始端特徴依存連続DPを用いた鼻歌入力による楽曲信号のスポッティング検索の高速化

西村 拓一† **橋口 博樹**‡ **関本 信博**‡ **張 建新**∗ **後藤 真孝**∗∗ **岡 隆一**‡
† **新情報処理開発機構 / 産業技術総合研究所** E-mail:taku@ni.aist.go.jp
‡ **新情報処理開発機構**
∗ **(株) メディアドライブ**
∗∗ **科学技術振興事業団さきがけ研究 21 「情報と知」領域 / 産業技術総合研究所**

あらまし　我々は，音楽音響信号のデータベースから，鼻歌のメロディーに類似した区間を見つけ出す音楽検索システムを提案している．本システムでは，個人的に収集したビデオデータからの検索も可能である．しかし，このシステムで用いていた「Model driven path 連続 DP」と呼ぶマッチング手法は，クエリーの時間軸，データベースの時間軸，音高の軸からなる 3 次元空間中において，局所類似度を連続 DP に基づいて累積し，極大となる累積類似度を計算するもので，その計算量が大きい．そこで，クエリーの始端周辺の音高が正しく推定できたと仮定することで，音高軸を削減した 2 次元空間における局所類似度の累積に基づく「始端特徴依存連続 DP」を提案する．本稿では，ポピュラー音楽 20 曲について鼻歌検索実験を行い，約 7 割の検索率を維持しつつ，計算量を従来法の約 1/40 に低減できることを示す．

Music Signal Spotting Retrieval by a Humming Query Using Start Frame Feature Dependent Continuous Dynamic Programming

Takuichi Nishimura†, **Hiroki Hashiguchi**‡, **Nobuhiro Sekimoto**‡, **J.Xin Zhang**∗, **Masataka Goto**∗∗ **and Ryuichi Oka**‡
†**Real World Computing Partnership / National Institute of Advanced Industrial Science and Technology** E-mail:taku@ni.aist.go.jp
‡ **Real World Computing Partnership**
∗ **Mediadrive Corp.**
∗∗ **“Information and Human Activity”, PRESTO, JST / National Institute of Advanced Industrial Science and Technology**

Abstract　We have developed a music retrieval method that takes a humming query and finds similar audio intervals (segments) in a music audio database. This method can also address a personally recorded video database containing melodies in its audio track. Our previous retrieving method took too much time to retrieve a segment: for example, a 60-minute database required about 10-minute computation on a personal computer. In this paper, we propose a new high-speed retrieving method, called start frame feature dependent continuous Dynamic Programming, which assumes that the pitch of the interval start point is accurate. Test results show that the proposed method reduces retrieval time to about 1/40 of present methods.

1　まえがき

近年，大規模かつ多様な音楽音響信号データがインターネット上で入手可能となってきているだけでなく，個人の計算機にも蓄積されるようになっている．しかし，このようなデータに対する検索は，予め信号に付加した曲名，歌手名，ジャンルなどのタグをテキストベースで検索するものが主となっている．そこで，データベースの音楽信号へのタグ付けを行わなくても，ユーザの鼻歌により希望の部分を検索することを本稿の目的とする．ここで,「鼻歌」は通常の歌詞の無い鼻歌だけでなく，ユーザが音高の変化を伴って自然に発する口笛や歌詞付きの歌も含むものとする．このような検索方法は，音響トラックに音楽を含むビデオデータに対しても有効である．日常的に収集する音楽音響信号データにタグを付加する手間を考えると，このような検索は重要であると考える．もし，データベースが MIDI など楽譜データから構成できるなら，相対音高，相対音長などによる記号ベースでロバストに検索できる．つまり，この場合には記号ベースの検索が効率的である．[1-4]

一方，音楽音響信号から抽出したメロディーは多くの誤りを含むため，このような記号ベースの検索手法は適用困難である．そこで，我々は，パターンベースの検索手法を提案した．[5] まず，音楽音響信号をフレーム（本実装では 64ms）ごとに分析し，音高ごとにメロディーの確信度を求めた時系列（本稿では，音高確信度時系列と呼ぶ）を作成する．鼻歌についてのみ，音高確信度時系列から各フレームごとに最も確信度の高い音高を取り出した音高時系列を作成し，さらに時間的な音高変化を求めた音程 (音高差) 時系列を求めて，これをモデルとする．そして，図 1 に示すように，クエリーの音高変化と類似した変化パターンを，データベースの音高-時間平面上にプロットした音高確信度時系列中から検出する．このとき，クエリーのパターンを時間方向に伸縮させかつ音高軸方向に移動させてマッチングするモデル依存傾斜制限型連続 DP(Model driven path Continuous DP : mpCDP) を提案した．時間方向の伸縮は 1/2〜2 倍を許容しているが，これは鼻歌のテンポが時間的に変動する場合に対処するためである．また，音高軸方向

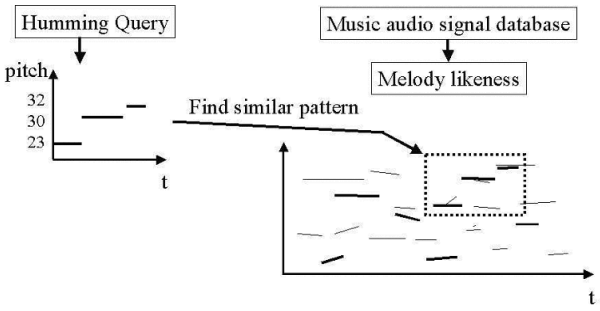

図 1: 鼻歌による音楽のパターンベースの検索

の移動は，ユーザが楽曲と異なる高さで歌唱した場合に対処するために必要である．本手法は，口笛のように音の切れ目が不明瞭なため安定した採譜が困難なクエリーでも適用可能である．

しかし，従来用いていた mpCDP では，クエリの時間軸，データベースの時間軸，音高軸からなる 3 次元空間中の局所類似度を累積して極大となる累積類似度を求めているため，計算量が大きい．そこで，クエリの始端周辺の音高が正しく推定できたと仮定することで，音高軸を削減した 2 次元空間を探索する始端特徴依存連続 DP (Start Frame Feature Dependent Continuous DP ： s-CDP) を提案する．この s-CDP では，従来の連続 DP と異なり，クエリ，データベースの 2 つの時間軸からなる平面内の各点の局所類似度を予め決定できない．平面内左下から漸化的に最適パスを伸ばし，該当の点に至る最適パスの始端が求まって初めて局所類似度が求まり，同時に累積類似度が求まる．本稿では，ポピュラー音楽 20 曲について鼻歌検索実験を行い，約 7 割の検索率を維持しつつ，計算量を従来法の約 1/40 に低減できることを示す．

本稿の構成は，2 節にて，従来の鼻歌検索システムについて述べ，3 節にて始端特徴依存連続 DP を提案し，その概念や特徴を説明する．4 節では，本手法の評価実験を行い，5 節でまとめる．

2　従来の鼻歌検索システム

連続 DP[7] は，音声やジェスチャの認識のために時間単調性を導入した動的計画法（DP) を改良した手法である．連続 DP により，クエリ

と類似した区間を DB 中から切り出し検索することができる．しかし，移調した鼻歌と類似した区間を検索することは困難であった．そこで，音高軸を追加したモデル依存傾斜制限型連続 DP(mpCDP) を提案した．今後，各種連続 DP においては，鼻歌などの検索入力であるクエリから得られる時系列パターンを参照パターン，データベースから得られる時系列パターンを入力パターンと呼ぶものとする．

2.1 連続 DP

連続DPは，図2(a)に示すように，T フレームの特徴ベクトルからなる参照パターン R_τ $(1 \leq \tau \leq T)$ と入力パターン $I_t (0 \leq t < \infty)$ との累積類似度を入力軸方向に連続して求める．ただし，入力，参照の時間軸を t, τ と区別した．累積類似度は，以下の手順で漸化的に求める．まず，I_t と R_τ との局所類似度を $s(t, \tau)$ と表記する．次に，図2(c)に示すような3個のパスについて，類似度が最も高いパスを選択して累積類似度 $S(t, \tau)$ を更新する．
境界条件 $(1 \leq \tau \leq T, 0 \leq t)$：

$$S(t, 0) = S(t, -1) = 0 \tag{1}$$

$$S(-1, \tau) = S(0, \tau) = 0 \tag{2}$$

漸化式 $(1 \leq t)$:

$$S(t, 1) = 3 \cdot s(t, 1) \tag{3}$$

$$S(t,\tau) = \max \begin{cases} S(t-2, \tau-1) + 2 \cdot s(t-1, \tau) \\ \quad + s(t, \tau) \\ S(t-1, \tau-1) + 3 \cdot s(t, \tau) \\ S(t-1, \tau-2) + 3 \cdot s(t, \tau-1) \\ \quad + 3 \cdot s(t, \tau). \end{cases} \quad (2 \leq \tau \leq T) \tag{4}$$

図2(c)中の各格子点近傍の数字は，局所類似度に対する重みであるが，どのパスを通っても参照軸方向に1フレーム上がるごとに3の重みがかかるため，累積類似度 $S(t, T)$ を重みの和 $3T$ で割れば正規化できる．

次に，図2(b)に示すように，累積類似度 $S(t, T)$ が入力軸 t において一定のしきい値 α 以上で極大点となる点を求め，これを終端とする入力中

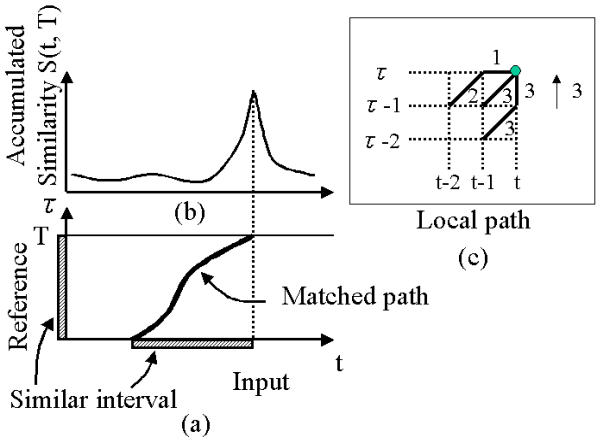

図 2: 連続 DP

の類似区間を参照パターンの類似区間として切り出す．

つまり，局所的なパスの選択処理を，参照，入力平面内の格子点すべてに対して左下から右上へ順次行うことで，最大の累積類似度およびそのときの最適パスを求めている．参照パターン全体との累積類似度 $S(t, T)$ は，$\frac{1}{2}$〜2倍の時間伸縮を許した入力と参照パターンとの最大の累積類似度となっている．

2.2 mpCDP による鼻歌検索

本節では，mpCDP を用いた従来の鼻歌検索システムを，図3を用いて説明する．まず，鼻歌から各フレームにおける音高時系列を求め，この時間差分である音程時系列を mpCDP のモデルとする．音楽からは音高確信度時系列を求め，この確信度を mpCDP の局所類似度とする．累積類似度を求めるための局所パスは，連続 DP と同様に時間伸縮を許容する．しかし，モデルと同じ音高変化パターンを検出するために，モデルの音程に従って音高軸方向へ局所パスをシフトさせる．局所パスの選択を漸化的に行うことにより，図3右上のように，点 (t, T, x)（音高軸を x とする）を終点とする累積類似度は参照パターン終点に対応する音高 x において極大となる．さらに，連続 DP と同様に，極大累積類似度 $\max_x S(t, T, x)$ の入力軸 t 方向の変化を調べ，一定のしきい値 α 以上の極大点を終端とする部分区間を参照パターンの類似区間として出

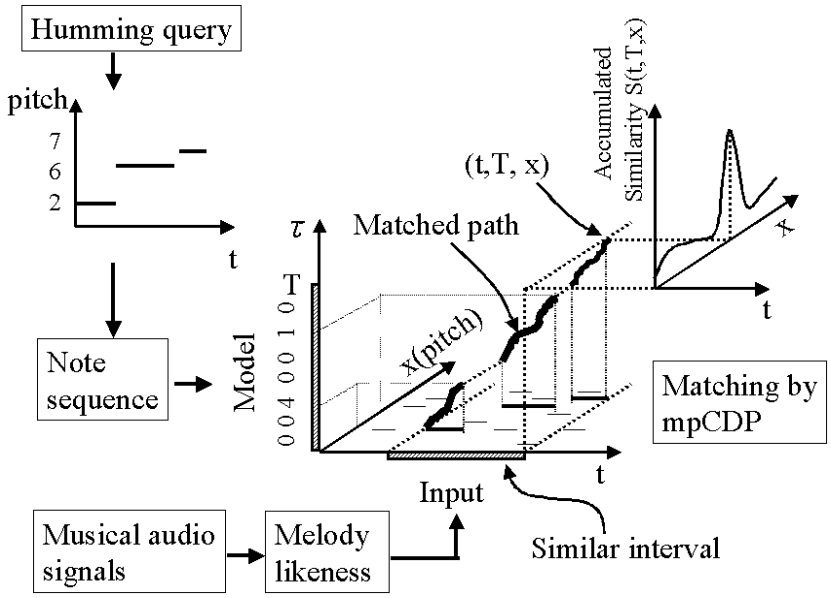

図 3: 従来の mpCDP を用いた鼻歌検索システム

力する．

3 始端特徴依存連続 DP の提案

3.1 始端特徴依存連続 DP

本稿で提案する始端特徴依存連続DP(s-CDP)とは，連続 DP で定義した I_t と R_τ との局所類似度 $s(t,\tau)$ が，点 (t,τ) に至る最適パスの始端 $(p(t,\tau),1)$ の特徴によって定義される連続DPのことである．つまり，s-CDP での局所類似度は，ある関数 $f()$ を用いて $s(t,\tau)=f(R_\tau,I_t,R_1,I_{p(t,\tau)})$ と記述することができる．(従来の連続 DP では，$s(t,\tau)=f(R_\tau,I_t)$) ．

従来の連続 DP では，図 4(a) に示すように予め局所類似度が定義でき，これを累積している．しかし，ｓ-CDP では該当の格子点に至る最適パスの始端の情報が無ければ局所類似度を決定できない．従って，図 4(b) に示すように，平面内の左下から局所類似度が順次決定される．始端によって局所類似度が異なるため，式 (4) の 3 個のパス毎に局所類似度を求める必要がある．ｓ-CDP の漸化式を定式化すると式 (4) は以下のよう書き直される．

漸化式 $(1 \le t)$:

$$S(t,1)=3\cdot s_2(t,1). \tag{5}$$

$$S(t,\tau)= \max\begin{cases} S(t-2,\tau-1)+2\cdot s_2(t-1,\tau) \\ \quad +s_1(t,\tau) \\ S(t-1,\tau-1)+3\cdot s_2(t,\tau) \\ S(t-1,\tau-2)+3\cdot s_2(t,\tau-1) \\ \quad +3\cdot s_3(t,\tau). \end{cases} \quad (2\le\tau\le T) \tag{6}$$

ここで，図 2(c) の局所パスを，左上から順に path1, path2, path3 とし，$s_1(t,\tau)$,$s_2(t,\tau)$,$s_3(t,\tau)$ はそれぞれ，path1, path2, path3 の局所パスを取ったときの局所類似度とする．これらは，前述の関数 $f()$ を用いて

$$\begin{aligned} s_2(t,\tau)&=f(R_\tau,I_t,R_1,I_{p(t,\tau)})(\tau=1)\\ s_1(t,\tau)&=f(R_\tau,I_t,R_1,I_{p(t-2,\tau-1)})(2\le\tau\le T)\\ s_2(t,\tau)&=f(R_\tau,I_t,R_1,I_{p(t-1,\tau-1)})(2\le\tau\le T)\\ s_3(t,\tau)&=f(R_\tau,I_t,R_1,I_{p(t-1,\tau-2)})(2\le\tau\le T) \end{aligned}$$

と記述することができる．ここで，$s_2(t,\tau)$ のみ $\tau=1$ となることがあり，このときは始端と同一の点となるため例外処理を行っている．式 (6) における path1,path3 の第 2 項が $s_2(\cdot,\cdot)$ となっている理由は，どちらも平面上で相対的に (-1, -1) の点からのパスとなっているためである．

類似区間は，従来の連続 DP と同様に累積類

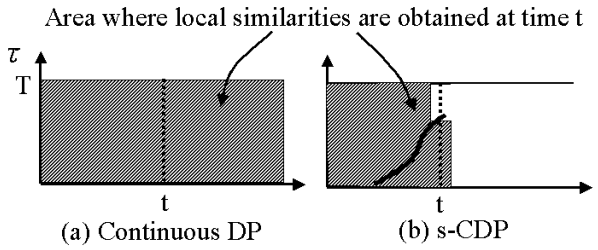

図 4: 連続 DP と s-CDP の局所距離決定手順の違い

似度の極大値で求めるが，同時に始端の位置および特徴も得られる．

以下に始端の入力軸の値 $p(t,\tau)$ の算出方法を示す．まず，$\tau = 0, 1$ において入力軸の時刻で初期化する．

$$\mathrm{p}(t,0) \quad = \mathrm{p}(t,1) = t. \tag{7}$$

次に，以下の式で，選択された局所パスの始点に保存されている最適パスの始端フレーム番号をコピーする．

$$\mathrm{p}(t,\tau) \quad = \max \begin{cases} \mathrm{p}(t-2,\tau-1) \ (\text{if} \ \ \text{path1}) \\ \mathrm{p}(t-1,\tau-1) \ (\text{if} \ \ \text{path2}) \\ \mathrm{p}(t-1,\tau-2) \ (\text{if} \ \ \text{path3}) \end{cases} \quad (2 \le \tau \le T) \tag{8}$$

以上の漸化式を連続 DP に追加することで，始端の入力軸の値 $p(t,\tau)$ を決定できる．

始端において特徴抽出に失敗した場合，R_1, $I_{p(t,\tau)}$ の値が誤ったものとなり，正確な局所類似度が得られない．従って，参照パターンでは，高い確信度で特徴抽出できたフレームを始端として切り出すなどの対策を行う必要がある．一方，入力パターンでは，切り出しを行うことが目的であるため，そのような対策は困難である．しかし，始端周辺のフレームのうちで一つでも正しい特徴を抽出できれば，連続 DP の時間伸縮で可能な限りパターンを変形することで，正しい始端特徴を自動的に選択できる．

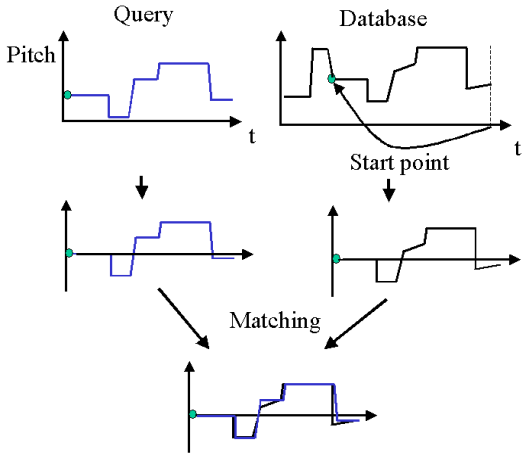

図 5: 移調に対応したメロディーのマッチングの概念

3.2 鼻歌検索システムへの適用

本節では，s-CDP を鼻歌検索システムに適用する．メロディーは音高時系列で表される．鼻歌が DB 中の楽曲と同じ音の高さであれば，従来の連続 DP で検出可能であるが，通常ユーザーは DB 中の楽曲と異なる音の高さで鼻歌を歌うと考えられる．そこで，始端（歌い始め）の音高が正しく推定できたと仮定して s-CDP を適用することで移調へ対応することを考える．クエリについては，図 5 に示すように始端の音高を基準とした音高時系列へ移調する．一方，DB についても当該の格子点へ至る最適パスの始端が求まった時点で，この始端の音高を基準とした音高を求める．これらの処理により，両者の音高が一致するか調べるだけで移調した類似メロデーの検索が可能となる．

具体的には，図 6 に示すように，鼻歌から各フレームにおける音高時系列を求め，この始端の音高との差を参照パターン $R'_\tau = R_\tau - R_1$ とする．音楽からは音高確信度時系列を求め，この確信度の高い順に N 個の候補を取り出した時系列を入力パターン $I_t(k)(k = 1, \cdots, N)$ とする．

本稿の実験では，局所類似度 $s(t,\tau)$ を以下の式で定義した．

$$s(t,\tau) \quad = \begin{cases} 1 \ (D_{t,\tau} = 0) \\ 0 \ (\text{else}) \end{cases} \tag{9}$$

ただし，$D_{t,\tau} = \min_k abs[R'_\tau - \{I_t(k) - I_{p(t,\tau)}(1)\}]$

とする．この式により，入力パターンの時刻 t における N 個の音高候補から始端における最も確信度の高い音高を引いた値が，一つでも参照パターンと一致すれば局所類似度は 1 となる．そうでない場合は，局所類似度は 0 となる．以下の実験では，$N = 5$ とした．

4 実験

4.1 実験方法

音楽データベースとして，ポップス系 10 曲，童謡 8 曲，アニメソング 1 曲，演歌 1 曲の合計 20 曲（約 80 分）を用意した．音楽の内訳は，男性ボーカル曲が 4 曲，女性ボーカルが 16 曲である．この 20 曲について，各曲の主にさびの部分を 5 人の人物（男性 3 人，女性 2 人）が約 20 秒歌って鼻歌クエリーを作成した．歌詞を伴って歌うかどうかは自由とし，歌詞のある区間と無い区間が混在するクエリーもあった．音響信号は，16 kHz · モノラルでサンプリングして計算機に取り込んだ．本実験では，20 曲のデータベースから 100 個のクエリーすべての検索実験を行い，本手法の有効性を示す．

特徴抽出法は，メロディーの基本周波数が，最低周波数 $\mathrm{f_b}$[Hz] (本実装では 55Hz) から最高周波数 $2^{X/12}\mathrm{f_b}$[Hz] ($X = 60$) の帯域にあるものと仮定し，入力の音楽音響信号から周波数 $2^{x/12}\mathrm{f_b}$[Hz] の音がメロディーである確信度を求めた．mpCDP での音高は，男女の音高を考慮して 5 オクターブ ($X = 60$ ステップ) とした．

クエリーから参照パターンを切り出すために，2 つのしきい値を用いた．信号の二乗和を平均した値 P を用いて，初めてしきい値 $0.5P$ 以上となる時刻を始端とした．また，クエリーの終わりから調べ，しきい値 0.1 以上となった時刻を終端とした．始端の音高は正しい必要があるため，始端を決定するしきい値を大きくした．これにより，始端周辺の音高は 9 割以上正しく推定できた．

実験では，適合率 N_C/N_D(precision rate) と再現率 N_C/N_T(recall rate) を求めた．ここで，N_C, N_D, N_T は，それぞれ検出区間のうちで正しい個数，検出結果として出力された検出区間の総個数，真の区間の個数を表す．検出区間が正しいと判断する基準は，真の区間と検出区間の重複率

$$\text{重複率} = \frac{\text{真の区間} \cap \text{検出区間}}{\text{真の区間} \cup \text{検出区間}} \qquad (10)$$

が 0.5 以上とした．この 0.5 というしきい値では，真の区間と検出区間の長さが等しいとき，重複区間の長さが検出区間の長さの約 7 割となる．

本稿では，しきい値 α （s-CDP では，累積類似度が α 以上の極大点となる区間を出力）を変化させ，再現率と適合率の平均値の最大値を求めてこれを検索率とした．さらに，100 個のクエリーに対して，検索率の平均（平均検索率）を求める．

実験は，従来法 mpCDP と提案手法 s-CDP について行い，本実験では，OS: windows2000, CPU: PentiumIV 1.5GHz の計算機を用いた．

表 1: 実験結果

検索手法	mpCDP	s-CDP
平均検索率	89.1%	73.3%
計算時間	532s	14.1s

4.2 実験結果

実験の結果，表 1 のようになった．提案手法により，計算時間が約 1/40 に削減できている．mpCDP で用いた音高軸のステップ数が 60 であったため，局所パスの選択回数は，s-CDP では音高軸が 1 になるため，1/60 になる．しかし，始端の計算や局所距離の計算量が増大したため約 1/40 の効果にとどまったと考えられる．

また，検索率は約 16%低下したものの，7 割以上の検索率が得られた．検索率低下の主な原因は，データベースの音高推定において類似区間の始端周辺の第一位の音高推定をすべて誤ったためと考える．

4.3 検索システム

上記の実験をもとに，検索システムを試作した．図 7 に，モニタ画面を示す．左上から順にクエリの音高時系列，データベースの各時刻に

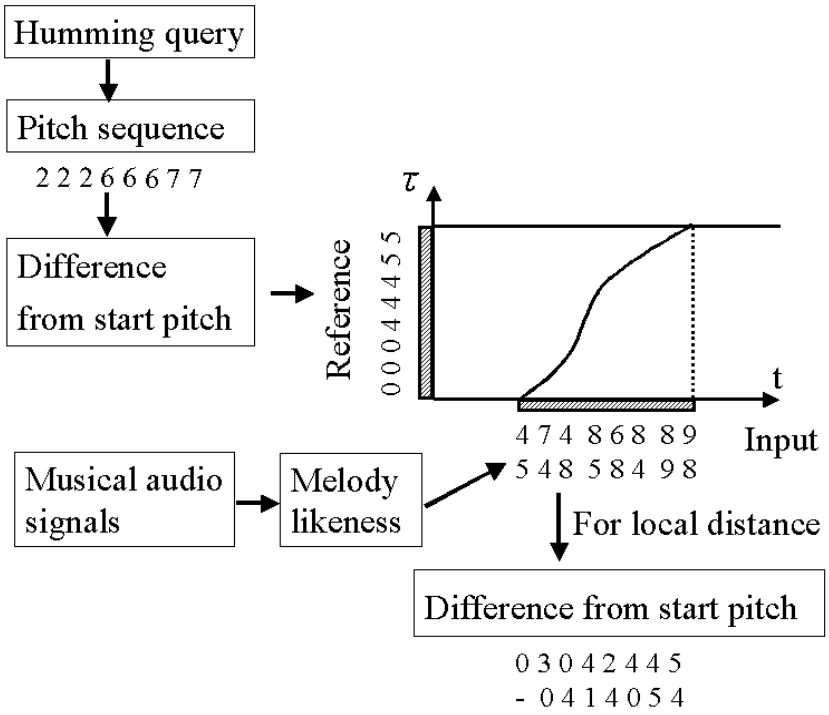

図 6: s-CDP を用いた鼻歌検索システム

おける類似度グラフ，リスト表現の出力結果となっている．波形のピークやリスト上をクリックすると音楽やビデオ中の該当個所が再生される．10 秒のクエリー，60 分のデータベースであれば，５秒程度で検索結果が得られる．図 7 では，同じ曲をビデオとラジオから取り込みデータベースとしているために，複数箇所に類似区間が検出されている．他に，同じ曲を異なるキーで別人が歌った音楽の検索も成功している．

5　まとめ

本稿では，クエリの始端周辺の音高が正しく推定できたと仮定することで，音高軸を削減した 2 次元空間における局所類似度の連続 DP による累積を行なう「始端特徴依存連続 DP」を提案した．本稿では，ポピュラー音楽 20 曲について鼻歌検索実験を行い，約 7 割の検索率を維持しつつ，計算量を従来法の約 1/40 に低減できることを示した．今後は, 音高確信度の高精度化 [6] およびデータベース中の類似したメロディーを検出して圧縮し検索時間を短縮することが課題である．

参考文献

[1] 蔭山哲也, 高島洋典, “ハミング歌唱を手掛かりとするメロディ検索”, 信学論 (D-II), vol. J77-D-II, no. 8, pp. 1543–1551, 1994.

[2] Asif Ghias and Logan J., “Query By Humming - Musical Information Retrieval in an Audio Database”, ACM Multimedia ’95, Electronic Proc., 1995.

[3] 園田智也, 後藤真季, 村岡洋一, “WWW 上での歌声による曲検索システム”, 信学論 (D-II), vol. J82-D-II, no. 4, pp. 721–731, 1999.

[4] Kosugi N., Nishihara Y., Sakata T., Yamamuro M., and Kushima K., “A practical Query-by-Humming system for a large music database”, ACM Multimedia 2000, 333–342.

[5] Hashiguchi H., Nishimura T., Takita J., Zhang J. X., and Oka R., “Music Signal Spotting Retrieval by Humming Query Using Model Driven Path Continuous

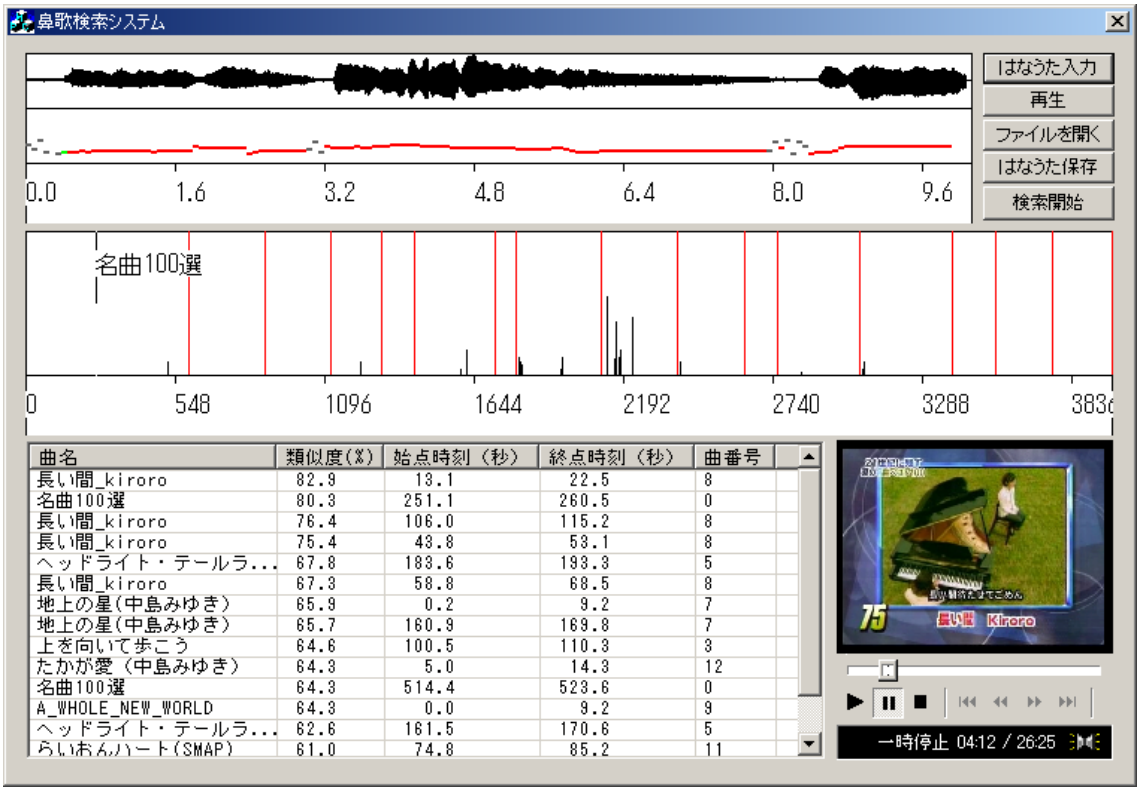

図 7: 検索システムのモニタ画面

Dynamic Programming”, SCI2001, Proc. Vol. VII, Comp. Sci. Eng. Part1, pp 280–284, 2001.

[6] Goto M., “A Predominat-F0 Estimation Method for CD Recordings: MAP Estimation using EM Algorithm for Adaptive Tone Models”, Proc. of ICASSP 2001.

[7] 岡, 隆一, “連続 DP を用いた連続単語認識”, 日本音響学会音声研資料, S78-20, pp. 145–152, 1978.