間長さＴの分析フレームにハミング窓をかけて取り出す。以下、同様に、混合音の全ての領域において、混合音の波形を分析フレーム毎に取り出す。
【００２５】
（クロマスペクトル）
次に、図３および図４を参照して、クロマスペクトルについて説明する。図３では、横軸は、周波数を表し、縦軸は、フーリエスペクトルの振幅を表す。所定の時間長さＴの分析フレーム毎に取り出された混合音（図２参照）について、フーリエスペクトル（図３の実線）が算出される。そして、フーリエスペクトルから各帯域窓の出力エネルギ（クロマスペクトル）が算出される。具体的には、ピアノの鍵盤に対応する各帯域窓（図３の三角形の点線で囲まれた領域、フィルタバンク）を設定する。なお、帯域窓は、周波数が高くなるほど、幅が広い三角形になる。そして、各帯域窓に含まれるフーリエスペクトルを積分することにより、図４に示すように、出力エネルギ（クロマスペクトル）が算出される。なお、図４では、横軸は、周波数を表し、縦軸は、クロマスペクトルの大きさを表す。
【００２６】
ここで、１オクターブの音程には、ピアノの鍵盤に対応するように、１２個の帯域窓（Ａ、Ａ＃、Ｂ、Ｃ、Ｃ＃、Ｄ、Ｄ＃、Ｅ、Ｆ、Ｆ＃、Ｇ、Ｇ＃）が存在する。本実施形態では、６オクターブ分の帯域窓（７２個＝１２×６）について、クロマスペクトル（特徴量）を算出する。これにより、混合音の特徴量ベクトルｋ（ｎ）は、７２次元の次元Ｄを有する。
【００２７】
（２値化手段）
ここで、本実施形態では、２値化手段２０は、抽出した特徴量を２値化するように構成されている。具体的には、図４に示すように、２値化手段２０は、混合音の特徴量（クロマスペクトル）が、１分析フレーム毎における特徴量の所定の基準値（具体的には、平均値）（図４の点線参照）以上の場合に特徴量を１とし、１分析フレーム毎における特徴量の所定の基準値未満の場合に特徴量を０とするように構成されている。すなわち、２値化手段２０は、スカラー量の要素を有するクロマスペクトルを、２値化するように構成されている。
【００２８】
具体的には、時刻ｔのクロマスペクトルをｃ（ｔ）＝［c_1（ｔ），c_2（ｔ），．．

．，ｃ$_D$（ｔ）］^{T}とする。ここで、Ｄは、ベクトルの次元数を表し、^{T}は、ベクトルの転置を表す。そして、２値化されたクロマスペクトルｂ（ｔ）＝［ｂ$_1$（ｔ），ｂ$_2$（ｔ），．．．，ｂ$_D$（ｔ）］^{T}は、下記の式（３）により表される。

【数３】

$$b_d(t)=\begin{cases}1, & c_d(t)\geq\dfrac{1}{D}\sum_{i=1}^{D}c_i(t)\\ 0, & c_d(t)<\dfrac{1}{D}\sum_{i=1}^{D}c_i(t)\end{cases}\quad\cdots(3)$$

$$d=1,2,...,D$$

【００２９】

（検索手段）

検索手段３０は、２値化手段により２値化された混合音の特徴量を検索キーとして、音源データベース５０から音源を検索するように構成されている。具体的には、検索手段３０は、複数（Ｐ個）の分析フレーム（累積分析フレーム）に対応する２値化された混合音の特徴量を検索キーとして、音源データベース５０から音源を検索するように構成されている。すなわち、音源の検索は、Ｐ個の特徴量ベクトルの列を検索キーとした類似パターン検索問題に帰着する。具体的には、ｎ番目からｎ＋Ｐ－１番目の分析フレームの特徴量ベクトルは、下記の式（４）～式（６）により表される。

【数４】

$$\mathbf{K}(n)=[\mathbf{k}(n),...,\mathbf{k}(n+P-1)]\quad\cdots(4)$$

$$\mathbf{S}_1(n)=[\mathbf{s}_1(n),...,\mathbf{s}_1(n+P-1)]\quad\cdots(5)$$

$$\mathbf{S}_2(n)=[\mathbf{s}_2(n),...,\mathbf{s}_2(n+P-1)]\quad\cdots(6)$$

【００３０】

ここで、Ｖ個の楽曲信号と、Ｗ個の音声信号とがあるとすると、Ｓ$_{1,v}$（ｎ）およびＳ$_{2,w}$（ｍ）を、ｖ番目およびｗ番目の特徴量とする。そして、ｖ番目の楽曲信号と、ｗ番目の音声信号とが混合された混合音の特徴量をＫ$_{v,w}$（ｎ）とすると、検索は、ｖ，ｗ，ｎ，ｍが未知の条件で、Ｋ$_{v,w}$（ｎ）から楽曲番号ｖ＊と、分析フレーム番号ｎ＊とを推定する問題となる。検索処理をｓｅａｒｃｈと表すと、検索は、下記の式（７）により表される。

【数５】

$$[v^*,n^*]=\underset{v,n}{search}(\mathbf{K}_{v,w}(n))\quad\cdots(7)$$

【００３１】

すなわち、検索処理ｓｅａｒｃｈは、検索の結果に該当する項目（特徴量ベクトル間の距離が最小の項目）を１組決定することになる。つまり、検索キーの特徴量ベクトルと検索対象の特徴量ベクトルとの間の距離が最小になる場合を検索結果とする。

【００３２】

（平均誤棄却率および平均誤検出率）

検索の性能は、誤棄却（Ｍｉｓｓ）と誤検出（Ｆａｌｓｅ　Ａｌａｒｍ）との２つの指標により評価される。誤棄却は、検索結果に、混合音を構成する楽曲信号に対応する［ｖ，ｎ］が含まれない場合に相当する。また、誤検出は、検索結果に、混合音を構成する楽曲信号以外の［ｖ，ｎ］（検索キーに無関係な楽曲）が含まれる場合に相当する。

【００３３】

Ｑ回検索する例において、平均誤棄却率と平均誤検出率とを説明する。ｑ回目の検索キーをＫ$_{v(q),w(q)}$（ｎ$^{(q)}$）とし、得られる集合をφ$^{(q)}$とし、Ｋ$_{v(q),w(q)}$（ｎ$^{(q)}$）の構成音源をＳ$_{1,v'(q)}$（ｎ'$^{(q)}$）とすると、平均誤棄却率は、下記の式（８）により表される。

【数６】

$$\overline{M} = 1 - \frac{1}{Q}\sum_{q=1}^{Q} c,$$

$$c = \begin{cases} 1, & if\left[v'^{(q)}, n'^{(q)}\right] \in \phi^{(q)} \\ 0, & otherwise \end{cases} \quad \cdots (8)$$

【００３４】

また、平均誤検出率は、下記の式（９）により表される。

【数７】

$$\overline{F} = \frac{1}{Q}\sum_{q=1}^{Q} \frac{|\phi^{(q)} \setminus [v'^{(q)}, n'^{(q)}]|}{I - 1} \quad \cdots (9)$$

ここで、Ｉは、音源データベース５０中の［ｖ，ｎ］が取り得る組の総数を表す。また、φ（ｑ）＼［ｖ′（ｑ），ｎ′（ｑ）］は、集合φ（ｑ）から、要素［ｖ′（ｑ），ｎ′（ｑ）］を取り除く処理を意味する。また、｜｜は、集合の要素数を求める処理を意味する。そして、平均誤棄却率および平均誤検出率は、共に、値が小さいほど、検索性能が高いことを表す。

【００３５】

（音源データベース）

音源データベース５０には、複数の楽曲が記憶されている。具体的には、上記の混合音と同様に、複数の楽曲の特徴量が２値化された状態で、音源データベース５０に記憶されている。

【００３６】

詳細には、図５に示すように、本実施形態では、音源データベース５０は、データベース用楽曲信号から特徴量を抽出する特徴量抽出手段５１と、抽出した特徴量を２値化する２値化手段５２と、２値化手段５２により２値化されたデータベース用楽曲信号の特徴量から音源データベース５０を構築する構築手段５３とを含む。なお、特徴量抽出手段５１と、２値化手段５２と、構築手段５３とは、たとえば、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）などの制御部により構成されている。なお、特徴量抽出手段５１および２値化手段５２は、それぞれ、本発明の「データベース側特徴量抽出手段」および「データベース側２値化手段」の一例である。

【００３７】

ここで、データベース用楽曲信号の特徴量は、データベース用楽曲信号のフーリエスペクトルに基づいて算出される各帯域窓の出力エネルギであるクロマスペクトルである。また、データベース用楽曲信号の特徴量は、所定の時間長さを有する１分析フレーム毎に抽出されている。そして、２値化手段５２は、データベース用楽曲信号の特徴量が、１分析フレーム毎における特徴量の所定の基準値（具体的には、平均値）以上の場合に特徴量を１とし、所定の基準値未満の場合に特徴量を０とするように構成されている。なお、データベース用楽曲信号からの特徴量の抽出、抽出した特徴量の２値化の詳細は、上記音源検索装置１００と同様である。

【００３８】

［音源データベースの構築方法］

次に、図６を参照して、本実施形態による音源データベース５０の構築方法を説明する。

【００３９】

まず、ステップＳ１１において、特徴量抽出手段５１に入力されたデータベース用楽曲信号から特徴量が抽出される。具体的には、１分析フレーム毎に混合音のフーリエスペクトルが算出された後、各帯域窓の出力エネルギであるクロマスペクトルが算出される。

【００４０】

次に、ステップＳ１２において、２値化手段５２により、抽出した特徴量が、上記式（３）に基づいて２値化される。そして、ステップＳ１３において、構築手段５３により、２値化されたデータベース用楽曲信号の特徴量から音源データベース５０が構築される。

【００４１】

［音源検索方法］

次に、図７を参照して、本実施形態による音源検索方法を説明する。

【００４２】

まず、ステップＳ１において、特徴量抽出手段１０に入力された楽曲信号と音声信号とを含む混合音から特徴量が抽出される。具体的には、１分析フレーム毎に混合音のフーリエスペクトルが算出された後、各帯域窓の出力エネルギであるクロマスペクトルが算出される。

【００４３】

次に、ステップＳ２において、抽出した特徴量が、上記式（３）に基づいて２値化される。

【００４４】

次に、ステップＳ３において、２値化された混合音の特徴量を検索キーとして、音源データベース５０から音源が検索される。具体的には、検索キーの２値化された特徴量ベクトルと、音源データベース５０に記憶されている楽曲の２値化された特徴量ベクトルとの間の距離が、最小の場合、混合音に合致する楽曲が検索（検出）されたと判断される。

【００４５】

次に、ステップＳ４において、検索結果の評価が行われる。具体的には、上記の式（８）および式（９）により、平均誤棄却率および平均誤検出率が算出される。さらに、算出された平均誤棄却率および平均誤検出率から、Ｆ値（調和平均）が算出される。

【００４６】

（クロマスペクトルが楽曲を表現するのに適しているか否かを確認する実験）

図８を参照して、クロマスペクトルが楽曲を表現するのに適しているか否かを確認するために行った実験について説明する。

【００４７】

まず、７１曲分の楽曲のクロマスペクトルをデータベース化した。そして、７１曲分の楽曲のうち、１０秒間分の長さの楽曲のクロマスペクトルをランダムに５００個選択して、検索キーとした。そして、５００個の検索キーの特徴量ベクトルと、７１曲分の楽曲の特徴量ベクトルとの間の距離分布を作成した。図８では、横軸は距離を表し、縦軸は、頻度を表している。この実験では、距離が０になる箇所が５００箇所あることが確認された。すなわち、１つの検索キーに対して、距離が０になる箇所が１箇所であることが確認された。これにより、楽曲（１０秒間分の長さの楽曲）の構造に数値的な繰り返し（同じ特徴量の繰り返し）が存在しないことが確認された。すなわち、クロマスペクトルが楽曲を表現するのに適している（楽曲の特徴量として適している）ことが確認された。

【００４８】

また、図８に示すように、頻度は、距離０から徐々に増加し、その後、頻度が急激に増加した後、頻度が急激に低下することが判明した。すなわち、頻度は、概ね１つの凸形状に形成されることが判明した。

【００４９】

（クロマスペクトルの２値化についての実験）

次に、図９～図１２を参照して、クロマスペクトルの２値化についての実験について説明する。なお、図９～図１２では、クロマスペクトルの値が大きい部分ほど、色が濃くなるように表されている。

【００５０】

図９は、２値化されていない１楽曲分のクロマスペクトルである。図１０は、図９に示された楽曲の信号波形のエネルギを相対的に１０ｄＢ高くした場合の、２値化されていない１楽曲分のクロマスペクトルである。図１０に示すように、楽曲の信号波形のエネルギ

を相対的に１０ｄＢ高くした場合では、全体的にクロマスペクトルの値が大きくなる（色が濃くなる）ことが判明した。すなわち、２値化されていないクロマスペクトル（特徴量）は、楽曲の信号波形のエネルギの変化に伴って変化することが確認された。

【００５１】

図１１は、２値化された１楽曲分のクロマスペクトルである。図１２は、図１１に示された楽曲の信号波形のエネルギを相対的に１０ｄＢ高くした場合の、２値化された１楽曲分のクロマスペクトルである。図１２に示すように、楽曲の信号波形のエネルギを相対的に１０ｄＢ高くした場合でも、２値化された１楽曲分のクロマスペクトルは、パターンが完全に一致することが判明した。すなわち、２値化されたクロマスペクトル（特徴量）は、楽曲の信号波形のエネルギの変化に対して不変であることが確認された。

【００５２】

（音源検索の実験）

次に、図１３および図１４を参照して、本実施形態による音源検索装置１００による音源検索の実験について、比較例による音源検索装置と比較しながら説明する。

【００５３】

比較例による音源検索装置では、特徴量抽出手段と検索手段とを備えている一方、本実施形態による音源検索装置１００のように２値化手段２０は備えていない。すなわち、比較例による音源検索装置では、特徴量は、クロマスペクトルの値そのもの（単精度浮動小数点数、３２ｂｉｔ）である。つまり、特徴量の次元は、６オクターブ分の７２次元×３２ｂｉｔである。一方、本実施形態による音源検索装置１００では、特徴量（クロマスペクトル）が２値化されているので、特徴量の次元は、６オクターブ分の７２次元×１ｂｉｔである。

【００５４】

図１３および図１４に示すように、音源検索の実験では、音声信号に対する楽曲信号の音圧の相対的な大きさ（音圧比）を、５ｄＢ小さくした混合音（混合比－５ｄＢ）と、互いに等しい混合音（混合比０ｄＢ）と、５ｄＢ大きくした混合音（混合比５ｄＢ）と、１０ｄＢ大きくした混合音（混合比１０ｄＢ）と、１５ｄＢ大きくした混合音（混合比１５ｄＢ）と、２０ｄＢ大きくした混合音（混合比２０ｄＢ）とを準備して、各々の混合音について、音源検索を実施するとともに、検索結果のＦ値を算出した。なお、たとえば、ナレーションの背景でＢＧＭが流れる場合の音圧比は、－５ｄＢ～０ｄＢに相当する。

【００５５】

また、音源検索の実験では、帯域窓（フィルタバンク）を、５５（Ｈｚ）～３５２０（Ｈｚ）とする６オクターブ（７２バンク）により構成した。また、信号のサンプリングレートを、１６０００Ｈｚとした。また、分析フレーム長を、１６．３８４（ｓ）とし、フレームシフト長を、１／１６（ｓ）とした。また、音源データベース５０には、市販のＣＤの７２曲（約２００，０００分析フレーム）分の楽曲を記憶した。

【００５６】

そして、累積フレーム数として、１０（ｓ）区間（１６×１０分析フレーム）と、２（ｓ）区間（１６×２分析フレーム）とを採用した。そして、これらを特徴量（検索キー）として、音源データベース５０に記憶された約２００，０００通りの候補から、連続する１６×１０分析フレーム（１６×２分析フレーム）がマッチする時刻を検索した。

【００５７】

楽曲信号に混合する音声信号は、ＪＮＡＳ（“ＪＮＡＳ：Ｊａｐａｎｅｓｅ ｓｐｅｅｃｈ ｃｏｒｐｕｓ ｆｏｒ ｌａｒｇｅ ｖｏｃａｂｕｌａｒｙ ｃｏｎｔｉｎｕｏｕｓ ｓｐｅｅｃｈ ｒｅｃｏｇｎｉｔｉｏｎ ｒｅｓｅａｒｃｈ”，Ｊ．Ａｃｏｕｓｔ．Ｓｏｃ．Ｊｐｎ（Ｅ）２０（３），ｐｐ．１９９－２０６，１９９９．）から、男女の発話を発話間ポーズを開けずに接続し準備した。そして、１つの楽曲に渡って発話がナレーションのようになるように混合した。

【００５８】

さらに、信号の振幅（相対値）を、１倍、０．５倍、２倍、０．１倍、１０倍にそれぞ

れ変化させた場合において、音源検索を実行した。

【００５９】

特徴量を２値化しない比較例による音源データベースのデータベースサイズ（３２、図１３参照）は、特徴量を２値化した本実施形態の音源データベース５０のデータベースサイズ（１、図１４参照）に比べて、３２倍の大きさになることが確認された。

【００６０】

また、比較例による音源検索装置では、相対処理時間（検索時間）が、２５０または１１３０であったのに対して、本実施形態による音源検索装置１００では、相対処理時間（検索時間）が、３４または１７０であった。これにより、特徴量の２値化を行うことにより、検索速度が高速化されることが確認された。

【００６１】

また、比較例による音源検索装置では、信号の振幅（１倍、０．５倍、２倍、０．１倍、１０倍）の変化に対して、Ｆ値の値が著しく変化していることが判明した。一方、本実施形態による音源検索装置１００では、信号の振幅の変化に対して、Ｆ値の値が不変であることが判明した。これは、図１１および図１２に示すように、２値化されたクロマスペクトル（特徴量）は、楽曲の信号波形のエネルギの変化に対して不変であることから、このような結果が得られたと考えられる。

【００６２】

また、信号の振幅が１倍の場合には、比較例による音源検索装置によるＦ値の方が高くなる場合がある一方、信号の振幅が１倍以外の０．５倍、２倍、０．１倍、１０倍では、全ての場合において、本実施形態による音源検索装置１００によるＦ値の方が高くなることが判明した。これにより、全体として、特徴量の２値化を行うことにより、検索精度が向上することが確認された。

【００６３】

[本実施形態の効果]

本実施形態では、以下のような効果を得ることができる。

【００６４】

本実施形態では、上記のように、抽出した特徴量を２値化する２値化手段２０を備えることによって、スカラー量（たとえば、単精度浮動小数点数、３２ｂｉｔ）の要素を有する特徴量を検索キーとして音源データベース５０から音源を検索する場合と比べて、特徴量が２値化される分、次元が小さくなる（１ｂｉｔ）ので、検索速度を高速化させることができる。

【００６５】

また、楽曲信号に音声信号を混合した場合、音声信号が混合される分、楽曲信号の包絡（形状）が変化する。そこで、本実施形態では、上記のように、抽出した特徴量を２値化する２値化手段２０を備えることによって、２値化後の特徴量のうち、「１」の部分は、音声信号が加法的に作用している限り、「１」のままである。一方、２値化後の特徴量のうち、「０」の部分に音声信号が加法的に作用しても、２値化するための基準値を超えない限り「０」のままである。なお、２値化するためのしきい値近傍では、音声信号が混合されることにより、２値化後の特徴量の「０」または「１」が反転する場合がある一方、音声信号の出力エネルギが大きい周波数（反転する可能性がある周波数）は基本周波数の整数倍の周波数近傍のみの比較的小さい範囲であるため、反転による影響は小さいと考えられる。その結果、２値化された特徴量は、楽曲信号の包絡（形状）を表しながら、混合される音声信号に対して頑強な特徴量となる。また、楽曲信号の音圧の変化に対しても、混合音の音量の変化に伴って所定の基準値も変化させることが可能であるので、混合音の特徴量の変化（特徴量が「０」であるか、または、「１」であるかの判断の変化）が防止される。これらによって、検索速度を高速化させ、かつ、検索精度を向上させることができる。

【００６６】

また、本実施形態では、上記のように、混合音の特徴量を、所定の時間長さＴを有する

１分析フレーム毎に抽出して、２値化手段２０を、混合音の特徴量が、１分析フレーム毎における特徴量の平均値以上の場合に特徴量を１とし、１分析フレーム毎における特徴量の平均値未満の場合に特徴量を０とするように構成する。これにより、１分析フレーム毎における特徴量の平均値に基づいて特徴量が２値化されるので、混合音の音量の変化に適切に対応させて、特徴量を２値化することができる。

【００６７】

また、本実施形態では、上記のように、混合音の特徴量は、混合音のフーリエスペクトルに基づいて算出される各帯域窓の出力エネルギであるクロマスペクトルであり、２値化手段２０を、クロマスペクトルを２値化するように構成する。これにより、クロマスペクトルを特徴量として、適切に音源を検索することができる。

【００６８】

また、本実施形態では、上記のように、混合音の特徴量は、所定の時間長さＴを有する１分析フレーム毎に抽出されており、検索手段３０を、複数の分析フレームに対応する２値化された混合音の特徴量を検索キーとして、音源データベース５０から音源を検索するように構成する。これにより、１つの分析フレームに対応する２値化された混合音の特徴量を検索キーとして検索する場合と比べて、検索キーの特徴量（情報量）が多くなるので、検索の精度を高めることができる。

【００６９】

また、本実施形態では、上記のように、検索手段３０を、１０×１６分析フレームまたは２×１６分析フレームに対応する２値化された混合音の特徴量を検索キーとして、音源データベース５０から音源を検索するように構成する。これにより、１０×１６分析フレームまたは２×１６分析フレームの比較的短い複数の分析フレームに対応する２値化された混合音の特徴量を検索キーとして検索が行われた場合でも、特徴量の２値化により、高速、かつ、高精度な検索を行うことができる。

【００７０】

また、本実施形態では、上記のように、音源データベース５０は、データベース用楽曲信号から特徴量を抽出する特徴量抽出手段５１と、抽出した特徴量を２値化する２値化手段５２と、２値化手段５２により２値化されたデータベース用楽曲信号の特徴量から音源データベース５０を構築する構築手段５３とを含む。これにより、音源データベース５０の特徴量が２値化されるので、スカラー量（たとえば、単精度浮動小数点数、３２ｂｉｔ）の要素を有する特徴量から音源データベース５０が構築される場合と比べて、特徴量が２値化される分、次元が小さくなる（１ｂｉｔ）ので、音源データベース５０のデータベースサイズを小さくすることができる。その結果、検索速度を高速化させることができる。

【００７１】

[変形例]

なお、今回開示された実施形態は、すべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は、上記した実施形態の説明ではなく特許請求の範囲によって示され、さらに特許請求の範囲と均等の意味および範囲内でのすべての変更（変形例）が含まれる。

【００７２】

たとえば、上記実施形態では、混合音およびデータベース用楽曲信号の特徴量が、特徴量の平均値以上の場合に特徴量を１とし、特徴量の平均値未満の場合に特徴量を０とすることにより、特徴量を２値化する例を示したが、本発明はこれに限られない。本発明では、特徴量の平均値以外の値を基準として、特徴量を１または０にしてもよい。たとえば、特徴量の中央値などを基準として、特徴量を１または０にしてもよい。

【００７３】

また、上記実施形態では、混合音およびデータベース用楽曲信号の特徴量は、所定の時間長さを有する１分析フレーム毎に抽出される例を示したが、本発明はこれに限られない。たとえば、混合音およびデータベース用楽曲信号の特徴を、複数の分析フレーム毎に抽

出してもよい。

【００７４】

また、上記実施形態（実験）では、１０×１６分析フレームまたは２×１６分析フレームに対応する２値化された混合音の特徴量を検索キーとして、音源データベースから音源を検索するように構成されている例を示したが、本発明はこれに限られない。たとえば、１０×１６分析フレームまたは２×１６分析フレーム以外の数の分析フレームに対応する２値化された混合音の特徴量を検索キーとして用いてもよい。

【００７５】

また、上記実施形態では、特徴量ベクトルが、６オクターブ分の次元（７２次元）を有する例を示したが、本発明はこれに限られない。たとえば、特徴量ベクトルが、６オクターブ以外の数のオクターブ分の次元を有するように構成されていてもよい。

【００７６】

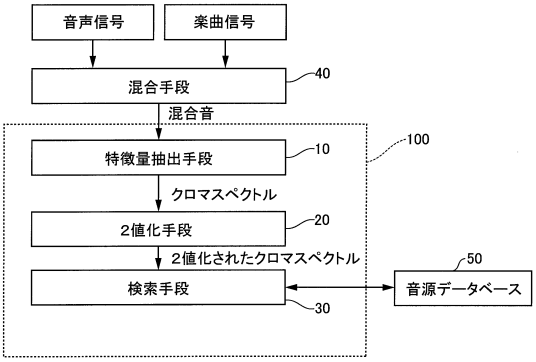
また、上記実施形態では、検索キーの特徴量ベクトルと検索対象の特徴量ベクトルとの間の距離が最小になる場合を検索結果とする例を示したが、本発明はこれに限られない。たとえば、２値化手段により２値化された混合音の特徴量と、音源データベースの音源の特徴量との差が、所定のしきい値未満の場合に、検索結果とするように構成してもよい。これにより、混合音を検索キーとする場合において、検索キーの特徴量ベクトルと検索対象（検索したい正解の音源）の特徴量ベクトルとの間の距離が最小にならない場合でも、検索したい音源が検索できなくなるのを防止することができる。

【符号の説明】

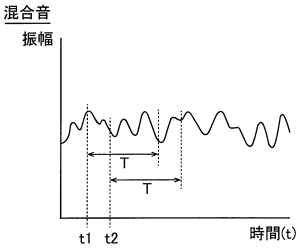
【００７７】

１０　特徴量抽出手段（検索装置側特徴量抽出手段）
２０　２値化手段（検索装置側特徴量抽出手段）
３０　検索手段
５０　音源データベース
５１　特徴量抽出手段（データベース側特徴量抽出手段）
５２　２値化手段（データベース側２値化手段）
１００　音源検索装置
Ｔ　（所定の）時間長さ

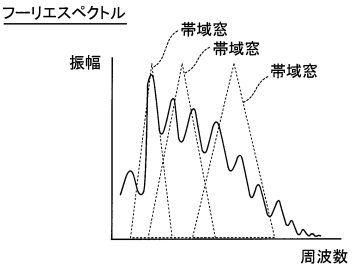
【図１】



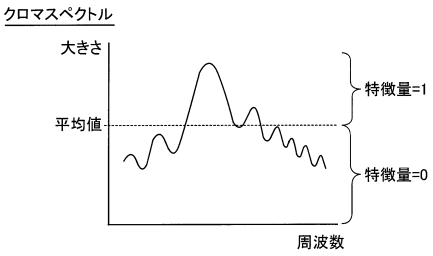
【図２】



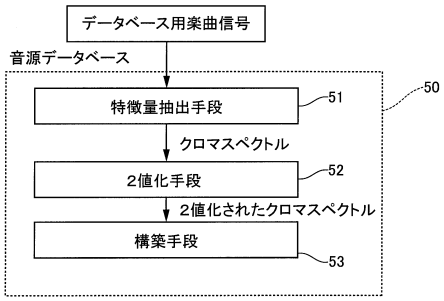
【図３】



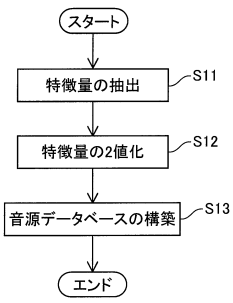
【図４】



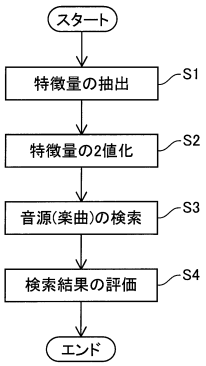
【図５】



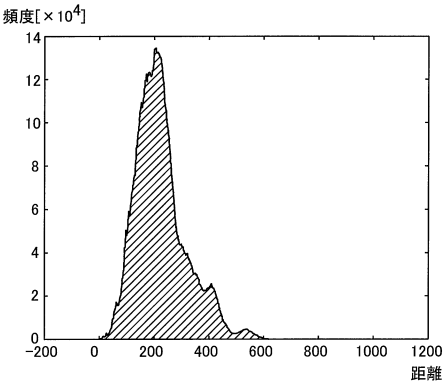
【図６】



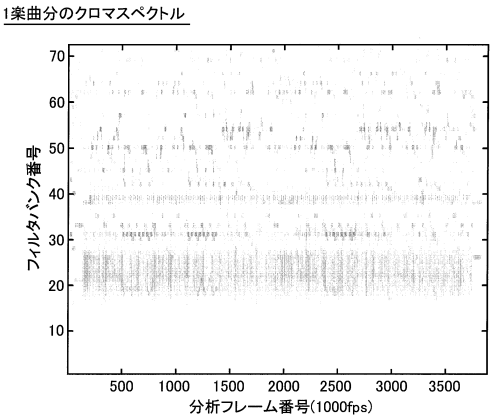
【図７】



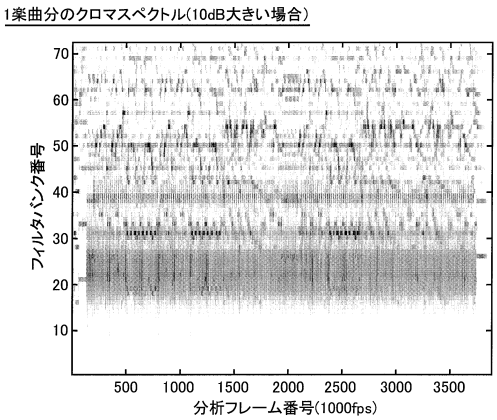
【図８】



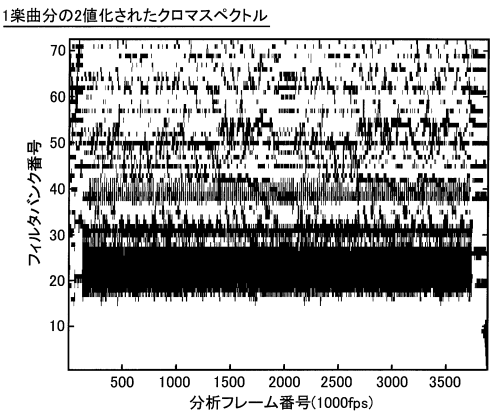
【図９】



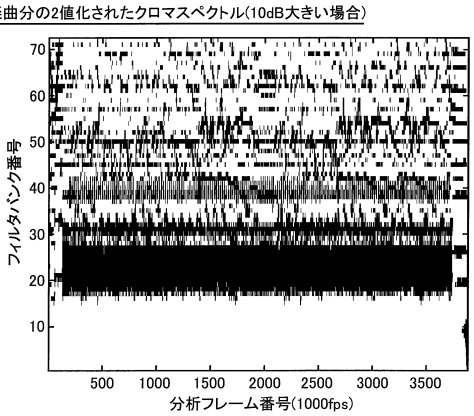
【図１０】



【図１１】



【図１２】



【図１３】

比較例による音源検索装置のF値

信号の振幅(相対値)	1倍		0.5倍		2倍		0.1倍		10倍	
累積フレーム数(秒)	10	2	10	2	10	2	10	2	10	2
混合比(-5)	0.88	0.43	0.07	0	0.58	0.13	0	0	0.01	0
混合比(0)	1	0.76	0.03	0	0.9	0.29	0	0	0.01	0
混合比(5)	1	0.91	0.02	0	0.98	0.47	0	0	0.01	0
混合比(10)	1	0.99	0	0	0.99	0.58	0	0	0.01	0
混合比(15)	1	1	0	0	0.99	0.69	0	0	0.01	0
混合比(20)	1	1	0	0	0.99	0.72	0	0	0.02	0
相対処理時間	1130	250	1130	250	1130	250	1130	250	1130	250
データベースサイズ(相対値)	32	32	32	32	32	32	32	32	32	32

【図１４】

本実施形態による音源検索装置のF値

信号の振幅(相対値)	1倍		0.5倍		2倍		0.1倍		10倍	
累積フレーム数(秒)	10	2	10	2	10	2	10	2	10	2
混合比(-5)	0.52	0.17	0.52	0.17	0.52	0.17	0.52	0.17	0.52	0.17
混合比(0)	0.95	0.5	0.95	0.5	0.95	0.5	0.95	0.5	0.95	0.5
混合比(5)	0.98	0.74	0.98	0.74	0.98	0.74	0.98	0.74	0.98	0.74
混合比(10)	1	0.88	1	0.88	1	0.88	1	0.88	1	0.88
混合比(15)	1	0.97	1	0.97	1	0.97	1	0.97	1	0.97
混合比(20)	1	1	1	1	1	1	1	1	1	1
相対処理時間	170	34	170	34	170	34	170	34	170	34
データベースサイズ (相対値)	1	1	1	1	1	1	1	1	1	1

フロントページの続き

(56)参考文献 国際公開第２００６／００６５２８（WO，Ａ１）
米国特許出願公開第２００７／０１４３１０８（ＵＳ，Ａ１）
特開２０１２－２２６０８０（ＪＰ，Ａ）
米国特許出願公開第２０１２／０２６６７４３（ＵＳ，Ａ１）
米国特許出願公開第２０１１／０３１４９９５（ＵＳ，Ａ１）
樋口　颯、高橋　徹，特徴量間の累積距離を用いた混合音からの音源検索システムの評価，電子情報通信学会技術研究報告　Ｖｏｌ．１１４　Ｎｏ．１９１，日本，一般社団法人電子情報通信学会，２０１４年　8月２１日，p.19～24

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／６３２
Ｇ１０Ｌ　２５／１８
Ｇ１０Ｌ　２５／５４