2006－MUS－67（2）
2006／10／27

音声だけでシームレスに ハミング検索と曲名検索が可能な楽曲検索システム

大石　康智† 　後藤　真孝†† 　伊藤　克亘††† 　武田　一哉†

† 名古屋大学大学院情報科学研究科
†† 産業技術総合研究所
††† 法政大学情報科学部
E-mail: †ohishi@sp.m.is.nagoya-u.ac.jp, kazuya.takeda@nagoya-u.jp,
††m.goto@aist.go.jp, †††itou@k.hosei.ac.jp

あらまし　メロディを歌っても，曲名を読み上げても検索可能な楽曲検索システムを提案する．このシステムは，歌声と曲名の読み上げ音声（朗読音声）を自動識別するため，ユーザはシステムの入力モードを切り替えるのではなく，入力音声の発話様式を切り替えるだけで，シームレスに楽曲を検索することができる．これまでに我々が提案した音声識別器を実装し，歌声と識別されれば，ハミング検索手法によってメロディから曲を検索する．一方，朗読音声と識別されれば，音声認識によって書き起こされた曲名から曲を検索する．大規模な歌声データベースを利用して提案システムの評価実験を行った結果，歌声と朗読音声の自動識別性能は 96.1%であった．さらに，検索キーのハミング検索，音声認識によって 100 曲中 10 位以内に正解の曲が含まれる平均検索率は，それぞれ 50.5%と 96.7%であった．
キーワード　歌声, 朗読音声, 音声識別器, 楽曲検索システム, ハミング検索手法, 音声認識

A Music Retrieval System with a Seamless Query Interface by Humming or Song Title

Yasunori OHISHI†, Masataka GOTO††, Katunobu ITOU†††, and Kazuya TAKEDA†

† Graduate School of Information Science, Nagoya University
†† National Institute of Advanced Industrial Science and Technology (AIST)
††† Faculty of Computer and Information Sciences, Hosei University

Abstract　We propose a music retrieval system that enables a user to retrieve a song by two different methods: by singing its melody or by saying its title. To allow the user to use those methods seamlessly without changing a voice input mode, a method of automatically discriminating between singing and speaking voices is indispensable. We therefore designed an automatic vocal style discriminator and built a music retrieval system that retrieves a song by query-by-humming for singing voice or by dictating the song title by automatic speech recognition (ASR) for speaking voice. Experimental results using a large music database that built for singing research show that our system is able to discriminate between singing and speaking voices with 96.1%. The average retrieval rates of correct songs in the top 10 of 100 songs by query-by-humming and ASR for song titles are 50.5% and 96.7% respectively.
Key words　Singing Voice, Speaking Voice, Automatic Speech Discriminator, Music Retrieval System, Query-By-Humming, Automatic Speech Recognition

1. はじめに

音声・言語コーパスや統計的学習を利用して，言葉の機械処理の基盤が確立されてきた．特に HMM を利用した音声認識では，数万語の語彙を 90%以上の精度で認識することができる [1]．しかし，音声認識システムが扱うことのできる発話の対象は，人間が日常のコミュニケーションに用いる多様な発話様式の中の，ごく一部（読み上げや講演など）に限定されている [2]．対象を拡大するために，多様な環境や発話様式，個人性に対応するための技術が必要である．従来の感情音声の特徴分析は，発話全体に着目するものであり，短時間信号特徴のような発話様式の違いを特徴付ける物理的あるいは信号的性質は明

らかでない [3]．そこで我々は，通常の話し声との違いを聞き分けやすい歌声を研究対象に取り上げ，話し声との違いを多様な観点から分析した．分析結果に基づき，人間の識別能力と同程度の性能で歌声と話し声を識別しうる信号特徴尺度を構築した．2s の音声信号に対して 87.3%の自動識別性能が得られた [4]．

本稿では，歌声と話し声の自動識別手法の応用として，メロディを歌っても，曲名を読み上げても，検索可能な楽曲検索システムを提案する（図 1）．本システムは，歌声と話し声を自動的に識別する技術が組み込まれるため，ユーザは，入力音声の発話様式を切り替えるだけで，シームレスに楽曲を検索することが可能である．これまでメロディの歌唱を検索キーとして音楽を検索するシステム，または書誌情報の音声発話を検索キーとして音楽を検索するシステムは数多く提案されている [5]〜[11]．しかし，システムの入力モードを切り替えることなく，メロディの歌唱と書誌情報の読み上げ音声をどちらも入力手段として利用可能な音楽検索システムはこれまで提案されていない．

また，本稿ではメロディの歌唱を検索キーとして曲を検索するハミング検索手法についての検討も行う．従来のようにメロディを離散化して記号ベースまたはパターンベースで検索を行うのではなく，フレーズや繰り返し構造のようなメロディの時間構造を特徴抽出し，その特徴ベクトルに基づいて曲を検索する方法を提案する．最後に，大規模な歌声データベースを利用して，提案する楽曲検索システムの評価実験を行う．

以下，2 章ではハミング検索と曲名検索が可能な楽曲検索システムの概要と処理の流れについて述べる．次に 3 章では，本システムにおいて代表的な 3 つのモジュール「音声識別器」「ハミング検索器」「音声認識器」の実装について述べる．4 章では評価実験を行い，各モジュールの動作と性能を確認し，5 章ではまとめを行う．

2. システムの概要

本稿で提案する楽曲検索システムは，歌声（ここで,「歌声」はユーザが音高の変化を伴って自然に発するハミングや歌詞付きの歌を含むものとする）と，曲名の読み上げ音声（以後，これらの音声を朗読音声と呼ぶ）で曲を検索することができる．図 2 の流れ図に示すように，入力音声は，まず音声識別器によって歌声であるのか，朗読音声であるのか識別される．入力音声が歌声であれば，ハミング検索手法にしたがって，メロディに基づいて曲が検索される．入力音声が朗読音声であれば，音声認識手法にしたがって，認識された曲名に基づいて曲が検索される．以下，各モジュールについて紹介する．

2.1 音声識別器

歌声と朗読音声の自動識別を実現するために，我々は，聴取実験を行うことにより，人間の識別能力を調査した．その結果，人間は，1s 程度の音声信号の聴取で，歌声と朗読音声を 100%識別可能であること，また，音声信号の短時間のスペクトル特徴，基本周波数（以後，F0 と呼ぶ）の時間変化を相補的に知覚して識別していることが明らかとなった．そこで，短時間のスペクトル特徴を表現する尺度として MFCC(Mel Frequency Cepstrum Coefficients) を，F0 の時間変化を表現する尺度と

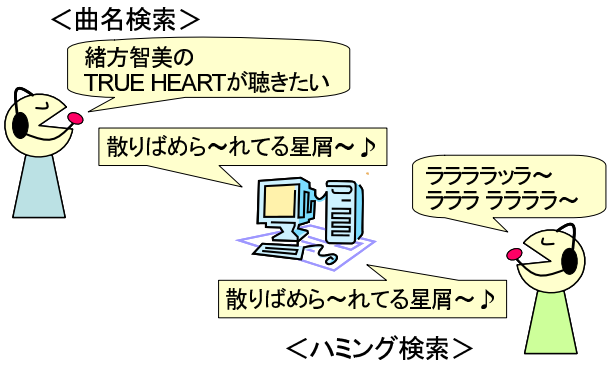

図 1　ハミング検索と曲名検索が可能な楽曲検索システムの概要

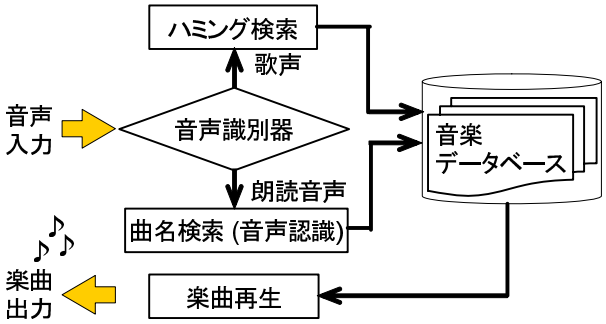

図 2　楽曲検索システムの流れ図

して ΔF0（F0 の時系列に対する線形回帰係数）を利用して特徴ベクトルを構成した．歌声と朗読音声から算出されるそれぞれの特徴ベクトルの分布を学習させた 2 つの混合ガウス分布（GMM）を用いて，自動識別実験を行ったところ，2s の音声信号に対して 87.3%の識別性能を実現した [4]．

2.2 ハミング検索器

歌声と識別された入力音声を検索キーとして，そのメロディを持つ曲を検索する．既に，蔭山ら [5]，園田ら [6]，西原ら [7] は，ハミングから曲名を検索する手法を提案している．これらでは，あらかじめ多数の曲のメロディデータを用意し，そのメロディの相対音高差，相対音長差を量子化してデータベースを作成した．ユーザのハミングが入力されると，メロディを抽出し，同様に量子化して検索キーを作成する．検索キーとデータベースとのマッチングは，この量子化された記号列同士で行われる．このとき，連続 DP を用いて時間伸縮や音符の挿入削除によるノイズに対処する方法も提案されている [6]．また，西村ら [9] は，音楽音響信号から抽出したメロディと，検索キーから抽出した音高時系列との類似した時間的な音高変化パターンを連続 DP によって求めるパターンベースの検索手法を提案した．

以上の検索手法はすべて，メロディを離散化して，記号列またはパターン同士のマッチングから類似区間を検出するものであった．したがって，連続したメロディのフレーズ構造，拍節構造，またそれらの繰り返しの構造のような，音楽を理解する上で重要なメロディの時間構造を検索の手がかりとして利用していない．そこで，本稿ではメロディを離散化するのではなく，メロディの時間構造を特徴抽出するために，異なる時点におけるメロディの時系列間の相関関係を検索に利用する手法を

提案する．この手法により，ユーザにとって曲がうろ覚えの状態であるため，音高が外れたり，調，テンポが原曲と異なる検索キーであっても，メロディの局所的・大局的なフレーズやその繰り返しの構造との類似性から曲が検索できることを目指す．

2.3 音声認識器

朗読音声と識別された入力音声を音声認識し，言語情報を書き起こす．「< アーティスト > の < 曲名 > を聞かせてください」「< 曲名 > が聞きたい」というように複数の認識用文法を作成し，認識された < 曲名 > に基づいて，曲を検索する．

3. 実　　装

提案する楽曲検索システムの実装について述べる．このシステムは，音声識別器，ハミング検索器，音声認識器の 3 つのモジュールからなる．以下，それぞれの処理について詳説する．

3.1 音声識別器

歌声と朗読音声の識別尺度として，短時間のスペクトル特徴と F0 の時間変化を利用する [4]．

3.1.1 短時間スペクトルに基づく尺度

MFCC とその時間変化 ΔMFCC を利用する．標本化周波数 16kHz，窓幅 32ms のハミング窓，フレームシフト 10ms，メルフィルタバンク数 24 個の分析条件のもとで計算される MFCC12 次までの係数を利用した．ΔMFCC は，5 つのフレーム（50ms）にわたって計算される回帰係数とした．

3.1.2 F0 の時間変化に基づく尺度

F0 は，後藤らの提案した有声休止検出のための F0 推定手法 [12] を利用して 10ms ごとに推定される．次に式（1）のように Hz で与えられる周波数の単位を cent に変換した．

$$f_{\mathrm{cent}} = 1200 \log_2 \frac{f_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}} \qquad (1)$$

ΔF0もΔMFCC と同様に，連続して推定された 5 つのフレーム（50ms）の F0 に対して計算される回帰係数とした．

3.1.3 識別モデルの学習

MFCC と ΔMFCC，ΔF0 から構成される特徴ベクトル（25 次元ベクトル）の分布が 16 混合 GMM によって学習される．歌声から抽出された特徴ベクトルで学習した歌声 GMM $\Lambda_{\text{歌声}}$ と朗読音声から抽出された特徴ベクトルで学習した朗読音声 GMM $\Lambda_{\text{朗読音声}}$，$\mathbf{X} = \{x_t | t = 1, \ldots, T\}$ を評価データから抽出された特徴ベクトル系列とすると，識別結果は以下の式（2）で計算される平均対数事後確率によって決定される．

$$\hat{d} = \underset{d=\text{歌声, 朗読音声}}{\operatorname{argmax}} \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{x}_t; \Lambda_d) \qquad (2)$$

3.2 ハミング検索器

検索キーのメロディ（F0）の時間構造を特徴抽出して検索に利用する手法を提案する．検索キーの F0 は，音声識別器において後藤らの手法により 10ms ごとに推定されたものを利用する．無音区間処理として，無声音のため F0 が推定されない箇所は，直前の推定されている F0 の値で置き換えた．楽曲データベースからもメロディを推定することが望ましいが，今回はメロディから曲を検索する手法に焦点を当てるため，あらかじめ正確なメロディデータが用意されていることを前提とする．

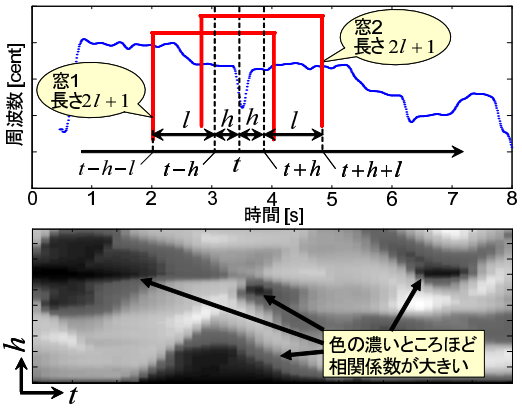

図 3　メロディの時間構造の特徴抽出：時刻 t において h だけ前後にシフトした長さ $2l+1$ の 2 つの窓に含まれる時系列間の相関係数を計算する．

3.2.1 メロディの時間構造の特徴抽出

メロディの時間構造を特徴抽出するために，時刻 t において h だけ前後にシフトさせた長さ $2l+1$ の 2 つの窓を考える（図 3）．各窓の時系列間の相関係数 $r_{t,h,l}$ を以下で定義する．

$$r_{t,h,l} = \frac{\sum_{\tau=-l}^{l} (x_{t-h+\tau} - \bar{x}_{t-h})(x_{t+h+\tau} - \bar{x}_{t+h})}{\sqrt{\sum_{\tau=-l}^{l} (x_{t-h+\tau} - \bar{x}_{x-h})^2 \sum_{\tau=-l}^{l} (x_{t+h+\tau} - \bar{x}_{t+h})^2}}$$

$$\bar{x}_{t-h} = \sum_{\tau=-l}^{l} x_{t-h+\tau}/(2l+1), \quad \bar{x}_{t+h} = \sum_{\tau=-l}^{l} x_{t+h+\tau}/(2l+1) \qquad (3)$$

ここで，x_t は時刻 t におけるメロディ（単位は cent）を示す．図 3 の下図は，上図のメロディに対して，各時刻の相関係数を式（3）にしたがって計算したものである．l は 1s として，縦軸 h は 20ms〜2s まで 20ms ずつ変化させた．色の濃い箇所（相関が高い）は，$2h$ だけ離れた 2 つの窓内のメロディの概形（音高・音長の変化のしかた）が似ていることを示す．すなわち離れた時点でのメロディ間の相関性が時間的にどのように移り変わるかが抽出される．以上の相関係数の系列を検索キーと楽曲データベースに対して算出し，それぞれ $\mathbf{q}_{t_1,l}$ と $\mathbf{d}_{t_2,l}$ の特徴ベクトル系列（t_1, t_2 は時刻を表す）として表現する．

3.2.2 検 索 方 法

検索キーから算出される特徴ベクトル系列 $\mathbf{Q} = \{\mathbf{q}_{t_1,l} | t_1 = 1, \ldots, T_1\}$ と楽曲データベースから算出される特徴ベクトル系列 $\mathbf{D} = \{\mathbf{d}_{t_2,l} | t_2 = 1, \ldots, T_2\}$ との局所類似度を計算する．図 4 に示すように特徴ベクトル $\mathbf{q}_{t_1,l}$ と $\mathbf{d}_{t_2,l}$ の局所類似度は，式（4）のコサイン距離で計算される．

$$s(t_1, t_2, l) = \frac{\langle \mathbf{q}_{t_1,l} \cdot \mathbf{d}_{t_2,l} \rangle}{||\mathbf{q}_{t_1,l}|| ||\mathbf{d}_{t_2,l}||} \qquad (4)$$

さらに検索キーが楽曲データベースのどの部分を歌ったものであるか特定するために，図 5 に示すように局所類似度の対角要素を足し合わせる．単純に足し合わせるのではなく，時間伸縮を考慮した整合窓を適用し，以下の漸化式に基づいて，時刻 t_2

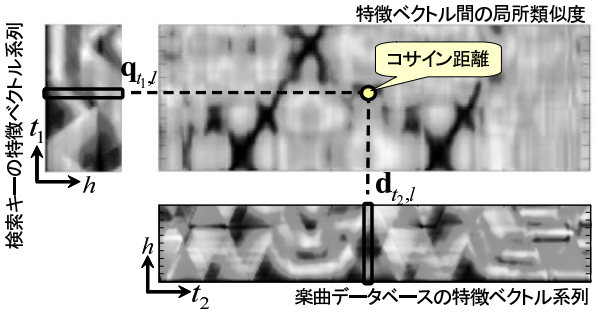

図 4　特徴ベクトル間の局所類似度の算出：$\mathbf{q}_{t_1,l}$ と $\mathbf{d}_{t_2,l}$ のコサイン距離を類似度とする．色の濃い部分ほど類似度が大きい．

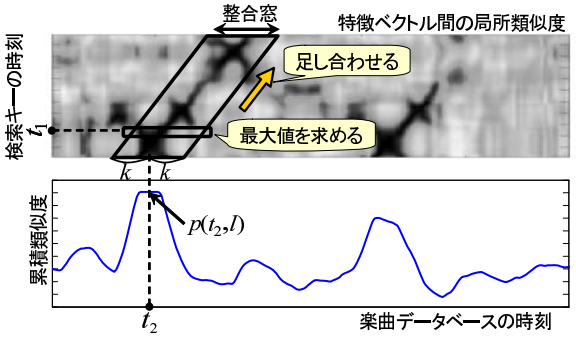

図 5　検索キーと楽曲データベースとの類似部分の探索：整合窓内の局所類似度の対角要素を足し合わせ，累積類似度を計算する．

における累積類似度 $p(t_2, l)$ を求める．整合窓は時刻 t_2 に対して前後 k の範囲で定義する．

漸化式（$1 \leqq t_2 \leqq T_2 - T_1$）：

$$t_3 = \underset{t_2-k+(t_1-1) \leqq t \leqq t_2+k+(t_1-1)}{\operatorname{argmax}} s(t_1, t, l) \tag{5}$$

$$p(t_2, l) \Leftarrow p(t_2, l) + s(t_1, t_3, l)$$

$$(1 \leqq t_1 \leqq T_1)$$

さらに窓幅を決定するパラメータ l を変化させて式（4）と式（5）の漸化式を繰り返し計算する．l を変化させることによって，メロディの短時間・長時間の時間構造を特徴抽出するためである．最終的に各 l に関して求められた累積類似度 $p(t_2, l)$ をすべて足し合わせ，累積類似度 $P(t_2)$ を算出する．

$$P(t_2) = \sum_{l} p(t_2, l) \tag{6}$$

$P(t_2)$ の最大値を検出することによって楽曲データベースにおいて検索キーと最も類似している部分が求められる．類似部分を含んでいる曲を上位 N 位まで提示する．

3.3　音声認識器

音声認識エンジンは，記述文法音声認識実行キット Julian-3.4.2 を利用した．Julian は，有限状態文法 (DFA) に基づいて，与えられた文法規則のもとで入力音声に対して最尤の単語系列を探し出す音声認識エンジンである．以下の音響モデルと言語モデルを利用して音声認識を行い，認識された曲名を提示する．

3.3.1　音響モデル

CSRC の標準日本語音響モデル [13] より，状態数 3000/129，性別非依存，64 混合，PTM triphone モデルを用いた．

3.3.2　言語モデル

言語モデルとして，図 6 に示す文のパターンを記述した認識用文法を作成した．ここで，「eps」はヌル遷移，「silB」は文頭，「silE」は文末を表すシンボルである．文法中の <Artist> はアーティスト名，<Music> は曲名を表すシンボルである．

3.4　音楽用通信プロトコル RMCP の利用

本システムを構成する音響信号の入出力，F0 の推定，音声識別器，ハミング検索器，音声認識器などの各モジュール間の通信は，音楽情報をネットワーク上で効率よく共有することを可能にする通信プロトコル RMCP(Remote Media Control

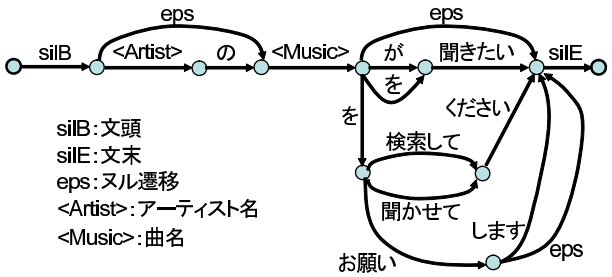

図 6　音声認識における認識用文法

Protocol) [14] を利用して実装した．これらのモジュールは，ネットワーク (LAN) 上の複数の計算機で分散して実行でき，リアルタイムに連携しながら処理が可能である。

4.　評 価 実 験

前章で提案した楽曲検索システムの各モジュールの動作と性能を確認するために評価実験を行う．

4.1　使用データ

4.1.1　楽曲データベース

「RWC 研究用音楽データベース：ポピュラー音楽」(RWC-MDB-P-2001) [15] の 100 曲を用意した．本システムのプロトタイプは，この 100 曲の中から曲を検索するシステムとする．

4.1.2　歌声データベース

AIST ハミングデータベース [16] の一部である，日本人歌唱者 75 名分（男性 37 名, 女性 38 名）の音声データを楽曲検索システムへの検索キーとして利用する．特別な歌唱訓練を受けていない一般的な歌唱者が，4.1.1 節の楽曲データベースから抜粋した合計 25 曲の歌唱の出だしの部分と一番代表的な盛り上がる主題の部分の二カ所を，うろ覚えの状態で歌詞付きで歌った歌声，ラーラーラーのような任意の音で歌詞は付けずにメロディを口ずさんだハミング，当該部分の歌詞を読み上げた朗読音声を用いた．1 名あたり計 150 サンプル（歌声：50 サンプル，ハミング：50 サンプル，朗読音声：50 サンプル）となる．音声サンプルの長さの平均は歌声で 12.0 秒，ハミングで 11.7 秒，朗読音声で 7.0 秒であった．

4.1.3　メロディデータベース

本稿で提案するハミング検索器は，検索キーから抽出される F0 とのマッチングのために，楽曲データベースのメロディが必

表 1　歌声，ハミング，朗読音声を検索キーとしたときの音声識別性能：9 割を超える識別性能が実現された．

	歌声	ハミング	朗読音声
識別率	96.2%	98.0%	94.2%

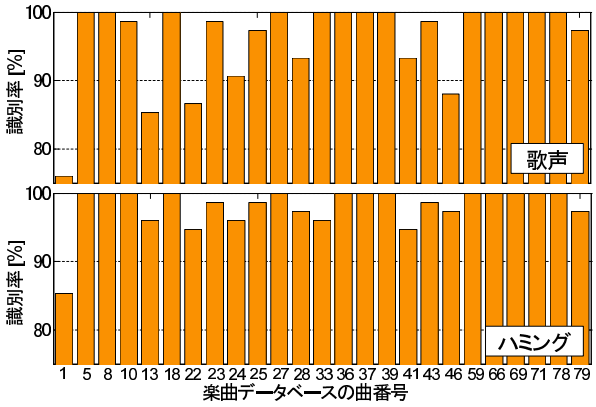

図 7　同一曲の同一箇所を 75 名が歌った検索キーごとの識別性能：曲によっては，ハミングで歌うことにより，識別性能が向上する．

要である．実用目的では、楽曲の音響信号からメロディを推定するのが望ましいが，今回は性能の上限を調べるために，4.1.1 節の楽曲データベースのメロディの F0 を手作業でラベル付けした結果 [17] を用いた．

4.1.4　曲名読み上げ音声の収録

曲名検索における音声認識性能を評価するために，被験者 6 名（男性 4 名，女性 2 名）に 4.1.1 節の楽曲データベース 100 曲の中から 10 曲を選ばせ，図 6 の認識用文法に基づいて発話させた曲名読み上げ音声を計 60 サンプル収録した．

4.2　音声識別性能の評価

4.2.1　実 験 条 件

歌声データベースにおける歌声，ハミング，朗読音声（計 11250 サンプル）を利用して，音声識別性能を評価する（歌声もしくはハミングと，朗読音声との 2 クラス識別．ここで朗読音声は，歌声データベースにおける歌詞の読み上げ音声である．曲名の読み上げ音声に対する識別性能は 4.4 節で議論する）．大石らの評価実験 [4] では，発声開始から 2s 間の音声サンプルに対する識別性能を評価していたが，ここでは，実際のシステムの利用時を考慮して，発声開始から終了までの入力音声に対する識別性能を評価する．11250 サンプルのうち，曲の出だしの部分の歌声・朗読音声（計 3750 サンプル）を利用して，歌声 GMM と朗読音声 GMM を学習し，曲の主題の部分の歌声・ハミング・朗読音声（計 5625 サンプル）を評価データとした．歌唱者に対してオープンデータで学習と評価を行うために，歌唱者 75 名を均等に 5 つのグループに分け，5 回のクロスバリデーションを行い，その平均値を識別率として算出する．

4.2.2　実 験 結 果

表 1 より，すべての音声に対して 9 割を超える識別性能が得られた．特に，歌声よりもハミングを検索キーとした方が，識別性能が高い．図 7 は，歌声とハミングによる識別性能を検索キーの曲ごとに示す．1，13，22，24，25，28，41，48 番はハミングを検索キーとした方が識別率が高く，59.6%識別誤りが改善された．曲によっては，歌詞付きで歌うよりも，ハミングで歌う方が，“歌声” と識別されやすいことが確認された．

表 2　2 つのハミング検索手法の性能比較：100 曲中上位 N 位以内（1，5，10 位以内）に正解の曲が含まれる平均検索率を示す．

	西村らによる手法 [9]		提案手法	
検索率	歌声	ハミング	歌声	ハミング
$N=1$	29.8%	29.9%	29.3%	28.5%
$N=5$	42.5%	44.4%	41.9%	41.4%
$N=10$	50.8%	52.1%	50.5%	49.3%

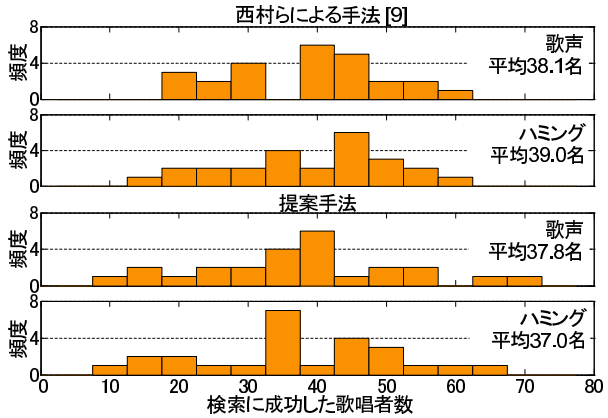

図 8　曲ごとの検索成功者数の分布：1 曲あたり 75 名による検索キーを評価する．10 位以内に正解の曲が含まれたら検索成功とする．

4.3　ハミング検索による楽曲検索性能の評価

4.3.1　実 験 条 件

歌声データベースにおける曲の主題の部分の歌声，ハミング（計 3750 サンプル）を検索キーとして，3.2 節で提案したハミング検索手法を評価する．パラメータ h を 20ms から 1s まで 20ms ずつ変化させた 50 次元の特徴ベクトル $\mathbf{q}_{t_1,l}$，$\mathbf{d}_{t_2,l}$ を 100ms ごとに求める．また，窓幅を決定するパラメータ l は 25ms から 150ms まで 25ms ずつ 6 段階に変化させた．整合窓のパラメータ k は 300ms とした．これらのパラメータは，すべて実験的に決定した．従来法として西村らの提案した始端特徴依存連続 DP を用いたハミング検索手法 [9] を実装し，検索性能を比較する．

4.3.2　実 験 結 果

表 2 は 2 つの検索手法による性能を比較したものである．利用した歌声データベースが，うろ覚えの状態で収録された歌声・ハミングであるため，全体的に検索性能は低い．検索に失敗したサンプルを聞いたところ，原曲とは大きく逸れたメロディを歌っている場合が多くみられた．以後，10 位以内に正解の曲が含まれる検索率をシステムの性能として議論をすすめる．提案手法の検索率は，数%ではあるが従来法に比べて劣っている．図 8 の頻度分布より，従来法の方が曲ごとにみた検索成功者数の平均値は大きく，その分散も提案手法に比べて小さい．すなわち，提案手法では検索に成功しやすい曲もあれば，検索できない曲も多くあることがわかる．また，図 9 の分布においても，

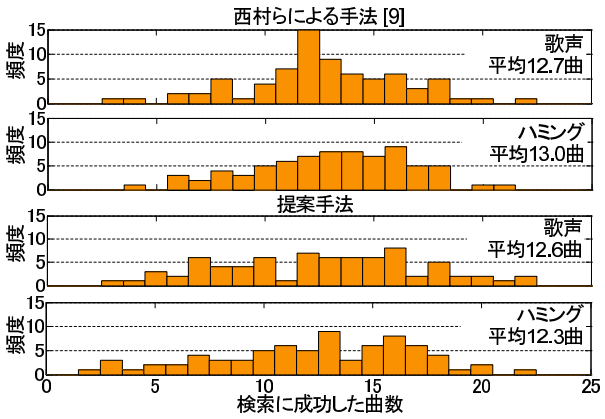

図 9　歌唱者ごとの検索に成功した曲数の分布：歌唱者あたり 25 曲の検索キーを評価する．10 位以内を検索成功とした．

表 3　曲名読み上げ音声を検索キーとしたときの性能評価

	音声識別性能	楽曲検索性能
識別率・検索率	100%	96.7%

以下の単語を認識誤りしたため，楽曲検索性能が識別性能に比べて低下した．小動物 (曲名) → So Long(曲名)，Cool Emotion(曲名) → Game of Love(曲名)

提案手法の方が分布の分散が大きい．すなわち，提案手法では，約 8 割検索に成功する歌唱者もいれば，全く検索に成功しない歌唱者もいるという偏りのある検索手法であることがわかる．

4.4　音声認識による楽曲検索性能の評価

4.4.1　実 験 条 件

4.1.4 節にしたがって収録した曲名読み上げ音声（計 60 サンプル）を検索キーとして，音声認識による楽曲検索性能を評価する．音声認識エンジンの言語モデルの辞書の語彙サイズは 142 単語（うちアーティスト 33 名，曲数 100 曲）である．

4.4.2　実 験 結 果

表 3 は，検索キーの音声識別性能と楽曲検索性能を示す．2 つの単語認識誤りのため，検索性能は 96.7%となった．さらに評価データを収録して評価実験を行うこと，また検索対象となる曲数を増やし，アーティスト名のようなその他の書誌情報からも曲が検索できるよう改良をすすめることが必要である．

5.　まとめと今後の展開

本稿では，歌声でも，曲名を読み上げた朗読音声でも検索可能な楽曲検索システムを提案した．歌声と朗読音声を自動識別し，歌声であれば，ハミング検索手法によりメロディから曲を検索する．朗読音声であれば，音声認識により書き起こされた曲名から曲を検索するシステムである．また，新しいハミング検索手法についても検討を行った．従来のように記号・パターンベースでマッチングを行うのではなく，自己相関関数を利用してメロディの時間構造（フレーズや繰り返しの構造）を特徴抽出し，検索に利用する手法を提案した．大規模な歌声データベースを利用して評価実験を行ったところ，十分な音声識別性能・音声認識性能は得られたものの，提案したハミング検索手法の性能は従来法に比べて低く，曲や歌唱者によって検索率のばらつきが大きい偏りのある検索手法であった．したがって，単純にメロディの時系列間の自己相関係数を計算するだけでは，時間構造をとらえきれないことが明らかとなった．歌声は，歌唱者の意図に基づいて原曲のメロディからどのように変化したものであるのか，複数の歌唱者による歌声からどのように本来の原曲のメロディを推定すればよいかという問題を考えることにより，多様な曲，歌の技術の個人差にも対応できる検索手法を検討し，従来の性能を向上させることが今後の課題である．

文　　献

[1] 李　晃伸，河原達也，武田一哉，鹿野清宏:phonetic tied-mixture モデルを用いた大語彙連続音声認識，電子情報通信学会論文誌 D-II，Vol. J83-D-II, No. 12, pp. 2517–2525 (2000).

[2] Furui, S., Nakamura, M., Ichiba, T. and Iwano, K.: Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese, *Speech Communication*, Vol. 47, pp. 208–219 (2005).

[3] Nicholson, J., Takahashi, K. and Nakatsu, R.: Emotion Recognition in Speech Using Neural Networks, *Neural Computing and Applications*, Vol. 9, No. 4, pp. 290–296 (2000).

[4] 大石康智，後藤真孝，伊藤克亘，武田一哉：スペクトル包絡と基本周波数の時間変化を利用した歌声と朗読音声の識別，情報処理学会論文誌，Vol. 47, No. 6, pp. 1822–1830 (2006).

[5] 蔭山哲也，高島洋典：ハミング歌唱を手がかりとするメロディ検索，電子情報通信学会論文誌 D-II，Vol. J77-D-II, No. 8, pp. 1543–1551 (1994).

[6] 園田智也，後藤真孝，村岡洋一：WWW 上での歌声による曲検索システム，電子情報通信学会論文誌 D-II，Vol. J82-D-II, No. 4, pp. 721–731 (1999).

[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*, pp. 117–124 (1998).

[8] 橋口博樹，西村拓一，張　建新，滝田順子，岡　隆一：モデル依存傾斜制限型の連続 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，電子情報通信学会論文誌 D-II，Vol. J84-D-II, No. 12, pp. 2479–2488 (2001).

[9] 西村拓一，橋口　博，後藤真孝，岡　隆一：時系列パターンマッチング手法を用いた鼻歌による音楽信号からの高速検索- 始点依存連続 DP および始点依存 RIFCDP の提案-，情処研報音楽情報科学，Vol. 2001, No. 103, pp. 57–62 (2001).

[10] 後藤真孝，伊藤克亘，秋葉友良，速水　悟：音声補完: 音声入力インタフェースへの新しいモダリティの導入，コンピュータソフトウェア (日本ソフトウェア科学会論文誌)，Vol. 19, No. 4, pp. 10–21 (2002).

[11] 原　直，白勢彩子，宮島千代美，伊藤克亘，武田一哉：音声対話による楽曲検索システム，情処研報音声言語情報処理，Vol. 2004, No. 103, pp. 31–36 (2004).

[12] 後藤真孝，伊藤克亘，速水　悟：自然発話中の有声休止箇所のリアルタイム検出システム，電子情報通信学会論文誌，Vol. J83-D-II, No. 11, pp. 2330–2340 (2000).

[13] 河原達也ほか：日本語ディクテーション基本ソフトウェア (98 年度版)，日本音響学会誌，Vol. 56, No. 4, pp. 255–259 (2000).

[14] 後藤真孝，根山　亮，村岡洋一：RMCP: 遠隔音楽制御用プロトコルを中心とした音楽情報処理，情報処理学会論文誌，Vol. 40, No. 3, pp. 1335–1345 (1999).

[15] 後藤真孝，橋口博樹，西村拓一，岡　隆一：RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol. 45, No. 3, pp. 728–738 (2004).

[16] 後藤真孝，西村拓一：AIST ハミングデータベース: 歌声研究用音楽データベース，情報処理学会 音楽情報科学研究会研究報告，Vol. 2005, No. 82, pp. 7–12 (2005).

[17] Goto, M.: AIST Annotation for the RWC Music Database, *ISMIR 2006* (2006).