

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2016-162411

(P2016-162411A)

(43) 公開日 平成28年9月5日(2016.9.5)

(51) Int.Cl.	F I	テーマコード (参考)
G 0 6 F 17/30 (2006.01)	G 0 6 F 17/30	1 7 0 E
G 1 0 L 25/54 (2013.01)	G 0 6 F 17/30	3 5 0 C
G 1 0 L 25/18 (2013.01)	G 1 0 L 25/54	
	G 1 0 L 25/18	

審査請求 未請求 請求項の数 3 O L (全 16 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)

(71) 出願人 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 官園 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

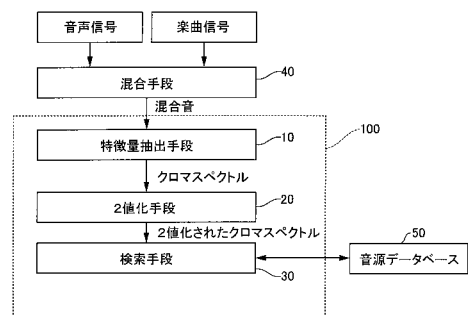
(54) 【発明の名称】 音源検索装置および音源検索方法

(57) 【要約】

【課題】混合音を検索キーとする音源検索装置において、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を提供する。

【解決手段】この音源検索装置100は、楽曲信号と音声信号とを含む混合音からクロマスペクトルを特徴量として抽出する特徴量抽出手段10と、特徴量の平均値以上の場合に特徴量を1とし、平均値未満の場合に特徴量を0とすることにより、抽出した特徴量を2値化する2値化手段20と、2値化手段20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検索手段30とを備える。

【選択図】 図1



【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ー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 とするよう構成されている、音源検索装置。

10

【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ー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 とするよう構成されている、請求項 1 に記載の音源検索装置。

20

30

【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ー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。

40

【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】**

50

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）。

【 0 0 0 3 】

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ー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、3 2 b i t）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。

10

【 先行技術文献 】

【 非特許文献 】

【 0 0 0 4 】

【 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、v o l . 1 1 4、n o . 1 9 1、p p . 1 9 - 2 4 .

【 発明の概要 】

【 発明が解決しようとする課題 】

【 0 0 0 5 】

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。

20

【 0 0 0 6 】

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 】

30

【 0 0 0 7 】

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ー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

40

【 0 0 0 8 】

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、3 2 b i t）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（1 b i t）ので、検索速度を高速化させることができる。

【 0 0 0 9 】

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

50

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ーが大きい周波数（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。

10

20

30

40

50

【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ー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ー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。

10

【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。

20

【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】

30

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)

40

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】

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

【0020】

ここで、 $j = 1, \dots, J$ 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 $s_1(t), \dots, s_J(t)$ を音源データベース50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、上記の式(1)は、下記の式(2)となる。

【数 2】

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

【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ー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、 $k(n) = [k(n, 1), k(n, 2), \dots, k(n, D)]^T$ 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 t_1 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 t_2 から一定の時間(フレームシフト長)経過後の時刻 t_2 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ー(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、出力エネルギー(クロマスペクトル)が算出される

10

20

30

40

50

。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。

10

【0028】

具体的には、時刻 t のクロマスペクトルを $c(t) = [c_1(t), c_2(t), \dots, c_D(t)]^T$ 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 $b(t) = [b_1(t), b_2(t), \dots, b_D(t)]^T$ は、下記の式(3)により表される。

20

【数3】

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

$$d = 1, 2, \dots, D$$

【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。

30

【数4】

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

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

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

40

【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】

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

【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 $\phi^{(q)}$ 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。

$$\bar{M} = 1 - \frac{1}{Q} \sum_{q=1}^Q c, \quad (8)$$

$$c = \begin{cases} 1, & \text{if } [v^{(q)}, n^{(q)}] \in \phi^{(q)} \\ 0, & \text{otherwise} \end{cases}$$

【0034】

また、平均誤検出率は、下記の式(9)により表される。

【数7】

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

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、 $\phi^{(q)} \setminus [v^{(q)}, n^{(q)}]$ は、集合 $\phi^{(q)}$ から、要素 $[v^{(q)}, n^{(q)}]$ を取り除く処理を意味する。また、 $|\cdot|$ 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。

10

20

30

40

50

【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ー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ー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ー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

10

20

30

40

50

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ーを相対的に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ーを相対的に10dB高くした場合には、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ー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ーを相対的に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ーを相対的に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ー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（

10

20

30

40

50

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ー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。

10

20

30

40

50

【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ーが大きい周波数（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。

10

【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。

20

【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ー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。

30

【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。

40

【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

50

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。

10

【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。

20

【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。

30

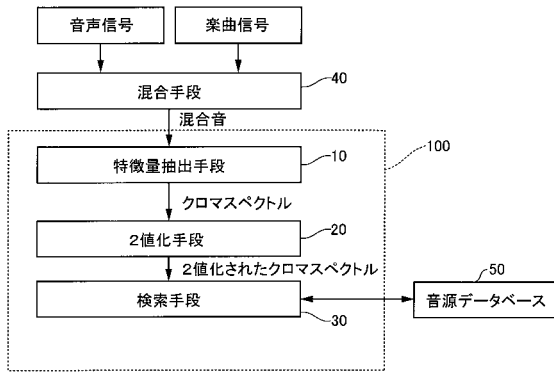
【符号の説明】

【0077】

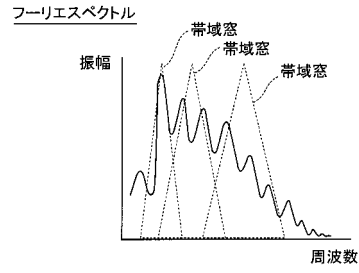
- 10 特徴量抽出手段（検索装置側特徴量抽出手段）
- 20 2値化手段（検索装置側特徴量抽出手段）
- 30 検索手段
- 50 音源データベース
- 51 特徴量抽出手段（データベース側特徴量抽出手段）
- 52 2値化手段（データベース側2値化手段）
- 100 音源検索装置
- T （所定の）時間長さ

40

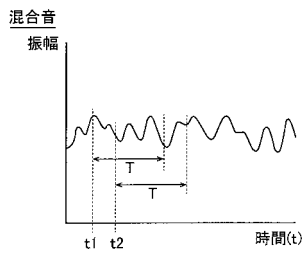
【 図 1 】



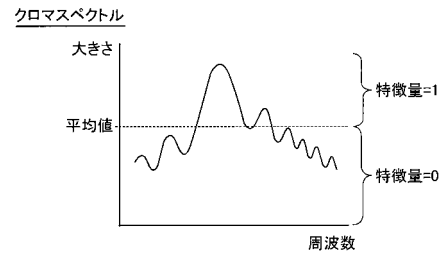
【 図 3 】



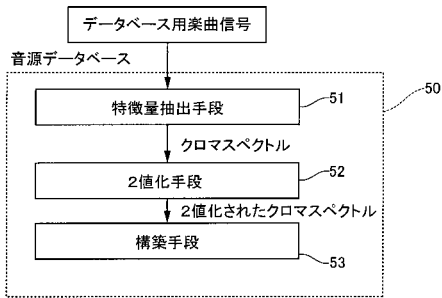
【 図 2 】



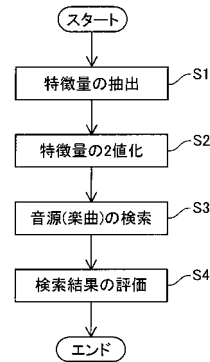
【 図 4 】



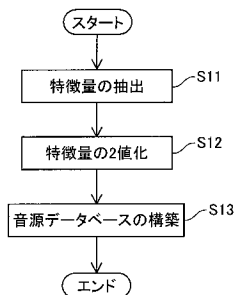
【 図 5 】



【 図 7 】



【 図 6 】



【 図 8 】

