社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MUS－74 (33)
2008－SLP－70 (33)
2008／2／9

曲内類似性，GMM，類似区間継続長を用いた曲境界の自動抽出方式の提案

吉田 拓真† 伊藤 慶明† 石亀 昌明† 小嶋 和徳†

† 岩手県立大学ソフトウェア情報学研究科 〒020–0193 岩手県滝沢村滝沢字巣子 152–52
E-mail: †g231e038@edu.soft.iwate-pu.ac.jp

あらまし 記憶媒体の大容量化に伴い，録画保存した番組を簡易に閲覧し，選択的に観られる機能が望まれる．筆者らは，音楽番組中の楽曲を抽出するため，一つの曲内の類似性と曲と曲の間の非類似性により，曲境界を抽出する曲内類似法を提案した [1]．また曲内類似法に加え，音響的特徴を頑強に識別可能な GMM を導入し，音声/音楽境界の高精度化を図った [2]．一つの楽曲内での楽器の切り替わりや転調時に起こる誤ったセグメンテーションをこれらの方法でも全て抑えることができない．本稿では，繰り返されている長めのメロディを認識し，その区間を一つの楽曲として捉えることで解決を図る．任意の長さで同一/類似区間を照合可能な Relay Continuous Dynamic Programing (Relay CDP) 法 [3] を導入し，一楽曲の纏まりを複数のメロディから捉える連続類似区間強調法を提案する．連続した音楽データ，音楽と会話が交互に現れるデータを用いて楽曲の境界抽出実験を行い，曲内類似法，GMM 判定方式と共に連続類似区間強調法を用いることによって境界抽出性能の向上を確認した．

キーワード 音楽情報処理，曲境界，GMM，DP，類似区間

A Proposal of Automatic Music Boundary Extraction Method using Similarity in a Music Selection, GMM, Duration of Similar Sections

Takuma YOSHIDA†, Yoshiaki ITOH†, Masaaki ISHIGAME†, and Kazunori KOJIMA†

† Iwate Prefectural University Sugo 152–52, Takizawa, Iwate, 020–0193 Japan
E-mail: †g231e038@edu.soft.iwate-pu.ac.jp

Abstract Along with recent increase of hard disc capacity, users want the function facilitating to watch recorded TV programs easily. We proposed a method detecting music boundaries using similarity in a music selection and non-similarity between music sections. We combined the method and GMM that enables to discriminate acoustic feature robustly. The method could not suppress the wrong segmentation caused by a change of musical instruments or modulation. To solve the problem, we introduce Relay-CDP to enable to extract continuous same/similar sections at arbitrary length at any locations. The section between these longer similar sections can regarded as within a music selection. We confirmed that the proposed method improves the accuracy of music boundary extraction.

Key words musical infomation processing, music boundary, GMM, dynamic programing, similarity section

1. はじめに

近年，大容量ハードディスクレコーダの普及により，自分の興味のある大量の番組を自動的に録画し，録画データの中で見たい場面のみを鑑賞するスタイルに変わってきており，視聴したい場面を迅速に提示する機能が求められている．本研究は容易な閲覧システムを目指し，音楽番組で一般的なポピュラーミュージックに着目し，音楽番組中の楽曲が流れている区間の自動的な抽出方式の研究を進めている．これにより楽曲の頭出しを行うことができ，自分が見たいアーティストや楽曲を容易に確認することができる．また既存のプレイヤーのシークバーと合わせて楽曲区間を提示することで，スキップしながら容易に閲覧できると考える．更に，楽曲部分に対しては音響特徴を分析し歌手の認識や楽曲ジャンルの分析，会話部分に対しては音声認識を行い会話情報の抽出等の応用も考えられる．

楽曲の構造を捉える手法として，繰り返されるサビ区間を検出し，サビ抽出や楽曲の内部情報を視覚的に表示する方法[4]や，歌声 (単独歌唱) と朗読音声の識別に関する研究[5]，動的計画法 (DP 法) により会話部と音楽部を識別しセグメンテーションする方法[6]がある．文献[6]に述べられているが，メロディの変化点や転調，楽器の交代する点で曲内での誤ったセグメンテーションが発生しやすいという問題点がある．

本稿では，楽曲の抽出及び曲境界の自動抽出を目的とし，楽曲内での誤ったセグメンテーションを抑えるために，以下三つの方法を統合した方法を提案する．

(1) 一定の長さの類似区間を抽出・利用し，一楽曲の纏まりを捉える方法 (曲内類似法)
(2) 音声/音楽の識別を行い境界を抽出する方法 (GMM 判定法)
(3) 繰り返される長めのメロディ等を認識しその区間を一つの楽曲として補正する方法 (連続類似区間強調法)

以下，各々の方法について簡単に説明する．

曲内類似法は，一般的な楽曲には類似したメロディの繰り返しが複数あることに着目し，類似音響が存在する区間を一つの楽曲として捉える方法である．繰り返されるメロディを類似区間として抽出するために，時系列データ内の任意位置，一定の長さの類似区間を抽出する Segmental CDP 法を用いる．ここでは楽曲の纏まりを捉えているため，推定した楽曲境界は時間的に精度が高くない．正確な境界を抽出するために，ここで推定した境界位置付近の音響的特徴の変化が大きい位置を再推定し，正確な曲境界を抽出する．

GMM 判定法では，音楽/音声を頑強に識別可能な GMM を導入する．音声と音楽の GMM をそれぞれ作成し，音響データを各 GMM に与えて音声/音楽どちらであるか尤度により判定する．この音声/音楽が切り替わっている部分が境界候補となるが，切り替わった部分全てを境界とすると，連続して境界と判定し，多くの誤ったセグメンテーションを生成するため，音声/音楽の継続性を考慮し境界検出を行う．

曲内類似法，GMM 判定法では，一つの楽曲内で楽器の切り替わりや転調時に起こる誤ったセグメンテーションを完全に抑えることはできなかった．そこで繰り返される長いメロディ等を抽出し，その区間内で転調や楽曲の切り替わりが起きても同一の楽曲区間として捉えることで解決を図る．任意の長さの同一/類似区間ペアを抽出可能な Relay CDP 法を用い，曲内類似法で抽出される類似区間よりも長いメロディ等の類似区間を抽出し，それらの区間を同一楽曲内として重視することで，楽曲内の誤ったセグメンテーションを訂正する方式である．

以下に，2. 章でシステムの概要および提案手法について述べ，3. 章で実験および考察を行い，最後にまとめを述べる．

2. 提案手法

2.1 システム概要

図 1 にシステム全体のフローチャートを示す．対象となる音楽番組の音響情報から特徴時系列パラメータを抽出する．抽出した特徴量を元に，楽曲の纏まりを捉えた曲マウンテングラフ (後述) の作成と，GMM による音声/音楽の判定を並列に行う．曲内の類似性に着目し，類似した音響区間を抽出することで曲マウンテングラフを作成する．連続類似区間強調法を併用して曲マウンテングラフの楽曲区間を補正し，GMM による音声/音楽結果と統合することで楽曲の纏まりを捉えた曲境界検抽出を行う．

2.2 曲内類似法

一楽曲中には，曲の一番二番といった繰り返される音響区間が存在し，異なる楽曲には類似音響が少ない．そこで，類似音響が多数存在する区間を一つの楽曲と捉えることができる．時系列データの中から自己相関的に類似音響区間を探索するには，時系列データの任意の位置，任意の長さの類似区間を照合・検出する必要がある．その任意の類似区間の探索を簡素化し近似的に照合する Segmental CDP 法を用いる．Segmental CDP 法の概念を図 2 に示す．Segmental CDP 法では，まず時系列データを一定の

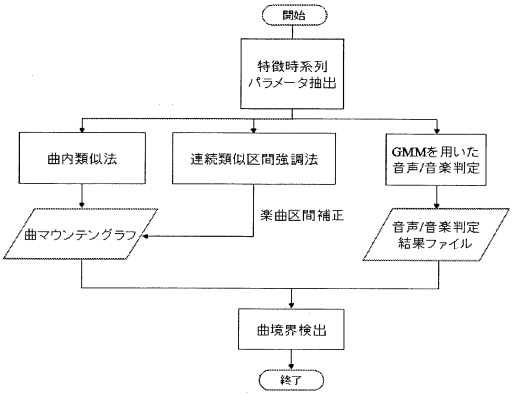

図 1　システムフローチャート

長さ (一定フレーム数) 毎にセグメントとして分割する．各セグメントに対して，それ以降の時系列データとの連続 DP による照合を行い，類似区間候補を上位数個保存しておき，全体の類似区間候補の中から類似性が高い区間を類似区間とする．

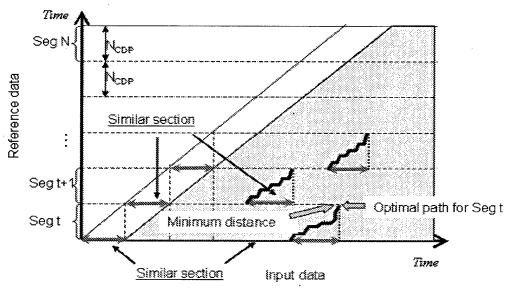

図 2　Segmental CDP 法の概念図

次に一楽曲の纏まりを捉えるために，類似区間のペアの出現位置・出現頻度を元に，楽曲の大局的な纏まりを表す曲マウンテングラフを作成する．図 3 に曲マウンテングラフの作成手順を示す．抽出された全ての類似区間ペアを直線で結び，各セグメント上の直線通過頻度を数える．各セグメント上の通過頻度をヒストグラム化したものを曲マウンテングラフと呼ぶ．曲マウンテングラフの山の頂上の前後には類似する区間が多く出現するため山部分は同一曲内，山が形成されていない区間は異なる楽曲の接続部分，あるいは楽曲以外の区間と捉えることができる．

曲境界位置は，曲マウンテングラフの山の始点と終点付近にあると考えられる．曲マウンテングラフの谷部分を検出するために，曲マウンテングラフに対して一定幅毎に窓掛けを行い，窓中から山の高さと深さを用いて曲境界候補を抽出する．

推定した曲境界候補は，楽曲の纏まりを大まかに捉えた曲マウンテングラフから推定しているため，曲境界を厳密に推定した訳ではない．そこで正確な境界を抽出するために，得られた境界位置付近の音響的特徴の変化が大きい位置を分析し，曲境界を再推定する．

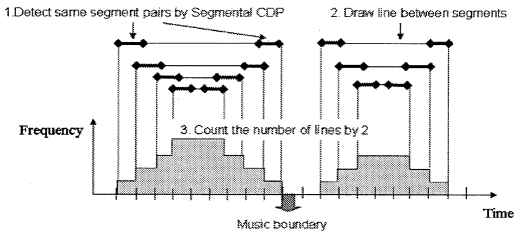

図 3　曲マウンテングラフの作成手順

2.3　GMM 判定法

曲内類似法では曲の類似区間を抽出するため，繰返し構造がない楽曲に対しては抽出精度が低化する．そのため，音響的特徴について識別能力が高い GMM を導入して音声と音楽の判定を行い，曲境界検出性能の向上を図る．音声と音楽の GMM をそれぞれ作成し，入力時系列データとの尤度を計算することにより，与えられたデータが音声と音楽どちらであるか判定を行う．

音声/音楽が切り替わっている部分が境界候補となるが，切り替わった部分を全て境界とすると，連続して境界と判定し誤ったセグメンテーションを起こすため，音声/音楽の継続性を考慮し，境界に重みを設定する．図 4 に示すように，重みはそれまでに継続した音声/音楽フレーム長と，切り替わってから次に切り替わるまでの音声/音楽フレーム長の和とする．重みが大きい部分は，音声，あるいは音楽が長く続いているため切り替わり場面としての重要度が高く，異なる場面間を捉えていると考えられる．全ての境界に対して重みを算出し，重みの大きい順に境界として検出する．

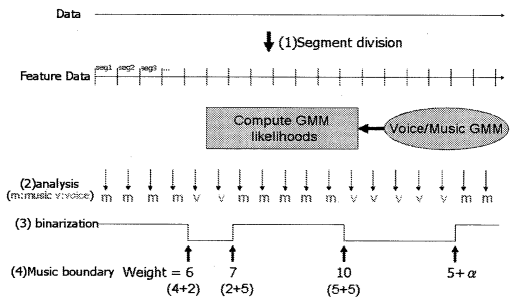

図 4　GMM 判定方法と重み付け

2.4 連続類似区間強調法

曲内類似法，GMM 判定法では，一つの楽曲内の誤ったセグメンテーションを完全に抑えることはできなかった．一方，サビ等の比較的長めの類似区間があれば，その類似区間の間に類似区間が無くても (曲マウンテングラフの山が形成されない場合も) 同一楽曲内とみなせる．そこで，繰り返される長いメロディ等を認識し，その類似した区間を一楽曲区間として捉える方法を導入する．任意の長さの同一/類似区間ペアを探索可能な Relay CDP 法を用い，曲内類似法よりも長い音響類似区間を検出し，その間の区間を同一楽曲内として重視することで楽曲区間の補正を行い，誤ったセグメンテーションを解消する．

Relay CDP 法は，Segmental CDP 法の任意のセグメントを始端として，連続する複数のセグメントとそれらと同一/類似である入力データ区間を抽出する．連続してマッチするセグメント数が N 以上であれば同一/類似区間ペアとみなして，その連続する参照セグメントと入力データの部分区間の位置情報を出力する．図 5 に Relay CDP 法の概念を示す．セグメント P_{i-1} の類似区間として t_a から t_b の区間が見つかった場合，次のセグメントである P_i と，見つかった区間の終わりの位置である t_b から始まる類似区間を探索する．この作業を繰り返し行い，連続したセグメント数を m とすると，その連続回数 m の値が大きい場合は，ヒストグラム作成の際のラインカウント数を大きく与えることで (例えば m^2) マウンテングラフの山が大きくなるよう補正する．

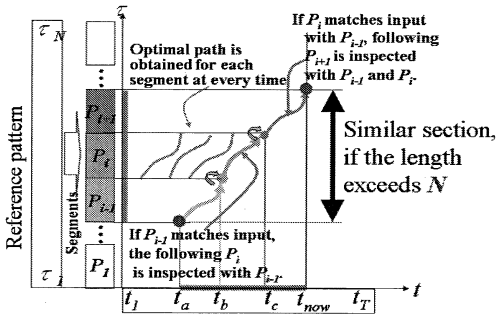

図 5　Relay CDP 法の概念

2.5 各手法の統合

曲内類似法から得られた曲マウンテングラフ，GMM 判定法の音声/音楽判定結果，連続類似区間強調法から得られた曲マウンテングラフを組み合わることで頑健な曲境界を検出する．

曲内類似法と連続類似区間強調法で形成される曲マウンテングラフを図 6 に示す．図は，異なる三つの楽曲が連続しているデータに対して両手法を実行したものである．二つ目の楽曲が存在する 400 秒から 700 秒の区間にかけて，曲内類似法の曲マウンテングラフでは山が大きく形成されていない．このような場合に楽曲内でのセグメンテーションが発生しやすいが，連続類似区間強調法の曲マウンテングラフを用い，その区間が楽曲であると補正することで問題を解決する．

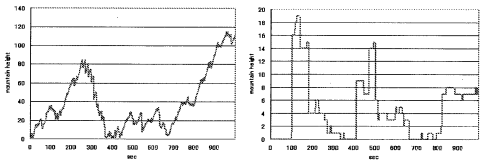

図 6　曲内類似法 (左) と連続類似区間強調法 (右) の曲マウンテングラフ比較

3. 評価実験

曲内類似法 (再推定前後)，GMM 判定法，曲内類似法と GMM 判定法の二つを統合した手法，曲内類似法と GMM 判定法と連続類似区間強調法の三つを統合した手法それぞれについて境界検出性能の評価実験を行った．

3.1 評価用データ

評価用のデータとして，音楽データには RWC 研究用音楽データベースのポピュラー音楽 100 曲 [8]，音声データには JNAS [9] の発話データを用いた．以降，RWC 音楽と JNAS 音声が交互に繋がれたデータのことを混合データ，RWC 音楽 100 曲を繋ぎ合わせたデータを音楽データと呼ぶ．音楽データは，実際の放送等で音楽が連続する場合を想定して，曲の最初と最後に付加されている無音部をパワーにより削除し繋ぎ合わせた．混合データは，JNAS の発話データから 1 分程度の長さの音声を楽曲間に挿入し，RWC 音楽と JNAS 音声が交互に現れるデータ (音楽・音声共に 100 個) で，全データ長は 30,355 秒である．音声データには，男女各 50 人ずつ発話を使用した．音楽データは最短の曲で 132 秒，最長の曲で 362 秒である．

3.2 実験条件

セグメント長は，事前研究により 1 秒とした場合に最も良い性能で類似区間抽出が行えたため，本稿でもセグメント長は 1 秒とした．音声/音楽 GMM の作成に用いる特徴パラメータを表 1 に示す．本稿では音声認識で一般的に用いられている，MFCC12 次元と MFCC ⊿12 次元，パワーと⊿パワーの計 26 次

表 1　特徴パラメータ

チャンネル	モノラル
サンプリング周波数	44.1kHz
フレーム長	25ms
フレーム周期	10ms
MFCC 次数	MFCC26 次元 (MFCC_E_D_Z)

元を用いた．GMM に識別させる音響時系列データの長さはセグメントと同様 1 秒とし，1 秒毎に音声/音楽の判定を行う．音楽データは楽曲間の継ぎ目，混合データは楽曲と音声データの継ぎ目を曲境界とし，全曲境界が音楽データ 99 箇所，混合データ 199 箇所とする．

実験では次の 6 つの手法について比較評価を行う．(1) 曲内類似法の曲マウンテングラフを用いた境界検出法 (曲内類似法)，(2) その曲境界位置から再推定した手法 (曲内類似法再推定)，(3)GMM の音声/音楽判定結果から境界を検出する手法 (GMM 判定法)，(4) 曲内類似法と GMM 判定法を統合した手法 (曲内類似法＋ GMM 判定法)，(5) 曲内類似法と連続類似区間強調法による曲マウンテングラフを加算し楽曲部分の推定を行い，GMM 判定法と統合した手法 (三手法の統合 (a))，(6) 曲内類似法と連続類似区間強調法の曲マウンテングラフからそれぞれ別途に楽曲部分の推定を行い，GMM 判定法と統合した手法 (三手法の統合 (b))．

正解判定の基準としては，各手法で推定した境界位置が，実際の正解位置の前後 2 秒以内の場合に正解とした．それぞれの結果について Precision Recall グラフ，及び F 値の最大値を用いて評価を行った．

3.3　結果および考察

各手法ごとの音楽データに対する実験結果を図 7 に，混合データに対する実験結果を図 8 に，両実験の F 値の最大値を纏めたものを表 2 に示す．

曲内類似法では楽曲の大局的な情報から曲境界を大まかに推定するため，境界検出精度は高くない．推定した境界位置に対して前後の音響的特徴の変化の大きい位置へ再推定する方法を用いることで境界検出の性能は向上したが，音楽データの境界検出精度が 79.4%に比べ，混合データの境界検出精度は 57.9%と低かった．楽曲と音声の境界検出のため，音声/音楽の識別が可能な GMM 判定法と統合することで，混合データにおいても 86.1%の検出精度が得られた．音楽データは，楽曲間の無音部を削除して繋ぎ合わせているため，GMM 判定法のみでは全て音楽と判定し，曲境界を抽出できなかった．

本稿で提案する連続類似区間強調法を用いた三手法の統合 (a)，(b) では，音楽データ，混合データ共に境界検出精度の向上が確認できた．

三手法を統合した手法では，音楽番組データの Recall の向上が見られた．当初，曲マウンテングラフが形成されにくい楽曲区間の改善により，Precision が向上すると期待していたが，これは図 9 に示すような楽曲の始まり部分に類似区間が少ない楽曲でも，連続類似区間強調法によって曲の出だしの長い類似音響を捉えられたと考える．三手法の統合 (a) では，三手法の統合 (b) より若干低い性能となっているが，統合した一つの曲マウンテングラフから楽曲区間を推定することができる．一方三手法の統合 (b) では，二つの曲マウンテングラフそれぞれで楽曲区間の推定を適切に調整する必要がある．

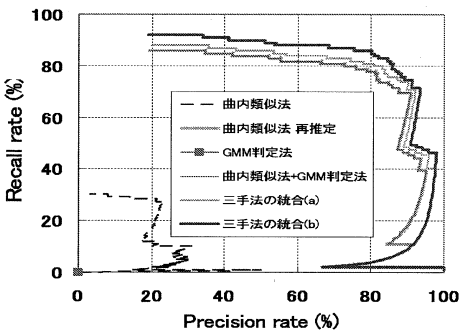

図 7　音楽データに対する境界検出実験

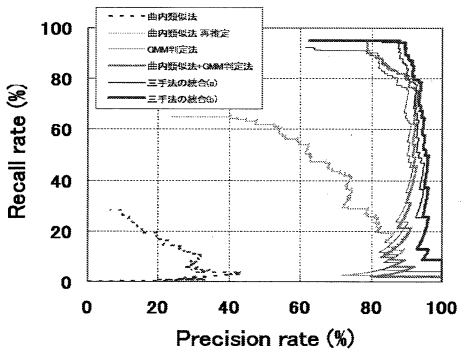

図 8　混合データに対する境界検出実験

また予備実験として，実際の音楽番組データを用いた曲境界検出実験を行った．本実験では，本学学生に実際に番組を視聴して番組構成を記述してもらい，その番組構成から楽曲前後の境界を正解とした．表 3 に用いた番組データを示す．番組 A はアーティスト特集番組，番組 B は演歌番組，番組 C はカウン

表 2　各手法の F 値の最大値 (%)

	音楽データ	混合データ
曲内類似法	25.0	11.7
曲内類似法再推定	79.4	57.9
GMM 判定法	0.0	85.9
曲内類似法＋ GMM 判定法	79.4	86.1
三手法の統合 (a)	81.5	90.8
三手法の統合 (b)	83.6	91.9

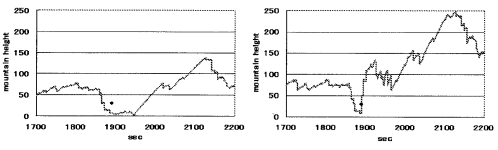

図 9　正解を検出できるようになった曲マウンテングラフ
(左:曲内類似法 右:二つの曲マウンテングラフの統合)

トダウン形式の番組で 30 分から 1 時間の番組データである．表 4 に，番組データを用いた場合の境界検出性能を示す．3.1 の実験データと比べ，境界検出性能の低下が見られた．実際の番組データの場合，曲の始まりや終わりに実際の演奏音以外の音が加えられる場合があるため，正確な境界が得られない傾向があった．また楽曲の境界が演奏と映像で異なっていることが多かった．正解境界位置が映像情報寄りに付与されているため，音響的に切り替わっている場面を推定するのに加え，映像情報を用いたカット点検出と組み合わせる必要があると考える．

表 3　番組データの詳細

番組名	チャンネル	放送日
アーティストスペシャル 吉井和哉 (番組 A)	Ch.269 (MusicJapanTV)	H18,11 月 5 日 16 時～17 時
さぶちゃんの歌仲間 (番組 B)	Ch.267 (第一興商スターカラオケ)	H18,10 月 31 日 11 時～11 時 30 分
邦楽チャート HOT 10 (番組 C)	Ch.265 (スペースシャワー TV)	H18,11 月 2 日 15 時～16 時

表 4　連続類似区間強調法の番組データの F 値 (%)

	正解範囲	
	前後 2 秒	前後 5 秒
番組 A	62.18	72.54
番組 B	46.04	69.97
番組 C	41.72	62.12

4. おわりに

本稿では，容易な音楽番組視聴のための楽曲の自動抽出方式を提案した．連続した類似区間を抽出し重視することで，曲内類似法と GMM では捉えることができなかった楽曲を捉え，より精密に曲境界を検出できた．

今後は，更なる性能向上を図るために適切な特徴量の検討，及び音声音楽以外の判定やポピュラー音楽以外のジャンルの識別を行いたい．また曲マウンテングラフの閾値のより詳細な検証や，実際の番組データに対する定量的な実験・評価を行いたい．

文　献

[1] 岩渕晃 他 : “曲内の類似性を利用した曲境界の検出”, 音講論, pp595-596 (2005-3).

[2] 吉田拓真, 伊藤慶明, 石亀昌明, 小島和徳, ”曲内の類似性と GMM を利用した曲境界判定方式の提案”, 音楽音響研究会,MA2006-34 (2006-9).

[3] Y. Itoh, K. Tanaka, S. Lee, “An algorithm for similar utterance section extraction for managing spoken documents”, Multimedia Systems, 10(5):432-443 (2005).

[4] 後藤真孝 : “リアルタイム音楽情報記述システム : サビ区間検出手法”, 情処研究報告, 2002-MUS-47-6,Vol.2002,No.100,pp.27-34 (2002-10).

[5] 大石 康智, 後藤 真孝, 伊藤 克亘, 武田 一哉: “スペクトル包絡と基本周波数の時間変化を利用した歌声と朗読音声の識別”, 情報処理学会論文誌, Vol.47, No.6, pp.1822-1830 (2006-6).

[6] M. Goodwin et al: A Dynamic Programing approach to audio segmentation and speech/music discrimination, ICASSP, pp.309-312, (2004).

[7] Y. Itoh, K. Tanaka, S. Lee, “Repeated Utterance Extraction by a New Algorithm for Labeling a Presentation”, Multimedia Information Retrieval, pp.179-185 (2003).

[8] M. Goto, H. Hashiguchi, T. Nishimura, R. Oka: “RWC Music Database: Popular, Classical, and Jazz Music Databases”, ISMIR, pp.287-288 (2002-10).

[9] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi. “JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research”, J. Acoust. Soc. Jpn, Vol. E20, No.3, pp. 199-206 (1999-5).

[10] D.A.Reynolds, R.C.Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models”, IEEE Trans. on Speech and Audio Processing, vol.3, no.1, pp.72-83 (1995-1).

[11] Y. Itoh, A. Iwabuchi, M. Ishigame, K. Kojima, K. Tanaka, S. Lee, “Music boundary detection using similarity in a music selection”, International Workshop on Multimedia Signal Processing, 4 pages (2007-10).

[12] K. Kashino, T. Kurozumi and H. Murase, “A quick search method for audio and video signals based on histogram pruning,” IEEE Trans. Multimedia, 5, 348-357 (2003).

[13] G. Tzanetakis, P. Cook, “Musical genre classification of audio signals,” IEEE Trans. Speech Audio Process., 10, 293-302 (2002).

[14] E. Scheirer, M. Slaney, ”Construction and evaluation of a robust multifeature speech/music discriminator”, ICASSP, (1997).