音楽情報科学　39－9
（2001. 2. 22）

鼻歌による音楽検索と歌詞音声検索の統合処理の検討

橋口 博樹† 西村 拓一† 矢部 博明† 赤坂 貴志‡ 岡 隆一†

†新情報処理開発機構
〒 305-0032 茨城県つくば市竹園 1-6-1 つくば三井ビル 13F
TEL:0298-53-1668, E-mail:hiro@rwcp.or.jp

‡メディアドライブ（株）

あらまし　近年，多量の音楽ディジタルデータを個人が所有するようになり，楽曲の検索ニーズが高まっている．これに伴い，著者らは鼻歌から音程を抽出し，音楽音響信号からは主旋律の候補を選定し，主旋律推定のあいまいさを考慮に入れた検索システムを開発した．この検索は，mp-CDP と呼ばれるマッチング手法により実現されている．本稿では，主旋律の検索機能に加え，歌詞を歌った場合の鼻歌を想定し音素認識に基づく歌詞検索機能も導入する．この場合，通常行なわれている音声認識の技術を単に採用するだけでは，歌のようにピッチが激しく変化するという状況を想定していないことと，ボーカル以外の BGM の影響により，音素認識は困難であると考えられる．そこで，本稿では，比較的認識しやすい母音の音素認識を取り上げ，さらに，楽曲からは，mp-CDP がたどった主旋律情報を音素認識に利用する方法を検討する．

キーワード　音楽検索，音階抽出，音素認識，連続 DP.

A study for integrating melody and phone retrieval functions

Hiroki Hashiguchi†, Takuichi Nishimura†, Hiroaki Yabe†,

Takashi Akasaka and Ryuichi Oka†

†Real World Computing Partnership
Tsukuba Mitsui Building 13F, 1-6-1 Takezono Tsukuba-shi, Ibaraki 305-0032, Japan
PHONE:+81-298-53-1668, E-mail:hiro@rwcp.or.jp

‡ Mediadrive Corp.

Abstract　The integration of rhythm and lyric recognition in a music retrieval system is the main purpose of this paper. In order to realize a music retrieval system based on rhythm extraction, we have already proposed a matching method called "Model driven path Continuous Dynamic Programming (mp-CDP)" to retrieve a part of music signal by a hamming query. This method detects several intervals in a music signal which are similar to a hamming query. This paper focuses on recognizing vowel categories in a song signal. Lyric recognition problem is out of scope of conventional speech recognition problems. Matching paths obtained by applying mp-CDP lead to enhance the features for recognizing vowel categories.

Key words　music retrieval, extraction of music scale, phoneme recognition, Continuous Dynamic Programming.

1 はじめに

近年，デジタル技術の進歩に伴い，音声，画像，動画，音楽，テキスト文書など，大量かつ多種多様な情報コンテンツがコンピュータ上に蓄積されている．これらのコンテンツを利用するための統合システムやマルチモーダルインターフェースの構築が盛んに研究されている．これらの研究の中では，マルチデータベースへのアクセスに応えるための，インターフェースの使い易さ，処理の高速性，正確性の実現が重要な課題となっている．特に，大量のデータベースの中からユーザが所望のコンテンツを高速かつ正確に取り出せる検索技術の開発が望まれている．

音楽と鼻歌などの音声を統合した既存の検索技術として，蔭山, 高島 [5]，園田ら [10]，西原ら [7], 橋口ら [3] がある．蔭山, 高島，園田ら，西原らは，採符処理処理された MIDI や採符が容易な鼻歌自身を楽曲データベースとしている．これに対し橋口らの特徴は，市販 CD などの音楽音響信号から主旋律の候補を抽出し，主旋律同定を行わず，鼻歌と類似する区間をスポッティング検索するという点にある．このスポッティング検索手法を mp-CDP として提案している．

本稿では，鼻歌に旋律と歌詞の情報が含まれる場合に，mp-CDP を用いた楽曲検索と歌詞音声からの検索を統合する方法について検討する．

2 鼻歌-楽曲検索システムの概要と検討すべき課題

ここで提案する鼻歌–楽曲検索システムは，鼻歌に旋律と歌詞がある場合を想定している．図 1 にシステム概要を示し，図中の実線は，それぞれ，主旋律検索と歌詞検索の一連の処理を示す．主旋律検索では，鼻歌から音程を抽出し，音楽音響信号からは主旋律の候補を選ぶという方法を採る．mp-CDP によるパターンマッチング処理 は，データベースのパワースペクトル中で，クエリーと同様なテンポで音程変化を行うパターン（一連のピーク列）を探し出す [3]. mp-CDP の一連の処理概要を図 2 に示す．

次に歌詞の抽出については，通常音声認識で行われている方法で歌詞を音素で抽出する．音声認識の手法を適用するにあたり，留意しなければならない点を以下に挙げる．

1. 人の歌はピッチ周波数が時間とともに大きく変化する．
2. 音が伸ばされることが多い

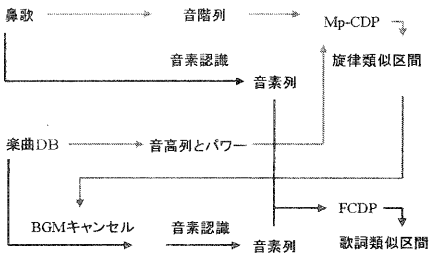

図 1: 鼻歌による旋律–歌詞検索システムの概要

3. 楽曲には，ボーカル以外に BGM (Background Music) がある．

これらの検討課題を解決するため，まず，比較的抽出しやすい母音を使った認識について考える．まず，1 については，学習の際にさまざまな周波数を含む音声から母音の音素を学習するという方法とる．また，会話でのスペクトル分析のフレーム長（約 10 msec）に比べ，長くとることを検討する．このとき，母音は子音よりも安定して取られる．3 番目の課題に対しては，mp-CDP で辿ったパスの旋律が，ボーカルのピッチ周波数の変化と類似していることに着目し，さらに母音には倍音構造があることに着目し，このピッチ周波数の倍音のパワースペクトルのみで，音素を認識してみる．

音素系列が抽出されたあとのマッチング処理は，高速連続 DP(FCPD: First Continious Dinamic Programing) [8] によって行なう．

3 主旋律検索

3.1 システム構成の概要

モデル依存傾斜制限型連続 DP(mp-CDP) は，鼻歌から得られる音程の時系列と，楽曲音響信号のスペクトル解析によって得られるパワースペクトルで決まる音階番号 x の寄与を求めた 2 次元時空間パターンとのマッチングを行う（図 2).

図 3 は，音階番号軸 x，鼻歌の時間軸 τ，楽曲の時間軸 t で構成される 3 次元空間内での mp-CDP の軌跡を示している．まず，図 3 (a) は，(x, τ)-平面においてクエリーである音程系列が指示するパスが，自由に横方向へ平行移動してもよいことを示している．これは，鼻歌の音程時系列が音階の絶対変化に依存せず，相対変化によって決まることから，鼻歌

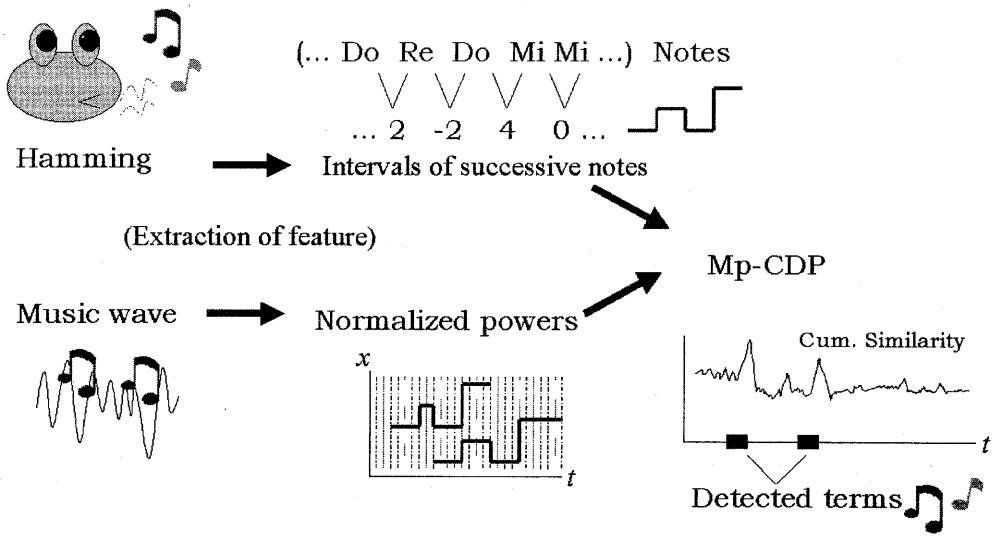

図 2: 鼻歌による主旋律検索システムの概要

が楽曲の音程よりも移調していたとしても柔軟に対応できることを示している．

図 3 (b) は，音階番号軸 x を (t,τ)-平面へ射影したときの図であり，この中のパスが連続 DP 同様に $1/2 \sim 2$ 倍の伸縮を許した中での最大パスとなっていることを示している．つまり，鼻歌のテンポが楽曲のテンポと多少違っていても許容することを意味している．

図 3 (c) は，(a) のパスの平行移動，(b) の時間伸縮を許容してマッチングされたパスを示している．これら図 3 (a), (b), (c) の定式化を次節で述べる．

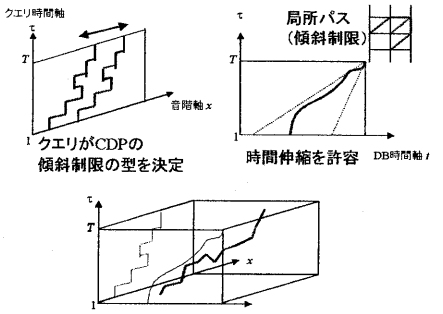

図 3: mp-CDP の概念図
Fig 3 Conceptual diagram of mp-CDP

3.2 定式化

まず，メロディーとして考えられる最も低音の周波数を f_b[Hz] とし，入力の音楽信号から周波数 $2^{1/12}f_b x$[Hz] の音が主旋律である確信度を周波数分析などを通じて求める．フレーム数 K の音高確信度時系列を，$I_t(x)$ $(t=1,\ldots,K, x=1,\ldots,X, 0 \leq I_t(x) \leq 1)$ と記述する．ただし，周波数 $2^{1/12}f_b X$[Hz] は，メロディーとして考えられる最も高い音の周波数とする．また $I_t(x)$ は，FFT 処理後のパワースペクトルを正規化して得られる量 [3] であり，1 に近いほど音高番号 x が主旋律を構成するであろうことを表している．

同様に，鼻歌の音楽信号に関しても $R_\tau(x)$ $(\tau = 1,\ldots,T, x=1,\ldots,X, 0 \leq R_\tau(x))$ を求める．このとき，$R_\tau(x)$ は，FFT 処理後のパワースペクトルに比例して決まる量 [3] であり，$R_\tau(x)$ が，閾値 ϵ より大きければ有音であり，以下であれば無音となる．各時刻において，確信度を最大とする音高番号とそのときの確信度をそれぞれ，

$$x_\tau = \arg \max_{1 \leq x \leq X} R_\tau(x)$$
$$R'(\tau) = R_\tau(x_\tau)$$

と書く．なお，一般性を失うことなく，入力鼻歌の

最初 ($\tau = 1$) と最後 ($\tau = T$) では，$R'(\tau) > \epsilon$ とすることができることに注意しておく．

鼻歌の音程時系列 $r(\tau)$ $(\tau = 1, \ldots, T)$ は，現時刻 τ 以前に確信度が閾値 ϵ を上回った最も現時刻に近い時刻を τ';

$$\tau' = \max_{\tau^* < \tau} \{\tau^* \mid R'(\tau^*) > \epsilon\}$$

とするとき，次式で求められる．

$$r(\tau) = \begin{cases} x_1, & \text{if } \tau = 1; \\ x_\tau - x_{\tau'}, & \text{if } \tau \geq 2 \text{ and } R'(\tau) > \epsilon; \\ 0, & \text{otherwize} \end{cases}$$

つまり，音程 $r(\tau)$ は，確信度が閾値 ϵ 以下の場合は 0，そうでない場合は，現時刻以前にある閾値 ϵ を超えた確信度を持つ最も近い音との音程である．

このとき，局所類似度 $d(x, \tau, t)$，累積類似度 $S(x, \tau, t)$ を次の (1) ～ (3) で定義する．

$$d(x, \tau, t) = \begin{cases} I_t(x), & \text{if } R'(\tau) > \epsilon \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$S(x, 1, t) = 3d(x, 1, t); \quad (2)$$

$$S(x, \tau, t) = \max \begin{cases} S(x - r(\tau), \tau - 1, t - 2) \\ \quad + 2d(x, \tau, t - 1) + d(x, \tau, t); \\ S(x - r(\tau), \tau - 1, t - 1) \\ \quad + 3d(x, \tau, t); \\ S(x - r(\tau) - r(\tau - 1), \tau - 2, t - 1) \\ \quad + 3d(x - r(\tau), \tau - 1, t) + 3d(x, \tau, t); \end{cases} \quad (3)$$

ただし，$d(x, t)$, $S(x, \tau, t)$ の境界条件は以下で与えられる．

$$d(x, \tau, t) = S(x, \tau, t) = 0. \ (x \leq 0 \text{ or } \tau \leq 0 \text{ or } t \leq 0)$$

mp-CDP の出力は，連続 DP 同様，$S(x, T, t)$ を重み和 $3T_v$ で 0 ～ 1 に正規化した値で求められる．ただし，T_v は，音程時系列の要素が有音と判定されたフレームの個数，つまり $R'(\tau) > \epsilon (\tau = 1, \ldots, T)$ となるフレームの個数である．類似区間の終端集合は，しきい値 α $(0 \leq \alpha \leq 1)$ として，(x, t) の集合:

$$\left\{ (x, t) \mid \frac{1}{3T_v} S(x, T, t) \text{ が} \alpha \text{より大かつ極大} \right\} \quad (4)$$

で求められる．

4 歌詞認識と検索

4.1 システム構成の概要

歌詞抽出は，歌では母音が伸びることが多いこと，また，母音には倍音構造が明確に現れることを積極的に利用し，今回は，母音のみを対象とした認識を行う．もちろん，倍音構造の有無を判定して，母音と子音を判別する方法も考えられる．

楽曲の場合は，mp-CDP で得られた鼻歌の旋律と類似した楽曲区間について，鼻歌の歌詞と類似するものをさらに選別する．このとき，mp-CDP で抽出された類似区間について，mp-CDP が楽曲中たどったパスの軌跡を，ボーカルのピッチ周波数とし，その倍音以外をカットすることで，BGM (Background Music) を抑える．なお，ステレオ録音されている場合は，同位相の部分に主旋律の歌詞が含まれるので，その部分を取り出す前処理を行なうことも考えられる．

歌詞認識では，音素ラベリングされたデータから，通常の音声認識と同様の方法で特徴ベクトルを抽出し学習を行う．A-D 変換された音楽音響信号の特徴ベクトルには，(メル) ケプストラム係数や方向性パターン [6] を用いる．

鼻歌からの音素系列と，楽曲からの音素系列のマッチングには (高速) 連続 DP ([8], [9]) を用いる．

4.2 BGM キャンセル

A-D 変換された音響信号からスペクトル解析と倍音処理 (詳細は [3]) を行い，最も確信度の高い周波数 (これがピッチ周波数となる) を求める．鼻歌については，上の方法でピッチ周波数を求めるが，楽曲からは，mp-CDP が辿ったパスをボーカルのピッチ周波数とする．BGM キャンセルは，ピッチ周波数の倍音およびその近傍のパワースペクトルを残し，それ以外は，ゼロとすることで行われる．

まず，サンプリング周波数 f_s の音響信号に対して，スペクトル解析により，時刻 t, 周波数 f のパワースペクトル $p(t, f)$ とピッチ周波数を f_{max} を求める．中心周波数を f_{max} の i 倍音 $i * f_{max}$ にとり，バンド幅 $W(if_{max}, J\Delta f)$:

$$W(if_{max}, J\Delta f) = [if_{max} - J\Delta f, if_{max} + J\Delta f] \cap [0, \frac{f_s}{2}] \quad (5)$$

を使って，BGM キャンセル処理としてパワースペクトル $p'(t, f)$ を

$$p'(t, f) = \begin{cases} p(t, f) & f \in \bigcup_{i \geq 1} W(if_{max}, J\Delta f) \\ 0 & \text{otherwise} \end{cases}$$

と変換する．

4.3 フレームの音素認識方法

ここでは，音素ごとにカテゴライズされたカテゴリーについて，ベイズ識別関数を使って特徴ベクトル $\boldsymbol{y}$ の識別を行なう．音響信号から抽出される特徴ベクトルには，(メル) ケプストラム係数，Δ,$\Delta\Delta$ (メル) ケプストラム係数，方向性パターンなどが使われるが，以下の実験では，メルケプストラム係数を特徴ベクトルとした．

今回は母音 (a i u e o) のみを対象とし，音素識別のための，それぞれのベイズ識別関数は，学習データから累積寄与率 0.95 で k_ℓ 次元縮約した

$$\mathrm{Bayes}_\ell(\boldsymbol{y}) = \sum_{i=1}^{k_\ell} \frac{[\boldsymbol{\phi}_\ell^{(i)\,t}(\boldsymbol{y}-\boldsymbol{\mu}_\ell)]^2}{\lambda_\ell^{(i)}} + \ln\prod_{i=1}^{k_\ell}\lambda_\ell^{(i)} - 2\ln p_\ell \qquad (6)$$

を用いる．ただし，$\boldsymbol{\mu}_\ell$, $\boldsymbol{\Sigma}_\ell$ は，音素 ℓ の学習サンプルでの平均ベクトルと共分散行列を表し，$\lambda_\ell^{(i)}$, $\boldsymbol{\phi}_\ell^{(i)}$ は $\boldsymbol{\Sigma}_\ell$ の第 i 番目の固有値，固有ベクトルを表し，p_ℓ は，学習サンプル内での音素 ℓ の出現頻度（確率）を表す．

音素識別は，$\mathrm{Bayes}_\ell(\boldsymbol{y})$ $(\ell = 1,\ldots,5)$ が最小となる音素を割り当てることになる．

5 音素認識に関する予備実験

5.1 予備実験（母音認識）の概要

mp-CDP の検索率に関する実験は，橋口ら [4] で行われており，その有効性が示されている．ここでは，歌詞音声検索のための事前準備として音素認識の評価実験を行う．しかし，歌に音素ラベリングされたデータセットがないため，学習サンプル用のラベリングデータを作成することから始める．まずは，ある個人の歌で，学習サンプルと評価用のテストデータを作る．本予備実験の目的は，音素認識率によって BGM キャンセルの効果を調べることにある．

5.2 スペクトル分析とメルケプストラム分析

音楽音響信号はサンプリング周波数 16 kHz を扱い，ピッチ周波数の分析には，ガウス窓によるフィルターをかけた，窓幅 2048 点，シフト幅 1024 点の FFT を用いた．したがって，FFT の 1 フレームの時間は 128 msec，シフト時間は 64 msec である．

音素認識のための特徴ベクトルは，LPC メルケプストラム分析により求めた．自己回帰の窓幅，シフト幅もピッチ周波数を求めたときと同じ，2048 点，シフト幅 1024 点とした．また，自己回帰の次数は 22 次，メルケプストラムの次元（特徴ベクトルの次元）は 19 次元とした．

5.3 学習サンプルと評価用データの作成

歌に音素ラベリングされたデータセットがないため，学習サンプル用データを作成する必要がある．これまでに音素ラベリングされている RWCP の発話データセットなどは，歌ではピッチの変化が大きいことや，比較的母音が長く続くことなど，歌と自然な発話には大きな違いがあるという点で，歌の学習サンプルには不適切である．

そこで，まず，学習用データとして，ある一人に，さまざまなピッチ周波数を含む「あー」を 6 秒発声してもらい，同様に，「いー」，「うー」，「えー」，「おー」を発声してもらった．これらの「あー」，「いー」，「うー」，「えー」，「おー」からメルケプストラム分析により特徴を抽出し学習モデルを作った．

上のように学習サンプルを作成した場合，学習サンプル内の 5 つの音素の頻度には，事前確率 $\{p_\ell\}$ に差をつけるだけの情報がないと考えられる．そこで，認識時には，事前確率 $\{p_\ell\}$ を無視し，(6) 式の識別関数に

$$\mathrm{Bayes}_\ell(\boldsymbol{y}) = \sum_{i=1}^{k_\ell} \frac{[\boldsymbol{\phi}_\ell^{(i)\,t}(\boldsymbol{y}-\boldsymbol{\mu}_\ell)]^2}{\lambda_\ell^{(i)}} + \ln\prod_{i=1}^{k_\ell}\lambda_\ell^{(i)} \qquad (7)$$

を用いる．

BGM キャンセルの効果を調べるための評価用データとして，次の 2 種類を用意した．一つは学習サンプルを構成する学習データに BGM を合成した音響信号と，他方は，同一人物に違う曲調で「あ」，「い」，「う」，「え」，「お」を歌ってもらい，BGM を合成した音響信号である．合成方法は，元の音響信号の音量と BGM の音量を同じにし，BGM 音量の α % を元の信号に合成した．なお，BGM には「ラブマシン」(モーニング娘) のサビの部分を用いた．

- (評価用データ 1): 学習用データ（単母音の孤立発声）に BGM を α % ($\alpha = 0, 10, 20$) 合成した歌データを作った．
- (評価用データ 2): 「上をむいて歩こう」(坂本九) の出だし約 30 秒の主旋律を「あ」だけで，同様に「い」，「う」，「え」，「お」だけでそれぞれ歌ってもらったデータを作成した．評価用デー

表 1: 母音認識率 (%) の平均 (平均認識率)

評価用データ 1 の平均認識率
(母音の孤立発声)

BGM 音量レベル (%)	$\alpha = 0$	$\alpha = 10$	$\alpha = 20$
$J = 0$	52	45	40
$J = 2$	74	72	59
$J = 4$	81	76	54
BGM キャンセルなし	97	83	52

評価用データ 2 の平均認識率 (%)
(「上を向いて歩こう」の母音のみによる発声)

BGM 音量レベル (%)	$\alpha = 0$	$\alpha = 10$	$\alpha = 20$
$J = 0$	41	42	32
$J = 2$	60	56	38
$J = 4$	58	52	34
BGM キャンセルなし	63	48	29

(J は BGM キャンセル処理のバンド幅パラメータ)

タ 1 と同じ BGM を α % ($\alpha = 0, 10, 20$) 合成した.

5.4 予備実験結果と考察

式 (5) で定義される, ピッチ周波数を中心とするバンド幅 $W(if_{max}, J\Delta f)$ において, J を変化させて BGM キャンセル処理の効果を調べた. 例えば, $J = 0$ はピッチ周波数の倍音のみのパワースペクトルだけから, $J = 2$ は $f = if_{max} \pm \Delta f, \pm 2\Delta f$ $(i = 1, 2 \ldots)$ のパワースペクトルだけから特徴ベクトルを求めている. 各 J, α について, 母音 5 音素の認識率を平均した値 (平均認識率) を表 1 にまとめている.

バンド幅を小さく (J を小さく) とると平均認識率が低下する. しかし, BGM レベルが上がっても, BGM キャンセル処理をしない場合に比べて, 平均認識率は急激には低下しない. この点では, BGM キャンセル処理が頑健な方法と思われる. $\alpha = 20$ のときは両評価データともに, $J = 0 \sim 5$ の中で $J = 2$ で平均認識率が一番高かったことを確認している.

6 まとめと今後の課題

鼻歌の音程時系列をクエリーとした音楽音響信号の検索手法である mp-CDP を述べ, さらに歌詞つきの鼻歌について, mp-CDP と歌詞音声検索の統合方法について検討した. しかし, 今回は, 極めて小規模な実験にとどまり, 複数歌い手を対象とした十分な実験が行われていない. 今後は, まず, 歌のようにピッチ変化の激しい音声について, 特徴空間を詳細に調べる必要がある. 認識率の向上と BGM に関する頑健性という観点から, BGM キャンセル処理方法についても検討する必要がある. また, 歌詞検索実験も今後の課題である.

参考文献

[1] 赤坂貴志, 伊原正典, 張建新, 岡隆一 "音声検索における音素フレームラベリングのための音素認識手法" 第 11 回情報統合研究会, pp. 34–39, 1999.

[2] 橋口博樹, 西村拓一, 高橋裕信, 岡隆一. "モデルに依存する傾斜制限型をもつ連続 DP によるハミングを用いた楽曲信号のスポッティング検索", 第 13 回情報統合研究会, pp. 1–4, 2000. (http://www.rwcp.or.jp/lab/mmtl/cii.htm)

[3] 橋口博樹, 西村拓一, 張健新, 高橋裕信, 岡隆一. "モデル依存の傾斜制限型をもつ連続 DP によるハミングを用いた楽曲信号のスポッティング検索" 信学技報, PRMU-2000-66 (2000-09), pp. 35–40, 2000.

[4] 橋口博樹, 西村拓一, 赤坂貴志, 岡隆一. "鼻歌の旋律と歌詞をクエリーとする楽曲信号のスポッティング検索" 信学技報, PRMU-2000-118 (2000-11), pp. 79–86, 2000.

[5] 薗山哲也, 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.

[6] 松村 博, 岡 隆一, 木暮一也, 小島有里江, " スペクトルベクトル場の方向性パターンを用いた不特定話者の単語音声認識, " 電子情報通信学会論文誌, Vol.J72-D-II, No.4, pp.487-498, 1989.

[7] 西原祐一, 梅田昌義, 紺谷精一, 山室雅司, 福本誠, "大規模音楽 DB に対する高速ハミング検索方式" ADBS, Tokyo, pp. 117–124, 1998.

[8] 西村拓一, 関本信博, 張建新, 伊原正典, 赤坂貴志, 高橋裕信, 岡隆一, "高速連続 DP による時系列データの検索", 第 11 回情報統合研究会, pp. 14–15, 1999.

[9] 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.

[10] 園田智也, 後藤真孝, 村岡洋一 (1999). "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.