Vol.2011-MUS-91 No.17
2011/7/29

トーン構造記述子を用いた高速背景音楽検索

澁谷 崇[†1] 東山恵祐[†1]
安部素嗣[†1] 西口正之[†1]

本稿では，テレビや映画などで用いられる背景音楽を音楽データベースと高速に一致検索する手法を提案する．我々は背景音楽よりも大きい前景音に対してロバストかつ高速に一致検索を行うために，音楽の持続性トーン成分に着目し，それを用いた特徴量 “トーン構造記述子” を提案する．トーン構造記述子を用いた実験では，S/N 比 −20 dB においても再現率が 96 %以上で， かつパーソナルコンピュータを用いてもリアルタイムに 10 万曲以上検索可能であることを示す．

Fast Background Music Retrieval using Tonal Structure Descriptor

Takashi Shibuya,[†1] Keisuke Toyama,[†1]
Mototsugu Abe[†1] and Masayuki Nishiguchi[†1]

This paper presents an extremely fast method for identifying background music with a piece of music in large database. We focus on continuous tonal components, which make the identification robust to loud foreground sounds, and propose a feature based on continuous tones, “Tonal Structure Descriptor”. In the experiments, we demonstrate that our descriptor enables a personal computer to compare background music with more than 100,000 tracks in real time, and realize more than 96% Recall at −20dB S/N Ratio.

1. はじめに

近年，ネットワークの高速化やクラウドの普及などに伴い，一般の人がマルチメディアコンテンツを扱う機会が増えている．それに伴って，簡単にコンテンツの検索を行い，アクセスするための技術が多くの関心を集めている．音楽の検索技術に関しても例外ではない．音楽検索技術は，テレビ放送・ラジオ放送のモニタリングやジングル検出，電子ファイルの楽曲同定などの応用があることから，研究が盛んに行われている[1)–8)]．特にここ数年は，一般の消費者が携帯電話のアプリケーション等で Gracenote 社の MusicID[9)] や Shazam[3)] などの音楽認識サービスを利用できる環境も整いつつある．

このような時代背景において，放送コンテンツの背景音楽を検索する需要も高まっている．これは背景音楽をキーとしたネットワークサービスや音楽の著作権管理などへの展開が目的である．しかしながら，背景音楽検索に関しては下記の問題がある．

- 音楽には前景音が重畳されている．前景音の多くは人の音声で，S/N 比[⋆1] は概ね −20 dB 程度でもマッチングする必要がある．
- 前景音に合わせて，1 曲の中でも頻繁にイコライジングやボリュームコントロールが行われることがある．
- 100 万曲以上のデータベースからリアルタイムに検索しなければならない．

すなわち，背景音楽検索に求められる要件は前景音に対するロバスト性と検索の高速性であると言える．

そこで，本稿では前景音にロバストで高速な背景音楽検索手法を提案する．特に，放送コンテンツでは前景音の多くが人の音声であるため，人の音声へのロバスト性に主眼を置き，S/N 比が低い場合においても高精度に検索できる手法を目指す．検索速度については，1 台のパーソナルコンピュータでリアルタイムに 10 万曲以上の検索することを目標とする．1 台で 10 万曲以上検索できれば，少数のクラスタ構成によって 100 万曲以上の検索を行うことも可能であると考える．

本稿の構成は以下の通りである．第 2 章で，前景音が重畳された背景音楽をどのように検索するかの方針を述べる．第 3 章で，提案特徴量であるトーン構造記述子の生成方法を説明する．第 4 章で，トーン構造記述子同士の比較方法を説明する．第 5 章で，提案手法に関する評価実験を行う．第 6 章で，本稿のまとめを述べる．

2. 背景音楽検索へのアプローチ

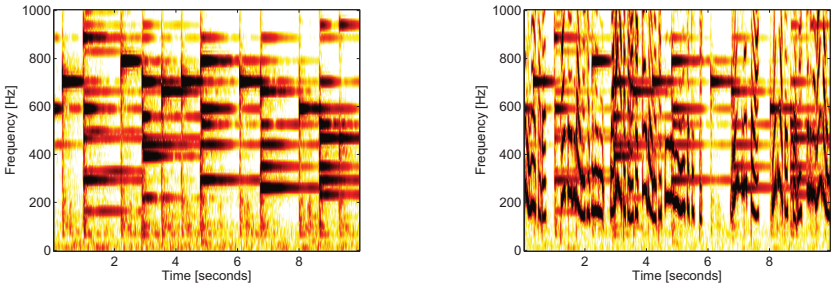
2.1 背景音楽と持続性トーン

図 1 に二つのスペクトログラムを示す．図 1 (a) はある音楽から生成したスペクトログラ

†1 ソニー株式会社 (Sony Corporation)

⋆1 本稿では，背景音楽を信号成分（S），前景音（音声など）を雑音成分（N）と見なす．

Vol.2011-MUS-91 No.17
2011/7/29



(a) 音楽のみ（参照信号）　(b) 音声と音楽の混合音（入力信号）

図 1　背景音楽のスペクトログラム

ム，図 1 (b) は (a) の音楽に音声を重畳したスペクトログラムである．図 1 (a) に見られるように，音楽の音は音声に比べトーン性の音 [*2]が持続する傾向がある．音声が重畳された図 1 (b) においても，この持続性トーン[*3]のスペクトルピークによって背景音楽の存在が確認できる．これは全体の S/N 比が低い場合においても，時間周波数領域では背景音楽と前景音のパワーがスパースに表現され，持続性トーンが局所的に優位に残るためである．本稿ではこの持続性トーンが背景音楽を検索するための鍵であると考え，持続性トーンを用いた特徴量を設計・提案する．持続性トーンを音楽検出に用いた研究もある[10),11)]．

2.2　記述子の設計方針

持続性トーンについて，図 1 のようなスペクトログラムから得られる情報は時間，周波数，振幅の 3 つである．本稿では，1 章で述べたボリュームの大きい前景音やイコライジング／ボリュームコントロールの影響を可能な限り軽減するために振幅情報は用いず，時間と周波数の情報を用いる．スペクトログラムのような時間周波数空間において，持続性トーンの存在する時間と周波数（持続性トーンの時間周波数分布）を表現する特徴量を提案する．

時間と周波数の情報を持つ特徴量の設計で問題となるのが分解能である．分解能を上げることで，楽曲の内容を細かく表現することができるが，特徴量同士の比較の際に細かい比較が必要となり検索速度に悪影響を及ぼす．また，分解能は特徴量のデータサイズにも影響を与えるが，高速検索においてはデータベース内の楽曲を可能な限りメモリ（主記憶装置）に置くことが望ましく，データサイズの観点からも分解能は低い方が良い．しかし，当然のことながら分解能を下げれば下げるほど個々の楽曲を判別できるほどの表現力は失われてしまう．高速な検索を実現するためには，この表現力が失われない程度に分解能を低くする必要がある．

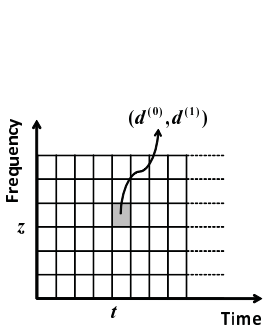


図 2　トーン構造記述子の概略

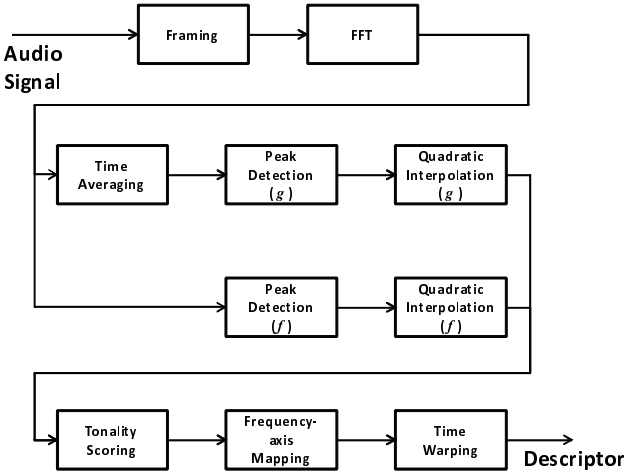


図 3　記述子生成処理の流れ

また，持続性トーンの時間周波数分布は持続性トーンの存在確率のような連続値ではなく，トーンが存在するか否かの 2 値で表現することを考える．これも特徴量のデータサイズを極力小さくするためである．ただし，トーンの有無だけでは同じ高さの音が連続して演奏された場合と，長い音が継続している場合の区別がつかなくなってしまうことが考えられる．そこで，持続性トーンの有無だけでなく．持続性トーンの開始点と終了点も表現する．

上記のことを踏まえ，高速背景音楽検索のための特徴量を提案する．**図 2** は提案特徴量 “トーン構造記述子” の概略である．トーン構造記述子は周波数軸（行成分） z と時間軸（列成分）t を持った 2 次元配列で，配列の各要素はベクトル $(d^{(0)}, d^{(1)})$ を持つ．$d^{(0)}$ は該当する時間周波数領域にトーンが存在するか否かを，$d^{(1)}$ は該当する時間周波数領域にトーンがあった場合のトーンの状態を示す成分である．具体的にはベクトル $(d^{(0)}, d^{(1)})$ は以下の 4 種類の値をとる．

- $(0, -)$: トーンがない状態（“ − ” は不定値を表す.）

⋆2 本稿では「トーン性の音」を，ある周波数にパワーのピークを持つような正弦波と見なせる音とする．
⋆3 本稿では「持続性トーン」を，周波数が音楽のメロディーを構成するのに十分な時間安定した音とする．

Vol.2011-MUS-91 No.17
2011/7/29

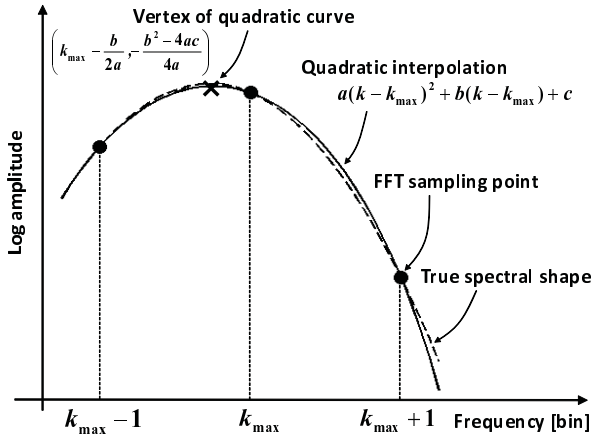


図 4 スペクトラルピークの 2 次補間

- $(1,+1)$：トーンの開始
- $(1,\ \ 0)$：トーンの持続
- $(1,-1)$：トーンの終了

2 次元配列のうち，第 z 行，第 t 列の要素を $(d^{(0)}_{\{z,t\}}, d^{(1)}_{\{z,t\}})$ と表記する．また，記述子の行数を Z とする（列数はオーディオ信号の長さに依存する）．

3. トーン構造記述子の生成方法

本章では，オーディオ信号からトーン構造記述子を生成する方法（**図 3**）について説明する．

3.1 持続性トーンらしさの定量化

持続性トーンの時間周波数分布を表現するために，まずスペクトログラム上の個々の音について，それがどれほど持続性トーンらしいかを表す “持続性トーンらしさ” の評価を行う．

持続性トーンらしさの定量化は，QIFFT 法（Quadratically Interpolated FFT 法）[12),13)] を応用して行う．QIFFT 法は対数振幅スペクトルからスペクトラルピークを抽出し，ピーク近傍 3 点を用いて 2 次補間を行うことで，正弦波パラメータ（周波数と振幅）[⋆4] を推定する手法である．2 次補間で得られた 2 次関数の頂点の座標は推定周波数と推定対数振幅となる（**図 4**）．また，得られた 2 次関数の 2 次係数はピーク近傍の曲率を表すが，これは窓関数の形状により定まる．この QIFFT 法は単一の FFT フレームにおいて正弦波パラメータを推定する手法であるが，持続性トーンらしさの定量化のために複数の FFT フレームをまたいだ形に拡張する．

まず，FFT によって得られた第 n フレームの近傍 $(2N_{\mathrm{a}}+1)$ フレームについて，対数スペクトルの時間平均 $g(k,n)$ を求める．

$$g(k,n)=\frac{1}{2N_{\mathrm{a}}+1}\sum_{m=n-N_{\mathrm{a}}}^{n+N_{\mathrm{a}}} f(k,m)\,. \tag{1}$$

ここで，$f(k,m)$ は第 k bin，第 m フレームの対数スペクトルである．

次に，第 n フレームの平均対数スペクトル $g(k,n)$ のうち，ピークとなる bin 番号群 $(k^{(1)}_{g\max},\cdots,k^{(j)}_{g\max},\cdots,k^{(J_n)}_{g\max})$ を検出する．J_n は $g(k,n)$ から検出されたピークの数，j は検出されたピークの番号である．ピークは $(2K_{\mathrm{p}}+1)$ -近傍の最大で検出する．[⋆5] そして，各ピーク $k^{(j)}_{g\max}$ の 3 近傍の対数振幅値の 2 次補間を行い，2 次の係数 $a^{(j)}_g$ およびトーンの周波数 $\omega^{(j)}_g$ を推定する．

ピーク検出と 2 次補間は平均を算出する前の単一フレームの対数スペクトル $f(k,n)$（瞬時スペクトルと呼ぶこととする）についても行う．ピーク $(k^{(1)}_{f\max},\cdots,k^{(i)}_{f\max},\cdots,k^{(I_n)}_{f\max})$ を検出し，2 次の係数 $a^{(i)}_f$ およびトーンの推定周波数 $\omega^{(i)}_f$ を算出する．

持続性トーンらしさは 2 次補間で得られたこれらの値を用いて定量化する．平均対数スペクトルの j 番目のピークについて，持続性トーンらしさ $\eta^{(j)}$ は次の式で定量化される．

$$\eta^{(j)}=\begin{cases}1-\dfrac{|a^{(j)}_g-a^{(h)}_f|}{|a^{(j)}_g|} & \text{if } |\omega^{(j)}_g-\omega^{(h)}_f|<\delta \text{ and } |a^{(j)}_g-a^{(h)}_f|<|a^{(j)}_g| \\ 0 & \text{otherwise}\end{cases} \tag{2}$$

$$\text{where } h=\arg\min_i |\omega^{(j)}_g-\omega^{(i)}_f|. \tag{3}$$

ここで，h は $\omega^{(j)}_g$ との差が最も小さい $\omega^{(i)}_f$ のピーク番号で，持続性トーンに周波数のブレが全くない場合は $\omega^{(h)}_f=\omega^{(j)}_g$ となる．また，δ は 2.1 節で述べたような周波数の変動をどれほど許容するかを表すパラメータで，サンプリング周波数および窓関数の形状，零詰め

⋆4 位相スペクトルを用いることで位相の推定も可能である．

⋆5 K_{p} の値は FFT で用いた窓関数のサイドローブ性ピークを検出しないように設定する．

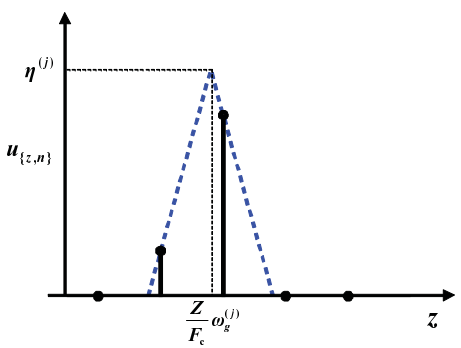


図 5 周波数軸離散化処理の概要

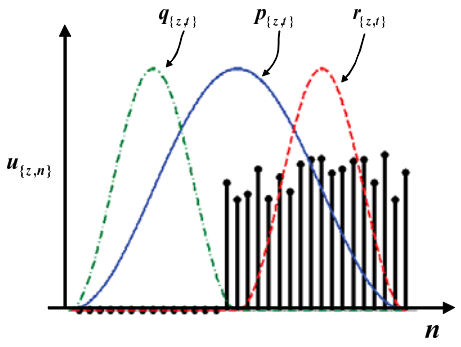


図 6 時間軸平滑化処理の概要

量に基づいて設定する．式 (2) は第 n フレーム $f(k,n)$ の h 番目のピーク近傍のスペクトル形状と，$(2N_a+1)$ フレームの平均スペクトル $g(k,n)$ の j 番目のピーク近傍のスペクトル形状との類似度を評価した値となっている．$\eta^{(j)}$ の取りうる値の範囲は $0 \le \eta^{(j)} \le 1$ で，$(2N_a+1)$ フレームの間でスペクトル曲線の形状が一定の場合は $\eta^{(j)}=1$ となる．

3.2 持続性トーンらしさに基いた記述子の生成

2.2 節で述べた記述子を作成するために，まず第 n フレームから得られたスペクトラルピークの推定周波数 $\omega_g^{(j)}$ を連続値から離散化する．J_n 個のピークの推定周波数と持続性トーンらしさ $((\omega_g^{(1)},\eta^{(1)}),\cdots,(\omega_g^{(j)},\eta^{(j)}),\cdots,(\omega_g^{(J_n)},\eta^{(J_n)}))$ を，持続性トーンらしさの周波数分布を表すベクトル $\boldsymbol{u}_n=(u_{\{1,n\}},\cdots,u_{\{z,n\}},\cdots,u_{\{Z,n\}})^T$ に変換する．ベクトル $\boldsymbol{u}_n$ の要素数 Z は最終的に得られる記述子の行数である．

$$u_{\{z,n\}}=\sum_{j=1}^{J_n} v\left(z-\frac{Z}{F_c}\omega_g^{(j)}\right)\eta^{(j)} \quad \text{where } v(x)=\begin{cases}1-|x| & \text{if } |x|<1\\ 0 & \text{otherwise.}\end{cases} \tag{4}$$

上記の式は，連続値の推定周波数 $\omega_g^{(j)}$ と組になっていた持続性トーンらしさ $\eta^{(j)}$ について，三角窓を用いて周波数の離散化を行っている (**図 5**)．$\boldsymbol{u}_n$ の各要素 $u_{\{z,n\}}$ は周波数 $(F_c/Z)z$ Hz における持続性トーンらしさを表す．F_c は $z=Z$ のときの周波数で，ここで F_c Hz 以下の周波数に帯域制限を行っている．

最後に，この $\boldsymbol{u}_n$ からトーン構造記述子の配列要素である $d_{\{z,t\}}^{(0)}$ と $d_{\{z,t\}}^{(1)}$ を算出する．

- $d_{\{z,t\}}^{(0)}$ の算出

2.2 節で述べたように，$d_{\{z,t\}}^{(0)}$ は該当する時間周波数領域にトーンが存在するか否かを表す 2 値の変数である．$d_{\{z,t\}}^{(0)}$ を求めるために Hann 窓を用いて時間軸方向に平滑化し（**図 6**），ダウンサンプリングを行う．

$$p_{\{z,t\}}=\sum_{n=1}^{N_w} w_0(n)u_{\{z,n+N_h(t-1)\}} \quad \text{where } w_0(n)=0.5-0.5\cos(2\pi n/N_w). \tag{5}$$

ここで，N_w は Hann 窓のサイズ，N_h はホップサイズである．そして，連続値の $p_{\{z,t\}}$ に対して閾値処理を行い，2 値化することで $d_{\{z,t\}}^{(0)}$ を得る．

$$d_{\{z,t\}}^{(0)}=\begin{cases}1 & \text{if } p_{\{z,t\}}>\alpha\\ 0 & \text{otherwise.}\end{cases} \tag{6}$$

ここでは 2 値化の閾値 α は固定値ではなく，統計値を用いる．具体的には，第 t 列の p の二乗平均平方根（RMS）の定数倍を 2 値化の閾値 α とする．

$$\alpha=\beta\sqrt{\sum_{z=1}^{Z}\left(p_{\{z,t\}}\right)^2}. \tag{7}$$

これは，音楽に音声が重畳されている場合，持続性トーンらしさの値が全体的に小さくなる傾向があるためである．上記のような統計値を閾値とすることで，全体的な傾向を補償し，局所的な “持続性トーンらしさ” を強調する効果を持たせる．

- $d_{\{z,t\}}^{(1)}$ の算出

$d_{\{z,t\}}^{(1)}$ は該当する時間周波数領域のトーンの状態を表す．$d_{\{z,t\}}^{(1)}$ の算出では，サイズが $N_w/2$ の Hann 窓を用いて 2 つの平滑化を行う（図 6）．

$$q_{\{z,t\}}=\sum_{n=1}^{N_w/2} w_1(n)u_{\{z,n+N_h(t-1)\}} \tag{8}$$

$$r_{\{z,t\}}=\sum_{n=1}^{N_w/2} w_1(n)u_{\{z,n+N_h(t-1)+(N_w/2)\}} \tag{9}$$

$$\text{where } w_1(n)=0.5-0.5\cos(4\pi n/N_w).$$

ともに t が 1 増えるごとに $\boldsymbol{u}_n$ の n が N_h 点ずつシフトするが，$r_{\{z,t\}}$ は $q_{\{z,t\}}$ の $N_w/2$ フレーム分先を平滑化したものである．これら 2 つの値は，トーンの状態によって次のよう

な関係性になることが期待される．

– トーンの開始 ($d^{(1)}_{\{z,t\}} = +1$)：$r_{\{z,t\}}$ が $q_{\{z,t\}}$ より有意に大きい

– トーンの持続 ($d^{(1)}_{\{z,t\}} = \ 0$)：$q_{\{z,t\}}$ と $r_{\{z,t\}}$ に有意な差がない

– トーンの終了 ($d^{(1)}_{\{z,t\}} = -1$)：$q_{\{z,t\}}$ が $r_{\{z,t\}}$ より有意に大きい

つまり，$p_{\{z,t\}}$ と $r_{\{z,t\}}$ を比較することで，トーンの状態を推定できると考えられる[⋆6]．ここでは $q_{\{z,t\}}$ と $r_{\{z,t\}}$（ともに非負値）の比を用いる．

$$d^{(1)}_{\{z,t\}} = \begin{cases} +1 & \text{if } \dfrac{r_{\{z,t\}}}{q_{\{z,t\}}} > \gamma \\ 0 & \text{if } \dfrac{1}{\gamma} \le \dfrac{r_{\{z,t\}}}{q_{\{z,t\}}} \le \gamma \\ -1 & \text{if } \dfrac{r_{\{z,t\}}}{q_{\{z,t\}}} < \dfrac{1}{\gamma}. \end{cases} \tag{10}$$

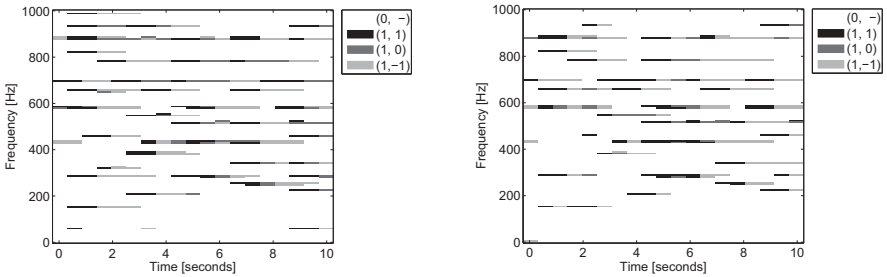
ここで，$q_{\{z,t\}}$ と $r_{\{z,t\}}$ を比較する際に差ではなく比を用いたのは，式（7）と同様，前景音がある場合，持続性トーンらしさの値が全体的に小さくなるためである．なお，$d^{(0)}_{\{z,t\}} = 0$ となった配列要素 $\{z,t\}$ については，$d^{(1)}_{\{z,t\}}$ の値を不定値 “–” とする．

3.3 トーン構造記述子の例

図 7 (a)，(b) はそれぞれ図 1 (a)，(b) のオーディオ信号から生成したトーン構造記述子である．縦軸は周波数を，横軸は時間を表す．また，色の濃い順に $(d^{(0)}, d^{(1)}) = (1,+1), (1,0), (1,-1), (0,-)$ を表している．図 1 (b) ではほとんどの時間周波数領域で音声が重畳されているにもかかわらず，図 7 (b) の記述子は図 7 (a) の記述子から大きな変化が見られない．これはトーン構造記述子が音声に対してロバストに音楽の特徴を表現していることを示している．

4. トーン構造記述子を用いたマッチング処理

本章では，前章で提案したトーン構造記述子について，背景音楽の含まれる入力信号から得られた記述子と，データベース中の参照信号から得られた記述子の類似度を算出する方法



(a) 音楽のみ（参照信号）　(b) 音声と音楽の混合音（入力信号）

図 7 背景音楽のトーン構造記述子

について述べる．

入力信号と参照信号から，それぞれ Z 行 $\times$ T 列のトーン構造記述子が得られたとする．トーン構造記述子は時間周波数領域におけるトーンの分布とその状態を表す特徴量であるが，比較の際には各時間周波数領域におけるトーンの有無が一致しているかの評価と，トーンがあった場合にその状態が一致しているかの評価を行う．これら 2 つの評価を総合した評価値を記述子の列ごとに算出した上で，T 列分の評価値の平均値を求め，記述子全体の類似度とする．

第 t 列におけるトーンの有無の一致度の評価には，相関指標を用いる．

$$S^{(0)}_t = \frac{\sum_{z=1}^{Z} d^{(0)}_{\mathrm{Q}\{z,t\}} d^{(0)}_{\mathrm{R}\{z,t\}}}{\lambda \left(\sum_{z=1}^{Z} d^{(0)}_{\mathrm{Q}\{z,t\}} \right) + (1-\lambda) \left(\sum_{z=1}^{Z} d^{(0)}_{\mathrm{R}\{z,t\}} \right)}. \tag{11}$$

ここで，d_{Q} は入力信号の記述子の要素，d_{R} は参照信号の記述子の要素である．$S^{(0)}_t$ の分母は入力信号に存在するトーンの数と参照信号に存在するトーンの数の加重平均を，分子は両者に存在するトーンの数を表している．λ は $0 \le \lambda \le 1$ の値の範囲をとるパラメータで，入力信号のトーンと参照信号のトーンのどちらに重きを置くかを表す．$\lambda = 1$ のとき，$S^{(0)}_t$ は入力信号に存在するトーンのうちのどれほどが参照信号にも存在するかの指標となる．この場合，参照信号にトーンが存在しない周波数領域に，入力信号のトーンが存在していても $S^{(0)}_t$ の値には影響しない．$\lambda = 0$ のときはその逆となる．

⋆6 $q_{\{z,t\}}$ と $r_{\{z,t\}}$ の比較は，持続性トーンらしさの時間変化量の評価を行っていることを意味する．

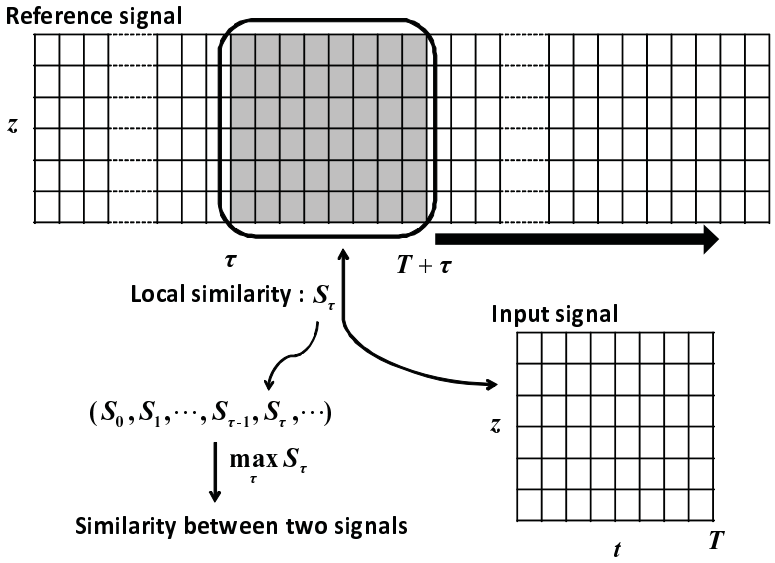


図 8 マッチング処理の概要

第 t 列におけるトーンの状態の一致度は，次の式で評価する．

$$S_t^{(1)} = \frac{\sum_{z=1}^{Z} I\left(d^{(1)}_{\mathrm{Q}\{z,t\}} = d^{(1)}_{\mathrm{R}\{z,t+\tau\}}\right)}{\sum_{z=1}^{Z} d^{(0)}_{\mathrm{Q}\{z,t\}} d^{(0)}_{\mathrm{R}\{z,t\}}}. \tag{12}$$

$I(\mathrm{cond.})$ は条件式 cond. が真であるときに 1 の値を返し，偽であるときに 0 の値を返す指示関数である．$S_t^{(1)}$ の分母は入力信号と参照信号の両方に存在するトーンの数を，分子はそれらの状態が一致した数を表している．

式（11）と式（12）の積 $S_t^{(0)} S_t^{(1)}$ を 2 つの記述子の第 t 列の類似度とし，記述子全体の類似度には T 列分の類似度の平均値を用いる．

$$S = \frac{1}{T}\sum_{t=1}^{T} S_t^{(0)} S_t^{(1)} = \frac{1}{T}\sum_{t=1}^{T} \frac{\sum_{z=1}^{Z} I\left(d^{(1)}_{\mathrm{Q}\{z,t\}} = d^{(1)}_{\mathrm{R}\{z,t+\tau\}}\right)}{\lambda\left(\sum_{z=1}^{Z} d^{(0)}_{\mathrm{Q}\{z,t\}}\right) + (1-\lambda)\left(\sum_{z=1}^{Z} d^{(0)}_{\mathrm{R}\{z,t+\tau\}}\right)}. \tag{13}$$

上記では入力信号と参照信号の長さ（記述子の列数）が等しいと仮定していたが，実際には入力信号は楽曲の断片であることが多く，参照信号の方が長い．本稿では上記の類似度を用いて，ブロックマッチングを行う．入力信号とマッチングを行う参照信号の記述子を時間方向にスライドさせていく（**図 8**）．時間シフトパラメータ τ を 1 ずつ増やしながら順次類似度計算を行う．

$$S_\tau = \frac{1}{T}\sum_{t=1}^{T} \frac{\sum_{z=1}^{Z} I\left(d^{(1)}_{\mathrm{Q}\{z,t\}} = d^{(1)}_{\mathrm{R}\{z,t+\tau\}}\right)}{\lambda\left(\sum_{z=1}^{Z} d^{(0)}_{\mathrm{Q}\{z,t\}}\right) + (1-\lambda)\left(\sum_{z=1}^{Z} d^{(0)}_{\mathrm{R}\{z,t+\tau\}}\right)}. \tag{14}$$

得られた類似度系列 $(S_0, \cdots, S_\tau, \cdots)$ のうち，その最大値を入力信号とその参照信号の最終的な類似度とする．

表 1 パラメータの値．

Sampling frequency	16000 Hz
Frame size	1024
Step size	256
FFT size	2048

N_a	4
K_p	2
δ	16000/2048
F_c	4000
Z	384
N_w	128
N_h	32
β	2.15
γ	1.2
λ	0.85

5. 評価実験

本章では，前章までで提案した記述子とそのマッチング方法の性能評価を行う．評価は検索の精度と速度の観点で行う．本章の実験では，**表 1** で示されるパラメータの値を用いる．ステレオ音源については， 2 ch の信号の和信号を入力信号として扱う．

5.1 実験 1

音楽に前景音として音声を重畳し，提案記述子の音声に対するロバスト性を調べた．実験には 1000 曲の楽曲[⋆7]を用いた．これらを参照信号のデータベースとして用い，さらに入力信号用に各々の楽曲から 10 sec を切り出した．切り出しは，基本的に参照信号からランダムに行ったが，10 sec が曲の終わり等のロングトーンにならないように切り出した．切り出された信号に 3 段階の S/N 比（−10 dB，−20 dB，−30 dB）で音声を重畳し，各 1000 本の入力信号を作成した．重畳した音声には 1 sec 以上の無言区間がないものを用いた．

⋆7 主に J-Pop，テレビドラマや映画のサウンドトラックから成る．

Vol.2011-MUS-91 No.17
2011/7/29

表 2 閾値 $\theta = 0.25$ における精度（実験 1）

S/N Ratio	F-measure	Recall	Precision
−10 dB	0.986	0.991	0.980
−20 dB	0.973	0.963	0.983
−30 dB	0.867	0.775	0.984

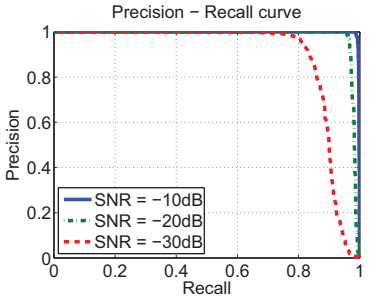


図 9 Precision-Recall 曲線（実験 1）

検索精度の評価は上記のデータセットから得られた（入力信号 1000 曲）×（参照信号 1000 曲）の類似度に，楽曲を判定するための閾値 θ を導入したときの Recall（再現率）と Precision（適合率），および F-measure（Recall と Precision の調和平均）によって行った．これらの評価指標は，閾値 θ を 0 から 1 まで 0.001 刻みに変化させながら算出した．Precision の算出については，データセットに含まれる同じ主旋律の曲（アレンジ違い等）が検出された場合は False-Positive（誤検出）にはカウントしないこととした．

実験結果を**表 2** と**図 9** に示す．表 2 は楽曲判定の閾値を $\theta = 0.25$ としたのときの F-measure，Recall，Precision を示したものである．図 9 は閾値 θ を変化させたときの Recall と Precision の変動をプロットしたグラフである． $\theta = 0.25$ のとき，S/N 比 −20 dB において Recall，Precision 共に 0.96 以上の値を示しており，これは提案手法が音声の重畳に対して頑健であることを示している．

S/N 比 −10 dB および −20 dB で検出漏れとなったのは，打楽器のみの音楽やディストーションなどのエフェクトが使われた音楽であった．打楽器音やディストーション音では持続性トーンらしさの指標が比較的低く，音声が重畳された場合は音声に埋もれてしまっていた．

5.2 実　験　2

実際のテレビ放送の音源を用いて実験を行った．録画したテレビ放送の中から，実験 1 で用いた参照信号の楽曲が使われていた区間を 10 sec だけ切り出し，実験 1 と同様の実験を行った．今回切り出したオーディオ信号ファイルは 1207 本である．評価方法は実験 1 と同じである．

実験結果を**表 3** と**図 10** に示す．閾値 $\theta = 0.25$ のとき，Recall，Precision 共に 0.9 以上

表 3 閾値 $\theta = 0.25$ における精度（実験 2）

F-measure	Recall	Precision
0.908	0.901	0.916

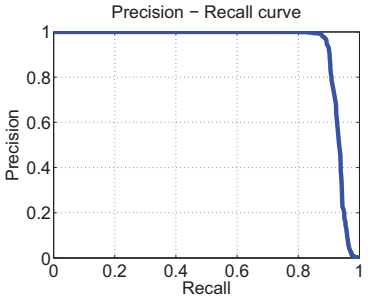


図 10 Precision-Recall 曲線（実験 2）

表 4 実験環境

OS	Windows 7 (64bit)
CPU クロック周波数	3.20 GHz
コンパイラ	Intel C++ XE 2011
	（並列化処理なし）

の値を示しており，10 sec の信号同士のマッチングとしては実用的な性能であると言える．

検出漏れがあった原因は大きく分けて 3 つあった．1 つ目は単純に S/N 比が相当に低い場合である．実際のテレビ放送では実験 1 と異なり，複数の前景音が重なっていることがほとんどだったため，持続性トーンの検出がうまくいかない場合があった．また，音声のみでなく，電話の着信音などのトーンを持つ電子音が記述子に反映され，Recall だけでなく Precision にも影響を与えていた例もあった．2 つ目は打楽器音やディストーション音で構成されている音楽の登場する頻度が高かったことである．実験 1 でも考察したように，このような音楽は他の音楽に比べ音声の重畳に弱いため，検出漏れを起こしていた．3 つ目は楽曲のミキシングに関する問題である．実際の放送では一つの楽器によるソロ演奏であったにもかかわらず，参照信号として用意された音源では放送音源と同じメロディーの後ろで，他の楽器による伴奏のある楽曲があった．つまり，参照信号と元の音源は同じであるものの，ミキシングの異なる楽曲が放送で用いられたという問題である．

5.3 検索速度について

実験 1 と実験 2 は，1 台のパーソナルコンピュータで，マルチコアの並列化を行わずにを行った．その他の実験環境は**表 4** に示す．1 つの入力信号（10 sec 分）を 1000 曲の参照信

号とマッチングするのに要した平均時間は 0.12 sec であった[⋆8]．これらの値から，10 sec 内で約 8.3 万曲の楽曲が検索可能である．これは 2 つのコアを用いるだけで 10 万曲以上の検索が可能であることを意味し，1 章で提示した目標に達したと言える．

6. おわりに

本稿では，音楽に音声などの前景音が重畳されているオーディオ信号から，その背景音楽を検索する手法を提案した．本手法のポイントは，音楽と音声の特徴的な違いである持続性トーンの有無のレベルに着目したことである．そして，持続性トーンの時間周波数分布を表現する記述子を提案した．実験では，本手法が S/N 比 −20 dB においても 96 %以上の再現率で，一般のパーソナルコンピュータで並列化を行わずに 8 万曲以上をリアルタイムに検索できることを示した．

今後については，まず音楽を構成する持続性トーン以外の要素を記述子に導入することが課題となる．実験で考察したように，S/N 比の低い場合に，打楽器音から持続性トーンを検出できない場合がある．そのため，持続性トーン以外のリズムやビートなどの要素を前景音に対してロバストに抽出し，マッチングに用いる手法を考える必要がある．

また，本稿の実験では 10 sec の入力信号から記述子を生成し，それのみを用いてマッチング処理を行ったが，一般にはより長い入力信号がストリームとして入力され，順次処理するものである．その場合，次々と算出される類似度を時系列処理することによって，認識精度を向上できる余地がある．例えば，5 章の実験 2 において検出漏れを起こした入力信号についても，その前後では正解曲が検出できていることが多く，前後関係も考慮することで検出精度をさらに高めることができると考えられる．さらに上位層となる時系列のマッチング方法の検討も今後の課題である．

参考文献

1) J. Haitsma and T. Kalker, “A Highly Robust Audio Fingerprinting System,” in *Proceedings of 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pp.107–115, 2002.
2) J. Pinquier and R. André-Obrecht, “Jingle Detection and Identification in Audio Documents,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.4, pp.329–332, 2004.
3) A. Wang, “The Shazam Music Recognition Service,” *Communications of the ACM*, Vol.49, No.8, pp.44–48, March 2006.
4) H. Nagano, K. Kashino, and H. Murase, “A Fast Search Algorithm for Background Music Signals based on the Search for Numerous Small Signal Components,” in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2003)*, Vol.1, pp.165–168, 2003.
5) K. Kashino, A. Kimura, H. Nagano, and T. Kurozumi, “Robust Search Methods for Music Signals based on Simple Representation,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.4, pp.1421–1424, 2007.
6) E. Dupraz and G. Richard, “Robust Frequency-based Audio Fingerprinting,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp.281–284, 2010.
7) C.-Y. Chiu, D. Bountouridis, J.-C. Wang, and H.-M. Wang, “Background Music Identification through Content Filtering and Min-Hash Matching,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp.2414–2417, 2010.
8) M. Abe and M. Nishiguchi, “Self-Optimized Spectral Correlation Method for Background Music Identification,” in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2002)*, Vol.1, pp.333–336, 2002.
9) http://www.gracenote.com/products/musicid
10) K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, “Video Handling with Music and Speech Detection,” *IEEE Multimedia*, Vol.5, No.3, pp.17–25, July 1998.
11) K. Seyerlehner, T. Pohle, M. Schedl, and G. Widmer, “Automatic Music Detection in Television Productions,” in *Proceedings of 10th International Conference Digital Audio Effects (DAFx-07)*, pp.221–228, 2007.
12) J. O. Smith III and X. Serra, “PARSHL: A Program for the Analysis/Synthesis of Inharmonic Sounds based on a Sinusoidal Representation,” in *Proceedings of International Computer Music Conference (ICMC 1987)*, pp.290–297, 1987.
13) M. Abe and J. O. Smith III, “AM/FM Rate Estimation for Time-Varying Sinusoidal Modeling,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.3, pp.201–204, 2005.

⋆8 実験に用いた参照信号の合計時間は 54.8 h である．また，オーディオ信号からトーン構造記述子への変換に要した平均時間は 0.11 sec であった．