音楽情報科学 16－9
（1996．7．28）

音響データベースのマルチモーダル検索の試み
― 音響によるデータ検索 ―

斉 海　　村松 太一　　橋本周司
早稲田大学理工学部応用物理学科

データベースを構築する際に考慮する第一の点は、データ間の関連付けや分類方法である。マルチメディアデータベースの構築においてはこれが複雑となる。人間は多様化した情報を様々な手段を用いて表現する。マルチメディア化した情報に対して、従来のテキストベースのデータベースシステムは、多様化した情報を検索するユーザーインターフェースとしては十分ではない。必要となるのは様々な情報を基にして検索を行なうマルチモーダルな検索システムである。本稿ではその第一段階として、音をキーとして音データを検索する音響データベースを提案する。このシステムでは、ユーザーは目的とする音を容易に検索することが出来る。なお、このシステムは音の加工機能を持っているため、データベース中に目的の音がなくとも音を加工することによって、目的の音にたどり着くことができる。

Sound Database System for Multimodal Data Retrieval
— Data Retrieval by Sound —

Hai Qi　　Taichi Muramatsu　　Shuji Hashimoto
School of Science and Engineering, Waseda University

When we build a database system, we should consider relationships between data and classification of data. These are, however, difficult for building multimedia database systems. Human expresses diversified informations by various ways. Text database systems may not provide useful user interface for retrieving diversified informations. We will need a database system with a multimodal data retrieval ability for various kinds of informations. As the first stage for multimodal database system, this paper proposes a new type of sound database system using a sound for data retrieval. By using this system, users are able to find an objective sound more easily. And this system can process sounds like a sound effector so that users are able to obtain the objective sound even when it is not included in the database.

1 はじめに

試行錯誤などによる人間の創造過程はデータベースの検索過程とよく似ている。してがって、データの加工機能を持ったデータベースシステムは、芸術的創造活動の支援に有効であると考えられる。例えば、音響データベースは作曲家、サウンドデザイナー、音響演出家など音を創り出す人々にとって、自分が想像している音を検索する際に有用なものである。しかしながら、音響のデータベースは文字ベースのものと異なり、データ間の関連づけや分類が容易ではない。そのため、音データの検索にキーワードを用いた場合わずかな時間で目的の音にたどり着くのは困難である。最近の研究では、いくつかの音検索システムが報告されている。その一つとして、データベースに音の分析やミキシングの機能を備えたものがある [Blum et al] [Keislar et al]。また、音の合成機能を備えたアナログデータベースも提案されている [Vertegaal & Bonis]。その他、ニューラルネットワークを用いてアコースティックな特質と知覚的な性質とのマッピングを行ない、音響データベースの自動索引生成を行なうシステムも提案されている [Feiten & Gunzel]。本稿では、マルチモーダルな検索システムを構築する第一段階として、音をキーとして音データを検索する音響データベースを提案する。これによって、ユーザーが探している音を容易に検索することが可能となる。

2 マルチモーダルな情報検索

従来のデータベースシステムは、主にキーワードを用いてテキストベースのデータ検索を行なっていた。しかし近年では、情報がマルチメディア化し、ユーザーがデータベースシステムを使用する際には多様化した情報をテキストのみで表現することが困難になっている。そのため、キーワードのみによるデータ検索では目的の情報を見つけるのに多大な労力を要する。そこで我々は、マルチモーダルな検索システムを提案する。これは、ある情報（画像、音楽など）を検索する時に、その情報に対する様々な要素（画像、音楽、言語、身振りなど）を従来のデータベースにおけるキーワードに代わるキー要素として構成し、これらを用いて検索を行なうものである。システム構成を図1に示す。このマルチモーダルな検索システムを用いると、ユーザーは

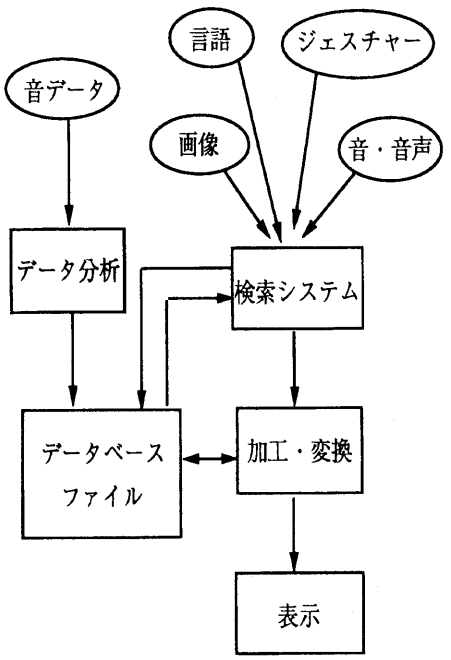

図 1: Multimodal data retrieval

検索したい情報を様々な要素を使って容易に表現することができる。また、図1に示すシステムは情報を加工、変換することができるため、ユーザーが探している情報がデータベース中になくとも新たにつくり出すことができる。このようなシステムは複雑な構造を必要とするが、新しいマンマシンインターフェースの構築の足掛かりになると考えている。本稿ではまず第一段階として、単音を対象として音により検索する音響データベースシステムを構築する試みについて報告する。

3 システム構成

本稿が提案する音響データベースシステムの構成を図2に示す。このシステムでは、音のデータから時間領域の特徴と周波数領域の特徴を抽出し、これらを分類して索引を作る。更にこれらの特徴を索引キーとしてデータベースを構築する。ユーザーがシステムにキーとなる入力音を与えると、同様な方法でこの音の特徴をとる。次に、キーとする音に

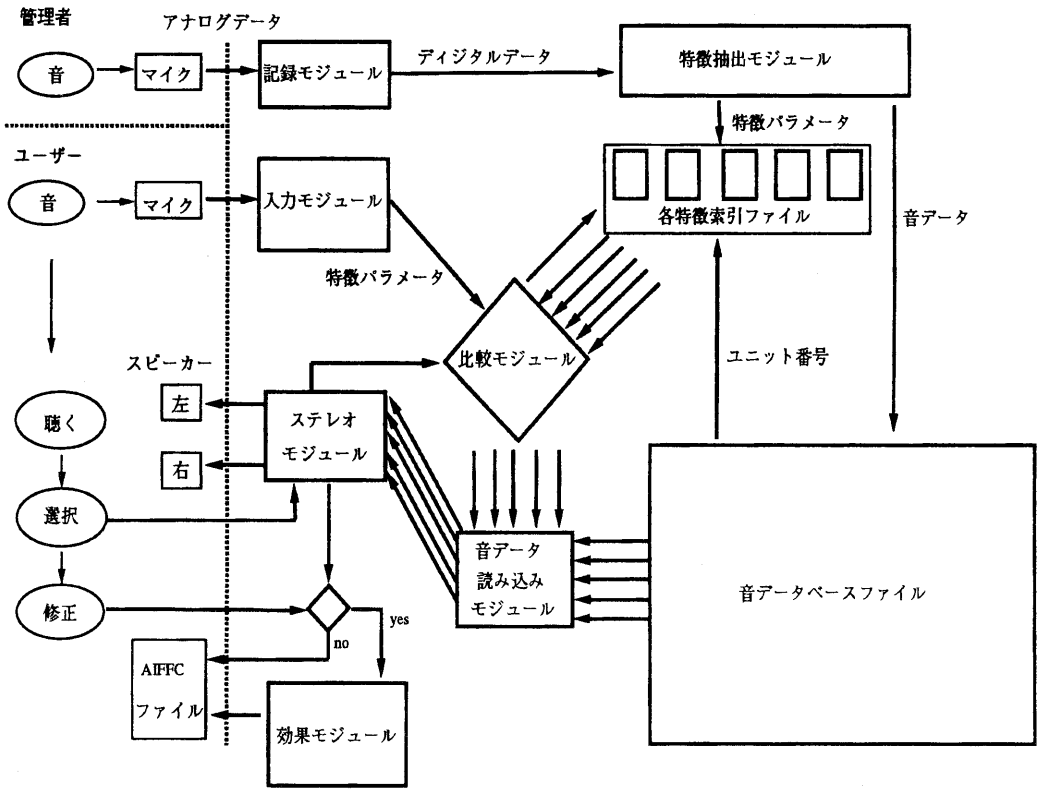

図 2: System overview

近い音をデータベース中から5つ選び、ステレオ出力を使い異なる方向から五つの音をユーザーに提示する。これにより音を5回提示するのに比べて時間が大幅に短縮できる。ユーザーはその中から最も相応しい音を選ぶ。そこで選んだ音に満足できなければ、さらにその選んだ音に近い5つの音を提示する。このとき、優先したいパラメータの比重を大きくすることもできる。この作業を繰り返すことによってユーザーは目的の音に容易にたどり着くことができる。

また、このシステムはデータ加工能力を持っており、最終的に選んだ音にも満足できない場合、ユーザーは自分で音の特徴パラメータを調整して目的の音に近付けることができる。このように、本システムではデータベースとしての機能だけでなく、音加工機能も実現しており、一種の音源として使用することもできる。

4 音の特徴抽出とデータ格納

4.1 記録モジュールと音データベースファイル

このモジュールでは入力音を記録しデジタルデータに変換した後、1つの音データユニットを構成する。このユニットは、サンプリングレート 44.1kHz で 16 bit 量子化された 65536 個のディジタル値からなっている。したがって、取り扱う音の持続時間は約 1.5 秒以下である。

図 3に示すように、記録モジュールで作られた音データユニットは一つずつ音データベースファイル

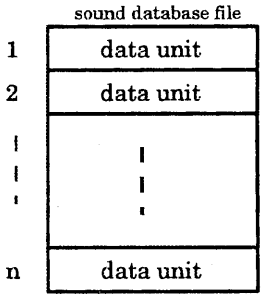

図 3: Sound database file

に格納される。各音データユニットは同じデータ数で構成されるため、ユニット番号からその先頭アドレスを計算することができる。例えば、仮に i 番目のユニットが音データベースファイル中にあるとすると、そのユニットは $(i-1) \times 65536 + 1$ というアドレスから始まっている。

4.2 特徴抽出モジュールと特徴パラメータ索引ファイル

特徴抽出は、正規化された波形データに対して時間領域と周波数領域について行なう。限られたパラメータによって単音の時間特性を正確に表現するためにエンベロープをモデル化して、パラメータを抽出する。ここでは、アタック時間 t_1, ディケイ時間 t_2, サステイン時間 t_3, リリース時間 t_4, 最大出力レベル L_1, 及び サステインレベル L_2 の 6 つとした。

一方、周波数特性はパワースペクトルを FFT で計算することによって解析する。図5のようなスペクトルから、まず第一に、スペクトル特性で最も重要なパラメータである基本周波数 F_1 を抽出する。次に包絡要素の抽出を行なう。スペクトル包絡には多くのピークがあるが、その中から大きい順に 3 つのピークを選び、それぞれのピークレベル (S_1,S_2,S_3) と周波数 (F_2,F_3,F_4) を抽出する。又、基本周波数及びその倍音のパワーと総信号パワーの比を周期性特徴 F_5 として計算する。

索引ファイルは各特徴パラメータごとに作る。そして、これらのパラメータは各索引ファイル中にレコードを構成するために格納され、データ検索

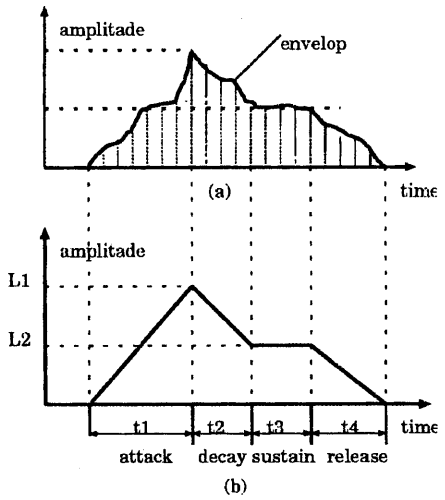

図 4: Temporal feature extraction

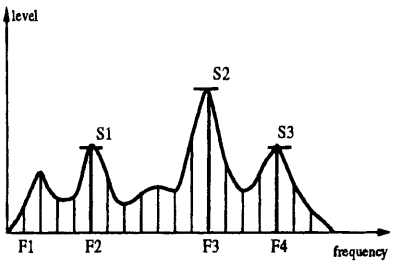

図 5: Spectrum feature extraction

に用いられる。各索引ファイルは全ての音データユニットに対して作成する。

5 音をキーとしたデータ検索

5.1 入力モジュール

このモジュールは音データ検索時に用いられるもので、ユーザーからの音入力を受け、特徴抽出モジュールと同様にその音の特徴を抽出し、音データ検索のキーとなるパラメータを各索引ファイル中にある音の特徴パラメータと同様の型で作成する。ここで抽出された特徴パラメータは比較モジュールに

渡される。以下では、検索パラメータは t_{kr}, L_{lr}, S_{mr}, F_{nr} $(k = 1,..,4, l = 1,2, m = 1,2,3, n = 1,...,5)$ とする。

5.2 比較モジュールと音データ読み込みモジュール

比較モジュールは参照データと保存データとの距離を次式より計算する。

$$E_i{}^2 = c_1(\sum_{k=1}^{4} a_{tk}(\frac{t_{kr} - t_{ki}}{t_{kr}})^2 + \sum_{l=1}^{2} a_{Ll}(\frac{L_{lr} - L_{li}}{L_{lr}})^2) + c_2(\sum_{m=1}^{3} a_{Sm}(\frac{S_{mr} - S_{mi}}{S_{mr}})^2 + \sum_{n=1}^{5} a_{Fn}(\frac{F_{nr} - F_{ni}}{F_{nr}})^2) \quad (1)$$

ここで、t_{ki}, L_{li}, S_{mi}, F_{ni}, は索引ファイル中の i 番目のレコードである。c_1, c_2 はデータ検索時に用いる係数であり、時間特徴とスペクトル特徴の重視度を決定するものである。なお、a は各特徴パラメータの重要性を比で表す係数である。これらの係数は、ユーザーが入力した音と最終的にデータベース中からんだ音との関係から、次式によって更新する。

$$a_i' = a_i^{x-1} + (\frac{1}{|f_{ri} - f_{si}| + 1}) \quad (2)$$

$$a_i^x = (\frac{a_i'}{\sum_k a_k'}) \quad (3)$$

$$(k = t_1, ..., t_4, L_1, L_2, S_1, .., S_3, F_1, ..., F_5)$$

f_{ri} はユーザーからの入力音の特徴パラメータ、f_{si} はユーザーが最終的に選んだデータベース中の音の特徴パラメータであり、x は検索の回数を表す。ここでは、式 (2) で a_i を更新し、式 (3) で規格化を行なっている。この a_i が更新されていくと、音の特徴を分析する際の各特徴の重要度が次第に導き出されていく。

$E_i < E_j$ の時、i 番目のレコードが指す音は j 番目のレコードが指す音よりも入力した音と似ているということを意味する。そして E を比較し音データ読み込みモジュールを使うことによってデータベースファイルから入力音と似た五つの音を選ぶことができる。

データベース中の各音データを比較していたのでは時間がかかってしまうため、このシステムでは比較する音データの個数を N と限定する。各索引ファイルそれぞれについて、索引ファイル中の特徴パラメータ値の中で検索キーとした音の特徴パラメータ値に近いものを選び、その選ばれた特徴パラメータに対する音ユニットを候補とするが、各索引ファイルから選ぶデータの個数を n_i とし、全体の個数 N に対する比 a_i から計算して選ぶ。つまり式で表すと、

$$n_i = N \times a_i \quad (4)$$

$$N = \sum_i n_i \quad (5)$$

$$(k = t_1, ..., t_4, L_1, L_2, S_1, ..., S_3, F_1, ..., F_5)$$

そして、N 個の音データについて E_i を比較し、E_i の小さい順に 5 つ選び、ユーザーに提示する。ユーザーは 5 つの音の中から 1 つを選び、その音が自分が探していた音ならばそこで検索は終るが、そうでなければさらに検索を続ける。そのとき、ユーザーはデータ比較に用いる時間特徴とスペクトル特徴との比率係数 $c_1 : c_2$ を自分で決めることができ、システムはユーザーが決めた比率係数を用いて検索を行なう。また、ユーザーが比率係数をシステムに与えなければ比率係数は 1 : 1 として検索を行なう。

5.3 ステレオモジュールと effect モジュール

ステレオモジュールは検索した結果として 5 つの単音データをステレオサウンドに変え、異なる位置から 5 つの音が聞こえるようにする。そのため、ユーザーは容易に出力音を比較し音の空間的な位置から目的の音に最も近い音を聞き分けることができると同時に提示時間を節約することができる。図 6に示すような単純なパワー配分で、この機能を実現している。各音は開始時間をずらして同時に提示

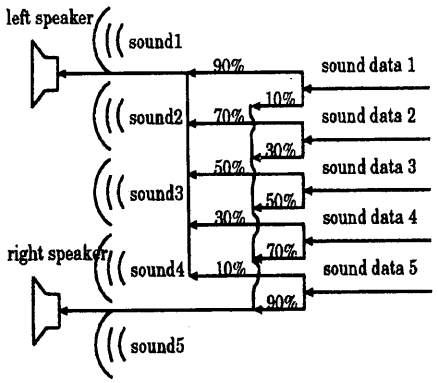

図 6: Stereo module

される。また、指定した音のみ、あるいは全音を同時に聞き直すことも可能である。

必要とする音はデータベース中にではなく、ユーザーの頭の中にあると考えられるため、データベース中の最も似た音でもユーザーの期待しているものであるとは限らない。この問題を解決するために、このシステムでは次のような機能持った効果モジュールを提供する。ユーザーはシステムから選んだ音の時間特徴パラメータを変え、効果モジュールは変わったパラメータを基にして、選んだ音データを修正する。さらに、時間軸変換とフィルタリングも可能である。このモジュールはユーザーが満足する音にたどり着くまで修正を繰り返す。そして、修正した音データをデータベースの保存データとして新しく格納することができる。従って、データベースを使っているうちに音の種類が増えることになる。

5.4 ファイル保存

このシステムでは、データベースから検索してきた音、または加工した音などをそれぞれ音データファイルとして保存することができる。このファイルは AIFFC フォーマットで、ファイル名はユーザー自身で決める。この機能は、音をファイルに保存することによって、他のシステム（AIFFC フォーマットが使えるシステム）でも利用可能にするためである。

6 終りに

本稿ではマルチモーダルな検索システムの第一段階として、音をキーとしてデータ検索を行なう新しい音響データベースを紹介した。また、目的とする音はデータベース中にはなくユーザーの頭の中にあることを踏まえ、データベースの新しい機能として、データベース中に探したいデータがなくとも、効果モジュールを用いて新しくデータを作成する機能も提案した。ここで提案したシステムは単なる音響データベースではなく、検索機能と音処理機能を用いて連想的に新しい音を探すインタラクティブな音作成システムの機能も備えているといえる。ただし、特徴パラメータの選定や音の加工方法には改良の余地が残されている。さらに細かい特徴の利用や複数音の混合による新しい音の生成などを考える必要がある。本稿では音を主体としたデータ検索方法を提案したが、さらにこのような問題を解決するとともに、画像、身振り、言葉なども検索の手段に用いるマルチモーダルなデータ検索システムの構築を考えて行きたい。

参考文献

[Blum et al] T. Blum, D. Keislar, J. Wheaton, and E. Wold, "Audio Database with Content-Based Retrieval", *Proc. IJCAI'1995, Workshop on Intelligence Multimedia Information Retrieval,* 1995.

[Keislar et al] D. Keislar, T. Blum, J. Wheaton, and E. Wold, "Audio Analysis for Content-Based Retrieval", *Proc. ICMC'1995,* pp.199-202, 1995.

[Vertegaal & Bonis] R. Vertegaal and E. Bonis, "ISEE: An Intuitive Sound Editing Environment", *Computer Music Journal,* Vol.18, No.2, pp.21-29,, 1994.

[Feiten & Gunzel] B. Feiten and S. Gunzel, "Automatic Indexing of a Sound Database Using Self-organizing Neural Nets.", *Computer Music Journal,* Vol.18, No.3, pp.53-65,, 1994.