

# 音楽や映像の 高速探索

## 8



柏野 邦夫

kunio@eye.brl.ntt.co.jp  
NTT コミュニケーション科学基礎研究所

### 音楽や映像をとりまく環境変化

多くの人が日常慣れ親しんでいる芸術作品として、音楽作品や、映画などの映像作品が挙げられる。これらは、作り手が、音や動画像といったメディア(媒体)に何らかの思いやメッセージを託して作品を制作し、それを受け手が受け取ることで成立している。中には受け手が共同参画することで作品が成り立つ場合などもあるが、今なお、最も基本的なスタイルは、作品を媒介とした作り手から受け手への思いの伝達、コミュニケーションによってよいだろう。

近年、このコミュニケーションをとりまく環境に変化が起きている。その第1は、情報流通手段、特にインターネットの拡大である。我が国では、9,000万を超える携帯電話契約のうちおよそ9割の契約でインターネット接続可能となっている<sup>☆1</sup>。また国内のブロードバンド接続(FTTH, DSL, CATVなど)の契約数は2,000万を超え、固定電話契約数の3分の1以上となった<sup>☆2</sup>。ネット上のメディア情報の量も急増している。たとえば、Web上のデータ量をファイルサイズで見ると、その2/3が画像、動画像、音のファイルであり、しかもそれらのデータ量はこの5年間で約16倍になったという。

環境変化の第2は、情報蓄積手段、特にハードディスク装置の進歩である。容量単価はこの20年で約10万分の1にもなり小型化も進んだ。そして今やその中には、音楽や映像などのメディア情報が盛んに蓄積されている。

<sup>☆1</sup> 電気通信事業者協会調べ(2006年2月現在)。

<sup>☆2</sup> 総務省資料(2005年9月現在)。

音楽CD1枚には1時間あまり(1曲を5分とすると十数曲分)の音楽を記録できる。このデータサイズは約600MBであるが、圧縮技術を用いるとほとんど音質を損うことなく1/10程度に圧縮できる。したがって、1TBの記憶媒体には、約20万曲の音楽が蓄積できる。数十TBの容量があれば、世界中で商品として流通している楽曲の大半を蓄積できることになる。

映像の蓄積には、標準テレビ放送並みの品質で記録した場合、1時間あたり2GBほどの容量を必要とする。つまり1TBの記憶媒体には映画約250本分の映像を蓄積できる。これはハリウッドのメジャースタジオで半年間に制作される映画の合計本数に匹敵する。

このように、メディア情報の流通と蓄積に関する急速な環境変化は、音楽や映像の世界にも影響を与えている。そこで本稿では、筆者らの研究チームの活動を中心にメディア探索技術の研究事例を紹介しつつ、メディア探索技術が、作品の作り手や受け手にとってどのような意味を持つかを考えてみたい。

### 音楽や映像を探す

#### ✧ 類似探索

情報の流通手段が拡大し膨大な情報が蓄積されたとすると、必要な時に的確かつ迅速に所望の情報を探し出す手段が重要になることは明らかである。

書店で本を探す場合に、本の題名、著者、ISBNなどが手がかりになるのと同様、音楽や映像の場合にも、タイトルなどの付加情報(音楽の場合、曲名、アーティスト名、カタログ番号など)で探すのが基本である。このため、

20万曲の楽曲といえども、それぞれにタイトルなどの付加情報(メタデータ)がつけて格納してあれば、それに基づいて探し出すことができる。

しかし、タイトルなどの情報に常に依存できるとは限らない。たとえば、問合せに用いるべきタイトルが分からない場合がある<sup>☆3</sup>。そこで、タイトルやファイル名などによらず、その内容によって音楽や映像を探す方法が検討されている。

内容検索の方法は、大きく2通りに分けて考えることができる。その第1は、記述レベルでの照合である。つまり、探したい内容を記述する単語やシンボルを問合せとする方法である。たとえば、「元気が出る曲」という記述を問合せとしたり、俳優の名前によってその登場場面を探したりといった場合である。このような形態を実現するためには、あらかじめ、蓄積された音楽や映像に対して、その内容の記述(上の例では感性情報の記述や俳優名)を得ておかなければならない。このようなメタデータは、手間やコストさえ厭わなければ人手で付与することもできるが、人手によらないとすれば、音楽や映像から自動的に内容記述を得る処理が必要となる。これらは一般に、自動採譜や音楽情景分析、人物認識やコンピュータビジョン、あるいは音声認識などといった、パターン認識の課題そのものに相当する。これらはチャレンジングな課題である。

内容検索の方法の第2は、対象レベルでの照合である。つまり、具体的な音楽(音)や映像を問合せとし、それに類似した個所を含む対象を探し出す方法である。このような機能は類似探索(similarity search)と呼ばれる<sup>☆4</sup>。

インターネットの検索サイトで、いくつか単語を投入し、それらに一致もしくは類似した文字列を含むWebページを探し出すことも、類似探索の一例と考えることができる。このような類似探索の方法論は、特にテキスト情報について進展が著しい。元をたどれば、1970年代の、Boyer-Moore法をはじめとする高速な文字列探索のアルゴリズムの開発に端を発すると考えられる。もちろんWebページの検索サイトでは、問合せが投入され

てからWebページに対する文字列探索が起動されるのではなく、数多くのWebページがあらかじめ収集・分析されて索引付けがなされており、何十億という数のページについての探索が瞬時に行える。また、速さばかりではなく結果の的確さについても工夫が積み重ねられている。

筆者らの研究チームでは、音楽や映像をはじめとするメディア情報の認識・探索技術の研究に取り組んでいる。我々は、テキストの場合と同様に、音楽や映像の探索においても類似探索は重要な基本的機能となるだろうと考え、近年では特に音や映像の類似探索技術に焦点を当ててきた。

類似探索といっても、似ているかどうかといった主観の問題には踏み込まず、元の情報から何らかの特徴を抽出し、特徴どうしの類似性をもって元の情報の類似性とみなすアプローチが一般的である。その最も基本となるのは、同一信号探索である。ここで同一信号とは、音楽であればある特定のCDの音、映像であれば映画の中のある場面といったように、特定の同一ソースに基づく音や映像のことである。このような探索は、特に一致探索と呼ばれることがある。

### ❖ 速度の追求

音や映像のような時系列信号を対象とする同一信号探索の基本的な考え方を図-1に示す。問合せとなるのは音や映像の断片である。その断片(たとえば15秒間)が、長時間(たとえば1,000時間)にわたって蓄積された音や映像の中のどの個所に出現するか、あるいはどこにも出現しないかを調べるのが探索の目的である。問合せに用いる音や映像の断片を目的信号と呼び、蓄積された音や映像を蓄積信号と呼ぶ。図-1では、まず、蓄積信号と目的信号から特徴ベクトルを抽出する。音や映像の時間変化をとらえるには、特徴ベクトルを1秒間あたり数十個~百個程度抽出して特徴ベクトルの時系列とする。次に、蓄積信号特徴の時系列に目的信号と同じ長さの照合窓を設定し、蓄積信号特徴の各部分と目的信号特徴との類似度を測りながら、照合窓を時間方向にずらしていき、類似度の高い部分を検出する。類似度を測る手段としては、特徴ベクトル間のユークリッド距離や相関などを用いることができる。

しかし、このような地道な方法では時間がかかりすぎるという問題がある。そこで、別の方法として、先に述べた高速文字列探索のアルゴリズムを適用することが考えられる。特徴ベクトルの時系列は、ベクトル量子化と

<sup>☆3</sup> CDショップの店員の話では、店頭で鼻歌を聞かせてこの曲が欲しいと訴える客が後を絶たないそうである。

<sup>☆4</sup> このほか、対象と記述の2つのレベルにまたがる照合もある。たとえば鼻歌(ハミング)を問合せにして音楽データベースを検索するハミング検索の研究では、まず問合せのハミングを音高情報の記号列に変換した上で、蓄積された音高記号列との照合を行う方法が多く試みられている。これは、問合せ(音響信号)を対象レベル、蓄積されたデータを対象の記述と見れば、対象レベルと記述レベルにまたがる照合と言える。しかし、照合処理自体は記号列どうしの照合であるから、ここで言う類似探索に該当する。

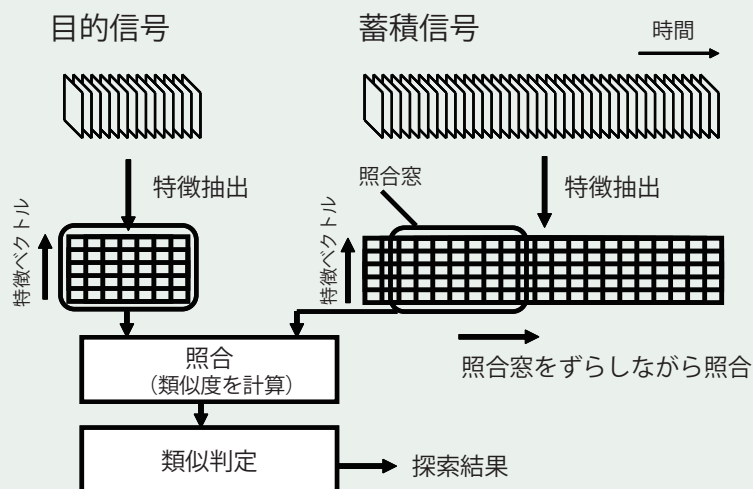


図-1 時系列メディア探索の基本

呼ばれる手法を用いると符号列に変換することができる。これにより同一信号探索は文字列探索に帰着されるので高速に処理できそうである。しかし、メディア情報の探索では、多少の時間伸縮やノイズ混入への耐性を持たせなければ実用性を欠くこともあり、高速な探索は依然として容易ではなかった。

そこで、筆者らは、ベクトル量子化された符号の頻度分布を信号の特徴として用いる手法を提案した<sup>1)</sup>。これを時系列アクティブ探索法と呼んでいる。

時系列アクティブ探索法では、まず、特徴ベクトルのベクトル量子化によって目的信号と蓄積信号のそれぞれに対し特徴符号系列を作る。次に、これらを直接比較するのではなく、頻度分布特徴を作って比較する。つまり、目的信号（たとえば15秒間）、および照合窓内の蓄積信号の双方について、ベクトル量子化符号それぞれの生起数を調べて頻度分布とし、頻度分布同士の類似度を測って信号の同一性を判定する。

頻度分布特徴自体は、たとえばカラーヒストグラムと呼ばれる特徴量に見られるように、静止画像の特徴量として従来から広く用いられていたものである。本手法ではそれを時間軸方向に用いたのが要点である。これは実はあまり自然な発想とは言えない。たとえば音声信号処理の感覚では、15秒間以上にもわたって特徴ベクトルを集積してしまうと、時間情報が失われる上に平均的な分布ばかりになってしまい、特徴として用をなさないようにも思われる。にもかかわらず、実際には高い判別能力を有しているというのは意外な事実であった。しかも、

特徴ベクトルを時間軸方向に集積するおかげで多少の時間伸縮やノイズ混入に対する耐性が備わるという、良い副作用もあった。

また、頻度分布特徴を用いることで、照合窓を時間軸方向にジャンプしながら探索できるというメリットが生まれる。一般に蓄積信号の特徴系列の上で照合窓を安易にジャンプさせると、検出すべき個所を見逃す可能性が生じてしまう。しかし本手法では、頻度分布特徴の特徴空間における軌跡が連続的にしか変化し得ないことを利用し、照合窓をジャンプさせずに特徴ベクトル1つずつずらした場合と同一の探索結果を保証しながら、その場その場で許される幅だけ照合窓のジャンプを行うことで、照合回数を大幅に削減できる。つまり、照合窓を、検出基準に達する可能性が決していない区間についてはスキップさせながら移動させることで、探索を高速化することができる。この点は、Boyer-Moore法をはじめとする高速文字列探索法の本質に通じるものがあると言えるかもしれない。

このような手法により探索に必要な照合回数を減らすことができたが、さらなる高速化のため、照合1回に必要な計算量を減らすことはできないかという発想が生まれた。これは、照合回数と照合1回あたりの計算量の積が全体の計算量になるからである。そこで提案したのが「区分線形写像法」である。この手法は木村が中心になって考案した<sup>2)</sup>。

一般に特徴ベクトルの次元数は少ない方が高速に照合できる。次元数を減らすためには、主成分分析などを



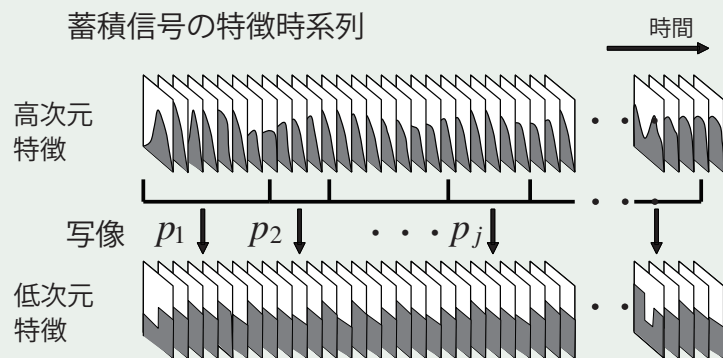


図-2 区分線形写像法の仕組み

用いることが考えられる。蓄積信号の長さ全体にわたって特徴ベクトルの次元数を下げることができればそれに越したことはないが、たとえば頻度分布特徴の場合には、元々無駄な次元があるべくできないようにベクトル量子化の符号帳を決めるので、全体の次元数を下げるアプローチにはあまり期待できない。ところが、メディア情報の特徴として、時間的に近い部分どうしは類似している場合が多いという性質がある。たとえばニュース番組やスポーツ中継の映像を想定すると、数秒間にわたって大きな変化がないことも多い。

区分線形写像法では、この性質に着目し、特徴時系列を時間的に区分した上で、それぞれの区間を低次元の特徴ベクトルで表現することを特徴とする。この概念図を図-2に示す。目的信号の特徴ベクトルも、区間ごとにそれぞれの特徴空間に写像した上で蓄積信号の特徴ベクトルと照合する。これによって、写像のコストはかかるが、それ以上に照合のコストを減らすことができ、結果として探索を1桁近く高速化することができた。この際、元の特徴空間で探索を行った場合と同じ精度を保証しながら効率的な照合を行えるように工夫されている。

これらの研究努力により、メディア探索速度は向上してきている<sup>☆5</sup>。時系列アクティブ探索法では、1998年の提案当時に、15秒間の目的信号に対しPC1台で1秒間に約3時間分の蓄積信号特徴を探索することができた。

☆5 この章で述べている探索所要時間は、目的信号が与えられてから探索が完了するまでの時間であり、蓄積信号に対する特徴抽出など、蓄積信号に対し蓄積時に一度だけ行っておけばよい処理（前処理）に要する時間は含まれていない。ただし前処理に必要な時間が過大にならないことも重要な条件である。たとえば、前処理の所要時間が実時間以下である（例、100時間分の蓄積信号に対する前処理が100時間以内に終了する）ことが実用性の1つの目安になると考えられる。本章で紹介した手法の前処理時間はいずれも実時間以下である。

最近では、研究の進展による進歩とPCの性能向上の効果が加わり、その数百倍以上の探索実行速度が得られている<sup>3)</sup>。

なお、上で紹介したのは、大規模な構造を用いることなく探索を行うことを目的としたアプローチの例である。このほか、関連する研究として、メディア探索を目的とした索引構造の研究も近年盛んに行われている。メディア情報を対象とする索引構造においては、特徴ベクトルが高次元であることに由来する効率低下の問題や、索引のためのデータ量が増大する問題など、テキストデータに対する索引構造の場合とは異なった高いハードルが存在する。これらの研究課題へのアプローチについては過去に本誌でも解説記事が特集されている<sup>4)</sup>。いずれにしても、今後、メディア情報の持つ特有の性質をさらに活かし、特徴空間内での特徴の分布を考慮した高速探索手法の開発が期待されている。

## ❖ 精度の追求

メディア情報の特性の1つは、容易に雑音や歪みが加わることである。メディア探索では、多くの場合、デジタル的な信号の完全一致だけを検出するのではなく、雑音や歪みが多少加わっても同一ソースの情報は同一ソースとして検出することが期待される。その方が人間の感覚に合っているからである。

筆者らの研究チームにおいても、高速化手法の研究とともに、音や映像の品質変化の影響をなるべく受けずに同一信号探索を精度よく行う方法について研究してきた。以下では、この中から、主に音響信号を対象とする場合について述べる。

雑音や歪みには加法的なものや乗法的なものがあるが、このうち加法性雑音への対処を目的とした手法が、永野

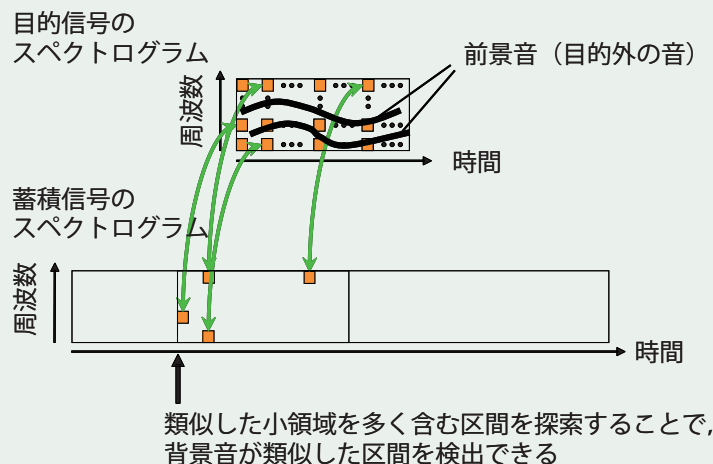


図-3 分割一致探索法の仕組み

を筆頭に開発した「分割一致探索法」である。加法性雑音の中にも帯域雑音と線スペクトル構造を持つ雑音（目的外の音）とがあるが、本手法は、線スペクトル構造を持つ加法性雑音の影響を低減することを目的としている。たとえば、人の声のパワーの大部分を占める有声音は線スペクトル構造を有する。つまり本手法は、ナレーションの背後に流れている音楽を探索する場合などを想定したものである。

本手法の原理を図-3に示す。線スペクトル構造を持つ音は、スペクトログラム上では時間方向に連なったスペクトルのピークの集まりとして表され、線スペクトルどうしの間には隙間がある。隙間には、元の音（背景音楽）が垣間見えている。そこで、スペクトログラムを時間と周波数に関して小領域に分割し、そのそれぞれに対して探索を行い、その結果を集計することで背景音楽の探索を行うことができる。小領域のそれぞれに対する探索には、前述の時系列アクティブ探索法を用いることも可能であるが<sup>5)</sup>、それに限られるものではない。現在では、前景音（ナレーション）に比べて-15dBという小さな音量の背景音楽に対しても高精度な探索が可能になっている。

さらに、乗法性雑音への対処をも目的とした手法は、黒住が中心になり考案された<sup>6)</sup>。この手法は、携帯電話でとらえた音楽の断片を問合せとして音楽データベースを検索することを想定したものである。

携帯電話でとらえた音には、周囲環境、端末の音響的特性、および符号化や電波状況に起因する伝送特性などにより、品質変化が生じる。この一例を図-4に示す。正しい探索を行うためには、これらの品質変化を吸収す

ると同時に、何十万曲と蓄積された楽曲特徴のうちのどの曲のどの部分かを正確に区別しなければならない。本手法では、特徴と照合法の双方の工夫により、問合せの音に音楽成分がある程度以上含まれていて、かつその曲の特徴データがデータベースに存在していれば、携帯電話を介してもほぼ確実に曲名を判定できる頑健性を実現することができた。

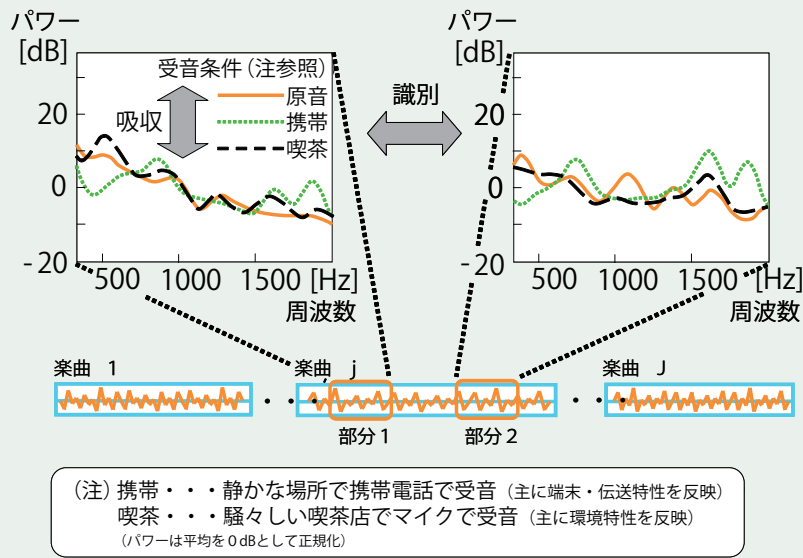
## 作品の作り手・受け手とメディア探索

このようなメディア探索技術が今後発展すると、音楽や映像の創作や鑑賞にどのような影響を与えるだろうか。

まず、作品の作り手においては、編集などの制作過程での使用が考えられる。音楽や映像の制作では、部分的に別の音楽や映像の一部を編集して用いることがある。このとき、編集された結果の作品から編集元への参照が、同一信号探索により可能になると考えられる。

次に、作品を使用する場合においては、権利処理への適用が考えられる。音楽や映像を使用する場合、あるいは音楽や映像が使用されたコンテンツを再使用する場合などには、法律上、例外規定に該当する場合を除き権利者からの許諾が必要である。このとき、タイトルなどだけではなく、現に使用する（使用されている）音楽や映像そのものでデータベース検索等ができれば便利だろう。

一方、作品の受け手においては、再生の利便性が高まることが期待される。少し前まで音楽はカセットテープやCDなどをかけ替えながら聴くものだったが、最近ではメモリやハードディスクに大量に蓄積して聴くのが一般的になった。テレビ番組の録画も、番組を選んで録画



するという従来の方法に加え、数多くの番組をとりあえず蓄積しておいて後から検索して視聴するという方法も提案されている。このように、蓄積された多くのコンテンツの中から、望みのものを探して再生するというスタイルが今後広がりそうである。そのとき、タイトルやアーティスト名などでの選択に加え、内容に基づいて検索できればより便利だろう。たとえば、現在、携帯電話でとらえた音楽の断片によって曲名などを検索できるサービスがすでにいくつか行われているが、これと同様の方法で、街角で耳にした音楽で手元のハードディスクを検索し、その曲をもう一度良い音でじっくり聞くといった楽しみ方もできよう。また、先に述べたように、同一信号探索以外の関連技術としてハミングなどの問合せで音楽を探す技術も研究されている。このように音や映像を問合せとしてメディア情報を探すことで、鑑賞の仕方が広がるのではないだろうか。

さらに、音楽の分析ツールとしての利用も考えられる。一般に音楽作品には構造を持つものが多いが、構造の分析は従来主に楽譜情報によって行われていた。これに対し、音そのものに着目して楽曲の各部分相互の類似性を測ることで、楽譜とは違った観点から、また、楽譜情報がない場合にも、楽曲構造の分析ができる可能性がある。

今後の課題

本稿では、音楽や映像を対象とするメディア探索技術に関し、筆者らの研究チームにおける研究事例を紹介す

るとともに、音楽や映像の分野での応用の可能性を検討した。周知のように、テキスト情報の探索技術は、同一文字列の探索から出発して多様な発展を遂げ有用性を増している。メディア探索技術も、芸術分野に限らず、急速な拡大を続ける我々の情報環境を存分に活用するための有効なツールになるかもしれない。しかし、現状のメディア探索技術にはまだ改善の余地が大きいことも確かである。本稿では主に同一信号探索の技術に着目したが、望みのものを的確に、しかも素早く探すという目的のためには、技術の積み上げとさらなる発展が必要である。これは単に、これまでのメディア探索の枠組みに、経験則に依拠した類似性尺度を導入するといったような安易なアプローチでは明らかに不十分である。特徴空間や対象の的確なモデル化に基づく高速探索を実現することが今後の発展の鍵となるだろう。

参考文献

1) 柏野邦夫, ガビン・スミス, 村瀬 洋: ヒストグラム特徴を用いた音響信号の高速探索法—時系列アクティブ探索法—, 電子情報通信学会論文誌, Vol.J82-D-II, No.9, pp.1365-1373 (1999).  
2) 木村昭悟, 柏野邦夫, 黒住隆行, 村瀬 洋: 区分線形写像に基づく音や映像の高速探索—特徴系列の分割による次元削減の導入—, 第14回データ工学ワークショップ, 3-A-1 (2003).  
3) Kashino, K., Kimura, A. and Kurozumi, T.: A Quick Video Search Method Based on Local and Global Feature Clustering, Proc. of International Conference on Pattern Recognition (ICPR), Vol.3, pp.894-897 (2004).  
4) 特集: マルチメディア時代のデータベース索引技術, 情報処理, Vol.42, No.10, pp.951-987 (Oct. 2001).  
5) 永野秀尚, 柏野邦夫, 村瀬 洋: 多数の小領域スペクトログラムの探索に基づく背景音楽の高速探索法, 電子情報通信学会論文誌, Vol.J87-D-II, No.5, pp.1179-1188 (2004).  
6) Kurozumi, T., Nagano, H. and Kashino, K.: 特許出願 PCT/JP2005/012667 (WO/2006/006528) (2005).

(平成18年2月10日受付)