(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2008-134606
(P2008-134606A)
(43)公開日　平成20年6月12日(2008.6.12)

(51)Int.Cl.		F I			テーマコード（参考）
G10L 11/00	(2006.01)	G10L	11/00	402G	5D015
G10L 15/10	(2006.01)	G10L	15/10	200C	5D108
G10L 11/02	(2006.01)	G10L	11/02		
G10L 15/04	(2006.01)	G10L	15/04	300B	
G10L 15/18	(2006.01)	G10L	15/18	200E	

審査請求 未請求　請求項の数 13　ＯＬ　（全 34 頁）　最終頁に続く

(21)出願番号　特願2007-233682 (P2007-233682)
(22)出願日　平成19年9月10日 (2007.9.10)
(31)優先権主張番号　特願2006-289289 (P2006-289289)
(32)優先日　平成18年10月24日 (2006.10.24)
(33)優先権主張国　日本国(JP)

特許法第３０条第１項適用申請有り　２００６年８月７～８日　社団法人　情報処理学会発行の「情報処理学会研究報告　情処研報　Ｖｏｌ．２００６，ＮＯ．９０」に発表

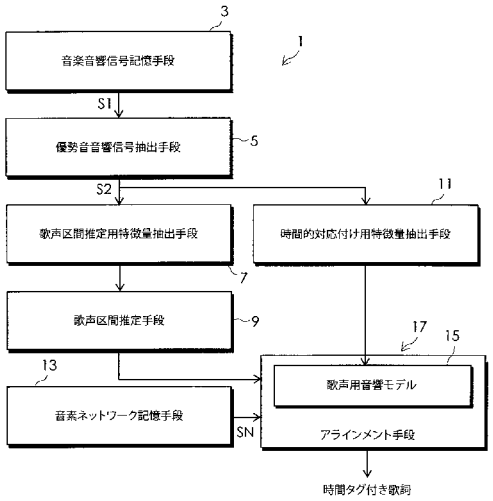
（出願人による申告）平成１８年度独立行政法人科学技術振興機構「音楽デザイン転写・音響信号理解に基づく音インタフェース」委託研究、産業活力再生特別措置法第３０条の適用を受ける特許出願

(71)出願人　504132272
国立大学法人京都大学
京都府京都市左京区吉田本町３６番地１
(71)出願人　301021533
独立行政法人産業技術総合研究所
東京都千代田区霞が関１－３－１
(74)代理人　100091443
弁理士　西浦　▲嗣▼晴
(72)発明者　藤原　弘将
京都府京都市左京区吉田本町　国立大学法人京都大学大学院情報学研究科内
(72)発明者　奥乃　博
京都府京都市左京区吉田本町　国立大学法人京都大学大学院情報学研究科内

最終頁に続く

(54)【発明の名称】音楽音響信号と歌詞の時間的対応付けを自動で行うシステム及び方法

(57)【要約】
【課題】　非歌声区間の影響により、時間的対応付けの精度が低下するのを抑制することができる音楽音響信号と歌詞の時間的対応付けを自動で行うシステムを提供する。
【解決手段】　アラインメント手段１７は、時間的対応付け用特徴量に対応する音素を推定する歌声用音響モデル１５を備える。アラインメント手段１７は、時間的対応付け用特徴量抽出手段１１から得た時間的対応付け用特徴量と、歌声区間推定手段９から得た歌声区間と非歌声区間に関する情報と、音素ネットワークＳＮとを入力として、少なくとも非歌声区間には音素が存在しないという条件の下で、アラインメント動作を実行する。
【選択図】図１



【特許請求の範囲】
【請求項１】
歌声と伴奏音とを含む楽曲の音楽音響信号から、各時刻において前記歌声を含む最も優勢な音の優勢音音響信号を抽出する優勢音音響信号抽出手段と、
前記各時刻における前記優勢音音響信号から前記歌声が含まれている歌声区間と前記歌声が含まれていない非歌声区間とを推定するために利用可能な歌声区間推定用特徴量を抽出する歌声区間推定用特徴量抽出手段と、
複数の前記歌声区間推定用特徴量に基づいて、前記歌声区間と前記非歌声区間を推定して、前記歌声区間と前記非歌声区間に関する情報を出力する歌声区間推定手段と、
各時刻における前記優勢音音響信号から、前記歌声の歌詞と前記音楽音響信号との間の時間的対応を付けるのに適した時間的対応付け用特徴量を抽出する時間的対応付け用特徴量抽出手段と、
前記音楽音響信号に対応する楽曲の歌詞に関して複数の音素とショートポーズとによって構成された音素ネットワークを記憶する音素ネットワーク記憶手段と、
前記時間的対応付け用特徴量に基づいて該時間的対応付け用特徴量に対応する音素を推定する歌声用音響モデルを備え、前記音素ネットワーク中の複数の音素と前記優先音音響信号とを時間的に対応付けるアラインメント動作を実行するアラインメント手段とを備え、前記アラインメント手段は、前記時間的対応付け用特徴量抽出手段から出力される前記時間的対応付け用特徴量と、前記歌声区間と前記非歌声区間に関する情報と、前記音素ネットワークとを入力として、少なくとも前記非歌声区間には音素が存在しないという条件の下で、前記アラインメント動作を実行することを特徴とする音楽音響信号と歌詞の時間的対応付けを自動で行うシステム。
【請求項２】
前記歌声区間推定手段は、予め複数の学習用楽曲に基づいて学習により得られた歌声と非歌声の複数の混合ガウス分布を記憶するガウス分布記憶手段を備え、
前記歌声区間推定手段は、複数の前記歌声区間推定用特徴量と前記複数の混合ガウス分布とに基づいて、前記歌声区間と前記非歌声区間を推定するように構成されている特徴とする請求項１に記載の音楽音響信号と歌詞の時間的対応付けを自動で行うシステム。
【請求項３】
前記歌声区間推定手段は、
前記各時刻における前記歌声区間推定用特徴量と前記混合ガウス分布とに基づいて、前記各時刻における歌声対数尤度と非歌声対数尤度とを計算する対数尤度計算手段と、
前記各時刻における前記歌声対数尤度と前記非歌声対数尤度との対数尤度差を計算する対数尤度差計算手段と、
前記音楽音響信号の全期間から得られる複数の前記対数尤度差に関するヒストグラムを作成するヒストグラム作成手段と、
前記ヒストグラムを、前記楽曲に依存した、歌声区間における前記対数尤度差のクラスと非歌声区間における対数尤度差のクラスに２分割する場合に、クラス間分散を最大とするような閾値を決定し、該閾値を楽曲依存のバイアス調整値と定めるバイアス調整値決定手段と、
前記バイアス調整値を補正するために、前記バイアス調整値にタスク依存値を加算して歌声区間を推定する際に用いる推定用パラメータを決定する推定用パラメータ決定手段と、
前記各時刻における前記歌声対数尤度及び前記非歌声対数尤度を前記推定用パラメータを用いて重み付けを行う重み付け手段と、
前記音楽音響信号の全期間から得られる、重み付けされた複数の前記歌声対数尤度及び重み付けされた複数の前記非歌声対数尤度を、それぞれ隠れマルコフモデルの歌声状態（s_V）の出力確率及び非歌声状態（s_N）の出力確率とみなして、前記音楽音響信号の全期間における前記歌声状態と前記非歌声状態の最尤経路を計算し、前記最尤経路から前記音楽音響信号の全期間における前記歌声区間と前記非歌声区間に関する情報を決定する最

尤経路計算手段とを備えている請求項２に記載の音楽音響信号と歌詞の時間的対応付けを自動で行うシステム。

【請求項４】

前記重み付け手段は、前記歌声状態（s_V）の出力確率$\log p(x|s_V)$及び前記非歌声状態（s_N）の出力確率$\log p(x|s_N)$を下記の式で近似し、

【数１】

$$\log p(x|s_V) = \log N_{GMM}(x;\theta_V) - \frac{1}{2}\eta$$

【数２】

$$\log p(x|s_N) = \log N_{GMM}(x;\theta_N) + \frac{1}{2}\eta$$

上記式において、$N_{GMM}(x;\theta_V)$は歌声の混合ガウス分布（ＧＭＭ）の確率密度関数を表し、$N_{GMM}(x;\theta_N)$は非歌声の混合ガウス分布（ＧＭＭ）の確率密度関数を表し、θ_V及びθ_Nは前記複数の学習用楽曲に基づいて予め学習により定められたパラメータであり、ηは前記推定用パラメータであり、

前記最尤経路計算手段は、前記最尤経路を下記の式を用いて計算し、

【数３】

$$\hat{S} = \underset{S}{\operatorname{argmax}} \sum_t \{\log p(x|s_t) + \log p(s_{t+1}|s_t)\}$$

上記式において、$p(x|s_t)$は状態s_tの出力確率を表し、$p(s_{t+1}|s_t)$は、状態s_tから状態s_{t+1}への遷移確率を表している請求項３に記載の音楽音響信号と歌詞の時間的対応付けを自動で行うシステム。

【請求項５】

前記アラインメント手段は、ビタビアラインメントを用いて前記アラインメント動作を実行するように構成され、

前記ビタビアラインメントの実行において、前記非歌声区間には音素が存在しないという条件として、少なくとも前記非歌声区間をショートポーズとする条件を定め、前記ショートポーズにおいては、他の音素の尤度をゼロとして、前記アラインメント動作を実行することを特徴とする請求項１に記載の音楽音響信号と歌詞の時間的対応付けを自動で行うシステム。

【請求項６】

前記歌声用音響モデルは、話し声用の音響モデルのパラメータを、歌声と伴奏音を含む楽曲中の前記歌声の音素を認識できるように再推定して得た音響モデルである請求項１に記載の音楽音響信号と歌詞の時間的対応付けを自動で行うシステム。

【請求項７】

前記音響モデルは、歌声だけを含む単独歌唱の適応用音楽音響信号と、該適応用音楽音響信号に対する適応用音素ラベルとを用いて、前記話し声用音響モデルのパラメータを、前記適応用音楽音響信号から前記歌声の音素を認識できるように再推定して得た単独歌唱用の音響モデルである請求項６に記載の音楽音響信号と歌詞の時間的対応付けを自動で行うシステム。

【請求項８】

前記音響モデルは、

歌声だけを含む単独歌唱の適応用音楽音響信号と、該適応用音楽音響信号に対する適応用音素ラベルとを用いて、前記話し声用音響モデルのパラメータを、前記適応用音楽音響

信号から前記歌声の音素を認識できるように再推定して得た単独歌唱用の音響モデルを用意し、
前記歌声に加えて伴奏音を含む適応用音楽音響信号から抽出した前記歌声を含む最も優勢な音の優勢音音響信号と、該優勢音音響信号に対する適応用音素ラベルとを用いて、前記単独歌唱用の音響モデルのパラメータを、前記優勢音音響信号から前記歌声の音素を認識できるように再推定して得た分離歌声用の音響モデルである請求項６に記載の音楽音響信号と歌詞の時間的対応付けを自動で行うシステム。
【請求項９】
前記音響モデルは、
歌声だけを含む単独歌唱の適応用音楽音響信号と、該適応用音楽音響信号に対する適応用音素ラベルとを用いて、前記話し声用音響モデルのパラメータを、前記適応用音楽音響信号から前記歌声の音素を認識できるように再推定して得た単独歌唱用の音響モデルを用意し、
次に前記歌声に加えて伴奏音を含む適応用音楽音響信号から抽出した前記歌声を含む最も優勢な音の優勢音音響信号と、該優勢音音響信号に対する適応用音素ラベルとを用いて、前記単独歌唱用の音響モデルのパラメータを、前記優勢音音響信号から前記歌声の音素を認識できるように再推定して得た分離歌声用の音響モデルを用意し、
次に前記時間的対応付け用特徴量記憶手段に記憶されている前記複数の時間的対応付け用特徴量と前記音素ネットワークに記憶されている前記音素ネットワークとを用いて、前記分離歌声用の音響モデルのパラメータを前記音響信号抽出手段に入力された前記音楽音響信号の前記楽曲を歌う特定の歌手の音素を認識できるように推定して得た特定歌手用の音響モデルである請求項６に記載の音楽音響信号と歌詞の時間的対応付けを自動で行うシステム。
【請求項１０】
音楽音響信号に時間的に対応付けられた歌詞を、表示画面上に表示させながら前記音楽音響信号を再生する音楽音響信号再生装置において、
請求項１に記載のシステムを用いて前記音楽音響信号に時間的に対応付けられた前記歌詞を前記表示画面に表示させることを特徴とする音楽音響信号再生装置。
【請求項１１】
歌声と伴奏音とを含む楽曲の音楽音響信号から、各時刻において前記歌声を含む最も優勢な音の優勢音音響信号を優勢音響信号抽出手段が抽出する優勢音音響信号抽出ステップと、
前記各時刻における前記優勢音音響信号から前記歌声が含まれている歌声区間と前記歌声が含まれていない非歌声区間とを推定するために利用可能な歌声区間推定用特徴量を歌声区間推定用特徴量抽出手段が抽出する歌声区間推定用特徴量抽出ステップと、
複数の前記歌声区間推定用特徴量に基づいて、前記歌声区間と前記非歌声区間を歌声区間推定手段推定して、前記歌声区間と前記非歌声区間に関する情報を出力する歌声区間推定ステップと、
各時刻における前記優勢音音響信号から、前記歌声の歌詞と前記音楽音響信号との間の時間的対応を付けるのに適した時間的対応付け用特徴量を時間的対応付け用特徴量抽出手段が抽出する時間的対応付け用特徴量抽出ステップと、
前記音楽音響信号に対応する楽曲の歌詞に関して複数の音素とショートポーズとによって構成された音素ネットワークを音素ネットワーク記憶手段に記憶する記憶ステップと、
前記時間的対応付け用特徴量に基づいて該時間的対応付け用特徴量に対応する音素を推定する歌声用音響モデルを備え、前記音素ネットワーク中の複数の音素と前記優先音音響信号とを時間的に対応付けるアラインメント動作をアラインメント手段が実行するアラインメントステップとからなり、
前記アラインメントステップでは、アラインメント手段が、前記時間的対応付け用特徴量抽出ステップで得られる前記時間的対応付け用特徴量と、前記歌声区間と前記非歌声区間に関する情報と、前記音素ネットワークとを入力として、少なくとも前記非歌声区間に

は音素が存在しないという条件の下で、前記アラインメント動作を実行することを特徴とする音楽音響信号と歌詞の時間的対応付けを自動で行う方法。
【請求項１２】
歌声と伴奏音とを含む楽曲の音楽音響信号と歌詞の時間的対応付けを行うためにコンピュータを、
前記音楽音響信号から、各時刻において前記歌声を含む最も優勢な音の優勢音音響信号を抽出する優勢音音響信号抽出手段と、
前記各時刻における前記優勢音音響信号から前記歌声が含まれている歌声区間と前記歌声が含まれていない非歌声区間とを推定するために利用可能な歌声区間推定用特徴量を抽出する歌声区間推定用特徴量抽出手段と、
複数の前記歌声区間推定用特徴量に基づいて、前記歌声区間と前記非歌声区間を推定して、前記歌声区間と前記非歌声区間に関する情報を出力する歌声区間推定手段と、
各時刻における前記優勢音音響信号から、前記歌声の歌詞と前記優勢音音響信号との間の時間的対応を付けるのに適した時間的対応付け用特徴量を抽出する時間的対応付け用特徴量抽出手段と、
前記音楽音響信号に対応する楽曲の歌詞に関して複数の音素とショートポーズとによって構成された音素ネットワークを記憶する音素ネットワーク記憶手段と、
前記時間的対応付け用特徴量に基づいて該時間的対応付け用特徴量に対応する音素を推定する歌声用音響モデルを備え、前記音素ネットワーク中の複数の音素と前記優先音音響信号とを時間的に対応付けるアラインメント動作を実行するアラインメント手段として機能させ、
前記アラインメント手段に、前記時間的対応付け用特徴量抽出手段から出力される前記時間的対応付け用特徴量と、前記歌声区間と前記非歌声区間に関する情報と、前記音素ネットワークとを入力として、少なくとも前記非歌声区間には音素が存在しないという条件の下で、前記アラインメント動作を実行させるための音楽音響信号と歌詞の時間的対応付け用プログラム。
【請求項１３】
請求項１２に記載のプログラムを記録したコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、歌声と伴奏音とを含む楽曲の音楽音響信号と歌詞との時間的対応付け（アラインメント）を自動で行うシステム及び方法並びに該システムで用いるプログラムに関するものである。
【背景技術】
【０００２】
コンパクトディスク（ＣＤ）などの記録媒体に記録されたディジタル音楽データ（音楽音響信号）のうち、特に、人の音声（例えば歌声）と人の音声以外の音（例えば伴奏音）とで構成されるディジタル音楽データを再生する際に、人の音声の発話内容（すなわち歌詞）を伴奏音と時間的に同期させながら視覚的に表示させる技術は、いわゆるカラオケ装置などでよく使用されている。
【０００３】
しかし、従来のカラオケ装置の場合、伴奏音とその歌手の歌声とは正確に同期しておらず、その音楽の歌詞が楽譜上で予定されているテンポで順次画面上に表示されているにすぎない。そのため、実際の発話のタイミングと画面上の表示とが大きくずれることも多い。しかも、伴奏音と歌声の同期作業は、人間の手を介して行われるものであり、かなりの人的労力を必要とする。
【０００４】
ところで、いわゆる音声認識技術に代表されるように、人の発話内容を解析する技術が知られている。この技術は、伴奏音がない歌声（これを「単独歌唱」という。）のディジ

タル音楽データからその発話内容（歌詞）を認識するというものである。これについてはいくつかの研究結果が報告されている。しかしながら、伴奏音の影響を一切考慮しない音声認識技術を、市販のコンパクトディスク（ＣＤ）またはインターネット等の電気通信回線を通じて配信されるディジタル音楽データにそのまま適用することは極めて困難である。

【０００５】

伴奏音を含む歌唱に関する研究としては、各音素の持続する時間長を学習し、歌声を複数の区間に割り振るものが知られている（下記非特許文献１参照）。この技術は、ビートトラッキングやさび部分の検出など高次の情報を利用する。しかしながら、この技術は音韻的な特徴（例えば、母音や子音など）を全く考慮していない。そのためため、正解率がそれほど高くないという問題がある。また、拍子やテンポなどについての制約が大きいため、多くの種類の楽曲に適用することができないという問題もある。

【０００６】

また特開２００１－１１７５８２号公報（特許文献１）には、カラオケ装置において、歌唱者（入力者）の歌声の音素列と特定の歌手の歌声の音素列とをアラインメント手段を利用して対応付けする技術が開示されている。しかしながらこの公報には、音楽音響信号と歌詞とを時間的に対応付ける技術は何も開示されていない。

【０００７】

また特開２００１－１２５５６２号公報（特許文献２）には、歌声と伴奏音とを含む混合音の音楽音響信号から、各時刻において歌声を含む最も優勢な音高の音高推定を行って優勢音音響信号を抽出する技術が開示されている。この技術を用いると、音楽音響信号から伴奏音を抑制した優勢音音響信号を抽出することができる。

【０００８】

そして藤原弘将、奥乃博、後藤真孝他が、「伴奏音抑制と高信頼度フレーム選択に基づく楽曲の歌手名同定手法」と題する論文［情報処理学会論文誌Vol.47 No.6（発表：2006.6）］（非特許文献２）にも、特許文献２に示された伴奏音を抑制する技術が開示されている。またこの論文には、歌声と非歌声を学習させた２つの混合ガウス分布（GMM）を用いて、優勢音音響信号から歌声区間と非歌声区間を検出する技術が開示されている。さらにこの論文には、歌声に関する特徴量としてLPCメルケプストラムを用いることが開示されている。

【非特許文献１】Ye Wang, et al.; LyricAlly: Automatic Synchronization of Acoustic Musical Signals and Textual Lyrics, Proceeding of the 12th ACM International Conference on Multimedia, October 10-15, 2004.

【非特許文献２】藤原弘将、奥乃博、後藤真孝他著の「伴奏音抑制と高信頼度フレーム選択に基づく楽曲の歌手名同定手法」と題する論文［情報処理学会論文誌Vol.47 No.6（発表：2006.6）］

【特許文献１】特開２００１－１１７５８２号公報

【特許文献２】特開２００１－１２５５６２号公報

【発明の開示】

【発明が解決しようとする課題】

【０００９】

人の音声（例えば歌声）と人の音声以外の音（例えば伴奏音）とで構成される音楽音響信号及び歌詞情報から、伴奏音と忠実に同期して歌詞を表示させるためには、時間情報を含む歌詞、換言すると、演奏の開始時刻から何秒後にその歌詞が発話されるのかという時間情報（本明細書ではこれを「時間タグ情報」という。）を伴う歌詞を得ることが必要となる。

【００１０】

歌詞自体はテキストデータ（テキスト形式のディジタル情報）として容易に入手することはできる。この「歌詞のテキストデータ」と、「その歌詞を発声する歌声を伴う音楽音響信号（ディジタル音楽データ）」とを用いて、「時間タグ付きの歌詞」を生成すること

を、実用可能な程度の精度（正解率）で完全自動化させる技術が切望されている。

【００１１】

伴奏音を含む音楽音響信号と歌詞とを時間的に対応させる上で音声認識技術は有用な技術である。しかしながら歌声が全く存在しない区間（本明細書ではこれを「無発声区間」または「非歌声区間」という。）の影響が、時間的対応付けの精度（正解率）を極端に低下させることが本件発明者らの研究により明らかとなった。

【００１２】

本発明の目的は、非歌声区間の影響により、時間的対応付けの精度が低下するのを抑制することができる音楽音響信号と歌詞の時間的対応付けを自動で行うシステム及び方法、並びにシステムに用いるプログラムを提供することにある。

【課題を解決するための手段】

【００１３】

本発明の音楽音響信号と歌詞の時間的対応付けを自動で行うシステムは、優勢音音響信号抽出手段と、歌声区間推定用特徴量抽出手段と、歌声区間推定手段と、時間的対応付け用特徴量抽出手段と、音素ネットワーク記憶手段と、アラインメント手段とを有する。

【００１４】

優勢音音響信号抽出手段は、歌声と伴奏音とを含む楽曲の音楽音響信号から、各時刻（例えば１０msec毎)において歌声を含む最も優勢な音の優勢音音響信号を抽出する。なおこの優勢音音響信号の抽出技術は、前述の特許文献２及び非特許文献２において使用されている抽出技術と同じである。

【００１５】

歌声区間推定用特徴量抽出手段は、各時刻（例えば１０msec毎)における優勢音音響信号から歌声が含まれている歌声区間と歌声が含まれていない非歌声区間とを推定するために利用可能な歌声区間推定用特徴量を抽出する。ここで利用可能な歌声区間推定用特徴量は、具体的な実施の形態では、１３次元の特徴量である。より具体的には、歌声状態と非歌声状態の識別のためのスペクトル特徴量として、ＬＰＣメルケプストラム及び基本周波数のＦ０の微分係数ΔＦ０を用いることができる。

【００１６】

歌声区間推定手段は、複数の歌声区間推定用特徴量に基づいて、歌声区間と非歌声区間を推定して、歌声区間と非歌声区間に関する情報を出力する。

【００１７】

また時間的対応付け用特徴量抽出手段は、各時刻における優勢音音響信号から、歌声の歌詞と前記優勢音音響信号との間の時間的対応を付けるのに適した時間的対応付け用特徴量を抽出する。具体的な実施の形態では、時間的対応付け用特徴量として、音素の共鳴特性等の２５次元の特徴量を抽出する。

【００１８】

なお歌声区間推定用特徴量抽出手段及び時間的対応付け用特徴量抽出手段により抽出した結果は、それぞれの手段に記憶部を設けておき、少なくとも１曲分を記憶部に記憶しておき、後の処理の際に利用するようにしてもよい。

【００１９】

音素ネットワーク記憶手段は、音楽音響信号に対応する楽曲の歌詞に関して複数の音素とショートポーズとによって構成された音素ネットワークを記憶する。このような音素ネットワークは、例えば、歌詞を音素列に変換し、その後、フレーズの境界を複数個のショートポーズに変換し、単語の境界を１個のショートポーズに変換することにより、日本語の歌詞であれば母音とショートポーズのみからなる音素列を用いて構成するのが好ましい。また英語の歌詞であれば、英語の音素とショートポーズのみからなる音素列を用いて音素ネットワークを構成するのが好ましい。

【００２０】

アラインメント手段は、時間的対応付け用特徴量に基づいて該時間的対応付け用特徴量に対応する音素を推定する歌声用音響モデルを備えている。そしてアラインメント手段は

、音素ネットワーク中の複数の音素と優先音音響信号とを時間的に対応付けるアラインメント動作を実行する。具体的には、アラインメント手段は、時間的対応付け用特徴量抽出手段から出力される時間的対応付け用特徴量と、歌声区間と非歌声区間に関する情報と、音素ネットワークとを入力として、歌声用音響モデルを用いて、少なくとも非歌声区間には音素が存在しないという条件の下で、アラインメントを実行して、音楽音響信号と歌詞の時間的対応付けを自動で行う。

【００２１】

本発明によれば、歌声区間及び非歌声区間の推定に用いるのに適した特徴量（歌声区間推定用特徴量）と、歌詞との時間的対応付けに用いるのに適した特徴量（時間的対応付け用特徴量）とを、優勢音音響信号からそれぞれ別個に抽出しているので、歌声区間及び非歌声区間の推定精度及び時間的対応付け精度を高くすることができる。特に、本発明によれば、アラインメント手段では、話し声用音響モデルを使用せずに、時間的対応付け用特徴量に対応する音素を推定する歌声用音響モデルを使用しているので、話し声とは異なる歌声の特徴を考慮したより精度の高い音素の推定を行うことができる。さらにアラインメント手段は、少なくとも非歌声区間には音素が存在しないという条件の下で、アラインメント動作を実行するので、非歌声区間の影響を極力排除した状態で、音素ネットワーク中の複数の音素と各時刻における優先音音響信号とを時間的に対応付けることができる。したがって本発明によれば、アラインメント手段の出力を用いて、音楽音響信号に同期した時間タグ付きの歌詞データを自動で得ることができる。

【００２２】

歌声区間推定手段の構成は、推定精度が高いものであれば、どのような構成のものでも任意である。例えば、歌声区間推定手段に、予め複数の学習用楽曲に基づいて学習により得られた歌声と非歌声の複数の混合ガウス分布を記憶するガウス分布記憶手段を設ける。そして、音楽音響信号から得た複数の歌声区間推定用特徴量と複数の混合ガウス分布とに基づいて、歌声区間と非歌声区間を推定するように、歌声区間推定手段を構成することができる。このように事前の学習により得られた混合ガウス分布に基づいて、歌声区間と非歌声区間とを推定すると、高い精度で歌声区間と非歌声区間とを推定することができ、アラインメント手段におけるアラインメント精度を高くすることができる。

【００２３】

このような歌声区間推定手段は、対数尤度計算手段と、対数尤度差計算手段と、ヒストグラム作成手段と、バイアス調整値決定手段と、推定用パラメータ決定手段と、重み付け手段と、最尤経路計算手段とから構成することができる。対数尤度計算手段は、音楽音響信号の最初から最後までの期間中の各時刻における歌声区間推定用特徴量と事前に記憶した混合ガウス分布とに基づいて、各時刻における歌声対数尤度と非歌声対数尤度とを計算する。そして対数尤度差計算手段は、各時刻における歌声対数尤度と非歌声対数尤度との対数尤度差を計算する。ヒストグラム作成手段は、推定に先立つ前処理において、優先音音響信号の全期間から得られる複数の対数尤度差に関するヒストグラムを作成する。そしてバイアス調整値決定手段は、作成したヒストグラムを、楽曲に依存した、歌声区間における対数尤度差のクラスと非歌声区間における対数尤度差のクラスに２分割する場合に、クラス間分散を最大とするような閾値を決定し、この閾値を楽曲依存のバイアス調整値と定める。また推定用パラメータ決定手段は、バイアス調整値を補正するため（アラインメントの精度を高めるため又は歌声区間を広げる調整のため）に、バイアス調整値にタスク依存値を加算して歌声区間を推定する際に用いる推定用パラメータを決定する。そして重み付け手段は、各時刻における歌声対数尤度及び非歌声対数尤度を推定用パラメータを用いて重み付けを行う。なおこの際に使用する歌声対数尤度及び非歌声対数尤度は、前処理の際に求めたものを使用してもよいが、あらたに計算をしてもよいのは勿論である。なお前処理の計算結果を利用する場合には、対数尤度計算手段に記憶機能を持たせておけばよい。最尤経路計算手段は、音楽音響信号の全期間から得られる、重み付けされた複数の歌声対数尤度及び重み付けされた複数の非歌声対数尤度を、それぞれ隠れマルコフモデルの歌声状態（S_V）の出力確率及び非歌声状態（S_N）の出力確率とみなす。そして最尤経

路計算手段は、音楽音響信号の全期間における歌声状態と非歌声状態の最尤経路を計算し、最尤経路から音楽音響信号の全期間における歌声区間と非歌声区間に関する情報を決定する。なお対数尤度差決定手段、ヒストグラム作成手段、バイアス調整値決定手段及び推定用パラメータ決定手段は、本発明のシステムで歌声区間を推定する前の前処理において、音楽音響信号に対して使用される。前処理により得た推定用パラメータを用いた重み付け手段による重み付けを、各時刻における歌声対数尤度及び非歌声対数尤度に対して行うと、後の最尤経路計算手段における歌声区間と非歌声区間の境界部の調整を、適切に調整することができる。なお推定動作時においては、歌声区間推定用特徴量抽出手段から各時刻において出力される歌声区間推定用特徴量から、対数尤度計算手段が計算した歌声対数尤度及び非歌声対数尤度に、直接重み付けを行って、最尤経路を計算することになる。このような前処理によって対数尤度差のヒストグラムを利用して、歌声対数尤度及び非歌声対数尤度のバイアス調整値（閾値）を決定すると、音楽音響信号に合ったバイアス調整値を決定することができる。このバイアス調整値（閾値）は、歌声状態と非歌声状態との境界部を決定する。そしてバイアス調整値により定めた推定用パラメータを用いて重み付けを行うと、楽曲ごとの音楽音響信号の音響的特性の違いによって現れる歌声区間推定用特徴量の傾向に合わせて、歌声状態と非歌声状態との境界部を中心にして歌声対数尤度及び非歌声対数尤度を調整することができ、歌声区間及び非歌声区間の境界を、個々の楽曲に合わせて適切に設定することができる。

【００２４】

なお最尤経路計算手段においては、以下のようにして、最尤経路を計算することができる。すなわち歌声状態（s_V）の出力確率$\log p(x|s_V)$及び非歌声状態（s_N）の出力確率$\log p(x|s_N)$を下記の式で近似する。

【数１】

$$\log p(x|s_V) = \log N_{GMM}(x;\theta_V) - \frac{1}{2}\eta$$

【数２】

$$\log p(x|s_N) = \log N_{GMM}(x;\theta_N) + \frac{1}{2}\eta$$

【００２５】

上記式において、$N_{GMM}(x;\theta_V)$は歌声の混合ガウス分布（ＧＭＭ）の確率密度関数を表し、$N_{GMM}(x;\theta_N)$は非歌声の混合ガウス分布（ＧＭＭ）の確率密度関数を表す。またθ_V及びθ_Nは複数の学習用楽曲に基づいて予め学習により定められたパラメータであり、ηは推定用パラメータである。最尤経路を下記の式を用いて計算すればよい。

【数３】

$$\hat{S} = \underset{S}{\operatorname{argmax}} \sum_t \{\log p(x|s_t) + \log p(s_{t+1}|s_t)\}$$

【００２６】

上記式において、$p(x|s_t)$は状態S_tの出力確率を表す。そして$p(s_{t+1}|s_t)$は、状態s_tから状態s_{t+1}への遷移確率を表している。

【００２７】

上記式を用いて最尤経路を計算すれば、音楽音響信号の全期間における歌声区間と非歌声区間に関するより正確な情報を得ることができる。

【００２８】

アラインメント手段は、ビタビアラインメントを用いてアラインメント動作を実行するように構成されたものを用いることができる。ここで「ビタビアラインメント」とは、音

声認識の技術分野において知られるもので、音響信号と文法（アラインメント用の音素列）の間の最尤経路を探索するビタビアルゴリズムを用いた最適解探索手法の一つである。ビタビアラインメントの実行においては、「非歌声区間には音素が存在しないという条件」として、少なくとも非歌声区間をショートポーズとする条件を定める。そしてショートポーズにおいては、他の音素の尤度をゼロとして、アラインメント動作を実行する。このようにするとショートポーズの区間においては、他の音素の尤度がゼロになるため、歌声区間情報を利用することができ、精度の高いアラインメントを行うことができる。

【００２９】

また使用する歌声用音響モデルとして、話し声用の音響モデルのパラメータを、歌声と伴奏音を含む楽曲中の歌声の音素を認識できるように再推定して（学習して）得た音響モデルを用いることができる。歌声用音響モデルとしては、歌声の発話内容（歌詞）に対してアラインメントを行うため、大量の歌声のデータから学習されたモデルを使用することが理想的である。しかしながら、現段階ではそのようなデータベースは構築されていない。そこで話し声用の音響モデルのパラメータを、歌声と伴奏音を含む楽曲中の歌声の音素を認識できるように再推定して（学習して）得た音響モデルを用いれば、話し声用の音響モデルを使用する場合よりも、高い精度で歌声の音素を認識することが可能になる。

【００３０】

なおこのような歌声用音響モデルとしては、歌声だけを含む単独歌唱の適応用音楽音響信号と、該適応用音楽音響信号に対する適応用音素ラベルとを用いて、話し声用音響モデルのパラメータを、適応用音楽音響信号から歌声の音素を認識できるように再推定して得た単独歌唱用の音響モデルを用いることができる。この音響モデルでは、伴奏音が無いかまたは伴奏音が歌声に比べて小さい場合に適している。

【００３１】

また歌声用音響モデルとしては、歌声に加えて伴奏音を含む適応用音楽音響信号から抽出した歌声を含む最も優勢な音の優勢音音響信号と、該優勢音音響信号に対する適応用音素ラベルとを用いて、前述の単独歌唱用の音響モデルのパラメータを、優勢音音響信号からの音素を認識できるように再推定して得た分離歌声用の音響モデルを用いることができる。このような分離歌声用の音響モデルは、歌声と同様に伴奏音が大きい場合に適している。

【００３２】

さらに歌声用音響モデルとしては、時間的対応付け用特徴量記憶手段に記憶されている複数の時間的対応付け用特徴量と音素ネットワークに記憶されている音素ネットワークとを用いて、前述の分離歌声用の音響モデルのパラメータを音響信号抽出手段に入力された音楽音響信号の楽曲を歌う特定の歌手の音素を認識できるように推定して得た特定歌手用の音響モデルを用いることができる。この特定歌手用の音響モデルは、歌手を特定した音響モデルであるため、アラインメントの精度を最も高くすることができる。

【００３３】

なお音楽音響信号に時間的に対応付けられた歌詞を、表示画面上に表示させながら音楽音響信号を再生する音楽音響信号再生装置において、本発明のシステムを用いて音楽音響信号に時間的に対応付けられた歌詞を表示画面に表示させると、再生される音楽と画面に表示される歌詞とを同期させて表示画面に表示することができる。

【００３４】

本発明の音楽音響信号と歌詞の時間的対応付けを自動で行う方法では、次のようにして、時間的対応付けを行う。まず歌声と伴奏音とを含む楽曲の音楽音響信号から、各時刻において歌声を含む最も優勢な音の優勢音音響信号を優勢音響信号抽出手段が抽出する（優勢音響信号抽出ステップ）。次に各時刻における優勢音音響信号から歌声が含まれている歌声区間と歌声が含まれていない非歌声区間とを推定するために利用可能な歌声区間推定用特徴量を歌声区間推定用特徴量抽出手段が抽出する（歌声区間推定用特徴量抽出ステップ）。そして複数の歌声区間推定用特徴量に基づいて、歌声区間と非歌声区間を歌声区間推定手段が推定して、歌声区間と前記非歌声区間に関する情報を出力する（歌声区間推定

ステップ）。また各時刻における優勢音音響信号から、歌声の歌詞と音楽音響信号との間の時間的対応を付けるのに適した時間的対応付け用特徴量を時間的対応付け用特徴量抽出手段が抽出する（時間的対応付け用特徴量抽出ステップ）。さらに音楽音響信号に対応する楽曲の歌詞の複数の音素が、該複数の音素の隣りあう二つの音素の時間的間隔が調整可能に繋がって構成された音素ネットワークを音素ネットワーク記憶手段に記憶する（記憶ステップ）。そして時間的対応付け用特徴量に基づいて該時間的対応付け用特徴量に対応する音素を推定する歌声用音響モデルを備え、音素ネットワーク中の複数の音素と優先音音響信号とを時間的に対応付けるアラインメント動作をアラインメント手段が実行する（アラインメントステップ）。このアラインメントステップでは、アラインメント手段が、時間的対応付け用特徴量抽出ステップで得られる時間的対応付け用特徴量と、歌声区間と非歌声区間に関する情報と、音素ネットワークとを入力として、歌声用音響モデルを用いて、少なくとも非歌声区間には音素が存在しないという条件の下で、アラインメント動作を実行する。

【００３５】

また本発明は、歌声と伴奏音とを含む楽曲の音楽音響信号と歌詞の時間的対応付けを行うためにコンピュータを利用する場合において、コンピュータを前述の優勢音響信号抽出手段と、歌声区間推定用特徴量抽出手段と、歌声区間推定手段と、時間的対応付け用特徴量抽出手段と、音素ネットワーク記憶手段と、アラインメント手段として機能させるプログラムとして特定することができる。なおこのプログラムは、コンピュータ読み取り可能な記録媒体に記録されていてもよいのは勿論である。

【００３６】

なお表示画面上に歌詞を表示させながら音楽ディジタルデータを再生するための音楽音響信号再生装置において、本発明に係る音楽音響信号と歌詞の時間的対応付けプログラムを実行させることができる。この場合には、予め時間情報を伴う歌詞を生成した後で表示画面上に歌詞を表示させる。そして表示画面上に歌詞を表示させた状態で、表示された歌詞の表示部分をポインタにより選択する。このようにすると、選択された歌詞の一部に相当する時間情報を元に、その部分から音響音楽信号の再生を行うように構成してもよい。また事前に本発明のシステムで予め生成した時間情報を伴う歌詞を音楽音響信号再生装置に設けたハードディスク等の記憶手段に記憶させておいてもよく、またネットワーク上のサーバーに記憶させておいてもよい。そして音楽音響信号再生装置による音楽ディジタルデータの再生と同期させて、記憶手段に記憶したまたはネットワーク上のサーバーから取得した時間情報を伴う歌詞を表示画面上に表示するようにしてもよい。

【発明を実施するための最良の形態】

【００３７】

以下図面を参照して、本発明の音楽音響信号と歌詞の時間的対応付けを自動で行うシステム及びその方法の実施の形態の一例について詳細に説明する。図１は、音楽音響信号と歌詞の時間的対応付けを自動で行うシステム１の実施の形態をコンピュータを用いて実現する場合に、コンピュータ内に実現される機能実現手段の構成を示すブロックである。また図２は、図１の実施の形態をプログラムをコンピュータで実行することにより実施する場合のステップを示すフローチャートである。このシステム１は、音楽音響御信号記憶手段３と、優勢音音響信号抽出手段５と、歌声区間推定用特徴量抽出手段７と、歌声区間推定手段９と、時間的対応付け用特徴量抽出手段１１と、音素ネットワーク記憶手段１３と、歌声用音響御モデル１５を備えたアラインメント手段１７とを備えている。

【００３８】

本発明は上記技術的課題を効果的に達成するための基本的なアプローチとして、大きく以下の３つのステップを実行する。

【００３９】

ステップ１：伴奏音抑制
ステップ２：歌声区間検出
ステップ３：アラインメント（時間的対応付け）

ステップ１を実行するために、音楽音響御信号記憶手段３には、対象とする歌声と伴奏音とを含む複数の楽曲の音楽音響信号が記憶されている。優勢音音響信号抽出手段５は、図３に示すフローチャートに従って、歌声と伴奏音とを含む楽曲の音楽音響信号Ｓ１から、各時刻（具体的には１０msec毎)において歌声を含む最も優勢な音の優勢音音響信号Ｓ２を抽出する。本実施の形態においては、優勢音音響信号とは、伴奏音が抑制された信号と見ることができる。優勢音音響信号の抽出技術は、前述の特開２００１－１２５５６２号公報（特許文献２）及び非特許文献２に示された抽出技術と同じである。歌声と伴奏音とを含む楽曲の音楽音響信号Ｓ１の信号波形は、例えば図４（Ａ）に示すような信号波形であり、優勢音音響信号抽出手段５が出力する伴奏音が抑制された優勢音音響信号Ｓ２の信号波形は、図４（Ｄ）に示すよう信号波形である。以下優勢音音響信号の抽出方法について説明する。

【００４０】

まず歌声と伴奏音とを含む楽曲（混合音）の音楽音響信号から、後述する歌声区間推定用特徴量及び時間的対応付け用特徴量［メロディ（歌声）の音韻的特徴を表す特徴量等］を抽出するためには、音楽音響信号から伴奏音の影響を低減させた優勢音音響信号を得ることが必要である。そこで優勢音音響信号抽出手段５では、図３に示す以下の３つのステップを実行する。

【００４１】

ＳＴ１：メロディ（歌声）の基本周波数Ｆ０を推定する。

【００４２】

ＳＴ２：推定された基本周波数に基づいて、メロディ（歌声）の調波構造を抽出する。

【００４３】

ＳＴ３：抽出された調波構造を優勢音音響信号に再合成する。

【００４４】

なお、優勢音音響信号には、間奏などの区間では歌声以外の音響信号（伴奏音や無音）を含んでいる場合がある。したがって本実施の形態では、伴奏音の「除去」ではなく伴奏音の「低減」と表現する。以下ステップＳＴ１乃至ＳＴ３について説明する。

【００４５】

（ＳＴ１：Ｆ０推定処理について）

メロディ（歌声）の基本周波数の推定方法には種々の方法が知られている。例えば、音源数を仮定しない音高推定手法（PreFEst）により、基本周波数を推定する方法を用いることができる（例えば、後藤 真孝著 "音楽音響信号を対象としたメロディとベースの音高推定"、電子情報通信学会論文誌 D-II, Vol.J84-D-II, No.1, pp.12-22, January 2001.参照）。ここで、PreFEstはメロディとベースの基本周波数Ｆ０を推定する手法として知られている。制限された周波数帯域において、各時刻で最も優勢な調波構造（つまり、最も大きな音）を持つ優勢音の基本周波数Ｆ０を推定する手法である。この音高推定手法（PreFEst）では、調波構造の形状を表す確率分布をあらゆる音高（基本周波数）に対して用意する。そして、それらの混合分布（加重混合＝重み付き和）として入力の周波数成分をモデル化する。

【００４６】

メロディ（歌声）は中高域の周波数帯域において、各時刻で最も優勢な調波構造を持つ場合が多い。そこで周波数帯域を適切に制限することで、メロディ（歌声）の基本周波数Ｆ０を推定することができる。以下、PreFEstの概要について説明する。なお、以下の説明で用いられるｘはcentの単位で表される対数周波数軸上の周波数であり、（ｔ）は時間を表すものとする。また、centは、本来は音高差（音程）を表す尺度であるが、本明細書では、４４０×２$^{\{(3/12)-5\}}$［Ｈｚ］を基準として、次式のように絶対的な音高を表す単位として用いる。

【数４】

$$f_{\text{cent}} = 1200\log_2 \frac{f_{\text{Hz}}}{440 \times 2^{\frac{3}{12}-5}}$$

【００４７】

パワースペクトル$\Psi_p^{(t)}$（ｘ）に対して、メロディの周波数成分の多くが通過するように設計された帯域通過フィルタ（Band Pass Filter）を用いる。例えば、4800cent以上の成分を通過させるフィルタを用いるのが好ましい。フィルタを通過後の周波数成分は、

ＢＰＦ（ｘ）・$\Psi_p^{(t)}$（ｘ）

と表される。但し、ＢＰＦ（ｘ）はフィルタの周波数応答である。以後の確率的処理を可能にするため、フィルタを通過後の周波数成分を確率密度関数（ＰＤＦ）として、以下のように表現する。

【数５】

$$p_{\Psi}^{(t)}(x) = \frac{BPF(x)\Psi_p^{(t)}(x)}{\int_{-\infty}^{\infty} BPF(x)\Psi_p^{(t)}(x)dx}$$

【００４８】

その後、周波数成分の確率密度関数ＰＤＦが、全ての可能な基本周波数Ｆ０に対応する音モデル（確率分布）の重み付き和からなる確率モデル：

【数６】

$$p(x|\theta^{(t)}) = \int_{\text{Fl}}^{\text{Fh}} w^{(t)}(F)p(x|F)dF,$$

$$\theta^{(t)} = \{w^{(t)}(F)|\text{Fl} \le F \le \text{Fh}\}$$

【００４９】

から生成されたと考える。

【００５０】

ここで、ｐ（ｘ｜Ｆ）は、それぞれのＦ０についての音モデルであり、Ｆｈは取りうるＦ０の上限値を表し、Ｆｌは取りうるＦ０の下限値を表すものとする。また、$w^{(t)}$（Ｆ）は音モデルの重みであり、

【数７】

$$\int_{\text{Fh}}^{\text{Fl}} w^{(t)}(F)dF = 1$$

【００５１】

を満たす。すなわち、音モデルとは典型的な調波構造を表現した確率分布である。そして、ＥＭ（Expectation Maximization）アルゴリズムを用いて$w^{(t)}$（Ｆ）を推定し、推定した$w^{(t)}$（Ｆ）を基本周波数Ｆ０の確率密度関数（ＰＤＦ）と解釈する。最終的に、$w^{(t)}$（Ｆ）の中の優勢なピークの軌跡をマルチエージェントモデルを用いて追跡することで、メロディ（歌声）のＦ０系列（Ｆ０　Estimation）を得る。図４は、このようにして取得したＦ０系列（Ｆ０　Estimation）を示している。

【００５２】

（ＳＴ２：調波構造抽出）

このようにして推定された基本周波数Ｆ０に基づいて、メロディの調波構造の各倍音成分のパワーを抽出する。各周波数成分の抽出には、前後ｒcentずつの誤差を許容し、この範囲で最もパワーの大きなピークを抽出する。ｌ次倍音（ｌ＝１，・・・，Ｌ）のパワーA_lと周波数F_lは、以下のように表される。

【数８】

$$F_l = \underset{F}{\operatorname{argmax}} |S(F)|$$
$$\left(l\overline{F}\cdot\left(1-2^{\frac{r}{1200}}\right) \le F \le l\overline{F}\cdot\left(1+2^{\frac{r}{1200}}\right)\right)$$
$$A_l = |S(F_l)|$$

【００５３】

ここで、Ｓ（Ｆ）はスペクトルを表し、Ｆの上部にバー（－）のある記号は、PreFEstによって推定された基本周波数Ｆ０を表す。本願発明者らの実験では、ｒの値として２０を用いて調波構造の抽出を実施し、後述のとおりその効果を確認した。図４（Ｃ）は、抽出した各周波数成分の調波構造を示している。

【００５４】

（ＳＴ３：再合成）

抽出された調波構造を正弦波重畳モデルに基づいて再合成することで、各時刻において歌声を含む最も優勢な音の優勢音音響信号を得る。ここで時刻ｔにおけるｌ次倍音の周波数を$F_l^{(t)}$とし、振幅を$A_l^{(t)}$と表す。各フレーム間（時刻ｔと時刻ｔ＋１との間）の周波数が線形に変化するように、位相の変化を２次関数で近似する。また、各フレーム間の振幅の変化は１次関数で近似する。再合成された優勢音音響信号ｓ（ｋ）は、以下のように表される。なお以下の式でθ_l（ｋ）は、ｌ次倍音の時刻ｋにおける位相であり、s_l（ｋ）は、ｌ次倍音の時刻ｋにおける波形である。

【数９】

$$\theta_l(k) = \frac{\pi\left(F_l^{(t+1)} - F_l^{(t)}\right)}{K}k^2 + 2\pi F_l^{(t)}k + \theta_{l,0}^{(t)}$$
$$s_l(k) = \left\{\left(A_l^{(t+1)} - A_l^{(t)}\right)\frac{k}{K} + A_l^{(t)}\right\}\sin\left(\theta_l(k)\right)$$
$$s(k) = \sum_{l=1}^{L} s_l(k)$$

【００５５】

ここで、ｋは時間（単位：秒）を表し、時刻ｔにおいてｋ＝０とする。また、Ｋは（ｔ）と（ｔ＋１）の時間の差、つまりフレームシフトを秒の単位で表す。

【００５６】

$\theta_{l,0}^{(t)}$は、位相の初期値を表し、入力信号の先頭のフレームでは、$\theta_{l,0}^{(t)}$＝０とする。以後のフレームでは、$\theta_{l,0}^{(t)}$は、前フレームのｌ次倍音の周波数$F_l^{(t-1)}$と、初期位相$\theta_{l,0}^{(t-1)}$とを用いて

【数１０】

$$\frac{\pi\left(F_l^{(t)} - F_l^{(t-1)}\right)}{2K} + \theta_{l,0}^{(t-1)}$$

【００５７】

で与えられる。

【００５８】
図１に戻って、歌声区間推定用特徴量抽出手段７は、各時刻（具体的には、１０ｍsec毎)における優勢音音響信号から歌声が含まれている歌声区間と歌声が含まれていない非歌声区間とを推定するために利用可能な歌声区間推定用特徴量を抽出する。本実施の形態では、１２次元のＬＰＣメルケプストラム（ＬＰＭＣＣ）と１次元の基本周波数Ｆ０の微分係数（ΔＦ０）をここで利用可能な歌声区間推定用特徴量として用いる。本実施の形態では、歌声区間推定用特徴量抽出手段７は、歌声と非歌声を識別するために、歌声区間推定用特徴量（スペクトル特徴量）として、下記の二種類の特徴量を抽出する。
【００５９】
・ＬＰＣメルケプストラム（ＬＰＭＣＣ）
第１の種類のスペクトル特徴量として、１２次元のＬＰＣメルケプストラム（ＬＰＭＣＣ）を用いる。ＬＰＭＣＣはＬＰＣスペクトルから計算されたメルケプストラム係数である。本願発明者らの実験によると、この特徴量は、メル周波数ケプストラム係数（ＭＦＣＣ）と比べて、歌声の特徴をよく表現することを確認している。本実施の形態では、ＬＰＣスペクトルからメル周波数ケプストラム係数ＭＦＣＣを計算することでＬＰＣメルケプストラムＬＰＭＣＣを抽出するものとした。
【００６０】
・$\Delta F0_s$
第２の種類のスペクトル特徴量として、基本周波数Ｆ０の微分係数（ΔＦ０）を用いる。これは、歌声の動的な性質を表現するのに役立つ。歌声は他の楽曲と比較して、ビブラートなどに起因する時間変動が多いので、基本周波数Ｆ０の軌跡の傾きを表す微分係数ΔＦ０は、歌声と非歌声の識別に適していると考えられるからである。
【００６１】
ΔＦ０の計算には、次式のように５フレーム間の回帰係数を用いた。
【数１１】

$$\Delta f[t] = \frac{\sum_{k=-2}^{2} k \cdot f[t+k]}{\sum_{k=-2}^{2} k^2}$$

【００６２】
ここで、ｆ［ｔ］は、時刻ｔにおける周波数（単位：cent）である。
【００６３】
そして前述のステップ２を実行するために、歌声区間推定手段９は、各時刻で抽出した複数の歌声区間推定用特徴量に基づいて、歌声区間と非歌声区間を推定して、歌声区間と非歌声区間に関する情報を出力する。本実施の形態の歌声区間推定手段９は、図５に示す構成を有している。図５に示した歌声区間推定手段９では、図２に示すように、予め複数の学習用楽曲８に基づいて学習により得られた歌声と非歌声の複数の混合ガウス分布を記憶するガウス分布記憶手段９１を備えている。歌声区間推定手段９は、１曲の音楽音響信号Ｓ１の全期間において、複数の歌声区間推定用特徴量と複数の混合ガウス分布とに基づいて、歌声区間と非歌声区間を推定し、その情報を出力する。そこでこの歌声区間推定手段９は、さらに対数尤度計算手段９２と、対数尤度差計算手段９３と、ヒストグラム作成手段９４と、バイアス調整値決定手段９５と、推定用パラメータ決定手段９６と、重み付け手段９７と、最尤経路計算手段９８とを備えている。対数尤度差計算手段９３と、ヒストグラム作成手段９４と、バイアス調整値決定手段９５と、推定用パラメータ決定手段９６とは、歌声区間の推定を行う前の前処理において使用される。図６は、図５に示した歌声区間推定手段９をプログラムによりコンピュータで実現する場合のフローチャートを示している。また図７には、歌声区間の検出をプログラムで実現する際のフローチャートを

示している。図７は、図６のステップＳＴ１１とステップＳＴ１６の詳細に相当する。

【００６４】

対数尤度計算手段９２は、音楽音響信号Ｓ１の最初から最後までの期間中の各時刻にいて、歌声区間推定用特徴量抽出手段７が抽出した歌声区間推定用特徴量（ステップＳＴ１１）と、事前に前処理によりガウス分布記憶手段９１に記憶した混合ガウス分布とに基づいて、各時刻における歌声対数尤度と非歌声対数尤度とを計算する。

【００６５】

そして対数尤度差計算手段９３は、各時刻における歌声対数尤度と非歌声対数尤度との対数尤度差を計算する（ステップＳＴ１２）。この計算は、入力された音楽音響信号から抽出された歌声区間推定用特徴量（特徴ベクトル列）に対して、次式のように歌声対数尤度と非歌声対数尤度の対数尤度差ｌ（ｘ）を計算する。

【数１２】

$$l(x) = \log \mathrm{N}_{\mathrm{GMM}}(x;\theta_V) - \log \mathrm{N}_{\mathrm{GMM}}(x;\theta_N)$$

【００６６】

上記式の前方の関数が歌声対数尤度を示し、後者の関数が非歌声関数尤度を示す。ヒストグラム作成手段９４は、音楽音響信号の全期間から抽出した優先音音響信号から得られる複数の対数尤度差に関するヒストグラムを作成する（ステップＳＴ１３）。図６には、ヒストグラム作成手段９４が作成したヒストグラムの例が例示してある。

【００６７】

そしてバイアス調整値決定手段９５は、作成したヒストグラムを、楽曲に依存した、歌声区間における対数尤度差のクラスと非歌声区間における対数尤度差のクラスに２分割する場合に、クラス間分散を最大とするような閾値を決定し、この閾値を楽曲依存のバイアス調整値$\eta_{dyn.}$と定める（ステップＳＴ１４）。図６には、この閾値を図示してある。また推定用パラメータ決定手段９６は、バイアス調整値$\eta_{dyn.}$を補正するため（アラインメントの精度を高めるため又は歌声区間を広げる調整のため）に、バイアス調整値$\eta_{dyn.}$にタスク依存値η_{fixed}を加算して歌声区間を推定する際に用いる推定用パラメータη（＝$\eta_{dyn.}$＋η_{fixed}）を決定する（ステップＳＴ１５）。混合ガウス分布（ＧＭＭ）の尤度には、楽曲によってバイアスがかかるため、全ての楽曲に適切な推定用パラメータηを定めるのは困難である。そこで、本実施の形態では、推定用パラメータηをバイアス調整値$\eta_{dyn.}$とタスク依存値η_{fixed}とに分割することとした。なおこのタスク依存値η_{fixed}は、楽曲の種別等を考慮して予め手動で設定する。一方、バイアス調整値$\eta_{dyn.}$は前述のステップを経てまたは公知の閾値自動設定法を用いて楽曲毎に自動的に設定してもよいし、楽曲の種別に応じて、代表的な学習用音楽音響信号に基づいて予め設定してもよい。

【００６８】

そして重み付け手段９７は、各時刻における歌声対数尤度及び非歌声対数尤度を推定用パラメータηを用いて重み付けを行う（図７のステップＳＴ１６Ａ）。なおこの例では、ここで使用する歌声対数尤度及び非歌声対数尤度として前処理の際に計算したものを用いる。すなわち重み付け手段９７は、歌声対数尤度及び非歌声対数尤度の出力確率を、次式のように近似する。

【数１３】

$$\log p(x|s_V) = \log \mathrm{N}_{\mathrm{GMM}}(x;\theta_V) - \frac{1}{2}\eta$$

【数１４】

$$\log p(x|s_N) = \log \mathrm{N}_{\mathrm{GMM}}(x;\theta_N) + \frac{1}{2}\eta$$

【００６９】
ここで、N_{GMM}（ｘ；θ）は混合ガウス分布（ＧＭＭ）の確率密度関数を表す。また、ηは正解率と棄却率の関係を調整する推定用パラメータである。歌声ＧＭＭのパラメータθ_Vと非歌声ＧＭＭのパラメータθ_Nはそれぞれ学習データの歌声区間と非歌声区間とを用いて学習する。本願発明者らの実験では、混合数６４のＧＭＭを用いて実施し後述のとおりその効果を確認した。
【００７０】
最尤経路計算手段９８は、音楽音響信号の全期間から得られる、重み付けされた複数の歌声対数尤度及び重み付けされた複数の非歌声対数尤度を、それぞれ隠れマルコフモデルの歌声状態（S_V）の出力確率及び非歌声状態（S_N）の出力確率とみなす（図のステップＳＴ１６Ｂ）。そして最尤経路計算手段９８は、音楽音響信号の全期間における歌声状態と非歌声状態の最尤経路を計算し（図７のステップＳＴ１６Ｃ）、最尤経路から音楽音響信号の全期間における歌声区間と非歌声区間に関する情報を決定する。すなわち歌声の検出には、図８に示すように、歌声状態（S_V）と非歌声状態（S_N）を行き来する隠れマルコフモデル（ＨＭＭ）を用いることとする。歌声状態とは、文字通り「歌声が存在する状態」を表し、「非歌声状態」は歌声が存在しない状態を表している。最尤経路計算手段９８は、次式のように、入力音響信号から抽出された特徴ベクトル列に対して、歌声・非歌声状態の最尤経路
【数１５】

$$\hat{S} = \{\hat{s_1}, \cdots, \hat{s_t}, \cdots\}$$

【００７１】
を検索する。
【数１６】

$$\hat{S} = \underset{S}{\operatorname{argmax}} \sum_t \{\log p(x|s_t) + \log p(s_{t+1}|s_t)\}$$

【００７２】
上記式において、ｐ（ｘ｜s_t）は状態の出力確率を表し、ｐ（s_{t+1}｜s_t）は状態s_{t+1}から状態s_tへの遷移確率を表している。
【００７３】
この歌声区間推定手段９では、前処理以外の通常の推定動作時においては、歌声区間推定用特徴量抽出手段７から各時刻において出力される歌声区間推定用特徴量から、対数尤度計算手段９２が計算した歌声対数尤度及び非歌声対数尤度に、直接重み付けを行って、最尤経路を計算することになる。このような前処理によって対数尤度差のヒストグラムを利用して、歌声対数尤度及び非歌声対数尤度のバイアス調整値η_{dyn}（閾値）を決定すると、音楽音響信号に合ったバイアス調整値η_{dyn}を決定することができる。そしてバイアス調整値η_{dyn}により定めた推定用パラメータηを用いて重み付けを行うと、楽曲ごとの音楽音響信号の音響的特性の違いによって現れる歌声区間推定用特徴量の傾向に合わせて、歌声状態と非歌声状態との境界部を中心にして歌声対数尤度及び非歌声対数尤度を調整することができ、歌声区間及び非歌声区間の境界を、楽曲に合わせて適切に調整することができる。
【００７４】
図１に戻って、時間的対応付け用特徴量抽出手段１１は、各時刻における優勢音音響信号から、歌声の歌詞と優勢音音響信号との間の時間的対応を付けるのに適した時間的対応付け用特徴量を抽出する。具体的な実施の形態では、時間的対応付け用特徴量として、音素の共鳴特性等の２５次元の特徴量を抽出する。この処理は、次のアラインメント処理において必要な前処理に当たる。詳細については図９に示すビタビアラインメントの分析条

件を参照して後述するが、本実施の形態で抽出する特徴量は、１２次元ＭＦＣＣ、１２次元ΔＭＦＣＣ及びΔパワーの２５次元とする。

【００７５】

音素ネットワーク記憶手段１３は、音楽音響信号に対応する楽曲の歌詞に関して複数の音素によって構成された音素ネットワークＳＮを記憶する。このような音素ネットワークＳＮは、例えば、日本語の歌詞であれば、歌詞を音素列に変換し、その後、フレーズの境界を複数個のショートポーズに変換し、単語の境界を１個のショートポーズに変換することにより、母音とショートポーズのみからなる音素列を用いて構成するのが好ましい。与えられた歌詞のテキストデータを元に、アラインメントに用いる文法（これを「アラインメント用の音素列」と定義する。）を作成する。

【００７６】

日本語の歌詞のためのアラインメント用の音素列は、ショートポーズ（sp）すなわち空白と母音と子音のみから構成される。これは、無声子音は調波構造を持たず、伴奏音抑制手法で抽出できないこと、有声子音も発声長が短いため安定して基本周波数Ｆ０を推定するのが難しいことなどがその理由である。具体的な処理としては、まず歌詞をそのまま音素列に変換（実質的には、歌詞を音読したものをローマ字に変換する作業に等しい）し、その後、以下の２つの規則（日本語用の文法）に従って、アラインメント用の音素列に変換する。

【００７７】

ルール１：歌詞中の文やフレーズの境界を複数回のショートポーズ（sp）に変換する。

【００７８】

ルール２：単語の境界を一回のショートポーズに変換する。

【００７９】

図１０は、日本語の歌詞からアラインメント用の音素列（音素ネットワーク）への変換の例を示している。まずオリジナルの歌詞のフレーズを表すテキストデータＡが音素列（Sequence of the phonemes）Ｂに変換される。音素列Ｂに上記「文法」を当てはめることにより、母音と子音とショートポーズ（sp）のみから構成される「アラインメント用の音素列」Ｃに変換される。

【００８０】

この例では、日本語の歌詞「立ち止まる時　また　ふと　振り返る」という歌詞Ａが、「tachidomaru toki　mata futo furikaeru」という音素列Ｂに変換され、さらに、母音と子音とを含む音素とショートポーズ（sp）からなるアラインメント用の音素列Ｃに変換される様子が示されている。このアラインメント用の音素列Ｃが、音素ネットワークＳＮである。

【００８１】

図１に戻って、前述のステップ３を実行するために、アラインメント手段１７は、前述の時間的対応付け用特徴量に基づいて該時間的対応付け用特徴量に対応する音素を推定する歌声用音響モデル１５を備えている。そしてアラインメント手段１７は、音素ネットワーク中の複数の音素と優先音音響信号とを時間的に対応付けるアラインメント動作を実行する。具体的には、アラインメント手段１７は、時間的対応付け用特徴量抽出手段１１からの時間的対応付け用特徴量と、歌声区間推定手段９からの歌声区間と非歌声区間に関する情報と、音素ネットワーク記憶手段１３からの音素ネットワークとを入力として、歌声用音響モデル１５を用いて、少なくとも非歌声区間には音素が存在しないという条件の下で、アラインメントを実行して、音楽音響信号と歌詞の時間的対応付けを自動で行う。

【００８２】

本実施の形態のアラインメント手段１７は、ビタビアラインメントを用いてアラインメント動作を実行するように構成されている。ここで「ビタビアラインメント」とは、音声認識の技術分野において知られるもので、音響信号と文法（アラインメント用の音素列すなわち音素ネットワーク）との間の最尤経路を探索するビタビアルゴリズムを用いた最適解探索手法の一つである。ビタビアラインメントの実行においては、非歌声区間には音素

が存在しないという条件として、少なくとも非歌声区間をショートポーズ（ｓｐ）とする条件を定める。そしてショートポーズ（ｓｐ）においては、他の音素の尤度をゼロとして、アラインメント動作を実行する。このようにするとショートポーズ（ｓｐ）の区間においては、他の音素の尤度がゼロになるため、歌声区間情報を利用することができ、精度の高いアラインメントを行うことができる。

【００８３】

図１１は、「フレーム同期ビタビ探索」と呼ばれるビタビアラインメントを用いて、アラインメント手段１７をプログラムによりコンピュータで実現する場合のプログラムのアルゴリズムを示すフローチャートである。なお以下のアラインメント動作の説明では、歌詞が日本語の場合を例として説明する。ステップＳＴ１０１のｔ＝１は最初の時間的対応付け用特徴量（以下図１１の説明においては、単に特徴量と言う）が入力されるフレームである。ステップＳＴ１０２では、スコア０で空の仮説を作成する。ここで「仮説」とは、今の時刻までの「音素の並び」を意味する。したがって空の仮説を作成するとは、何も音素がない状態にすることを意味する。

【００８４】

次にステップＳＴ１０３では、ループ１として、現在持っているすべての仮説に対して処理をする。ループ１は、前のフレームでの処理が終わった時点で持っている仮説それぞれについてスコアの計算処理を行うループである。例えば、「ａ－ｉ－ｓｐ－ｕ－ｅ・・・」という音素ネットワークとの間の時間的対応を付けると仮定する。この場合に、６フレーム目（６音素目）まで来たときのあり得る仮説（音素の並び）には、「ａａａａａａ」という仮説や、「ａａａｉｉｉ」という仮説や、「ａａｉｉspu」という仮説等の様々な仮説が考えられる。探索の途中では、これら複数の仮説を同時に保持して計算処理が実行される。なおこれらの複数の仮説は、すべて自分のスコアを持っている。ここでスコアとは、６フレームまであるとしたとき、１～６フレームまでの特徴量それぞれが、例えば「ａａａｉｉｉ」という音素の並びであった可能性（対数尤度）を、特徴量と音響モデルとを比較することにより計算したものである。例えば、６フレーム目（ｔ＝６）の処理が終わり、７フレーム目の処理が始まると、現在保持しているすべての仮説に対して計算処理が行われる。このような処理をすることがループ１の処理である。

【００８５】

次にステップＳＴ１０４では、音素ネットワークを元に仮説を「１フレーム展開」する。ここで「１フレーム展開」するとは、仮説の長さを１フレーム延ばすことを意味する。そして展開した場合には、一つ次の時刻のフレームまで考慮に入れることにより、１つの仮説に新たな音素が続いて複数の新たな仮説ができる可能性がある。次に続く可能性のある音素を見つけるために、音素ネットワークが参照される。例えば、「ａａａｉｉｉ」という仮説については、音素ネットワークを参照すると、次のフレームでは「ａａａｉｉｉｉ」というように「ｉ」が続く場合と、「ａａａｉｉｉsp」というようにショートポーズspに移る場合の２通りの新しい仮説が考えられる。この場合には、１つの仮説を「１フレームに展開」すると次の時刻のフレームまで考慮した新しい２つの仮説が出ることになる。ステップＳＴ１０５では、ループ２として、すべての仮説について１フレーム展開されて発生した新たなすべての仮説に対して、スコアを計算する。スコアの計算は、ループ１におけるスコアの計算と同じである。ループ２は、保持しているそれぞれの仮説から新たに幾つかの仮説が展開されるので、その新しく展開されたそれぞれの仮説についてスコア計算の処理を行うループである。

【００８６】

次にステップＳＴ１０６では、歌声区間推定手段９からの歌声区間情報を入力として、ｔ番目のフレームが歌声区間であるか又は音素がショートポーズ(sp)であるか否かの判定が行われる。例えば、７フレーム目は非歌声区間であるという歌声区間情報があるとする。この場合に、７フレーム目で仮説を展開した時点で、「ａａａｉｉｉsp」という仮説はあっても、「ａａａｉｉｉｉ」という仮説はあり得ないことになる。このようなあり得ない仮説は、ステップＳＴ１０７で棄却される。このように歌声区間情報があると、ステッ

プＳＴ１０６及び１０７を経て、あり得ない仮説が棄却できるため、アラインメントが容易になる。ステップＳＴ１０６において、Ｙｅｓの判定がなされると、ステップＳＴ１０８へと進む。

【００８７】

ステップＳＴ１０８では、入力された特徴量と音響モデルとを用いて、ｔ番目の特徴量の音響スコアを計算し、仮説のスコアに加算する。すなわちｔ番目の特徴量を音響モデルと比較して、対数尤度（スコア）を計算し、計算したスコアを仮説のスコアに加算する。結局、スコアの計算は、特徴量と音響モデルとを比較し、特徴量が音響モデル中にある複数の音素についての情報にどの程度似ているのかを計算していることになる。なおスコアは対数で計算するため、全く似ていないといった場合には、その値は－∞となる。ステップＳＴ１０８では、すべての仮説についてスコアの計算が行われる。ステップＳＴ１０８での計算が終了すると、ステップＳＴ１０９へと進み、仮説とスコアとが保持される。そしてステップＳＴ１１０ではステップＳＴ１０５に対応したループ２が終了する。ステップＳＴ１１１ではステップＳＴ１０３に対応したループ１が終了する。その後、ステップＳＴ１１２で、現在の処理対象時刻を１増加させ（ｔ＋１）、次のフレームに進む。そしてステップＳＴ１１３で、フレームが入力されてくる複数の特徴量の終端であるか否かの判断がなされる。すべての特徴量が入力されるまでは、ステップＳＴ１０３からステップＳＴ１１２までの各ステップが繰り返し実行される。すべての特徴量について処理が終了すると、ステップＳＴ１１４へと進む。この時点では、特徴量と音響モデルとの比較は、音素ネットワークの終端に達している。そして音素ネットワークの終端に達している複数の仮説の中から合計スコアが最大の仮説（音素の並び）を最終決定された仮説として選ぶ。この最終決定された仮説すなわち音素の並びは、時刻と対応している特徴量を基準にして定められている。すなわちこの最終決定された音素の並びは、音楽音響信号と同期した音素の並びになっている。したがってこの最終決定された音素の並びに基づいて表示される歌詞のデータが、時間タグ付きの（音楽音響信号と同期するための時刻情報が付いた）歌詞となる。

【００８８】

図１２（Ａ）は、ビタビアラインメントを利用して、時刻において音楽音響信号から抽出した優勢音音響信号の信号波形Ｓ′（伴奏音が抑制された音響信号の音声波形）に対して、音素ネットワーク（文法）を時間的に対応付けた様子を示している。アラインメントが完了した後は、時間情報を伴ったアラインメント用の音素列（文法）から逆に歌詞に戻すことで、最終的に、時間情報を含む「時間タグ付き歌詞データ」が得られる。図１２（Ａ）では図示を簡単にするために母音のみを示してある。

【００８９】

図１２（Ｂ）は、アラインメントが完了した後、音素列（文法）から歌詞に戻すことによって伴奏音を含む混合音の音楽音響信号Ｓと歌詞の時間的対応付けが完了した様子を示している。ＰＡ～ＰＤは、それぞれ歌詞のフレーズである。

【００９０】

次にアラインメント手段１７で使用する歌声用音響モデル１５について説明する。使用する歌声用音響モデル１５としては、歌声の発話内容（歌詞）に対してアラインメントを行うため、大量の歌声のデータから学習された音響モデルを使用することが理想的である。しかしながら、現段階ではそのようなデータベースは構築されていない。そこで本実施の形態では、話し声用の音響モデルのパラメータを、歌声と伴奏音を含む楽曲中の歌声の音素を認識できるように再推定して（学習して）得た音響モデルを用いる。

【００９１】

話し声用の音響モデルをベースにして歌声用音響モデルを作る手法（適応：adaptation）は、以下のように３段階ある。なお事前の作業として、「話し声用の音響モデル」を準備するステップが必要であるが、この点は公知であるので省略する。

【００９２】

（１）話し声用の音響モデルを単独歌唱の歌声に適応させる。

【００９３】
（２）単独歌唱用の音響モデルを伴奏音抑制手法によって抽出された分離歌声に適応させる。
【００９４】
（３）分離歌声用の音響モデルを入力楽曲中の特定楽曲（特定歌手）に適応させる。
【００９５】
これら（１）乃至（３）段階は、いずれも図２における「学習時」の処理に対応するものであり、実行時よりも前に行うものである。
【００９６】
（１）段階の適応では、図２に示すように、話し声用音響モデル１０１を音素ラベル１０２（教師情報）及び伴奏音を伴わない歌声だけのすなわち単独歌唱の歌声１０３に適応させて単独歌唱用の音響モデル１０４を生成する。（２）の適応では、単独歌唱用の音響モデル１０４を、伴奏音抑制手法によって抽出された優勢音音響信号からなる歌声データ１０５及び音素ラベル１０２（教師情報）に適応させて、分離歌声用の音響モデル１０６を生成する。（３）の適応では、分離歌声用の音響モデル１０６を、入力楽曲の特定楽曲の音素ラベル（音素ネットワーク）と特徴量とに適応させて、特定歌手用音響モデル１０７を生成する。図２の例では、図１の歌声用音響モデル１５として、特定歌手用音響モデル１０７を用いている。
【００９７】
なお、（１）乃至（３）は必ずしも全て実施する必要はなく、例えば（１）のみを実施する場合（これを「１段階適応」という。）、（１）及び（２）を実施する場合（これを「２段階適応」という。）、及び（１）乃至（３）を全て実施する場合（これを「３段階適応」という。）、などのように、一つ又は複数を適宜組み合わせて、音響モデルの適応を実施することができる。
【００９８】
ここで、教師情報とは、各音素ごとの時間情報(音素の始端時間、終端時間)を指している。従って、単独歌唱データ１０３や音素ラベル１０２のような教師情報を用いて、話し声用の音響モデルを適応させる場合は、時間情報により正確にセグメンテーションされた音素データを用いて適応が行われる。
【００９９】
図１３は、時間情報を伴う日本語の歌詞の場合の適応用音素ラベル１０２の一例を示している。なお、図１３の音素ラベル１０２は手動で付与した。適応時のパラメータ推定には、最尤線形回帰ＭＬＬＲ（Maximum Likelihood Linear Regression）と最大事後確率ＭＡＰ（Maximum a Posterior）推定を組み合わせることができる。なお、ＭＬＬＲとＭＡＰを組み合わせるということの意味は、ＭＬＬＲ適応法で得られた結果を、ＭＡＰ推定法における事前分布（初期値のようなもの）として使用することを意味する。
【０１００】
以下さらに音響モデルの具体的な適応技術について説明する。図１４は、前述の１段階適応の詳細を示すフローチャートである。１段階適応では、歌声用音響モデル１５としては、歌声だけを含む単独歌唱のデータすなわち適応用音楽音響信号１０３を、適応用音楽音響信号１０３に対する適応用音素ラベル１０２を元に音素ごとに分割する。そして音素ごとに分割されたデータを用いて、話し声用音響モデル１０１のパラメータを、適応用音楽音響信号１０３から歌声の音素を認識できるように再推定して単独歌唱用の音響モデル１０４を得る。この音響モデル１０４は、伴奏音が無いかまたは伴奏音が歌声に比べて小さい場合に、適している。
【０１０１】
また図１５は、前述の２段階適応の詳細を示すフローチャートである。２段階適応では、歌声に加えて伴奏音を含む適応用音楽音響信号から抽出した歌声を含む最も優勢な音の優勢音音響信号１０５を適応用音素ラベル１０２を元に音素ごとに分割する。そして音素ごとに分割されたデータを用いて、単独歌唱用の音響モデル１０４のパラメータを、優勢

音音響信号１０５から歌声の音素を認識できるように再推定して得た分離歌声用の音響モデル１０６を得る。このような分離歌声用の音響モデル１０６は、歌声と同様に伴奏音が大きい場合に適している。

【０１０２】

さらに図１６は、前述の３段階適応の詳細を示すフローチャートである。３段階適応では、システムの実行時に入力された歌声と伴奏音とを含む音楽音響信号Ｓ１から伴奏音抑制法により伴奏音を抑制して得た優勢音音響信号Ｓ２を用いる。そしてシステムに入力された音楽音響信号から抽出した歌声を含む最も優勢な音の優勢音音響信号Ｓ２から時間的対応付け用特徴量抽出手段１１によって抽出された複数の時間的対応付け用特徴量と入力された音楽音響信号に対する音素ネットワークＳＮを用いて、分離歌声用の音響モデル１０６のパラメータを音楽音響信号の楽曲を歌う特定の歌手の音素を認識できるように推定して特定歌手用の音響モデル１０７を得る。この特定歌手用の音響モデル１０７は、歌手を特定した音響モデルであるため、アラインメントの精度を最も高くすることができる。

【０１０３】

なお音楽音響信号に時間的に対応付けられた歌詞を、表示画面上に表示させながら音楽音響信号を再生する音楽音響信号再生装置において、本発明のシステムを用いて音楽音響信号に時間的に対応付けられた歌詞を表示画面に表示させると、再生される音楽と画面に表示される歌詞とが同期させて表示画面に表示することができる。

【０１０４】

本発明の音楽音響信号と歌詞の時間的対応付けを自動で行う方法を、図１及び図２を用いて説明する。まず歌声と伴奏音とを含む楽曲の音楽音響信号Ｓ１から、各時刻において歌声を含む最も優勢な音の優勢音音響信号Ｓ２を優勢音響信号抽出手段５が抽出する（優勢音響信号抽出ステップ）。次に各時刻における優勢音音響信号Ｓ２から歌声が含まれている歌声区間と歌声が含まれていない非歌声区間とを推定するために利用可能な歌声区間推定用特徴量を歌声区間推定用特徴量抽出手段７が抽出する（歌声区間推定用特徴量抽出ステップ）。そして複数の歌声区間推定用特徴量に基づいて、歌声区間と非歌声区間を歌声区間推定手段が推定して、歌声区間と前記非歌声区間に関する情報を出力する（歌声区間推定ステップ）。また各時刻における優勢音音響信号Ｓ２から、歌声の歌詞と音楽音響信号との間の時間的対応を付けるのに適した時間的対応付け用特徴量を時間的対応付け用特徴量抽出手段１１が抽出する（時間的対応付け用特徴量抽出ステップ）。さらに音楽音響信号Ｓ１に対応する楽曲の歌詞の複数の音素が、該複数の音素の隣りあう二つの音素の時間的間隔が調整可能に繋がって構成された音素ネットワークＳＮを音素ネットワーク記憶手段１３に記憶する（記憶ステップ）。そして時間的対応付け用特徴量に基づいて該時間的対応付け用特徴量に対応する音素を推定する歌声用音響モデル１５を備え、音素ネットワークＳＮ中の複数の音素と優先音音響信号Ｓ１とを時間的に対応付けるアラインメント動作をアラインメント手段１７が実行する（アラインメントステップ）。このアラインメントステップでは、アラインメント手段１７が、時間的対応付け用特徴量抽出ステップで得られる時間的対応付け用特徴量と、歌声区間と非歌声区間に関する情報と、音素ネットワークＳＮとを入力として、歌声用音響モデル１５を用いて、少なくとも非歌声区間には音素が存在しないという条件の下で、アラインメント動作を実行する。

【０１０５】

一般に、歌声の検出は、正解率（hit rate）と棄却率（correct rejection rate）によって評価される。但し、正解率とは実際に歌声を含む領域のうち、正しく歌声区間として検出できた割合を指し、棄却率とは実際に歌声を含まない領域のうち、正しく非歌声区間として棄却できた割合を指すものとする。なお、本上記実施の形態で採用した歌声区間推定手段９は、正解率と棄却率のバランスを調整することができる仕組みとなっている。このような仕組みが必要になる理由は、正解率と棄却率の基準はいわばトレードオフの関係にあるからであり、適切な関係は例えば用途によっても異なるものだからである。歌声検出区間の推定は、ビタビアラインメントの前処理としての意味を持つため、正解率をある程度高く保つことによって歌声を含む可能性が少しでもあれば漏れなく検出できるように

することが一般的には望ましい。しかし、その一方で、歌手名の同定などの用途に用いる場合は、棄却率を高く保つことによって、確実に歌声を含む部分のみを検出するべきである。ちなみに、歌声の検出に関する従来技術では、正解率と棄却率のバランスを調整できるものはなかった。

【０１０６】

次に本発明を適用した実施の形態の評価結果について説明する。

【０１０７】

本発明に係る方法を実際に市販されているディジタル音楽データと歌詞データに適用し、再生と同期した歌詞の表示を実験により確かめた。その結果、本発明に係る方法によると、様々な伴奏音を含む実世界の音楽音響信号に対して頑健にその歌詞を時間的に対応付けることができることが確認された。以下、評価実験の方法について説明する。

【０１０８】

（実験方法）

公的な研究用音楽データベースの一つであるＲＷＣ研究用音楽データベースに登録されているポピュラー音楽データベース（ＲＷＣ－ＭＤＢ－Ｐ－２００１）から、１０歌手１０曲（男性歌手５曲・女性歌手５曲）をランダムに抽出した。

【０１０９】

楽曲の大半の部分は日本語で歌われているが、一部は英語で歌われている。本実験では、英語の音素は類似した日本語の音素の音響モデルを用いて近似した。これらの楽曲に対して、性別毎の５ｆｏｌｄ　ｃｒｏｓｓ－ｖａｌｉｄａｔｉｏｎ法で評価をした。つまり、ある歌手によって歌われている楽曲を評価する際は、その歌手と同じ性別の歌手によって歌われている他の楽曲を用いて音響モデルを適応させた。

【０１１０】

歌声区間検出手法の学習データには、ランダムに選ばれた１１歌手からなる１９曲を用いた。なお、これらの楽曲も“ＲＷＣ音楽データベース:ポピュラー音楽（RWC-MDB-P-2001）”から抽出した。

【０１１１】

また、これらの１１歌手は学習用のデータであるため、評価に用いられた１０歌手には含まれていない。歌声区間検出手法の学習データにも、伴奏音抑制手法は適用した。また、η_{fixed}の値は１５に設定した。

【０１１２】

前述の図９は、ビタビアラインメントの分析条件を示している。初期音響モデルとしては、ＣＳＲＣ ソフトウェア中の性別非依存モノフォンモデルを用いた。また、歌詞から音素列の変換には、日本語形態素解析システム茶筅（ChaSen）を実行し、その際に出力される読みの情報を用いた。音響モデルの適応には、Hidden Markov Toolkit （HTK）を用いた。

【０１１３】

評価は、フレーズ単位のアラインメントを元に行った。本実験では、フレーズとは、元歌詞中のスペースや改行で区切られた一節を意味するものとする。

【０１１４】

図１７は、評価基準を説明するための図である。まず、図１７に示すように、「正解していた区間」とは、正解ラベルと出力結果とが重複している時間を指し、その他を「不正解」とする。楽曲の全体長（正解区間と不正解区間の長さの総和）に対する、正解区間の長さの総和を「正解率」［＝正解区間の長さの総和（Length of "correct " regions)／楽曲の全体長さ(Total length of the song）］と定義した。例えば図１０の例であれば、「立ち止まる時」と「またふと振り返る」がそれぞれ、１フレーズを構成している。

【０１１５】

そして、全体の評価基準として、楽曲の全体長の中で、フレーズ単位のラベルが正解していた区間の割合を計算した。精度が９０％を超えていた場合に、その楽曲は正しくアラインメントされたと判断した。

【０１１６】
（システム全体の評価）
提案手法全体での性能を評価するため、発明に係る方法により実験を行った。
【０１１７】
図１８（Ａ）及び（Ｂ）は、本発明の効果を確認するための評価実験の結果を示している。図１８（Ａ）に示すとおり、＃００７と＃０１３の２曲を除き１０曲中８曲で９０％以上のアラインメントの正解率を達成した。また、図１８（Ｂ）はフレーズの開始時刻の平均誤差を楽曲別に示した結果を示す一覧表である。
【０１１８】
これらの結果は、本手法により１０曲中８曲について十分な精度で時間的対応を推定することができることを示している。また、男声の精度が女性の精度に比べて高いことが見て取れる。これは、女声は一般に男声よりも高いＦ０を持つため、ＭＦＣＣなどのスペクトル特徴量を抽出するのが困難であるからである。代表的な誤りは、歌詞に書かれていないハミング等が歌われている部分で発生していた。
【０１１９】
（音響モデル適応の効果の確認）
音響モデルを適応させた効果を確認することを目的として、以下の４つの条件でアラインメント実験を行った。
【０１２０】
（ｉ）適応なし：音響モデル適応を行わなかった。
【０１２１】
（ｉｉ）１段階適応：話し声用の音響モデルを直接分離歌声に適応させた。特定歌手への教師なし適応は行わなかった。
【０１２２】
（ｉｉｉ）２段階適応：まず、話し声用の音響モデルを単独歌唱音声に適応させた後、分離歌声に適応させた。特定歌手への教師なし適応は行わなかった。
【０１２３】
（ｉｖ）３段階適応（提案手法）：まず、話し声用の音響モデルを単独歌唱音声に適応させた後、分離歌声に適応させた。最後に、入力音響信号の特定歌手への教師なし適応を行った。なお、本実験では（ｉ）乃至（ｉｖ）全ての条件について伴奏音抑制（ステップ１）と歌声区間検出（ステップ２）を適用した。
【０１２４】
図１９（Ａ）及び（Ｂ）は、条件（ｉ）乃至（ｉｖ）とした場合の実験の結果を示している。このうち、図１９（Ａ）は、各楽曲に対するアラインメントの正解率をそれぞれの条件ごとに調べた結果を示している。また、図１９（Ｂ）は、その正解率を数値で一覧表にまとめたものである。
【０１２５】
これらの結果は、全ての楽曲で一定の効果があることを示している。特に、条件（ｉｖ）が最も正解率が高いことが分かる。この意味において、条件（ｉｖ）は発明を実施するための最良の形態であるということができる。
【０１２６】
（歌声区間検出の評価）
次に、ステップ２において説明した歌声区間検出の有効性を確認することを目的として、各楽曲に対する歌声区間検出の正解率（hit rate）と棄却率（correct rejection rate）を調べた。
【０１２７】
また、これと共に歌声区間検出自体の性能の評価も行った。これについては歌声区間検出を用いた場合と用いない場合の２通りの条件で実験した。本実験では、適応処理には全て３段階（ステップ１乃至ステップ３）の適応手法を使用した。
【０１２８】

図２０（Ａ）は、各楽曲に対する歌声区間検出の正解率（hit rate）と棄却率（correct rejection rate）を示している。また、図２０（Ｂ）は各楽曲に対するアラインメントの正解率を、歌声区間検出有りの場合と無しの場合の比較を示している。

【０１２９】

これらの結果から、平均的に見ると、歌声区間検出を適用することによってアラインメントの正解率が向上したと評価できる。特に、図２０（Ｂ）の結果から明らかなように、比較的精度が低い楽曲に歌声区間検出を適用したとき、特にアラインメントの正解率が向上していることがわかる。但し、＃００７と＃０１３に関しては、元々精度が低い楽曲に適用されたにもかかわらず、歌声区間検出手法の効果が薄い。この理由は、これらの楽曲は、図２０（Ａ）に見られるように、歌声区間検出の棄却率が高くないため非歌声区間を十分に除去できなかったからであると考えられる。

【０１３０】

また、＃０１２や＃０３７などのように、元々アラインメントの正解率が高い楽曲に歌声区間検出を行うと、正解率が僅かながら低下していることがわかる。これは、歌声区間検出で誤って除去（棄却）されてしまった歌声区間は、アラインメントの際には必ず不正解となるからと考えられる。

【０１３１】

なお、上述の通り、本発明では、日本語歌詞の楽曲を用いて実験を行い動作を確認した。しかし英語楽曲においては、英語の音素を発音が最も近い日本語の音素に変換して音素ネットワークを作成することで、英語の楽曲に対しても、比較的高い精度で時間的対応付けが推定できることを確認した。対象の楽曲の言語に応じて適切な音響モデルと音響モデル適応用データを準備することができれば、英語を含む他の言語の楽曲についても、より高い精度時間的対応付けが推定可能である。

【０１３２】

さらに、楽曲中に含まれる部分的な繰り返し部分やテンポなどの高次の楽曲構造情報を利用することで、より高度な音楽と歌詞の時間的対応付けが可能になると考えられる。

【０１３３】

本発明に係る音楽音響信号と歌詞の時間的対応付け方法は、現時点では各ステップがツールキットなどの形で配布されるそれぞれ独立したプログラムで構成されているが、用途に応じて適切にプログラミングすれば、一つのコンピュータプログラムの形で実施されることも考えられる。その具体的な本発明の応用例としては、以下のような適用事例が考えられる。

【０１３４】

（適用事例１）再生と同期した歌詞の表示

再生と同期した歌詞の表示を行うという用途である。本件発明者らは、時間タグ付き歌詞に基づき音楽の再生と時間的に同期して歌詞の色を変化させる音楽ディジタルデータ再生用ソフトウェアを同時に開発することで、再生中の歌声と時間的に同期して歌詞の色を変化させることに成功し、アラインメントの正解率は上記の通りであることを確認した。

【０１３５】

なお、表示されている画面上に歌詞が表示され、歌声と共に色が変化する動作は、一見するといわゆるカラオケのように見えるが、フレーズと歌詞の追随が極めて正確であり、楽曲の鑑賞が一層充実するという印象を得た。しかも、人間を介することなくプログラムによって自動的に対応付けされたものである点で、従来のものとは全く異質のものである。

【０１３６】

（適用事例２）歌詞を用いた楽曲の頭出し

本発明に係る方法によって歌詞に時間情報が得られる場合、予め歌詞を表示させておき、歌詞の一部をクリックするとそこから演奏が開始されるようにプログラミングすることも可能である。

【０１３７】

本件発明者らは，前記の本件発明者らが開発した音楽ディジタルデータ再生用ソフトウェアに機能を追加することで、歌詞をクリックすることで、そこから演奏が開始させることに成功した。この動作は、今までには実現されていなかった機能であり、ユーザの好みの部分を能動的に選択しながら楽曲を鑑賞出来るという点で新しい音楽鑑賞方法を実現したと言える。

【０１３８】

なお、上記適用事例１及び２においては，本件発明者らが独自に開発した音楽ディジタルデータ再生ソフトウェアを使用しているが，これに限定されずに他の音楽ディジタルデータ再生用ソフトウェアを用いてもよいのは勿論である．

【産業上の利用可能性】

【０１３９】

本発明は、音楽鑑賞支援技術或いは検索技術といった産業上の利用分野に適用されることが期待されるものであり、特に、近年のディジタル音楽データ配信サービスの普及に伴い、その重要性は一層増大しているものと考えられる。

【図面の簡単な説明】

【０１４０】

【図１】音楽音響信号と歌詞の時間的対応付けを自動で行うシステムの実施の形態をコンピュータを用いて実現する場合に、コンピュータ内に実現される機能実現手段の構成を示すブロック図である。

【図２】図１の実施の形態をプログラムをコンピュータで実行することにより実施する場合のステップを示すフローチャートである。

【図３】伴奏音抑制処理について、その処理手順を示す図である。

【図４】（Ａ）乃至（Ｄ）は、音楽音響信号から優勢音音響信号を抽出する仮定を説明するために用いる波形図である。

【図５】歌声区間推定手段の具体的な構成を示すブロック図である。

【図６】図５に示した歌声区間推定手段をプログラムにより実現する場合のフローチャートである。

【図７】歌声区間の検出をプログラムで実現する際のフローチャートである。

【図８】歌声状態（S_V）と非歌声状態（S_N）を行き来する隠れマルコフモデル（ＨＭＭ）を用いることを説明するために用いる図である。

【図９】ビタビアラインメントの分析条件を示す図である。

【図１０】歌詞からアラインメント用の音素列への変換の例を示す図である。

【図１１】アラインメント手段をプログラムによりコンピュータで実現する場合のプログラムのアルゴリズムを示すフローチャートである。

【図１２】（Ａ）はビタビアラインメントを利用して、時刻において音楽音響信号から抽出した優勢音音響信号の信号波形に対して、音素ネットワークを時間的に対応付けた様子を示す図であり、（Ｂ）はアラインメントが完了した後、音素列から歌詞に戻すことによって伴奏音を含む混合音の音楽音響信号と歌詞の時間的対応付けが完了した様子を示す図である。

【図１３】時間情報を伴う適応用音素ラベルの一例を示す図である。

【図１４】音響モデルを作成する場合の流れを示すフローチャートである。

【図１５】音響モデルを作成する場合の流れを示すフローチャートである。

【図１６】音響モデルを作成する場合の流れを示すフローチャートである。

【図１７】評価基準を説明するための図である。

【図１８】（Ａ）及び（Ｂ）は、本発明の効果を確認するための評価実験の結果を示している。

【図１９】（Ａ）及び（Ｂ）は、条件（ｉ）乃至（ｉｖ）とした場合の実験の結果を示している。このうち、図１９（Ａ）は、各楽曲に対するアラインメントの正解率をそれぞれの条件ごとに調べた結果を示している。図１９（Ｂ）は、その正解率を数値で一覧表にまとめたものである。

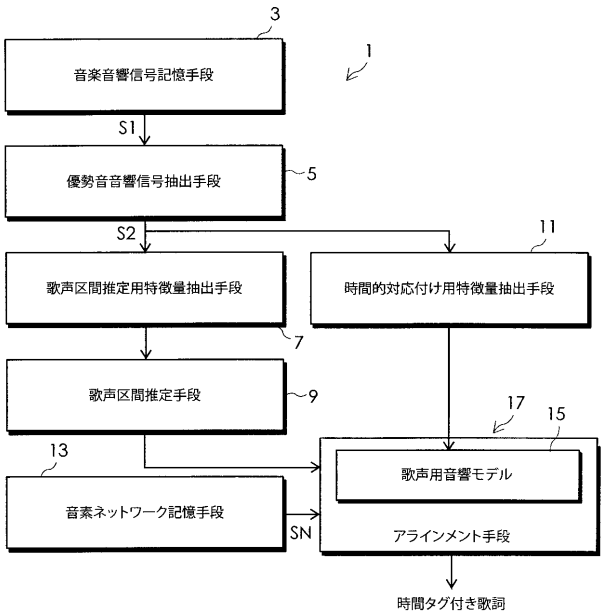
【図２０】（Ａ）は各楽曲に対する歌声区間検出の正解率（hit rate）と棄却率（correct rejection rate）を示している。（Ｂ）は楽曲に対するアラインメントの正解率を、歌声区間検出有りの場合と無しの場合の比較を示している。

【符号の説明】

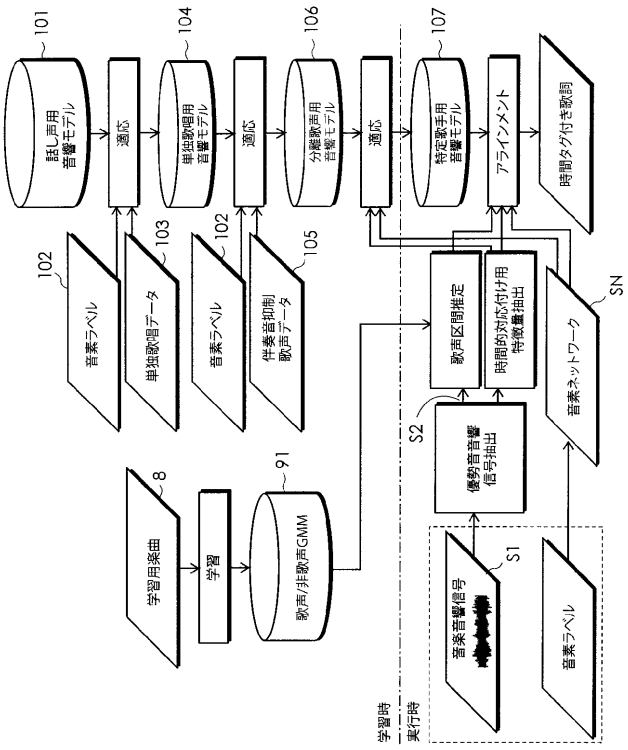
【０１４１】

１　音楽音響信号と歌詞の時間的対応付けを自動で行うシステム
３　音楽音響御信号記憶手段
５　優勢音音響信号抽出手段
７　歌声区間推定用特徴量抽出手段
９　歌声区間推定手段
１１　時間的対応付け用特徴量抽出手段
１３　音素ネットワーク記憶手段
１５　歌声用音響モデル
１７　アラインメント手段

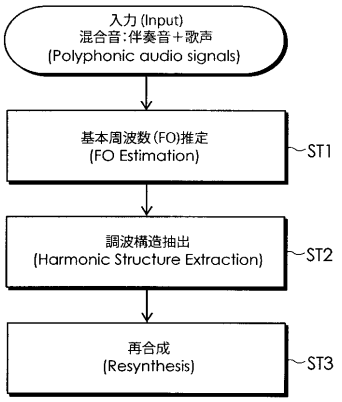
【図１】



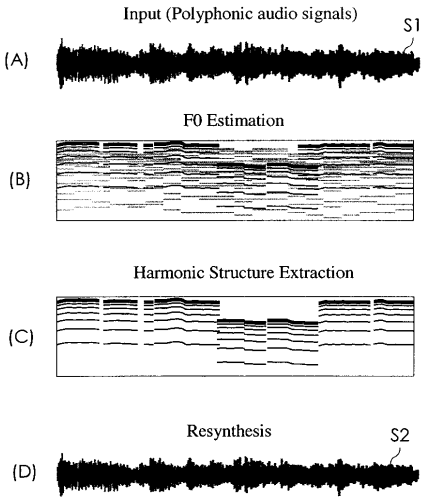
【図２】



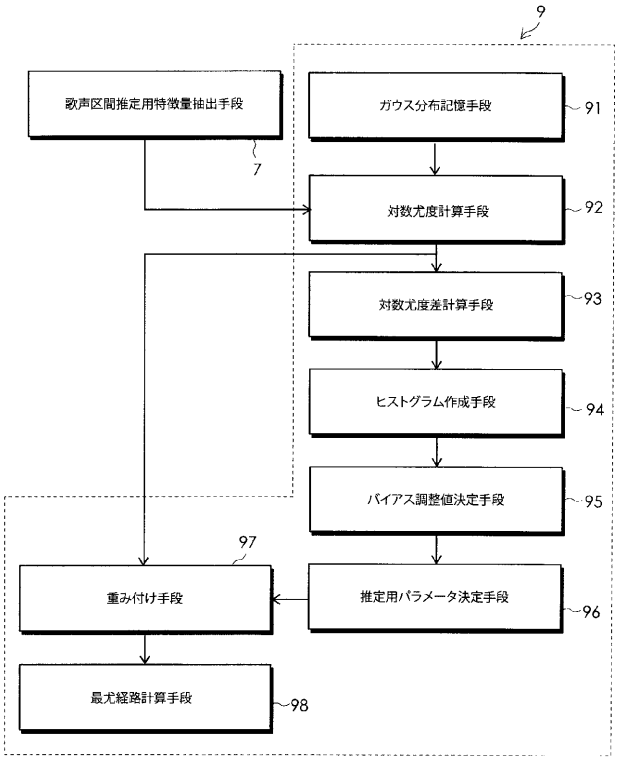
【図３】


入力 (Input)
混合音:伴奏音＋歌声
(Polyphonic audio signals)
基本周波数 (FO)推定
(FO Estimation)
ST1
調波構造抽出
(Harmonic Structure Extraction)
ST2
再合成
(Resynthesis)
ST3

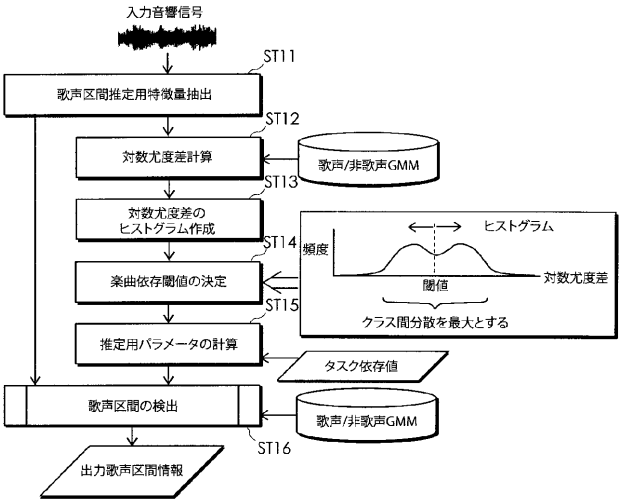
【図４】


Input (Polyphonic audio signals)
S1
(A)
F0 Estimation
(B)
Harmonic Structure Extraction
(C)
Resynthesis
S2
(D)

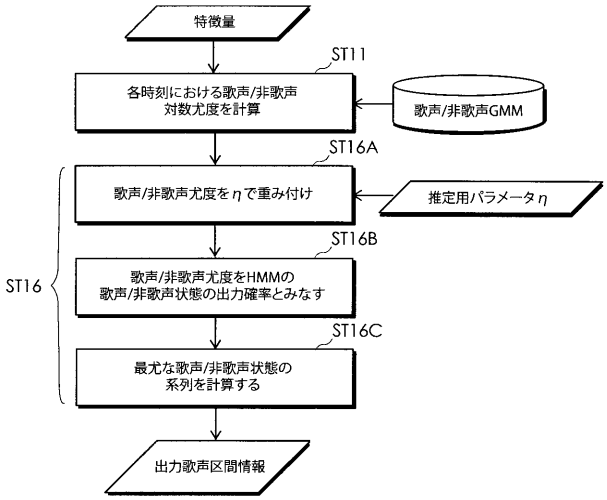
【図５】


9
歌声区間推定用特徴量抽出手段
7
ガウス分布記憶手段
91
対数尤度計算手段
92
対数尤度差計算手段
93
ヒストグラム作成手段
94
バイアス調整値決定手段
95
97
重み付け手段
推定用パラメータ決定手段
96
最尤経路計算手段
98

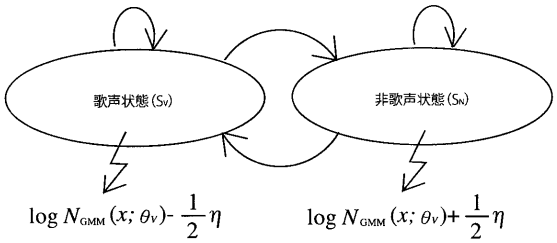
【図６】


入力音響信号
ST11
歌声区間推定用特徴量抽出
ST12
対数尤度差計算
歌声/非歌声GMM
ST13
対数尤度差の
ヒストグラム作成
ST14
楽曲依存閾値の決定
ST15
推定用パラメータの計算
タスク依存値
歌声区間の検出
歌声/非歌声GMM
ST16
出力歌声区間情報
ヒストグラム
頻度
対数尤度差
閾値
クラス間分散を最大とする

【図７】



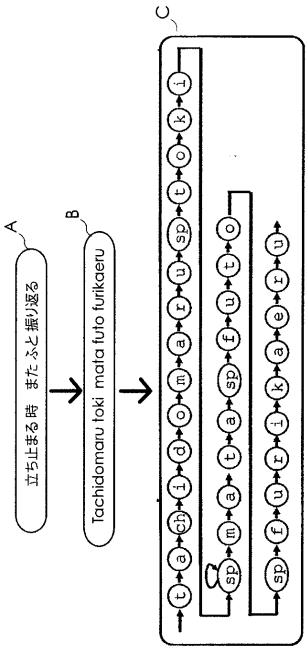
【図８】



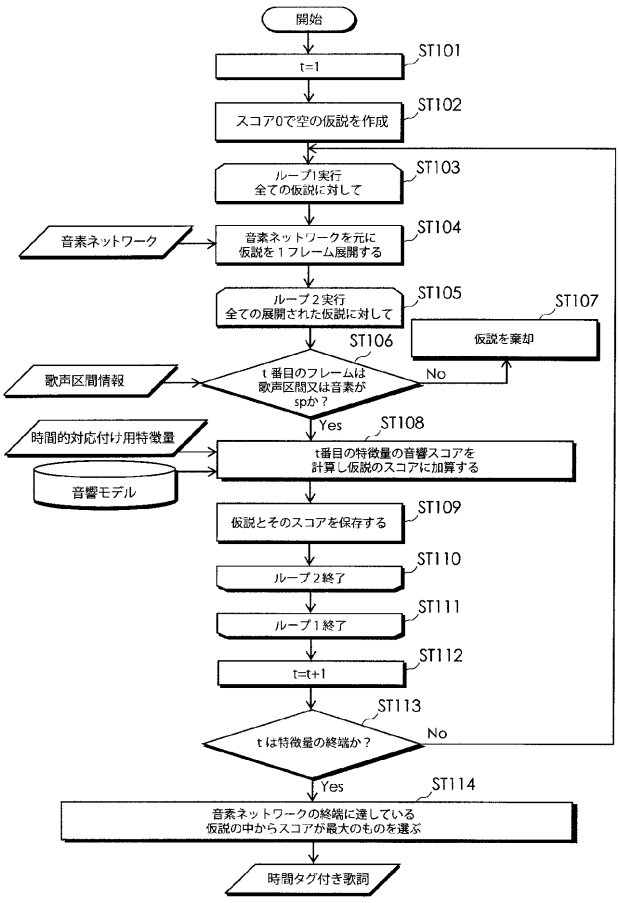
【図９】

サンプリング	16 kHz, 16 bit
窓関数	Hamming 窓
フレーム幅	25 ms
フレームシフト	10 ms
サンプリング	12 次元 (12th order MFCC) 12 次元 (12th order ΔMFCC) Δパワー (ΔPower)

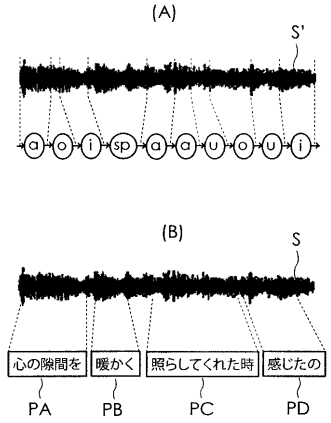
【図１０】



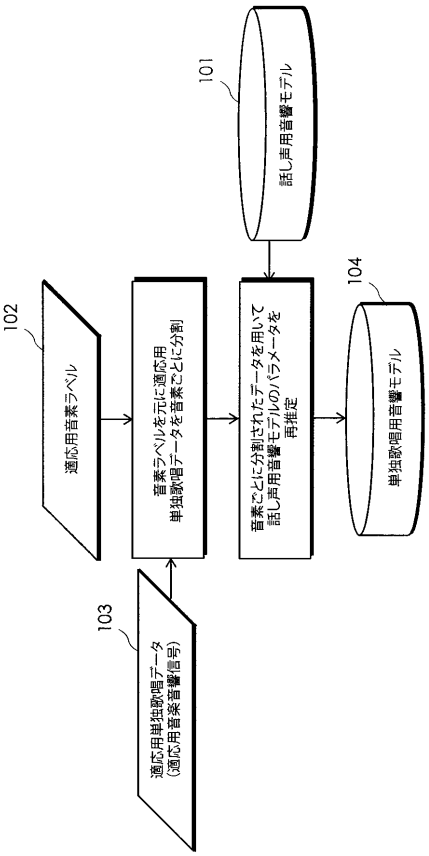
【図１１】



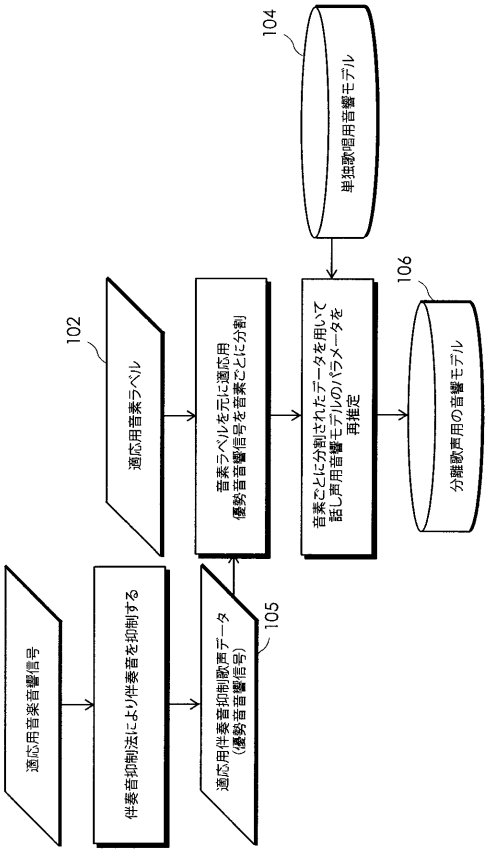
【図１２】



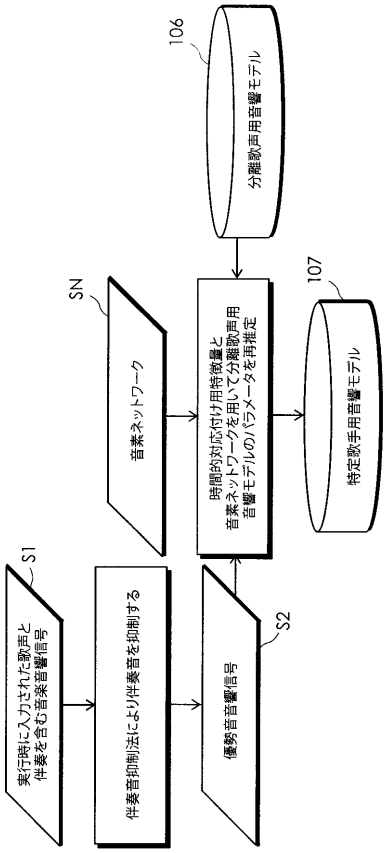
【図１４】



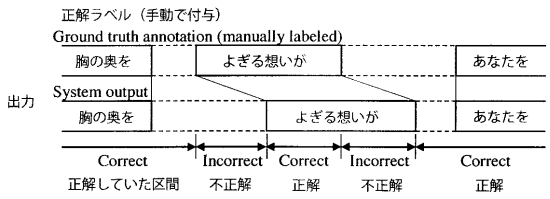
【図１５】



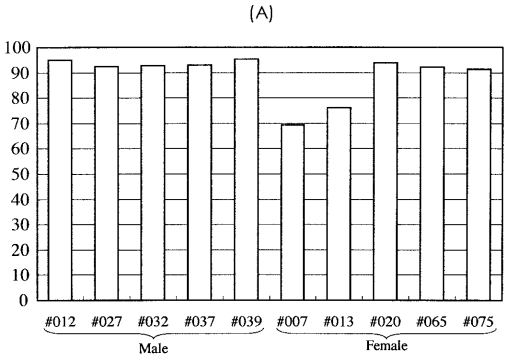
【図１６】



【図１７】



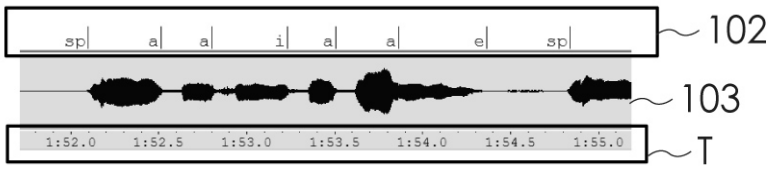
【図１８】



(B)

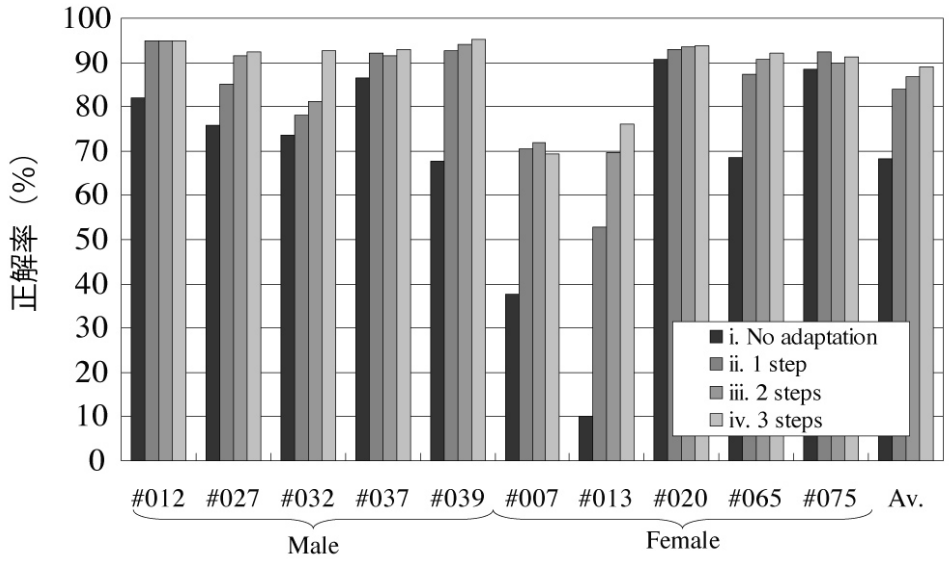
男声	#012	0.3868[sec]
	#027	0.2717[sec]
	#032	0.1292[sec]
	#037	0.1108[sec]
	#039	0.7783[sec]
女声	#007	1.9318[sec]
	#013	1.4954[sec]
	#020	0.2911[sec]
	#065	0.2578[sec]
	#075	0.1182[sec]

【図１３】



【図１９】

(A)

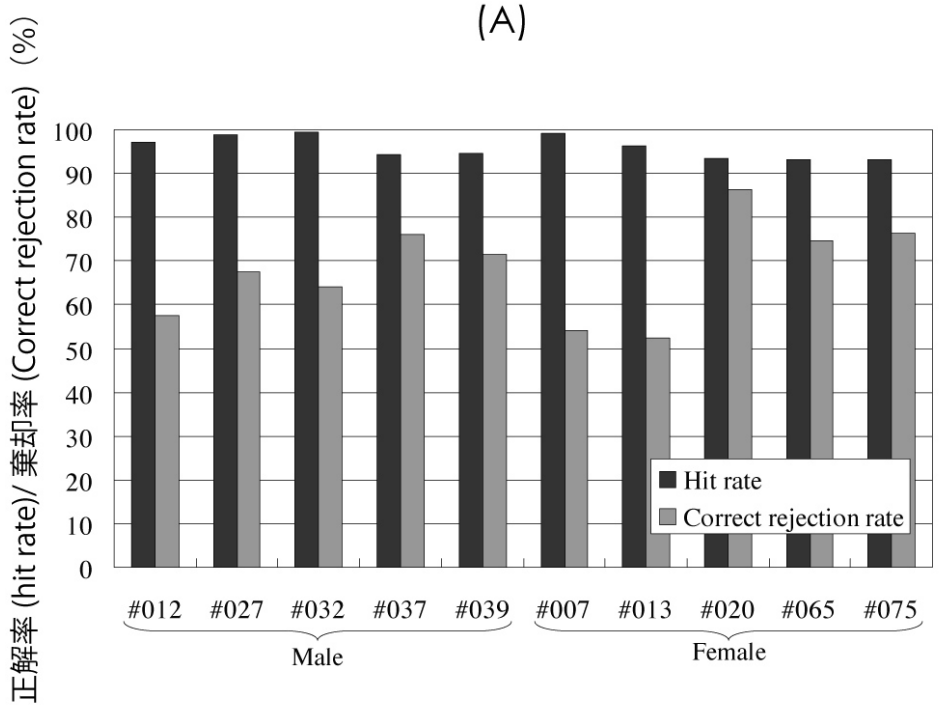


(B)

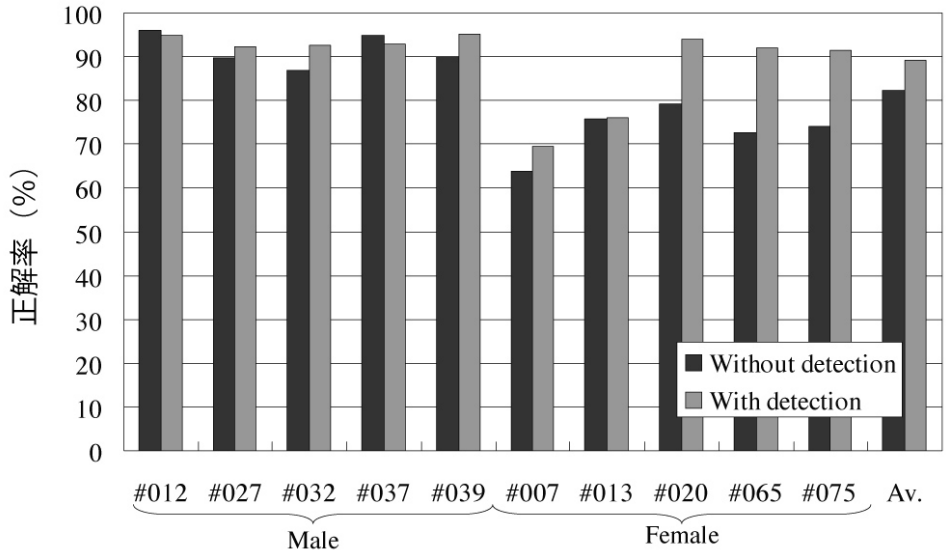
	単独歌唱への適用	分離歌声への適用	特定歌手への適用	正解率
条件ｉ	×	×	×	68%
条件ii	×	○	×	84%
条件iii	○	○	×	87%
条件iv	○	○	○	89%

【図２０】

(A)


正解率 (hit rate)/ 棄却率 (Correct rejection rate) (%)
100
90
80
70
60
50
40
30
20
10
0
Hit rate
Correct rejection rate
#012
#027
#032
#037
#039
#007
#013
#020
#065
#075
Male
Female

(B)


正解率 (%)
100
90
80
70
60
50
40
30
20
10
0
Without detection
With detection
#012
#027
#032
#037
#039
#007
#013
#020
#065
#075
Av.
Male
Female

フロントページの続き

(51)Int.Cl.	F I	テーマコード（参考）
Ｇ１０Ｋ　15/04　(2006.01)	Ｇ１０Ｋ　15/04　３０２Ｄ	

(72)発明者　後藤　真孝
茨城県つくば市東１－１－１　独立行政法人産業技術総合研究所つくばセンター内

Ｆターム(参考) 5D015 HH01
5D108 BD02 BD13