電話を用いたハミング検索システム

小杉尚子[†] 櫻井保志[††] 森本正志[†]

本論文では，電話を用いたハミング検索システムについて述べる．電話を用いたハミング検索で問題になるのは，電話回線を経由して収音したハミングデータの音声信号処理方法と，電話機を通して検索結果を知らせるための検索結果出力方法である．これらはそれぞれ，電話回線を経由することによる音声信号の帯域制限と，電話機のインタフェースとしての自由度の低さに起因する．本論文ではこれらの問題に対して，限られた周波数帯域の高調波成分を用いた発声音高の特定手法，および電話機を通してユーザに検索結果を提供するために，ハミングされた部分を正確に特定して再生する手法を提案する．またメロディデータの中に他の楽曲とよく似た音符の並びがある場合は正解曲の特定が困難になるので，それらに影響されない検索結果の調整方式を提案する．各提案手法は，我々が研究開発中のハミング検索システムである SoundCompass に実現し，定量的な評価を通してそれぞれが電話を用いたハミング検索サービスに対して有効であることを確認する．

Query-by-Humming System on the Phone

Naoko Kosugi,[†] Yasushi Sakurai[††] and Masashi Morimoto[†]

This paper describes a query-by-humming system that works over the phone. It describes the voice signal processing for hummings over the phone and how results are sent back to the receiver. The bandwidth limitation of telephone lines and inflexibility of the receivers as an interface are problems. For these problems, we propose a method to track the pitches of utterances from the voice signals of limited frequencies and a method to provide retrieval results for telephone receivers. In addition, there are similar note patterns that appear in many melodies, which make it difficult for the system to identify the correct song. Thus, a method to adjust the final result so that it will be without such influences is also proposed. All techniques have been implemented in our query-by-humming system, called SoundCompass. Quantitative evaluations with SoundCompass confirmed their effectiveness for query-by-humming on the phone.

1. はじめに

メロディの一部を歌唱してその楽曲を検索することをハミング検索という．ハミング検索は音楽内容検索の一種で，デジタル化された音楽データの利用が急速に増加する中で，より便利な利用法を実現するための重要な研究課題である．本課題に対し，我々は “SoundCompass” というハミング検索システムの研究開発を推進してきた[1)~3)]．SoundCompass は，2000 年 12 月から都内のカラオケボックスにおいて選曲システムの一部として導入された[4)]．当時のシステムのハミング入力のインタフェースはパソコンにつながれたマイクだった．

日本国内においてパソコンやインターネットの利用者は著しく増加しているが，声や音を入出力とするサービスを考えたとき，誰にとっても想像しやすい身近なインタフェースはやはり電話であろう．パソコンにつながれたマイクとインターネットよりも，電話機と電話回線のほうが多くの人にとって身近で，たくさんのユーザにサービス利用の機会を提供できると考えられる．実際に，携帯電話などを使って町なかに流れる楽曲を収音し，検索してユーザに曲名情報などを提供するサービスがイギリスやスペインなどで始まっている[5)~7)]．このような事情から，ハミングによる楽曲の検索サービスを電話を用いて提供することは，より多くの幅広いユーザの獲得や利用機会の拡大には必要不可欠である．弊社でも電話回線を利用したハミング検索サービスの必要性を認識しており，SoundCompass が電話を介して利用できるようになることが期待されている．

しかし，電話機・電話回線を用いてハミング検索サービスを提供する場合には，以下の課題を解決する必要

† NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories
†† NTT サイバースペース研究所
NTT Cyber Space Laboratories

がある．

課題 1 電話回線の周波数帯域制限に対応した音声信号処理方式の確立

電話では音声信号の圧縮により，復号可能な周波数帯域は約 300 Hz から 3.4 kHz に制限される．したがって，電話回線を通してハミングを収音した場合，発声の音高を確定するために必要な基本周波数の情報の多くが欠落すると同時に，基本周波数をより正確に推定するために必要な倍音の高周波数成分も欠落する．限られた周波数情報から正確に発声の基本周波数を同定することができる音声信号処理方式が必要である．

課題 2 自由度の低いインタフェースを有効に活用した情報の取得/提供方式の確立

パソコンと違って電話機には大きくて見やすい画面がないので，ユーザに対してサービス利用のための魅力的で分かりやすい GUI を使ったガイドなどをすることはできない．カラオケボックスで使用された SoundCompass では，ハミングを収録する際にはメトロノームを提供し，そのテンポ値を用いてハミングを時間正規化したり，ユーザが一定のテンポで歌えるようにガイドする機能を備えていた．また検索結果として，ハミングに似ている部分を持つ複数の曲をそれぞれ似ている順に点数をつけてリスト形式でユーザに提示する機能も備えていた．しかし電話機をインタフェースに用いた場合は，これらの機能をユーザに提供することはできない．したがって，メトロノームに依存しない音楽データの検索手法や，検索結果の新しい提供方式が必要である．

課題 3 検索精度の向上

上記 2 点に述べたように，ハミング検索サービスに電話機・電話回線を使用すると，パソコンとインターネットでサービスをする場合に比べて，扱う音声信号の品質の劣化や，入出力を担う機器類の自由度の低下は避けられない．したがって，実用可能なシステムとするには今まで以上に高い検索精度を達成し，上記の問題を補償するための新しい技術が必要である．

課題 4 電話機・電話回線を用いたハミング検索システムに対する検討

電話機・電話回線を用いたハミング検索については，これまでまったく報告されていない．電話機・電話回線を介した場合，従来技術による性能がどのように変わるのか，ハミング検索にどのような問題点があるのか，これらを明らかにする必要がある．

本論文の目的は，これらの課題を以下に示す基本方針に則って解決することにある．

基本方針 1 音声信号処理

ハミングの収録において，定常的に混入する雑音を除去し，また限られた帯域内の倍音構造を効果的に利用することで，マイク収音とほぼ同等の精度で，発声のタイミングと音高情報を抽出する（4.2 節）．

基本方針 2 入力/出力情報

入力情報として電話機から得られるのは，ユーザのハミングだけである．そこで，メトロノームに依存せずに音楽データを検索するため，文献 3), 8) の音楽データの自動時間正規化手法を新システムに実装する．出力に関しては，1 位に検索された曲を再生することでユーザに検索結果を提供する．その際，ユーザがハミングした部分を正確に特定して再生する（4.4 節）．検索結果を曲名のリストで提供する場合は，ユーザはその曲名リストとあわせて検索された楽曲を試聴するので，再生される部分がユーザがハミングした部分と必ずしも正確に一致しなくても，検索結果の正解/不正解を容易に確認することができた．しかし今回は検索された楽曲を聞くことでしか検索結果の正解/不正解を確認できないので，ユーザがハミングした部分はできるだけ正確に特定されて再生されなければならない．

基本方針 3 検索結果の調整

今まで蓄積してきた SoundCompass の検索結果を詳細に調査する中で，正解曲との距離が比較的近いにもかかわらず 1 位に検索されないケースがあることが分かった．そこで，一度取得した検索結果の中身を調べて，それに応じて最終結果へのマージ方法を変えることで検索精度を向上させる方式を導入する（4.3 節）．

基本方針 4 実システムの構築

実際に電話機・電話回線を使用したハミング検索を行うための環境を構築し（4.1 節），その環境でハミング検索を行う際の問題点を検証する．さらに，新システムの性能を多角的，定量的に評価したうえで，新技術がどのように問題点を解決したのかについて議論する（5 章）．

本論文の構成は以下のとおりである．2 章で関連研究について述べた後，最初に 3 章で SoundCompass における従来技術について説明する．4 章では電話を用いたハミング検索システムについて述べる．5 章

では本論文で提案する手法を実装した新しい SoundCompass を用いて定量的に評価し効果を議論する．最後に 6 章で本論文をまとめる．

2. 関連研究

日本はハミングを用いた楽曲検索システムの研究においては先駆的な役割を果たしてきた[9]．また，1998 年以降には数々の新技術が報告されてきた[10)~13)]．これらの文献においては，それぞれがそれぞれの問題意識を持って課題に取り組み重要な成果をあげてきた．Sonoda らは，特に WWW ベースのハミング検索システムの構築を目指して[14),15)]，WWW 上にデモシステムを公開している[16)]．また西村らは，より自然で柔軟な歌唱の入力を許すシステムの構築に力を入れており，DP マッチングを改良することで柔軟でかつ高速な歌唱を用いた音楽検索システムを提案している[17)~19)]．我々は，大規模なデータベースに対して精度高く高速に検索できるハミング検索システムの実現を目指して，多次元空間インデクスを使用した高速類似検索方式のための，効果的な特徴量や検索方式などを提案し，それらの技術を SoundCompass を用いて定量的に評価したり[1),2),13),20)~23)]，実用化を行うなどしている[24)~26)]．

3. SoundCompass システム

3.1 SoundCompass の処理の流れ

図 1 に SoundCompass[3)] の処理の流れを簡単に示す．左側にデータベースの構築処理の流れを示す．まず最初に，楽曲データの時間正規化と部分データ（子/孫シーケンス[3)]）への分割を行う（3.3 節）．部分データへの分割は，スライディング・ウィンドー方式[27)] を用いる．次に各部分データから特徴ベクトル（3.4 節）を作成する．特徴ベクトルは特徴量ごとにインデクス化する．

図 1 の右側に検索の処理を示す．集録した音声に対して各発声の開始時間と音高を特定し，それを基に時間正規化を行う（3.3 節）．次にデータベースと同様に，スライディング・ウィンド方式で部分データ（子シーケンス）に分割する．各部分データからは，データベース作成時と同じ特徴量の特徴ベクトルを作成し，それを用いてクエリを作成する．検索はそのクエリを用いて行う．

3.2 タタタ歌いの利点

SoundCompass ではユーザには「タ」を用いてはっきりハミングしてもらう．「タ」は無声破裂音の “t” と，広い周波数領域にわたってバランスのとれたエネルギーを持つ中舌母音 “a” の組合せである[28)]．無声破裂音が発声される直前には，ストップ・ギャップと呼ばれる無音区間が存在するので，発声開始のタイミングを抽出しやすくなるうえ，エネルギーのバランスがとれている母音と組み合わされることで，音高の抽出もより正確に行えるといった利点がある．しかし SoundCompass を様々なユーザが使用する中で，この「タタタ歌い」にはこれ以外にも重要な利点があることが分かってきた．

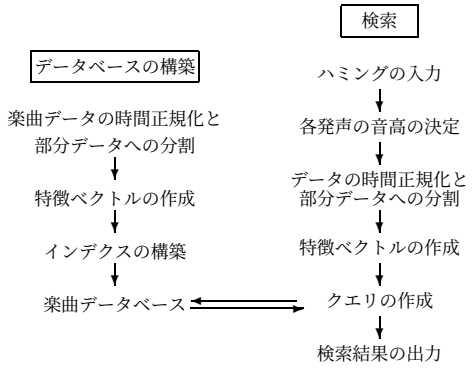

図 1 SoundCompass[3)] の処理の流れ
Fig. 1 Process flow for SoundCompass[3)].

まず最も重要な点として，タタタで歌うとユーザがメロディラインを再現することに集中できるので，より良いハミングデータを得ることができ，結果的に所望の検索結果を得られる回数が増えるということが分かった．歌詞で歌うと歌詞に気をとられて，音高や音長が正しく発声されなくなる傾向がある．また，メロディは分かっているのに歌詞を思い出せない部分があるとハミングの最中につまってしまい，結果的には検索ミスにつながる．

また「タタタ歌い」に慣れると，ハミング検索の際に歌詞で歌うのが恥ずかしくなったという感想が寄せられた．理由は人によって様々だが，歌詞で歌うとその歌を歌っている歌手と比較され，歌の上手い下手が容易にはっきりするというのも理由の 1 つである．

メロディを口ずさんで楽曲を検索するという目的を効率的に達成するためには，音高と音長をできるだけ正確に発声することに集中するのが望ましい．そのためには特定の文字だけで発声した方が良いし，その特定の文字としては音声信号処理において効率的に発声のタイミングと音高を抽出することができる無声破裂音とパワーの安定した母音の組合せが望ましい．したがって，タタタ歌いはハミング検索には非常に有効な方法である．

3.3 音楽データの自動時間正規化と部分データへの分割

本論文では，文献 3)，8) で提案した手法を用いてハミングデータも楽曲データも時間正規化し，部分データに分割する．

音楽は音高値の時系列データと考えることができる．時系列データは「ある一定の時間間隔ごとに計測される実数の列[29]」と定義される．しかし，音楽はゆっくり演奏されても速く演奏されても同じ曲であることが認識できるので，この「一定の時間間隔」を「秒」などで定義するのは，後の処理を考えると効率的ではない．そこで我々は「基底単位長」と呼ぶ情報量を定義し，音楽データを「一定の基底単位長ごとに計測される音高値の列」に変換する時間正規化手法を提案した．基底単位長は，ハミングデータについては，ハミングの中の各発声の発声時間の中の最頻値，また楽曲データについては，楽曲中および部分データの中の各音符の再生時間の中の最頻値とする．1 回の発声時間は，その発声の開始のタイミングから次の発声の開始のタイミングまでと定義する．同様に，再生時間は 1 つの音符の再生開始から次の音符の再生開始までを，その音符の再生時間と定義する．これを「オンセット情報処理[3]」と呼ぶ．

基底単位長を決定した後は，ハミングデータ/楽曲データの各発声時間/各再生時間をその基底単位長で除算（式 (1)）して相対値化する．

$$L'_{h_i} = L_{h_i}/U_h \qquad L'_{s_j} = L_{s_j}/U_s \qquad (1)$$

ハミング中の各発声時間を L_{h_i}，楽曲データ中の各再生時間を L_{s_j}，それぞれの基底単位長を U_h，U_s とする．相対値化した各発声時間/各再生時間を，限定した代表値にマップすることで多少のリズムの狂いに影響されない音高列に変換する．この代表値を音価レートと呼び以下の 8 種を定義する．

$$0.5, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0, 6.0 \qquad (2)$$

ハミング/楽曲データは，最終的にはこの 8 種の値だけで長さの情報が表現され，時間正規化される．

ただし，楽曲データについては楽曲全体を通して最頻出する再生時間が，楽曲のある部分について最頻出する再生時間と一致しているとは限らないので，2 段階の時間正規化を行う．まず各楽曲において最頻出する再生時間を仮基底単位長として，いったん楽曲データを時間正規化し，スライディング・ウィンド方式で部分データに分割する（子シーケンスの作成）．次に各部分データに対して，その中で最頻出する再生時間と仮基底単位長が等しいかどうかをチェックし，もしそうでない場合はその最頻出する再生時間を正基底単位長として定義し，それを用いて再度その部分データを時間正規化しなおす（孫シーケンスの作成）．

3.4 特徴ベクトル

本システムで作成する特徴ベクトルは，「固定スライド幅の音高値の時系列特徴ベクトル」，「可変スライド幅の音高値の時系列特徴ベクトル」と「音高差分布特徴ベクトル」である[3]．

「音高値の時系列特徴ベクトル」とは，1 音価レートごとの音高値を並べたベクトルで，メロディの類似性を判定するのに効果的である．音高値の時系列特徴ベクトルには，連続するベクトル間の先頭の間隔が一定となる「固定スライド幅の特徴ベクトル」と，連続するベクトル間の先頭の間隔が一定ではない「可変スライド幅の特徴ベクトル」の 2 種類がある．「可変スライド幅の特徴ベクトル」では楽曲データとハミングに共通のルールでベクトルの先頭を選択して設定するので，列データのマッチングで致命的な，比較するベクトルどうしの開始点と終了点が一致しない問題を軽減させることができるのが特徴である．なおデータベース内の楽曲とは異なるキーのハミングでも検索可能にするために，ベクトル内の各要素はある音高を基準としてシフティングする．

「音高差分布特徴ベクトル」は隣り合う音符の音高差の分布を示す特徴ベクトル[2] である．このベクトルは音符数や発声回数の違いを表現することができ，音高値の時系列特徴ベクトルに対する補助的な役割を果たす．

4. 電話を用いたハミング検索システム

4.1 実験システムの構成と処理の流れ

図 2 に電話を用いたハミング検索システムの構成を示す．また，**図 3** に電話を用いた SoundCompass の処理の流れを示す．図 1 と異なる処理を行う部分を下線を引いて示す．

実験システムは，電話機，ボイスモデム，クライアント PC，サーバマシンからなっている．クライアント PC とサーバマシンはネットワークでつながっており，この部分が SoundCompass である．

ユーザは，ハミング検索を行うために，検索受け付け専用の電話番号に電話をする．電話がつながると録音状態になるので，ユーザはハミングする．入力されたハミングは，ボイスモデムを通してクライアント PC のハードディスクに音声情報として直接保存する．収録時間は 15 秒である．保存されたハミングデータは，4.2 節に記載する方法で処理され，クエリとなってサーバマシンに送信される．サーバマシンでは類似検索を行い，検索結果は曲名と類似度のリストとして

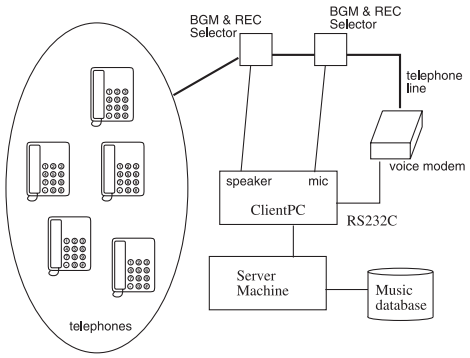

図 2 実験システムの構成
Fig. 2 Experimental system architecture.

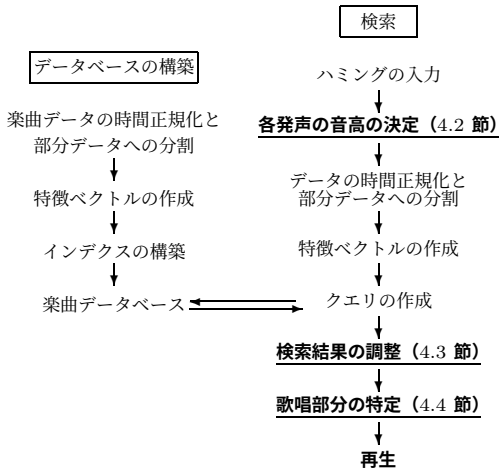

図 3 電話を用いた SoundCompass の処理の流れ
Fig. 3 Process flow for SoundCompass on the phone.

クライアント PC に送信される．クライアント PC では，曲名リストの中の 1 位でヒットした曲のヒットした部分を 4.4 節に述べる方法で切り出して再生する．再生後，通話を切断する．

4.2 ハミングの音声信号処理

本論文ではハミングデータの解析，および検索システムの精度評価実験用に，図 2 に記載したシステムを用いて 173 個のハミングを収集した．ほとんどは日本の歌謡曲であるが，12 曲は洋楽や演歌やアニメソングだった．日本の歌謡曲とは，多くは「宇多田ヒカル」や「SMAP」などの，いわゆる J ポップである．これらのハミングは，11 人（男性 9 人，女性 2 人）の被験者によってハミングされたもので，11 kHz/8 bit/モノラルの RAW 形式で収音した．

4.2.1 前 処 理

前処理では，収録したハミングデータから，主に定常的に混入する雑音を除去することを目的とする．電話を用いて収集したハミングデータには，ボイスモデム機器類による雑音や，ハミング収音時の周囲の雑踏や空調の音などが重畳している可能性がある．ハミングデータからの音高の抽出において，これらの雑音が，ホワイトノイズのように特定の周波数に片寄らない雑音の場合は影響はないが，特定の周波数を持つ音が定常的に混入した場合は発声の音高を特定する際に邪魔になる（4.2.2 項参照）．しかしこれらの定常雑音に対して，個別に対応するのは予測可能性，およびコストの観点から効率的ではない．また，ハミング全体を収録し終わってからこれらの定常雑音を推定する方がより正確に雑音除去を行えると考えられるが，収録し終わるまで待つのは効率的ではない．そこで，本論文ではハミングデータの収録が始まってから，実際のハミングが収録されるまでの無音区間の音響データ（高速フーリエ変換後の各周波数成分）をハミングデータ全体から減算することで雑音を除去する．

具体的には，録音開始後 n フレーム分の各周波数成分のパワーの平均値を，全フレームの各周波数成分のパワーから減算する．予備実験から，多くの人は録音開始から少なくとも約 1 秒弱は発声を開始しないが，稀に無音区間が約 0.2 秒程度しかとれない場合があることなどが分かった．また n が大きすぎると，対象区間内に突発的な雑音が混入する可能性が高くなることも分かった．そこで，本システムでは $n = 5$（約 0.12 秒分）とした．ただし，この 5 フレームの間に突発的な雑音が発生した場合は本方式は逆効果になる場合があるので，その場合はユーザは再度歌いなおし，検索し直す必要がある．

4.2.2 フレームごとの発声音高の確定

人間の声は音高を決める基本周波数を持つ基本波（基音）と，基本周波数の整数倍の周波数を持つ高調波（倍音）から構成される[30]．この事実を利用することで，かなり正確に基本周波数を推定できることが知られている[31]．歌唱によって発声される音高の基本周波数は，一般的な男声の場合☆は約 98 Hz（G2）から約 277 Hz（C#4）であり，一般的な女声の場合☆☆は約 175 Hz（F3）から約 466 Hz（A#4）である[32]．しかし，電話回線を通して収音すると 300 Hz 以下と 3.4 kHz 以上の周波数成分は失われてしまうので，男声についてはほぼ全域，女声の場合でも約半分については基本周波数の情報を直接得ることはできない．一方，高音が

☆ プロの声楽家の場合は約 82 Hz（E2）から約 494 Hz（B4）．
☆☆ プロの声楽家の場合は約 165 Hz（E3）から約 1047 Hz（C6）．

発声された場合は，基本周波数情報は失われないが高調波成分は失われてしまう．したがって，電話を介して収録された歌声から，どれだけ正確に基本周波数を推定できるかは，音声信号中の基本周波数成分の有無にかかわらず，300 Hz から 3.4 kHz 以内の周波数成分のピークの構成パターンをどれだけ正しく抽出できるかに左右されることになる．

なお，ハミングだけではなく，楽器や人の声が混ざった音響データに対して基本周波数の存在の有無に左右されずに基本周波数を推定する手法も提案されている[33]が，本システムではユーザに対して，ハミングを「タタタ」で入力してもらうことを前提にしているので，本論文では高速フーリエ変換（以後，FFT）を用いて各周波数成分のパワーを算出した後，大きなピークを示す周波数成分をピックアップし，それらとその周囲の周波数成分のパワーの値から基本周波数を割り出すという，処理時間などの実用性に照らし合わせてシンプルな手法を提案する．たとえば，120 Hz が発声された場合，電話で収音したハミングデータについては，360 Hz, 480 Hz, 600 Hz, 720 Hz, 840 Hz, ... の周波数成分にピークが出る．これに対し 240 Hz が発声された場合は，480 Hz, 720 Hz, 960 Hz, 1,200 Hz, ... の周波数成分にピークが出る．したがって 480 Hz にピークが存在した場合は，360 Hz や 600 Hz にもピークが出ていれば，基本周波数が 120 Hz の音高が発声されたと推測することができ，360 Hz や 600 Hz にはピークがなければ，基本周波数が 240 Hz の音高が発声されたことが推測できる．

具体的には，300 Hz 以上 3.4 KHz 以下の周波数の周波数成分のパワーの中で，そのフレーム内の平均パワーの z%を超えるパワーを持ち，かつピークを示しているパワーの周波数を周波数の低い方から n 個選出する（周波数：$f_i(1 \leq i \leq n)$）．次に，各周波数 f_i が基本周波数からある基本周波数の第 h 倍音周波数（h は正数）までのいずれかであることを仮定して，それぞれの場合の該当する h 個の高調波成分のパワーを合算し，その中で最も大きな値を示したケースを正解ケースとする．アルゴリズムを**図 4** に示す．周波数 f の周波数成分のパワーを $p(f)$ で表す．基本周波数は F_0 と表す．

選出したピークを示す周波数 f_i の多くは，お互いが倍音関係にある場合が多いので，この手法を用いて効率的に基本周波数を推定することができる．予備実験から $z = 25$，$n = 5$ とした．また，h は 7 としたときが，最も基本周波数の抽出精度が良かった．これは一般女声の最高周波数 466 Hz の約 7 倍が電話の周波数帯域の上限だからだと考えられる．この処理を全フレームに対して行い，以下の式 (3) を用いて各フレームの推定基本周波数 F_0 を音高値 M（MIDI の音高番号）に変換する．

begin
1: **for** i= 0 to n **do**
2: **for** $j = 0$ to $h - 1$ **do**
3: $P(i,j) = \sum_{k=j+1}^{h+j} p(\frac{f_i}{j+1} \times k)$
4: $P(i,j)$ を最大にする i と j を選出する．
5: $F_0 = \frac{f_i}{j+1}$
end

図 4 基本周波数 F_0 確定アルゴリズム
Fig. 4 Algorithm to determine F_0.

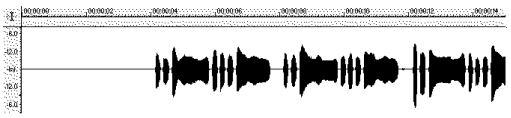

図 5 ハミングデータの音声信号
Fig. 5 Voice signal of humming.

$$M = \frac{12}{\log 2} \times \log \frac{F_0}{F_b} \tag{3}$$

F_b は $M = 0$ の基本周波数である（$F_b = 8.175799$）．

4.2.3 各発声の発声音高の確定

本論文では 1 回の発声に 1 つの音高を割り当てるという方針をとる．これは，人が「タ」でメロディを口ずさむ場合，音高が変化するときや同じ音高が続いていても歌詞の文字が変わるタイミングが,「タ」を発声しなおすタイミングとして自然だからである．

各発声の音高の決定には，有声フレームの音高情報のみを使用する．主に以下の 2 点の条件を満たすフレームを有声フレームと判定する．

- そのフレームの振幅の平均値が，直前のフレームの振幅の平均値の a 倍以上である．
- そのフレームの平均パワーが，全フレームの平均パワーの z%を上回る．

予備実験から $a = 2$，$z = 25$ とした．

図 5 に「タタタ」でハミングした際の音声波形信号を示す．時間は右に向かって進んでいる．図からハミングデータは長い無音区間と，それに続く複数の黒い塊で示される発声で構成されることが分かる．1 つの塊が 1 回の「タ」の発声で，発声と発声の間にはきれいに無音区間（ストップ・ギャップ[28]）を観察することができる．無音区間のフレームは無声フレームと判定し，発声している間のフレームは有声フレームと判定するので，無声フレームに囲まれた有声フレームの集合を 1 回の発声と考える．

各有声フレームの集合における音高値の分布の中で，以下の式 (4) が最大値を示す音高値 T をその発声の音高とする．音高値 t の有音フレーム数を $w(t)$ と表す．

$$T = \arg\max_{t} \frac{\sum_{i=-2}^{2} \alpha_i \cdot (t+i) \cdot w(t+i)}{\sum_{i=-2}^{2} w(t+i)} \quad (4)$$

$$2 \leq t \leq 125$$

$$\alpha_i = \begin{cases} 0.25 & if \quad i = \pm 2 \\ 0.5 & if \quad i = \pm 1 \\ 1.0 & if \quad i = 0 \end{cases}$$

4.3 特徴のある部分による検索結果のみを用いた最終検索結果の作成

文献 3) で，ウィンド長より長いハミングが入力された場合に複数の子シーケンスを作成し，それらすべてを検索に使用して，最終的な検索結果は各子シーケンスからの検索結果を OR 方式で統合して作成する手法について述べた．すなわち，ハミング h から m 個の子シーケンスができ，曲 s から n 個の子/孫シーケンスができたとき，ハミングの i 番目の子シーケンスと曲 s の j 番目の子/孫シーケンスとの距離を $d(h_i, s_j)$ と表すと，ハミング h と曲 s との距離 $D(h,s)$ は以下の式 (5) から算出する．

$$D(h,s) = \min_{1 \leq i \leq m, 1 \leq j \leq n} \{d(h_i, s_j)\} \quad (5)$$

この検索方式は，短く分割されたハミングデータを基に楽曲データとの類似性を判定するので，一部だけでもうまく歌えていれば検索することが可能になるという点で，ハミング検索には好都合である．しかし，今まで蓄積してきた検索結果を調査した結果，この方式には欠点もあることが分かった．

実際に正解楽曲を検索できるかどうかは，どれだけ正しく歌えているかとは別に，マッチングの対象となる子シーケンスがどれだけ特徴的であるかにも依存する．式 (5) を用いて，各子シーケンスからの検索結果をマージし，最終結果として l 曲が検索されたとき，ある子シーケンスが特徴的であるとはいえない場合は，多数の $D(h, s^k)$（k は検索された順位で，k 位で検索された楽曲を s^k と表す）において以下の式 (6) の $f(D(h, s^k))$ 値が同一になるうえ，多数の $D(h, s^k)$ の値もほぼ等しくなってしまう．

$$f(D(h, s^k)) = \arg\min_{i} \{d(h_i, s_j^k)\} \quad (6)$$
$$1 \leq k \leq l, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n$$

そこで，本論文では l 曲の検索結果の中で，$f(D(s^k, h)) = i$ である楽曲数を $scnt(i)$ と表すとき，以下の式 (8) の値が x を超える i に対しては，その子シーケンス (h_i) による検索結果を使用せずに最終的な検索結果を作成する方式を導入する．

$$scnt(i)/l \qquad 1 \leq i \leq m \quad (7)$$

この手法による有効性は，5.2 節で定量的に評価する．

4.4 ピッタリ再生機能

ユーザには 1 位にヒットした曲を再生することで検索結果を提供するので，楽曲の再生に際しては，ユーザが歌った部分が正確に特定されていることが望ましい．しかし，楽曲データは子/孫シーケンス単位でデータベースに登録するので，検索結果からは子/孫シーケンス単位でしか部分を特定できない．またハミングから複数の子シーケンスを作成できた場合，必ずしもつねに先頭の子シーケンスが正解曲を探し当てるとは限らない．そこで，本節ではマッチしたハミングの子シーケンスのどの部分と楽曲データの子/孫シーケンスのどの部分がマッチしたかという情報を基に，ユーザが歌い始めた音符をできるだけ正確に特定して再生する「ピッタリ再生機能」を提案しその実現方法を説明する．

ここでは，楽曲データとハミングの可変スライド幅の音高値の時系列特徴ベクトルがマッチした場合について説明する．子/孫シーケンスを SS で示し，可変スライド幅の特徴ベクトルは FV で示すとする．また，h はハミングを，s は楽曲データを表すとする．t_{h_i} はハミングの i 番目の発声の発声時間を表すとする．ここで，ハミングデータの k 番目（$0 \leq k \leq n_h$，n_h は整数）の子シーケンス SS_{h_k} から作成された可変スライド幅の特徴ベクトル FV_{h_k} が，楽曲データ s の l 番目（$0 \leq l \leq n_s$，n_s は整数）の子/孫シーケンス SS_{s_l} から作成された可変スライド幅の特徴ベクトル FV_{s_l} とマッチしたとする．SS_{h_k} の基底単位長は u_k 秒，SS_{s_l} の基底単位長は b_l ティック☆とする．また，FV_{h_k} はハミングの p 番目の音符を先頭として作られ，FV_{s_l} は，楽曲データ内の q ティックから再生される音符であるとする．このとき，楽曲内のハミングされたと推定される音符の再生開始ティックは，以下の式 (8) で算出されるティックから最も近いティックに再生される音符であると推定できる．

$$q - b_l \times \sum_{i=0}^{p-1} t_{h_i}/u_k \quad (8)$$

☆ MIDI データの分解能を表すための最小単位．MIDI データの分解能は，データの時間管理を行うために必要な情報である[34]．

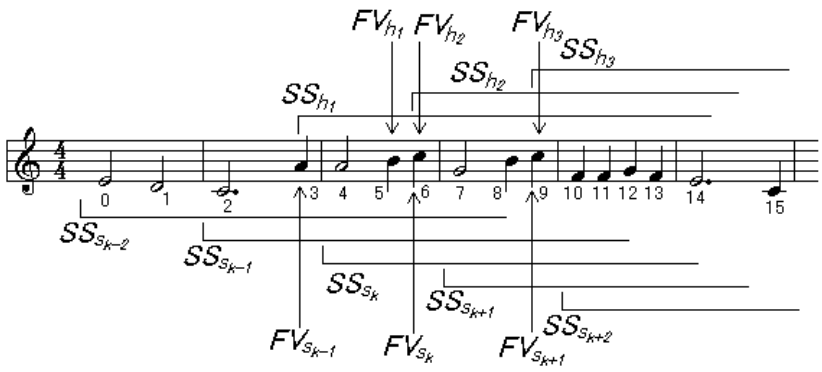

図 6 ピッタリ再生機能説明用楽譜例
Fig. 6 Example tune to describe the "just-play function".

表 1 図 6 の楽譜の各音符の再生ティックと発声時間例
Table 1 Example of times of each note and utterance starts.

音符番号	再生開始ティック	発声番号	発声時間（秒）
0	0	-	–
1	960	-	–
2	1920	-	–
3	3360	0	0.5
4	3840	1	1.01
5	4800	2	0.48
6	5280	3	0.51
7	5760	4	0.99

図 6 を用いて詳しく説明する．図 6 に示すのは楽曲の一部で，ユーザは 3 番目の音符からハミングしたとする．ハミングにおける各音符の発声時間と，楽曲データにおける各音符の再生開始ティックを**表 1** に示す．ハミングデータの基底単位長は 0.25 秒とする．簡単のため，この部分に関しては楽曲データの基底単位長は 8 分音符（240 ティック）であるとする．また，スライド幅は 8 音価レート（拍数に換算すると 4 拍）とする．楽曲データは，各小節の 1 拍めを先頭とする子シーケンスに分割されているとする．$P()$ は，子/孫シーケンス，および特徴ベクトルの先頭位置を示すとする．特に特徴ベクトルの先頭位置は図中では矢印で示す．したがって，たとえば SS_{h_2} はハミングの 2 番目の子シーケンスのことで，その先頭位置は $P(SS_{h_2})$ と表す．同様に，$FV_{s_{k+1}}$ は楽曲データの $k+1$ 番目の子/孫シーケンスである $SS_{s_{k+1}}$ から作成される可変スライド幅の特徴ベクトルのことで，その先頭位置は $P(FV_{s_{k+1}})$ と表す．可変スライド幅の特徴ベクトルは，子/孫シーケンスの先頭とは独立に選択するので，図 6 では文献 3) と同様に，スライド幅内で最も高い音が最初に出現する位置とする．したがって，この例では

$$P(SS_{h_1}) \neq P(SS_{s_{k-1}}) \quad P(FV_{h_1}) \neq P(FV_{s_k}) \quad (1)$$
$$P(SS_{h_1}) \neq P(SS_{s_k}) \quad P(FV_{h_1}) \neq P(FV_{s_{k-1}}) \quad (2)$$
$$P(SS_{h_2}) \neq P(SS_{s_k}) \quad P(FV_{h_2}) = P(FV_{s_k}) \quad (3)$$
$$P(SS_{h_2}) \neq P(SS_{s_{k+1}}) \quad P(FV_{h_2}) \neq P(FV_{s_{k+1}}) \quad (4)$$
$$P(SS_{h_3}) \neq P(SS_{s_{k+1}}) \quad P(FV_{h_3}) = P(FV_{s_{k+1}}) \quad (5)$$
$$P(SS_{h_3}) \neq P(SS_{s_{k+2}}) \quad P(FV_{h_3}) \neq P(FV_{s_{k+2}}) \quad (6)$$

となる．(3) と (5) で，ハミングデータと楽曲データの特徴ベクトルの先頭が一致しているが，ここでは (3) のケースを用いて説明する．(3) のケースでは，6 番目の音符が FV_{s_k} の先頭である．表 1 より，6 番目の音符の再生開始ティックは 5280 であることが分かるので，式 (8) より

$$5280-240\times(0.5+1.01+0.48)/0.25=3369.6 \quad (9)$$

を得る．3369.6 ティックに最も近い音符は音符番号 3 の音符なので，これが楽曲データにおけるハミングの正確な発声開始の音符であると推定できる．この手法の有効性は 5.3 節で定量的に評価する．

5. 評　　価

本章の評価実験で使用するハミングデータは，4.2 節で述べたものと同じである．また楽曲データには，MIDI 形式の 21,804 曲を使用する．ほとんどの曲は日本の歌謡曲であるが，約 6,000 曲は洋楽，童謡，民謡などである．

5.1 検索精度評価

本節では，電話を用いたハミング検索システムの検索精度を，「収音方式と音声信号処理方式」と「男声と女声」という 2 つの観点から定量的に評価する．本システムは，検索結果として 1 位にヒットした曲のみを再生するが，本節では検索精度をより詳細に評価するため，25 位までの精度を示す．本論文における検索精度とは，検索結果の特定の順位以内に正解が検索されたハミングの割合であると定義する．**図 7** と**図 8** に検索精度を示す．図の横軸は順位を示す．縦軸は検索

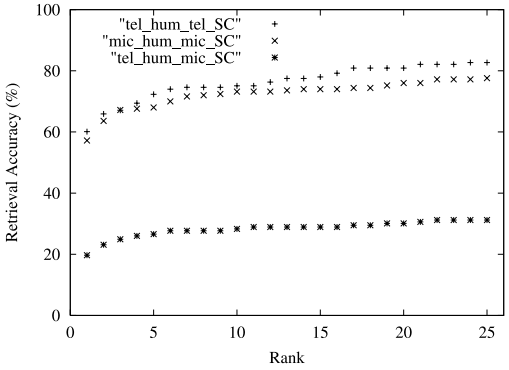

図 7 検索精度（収音方式と音声信号処理方式）
Fig. 7 Retrieval accuracy (recording and signal-processing).

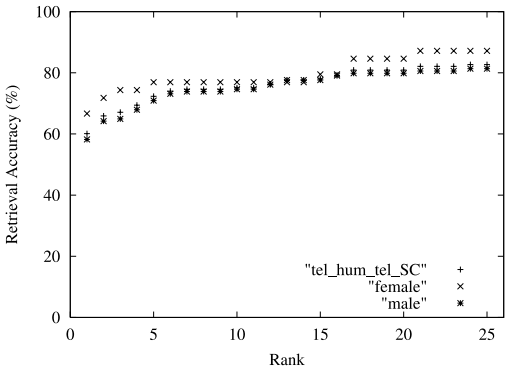

図 8 検索精度（男声と女声）
Fig. 8 Retrieval accuracy (male/female voice).

精度を示しており，その順位以内に検索されたハミングの割合を示している．

まず,「収音方式と音声信号処理方式」に関する実験結果について述べる（図 7 参照）．電話回線経由で収音した 173 個のハミングデータを，4.2 節で述べた方式で処理した場合の本システムの検索精度を “tel_hum_tel_SC” で表示する．また，同じハミングデータを従来のマイク収音用の音声信号処理方式で処理した場合の検索精度を “tel_hum_mic_SC” で表示する．なお，比較のため文献 3) に記載したマイクで収音した 250 個のハミングデータを，マイク収音専用の音声信号処理方式で処理した場合の検索精度を “mic_hum_mic_SC” として併記する．これらのハミングデータはすべてメトロノームを使用せずにハミングされたものである．

“tel_hum_tel_SC“と”tel_hum_mic_SC” の検索精度の差は，音声信号処理方式が電話回線経由専用かマイク収音専用かの違いだけに起因する．“tel_hum_mic_SC” において，電話回線経由で収音されたにもかかわらず，マイク収音用の音声信号処理を適用しても上位に検索されているハミングの半分以上は，女声のハミングである．女声は基本周波数が 300 Hz を超える場合もあるので，基本周波数の存在を仮定したマイク収音用の音声信号処理を適用しても，正しく基本周波数を同定し，問題なく検索できる場合があったと考えられる．しかしそれ以外の場合は，電話回線を経由して収音したハミングは，マイク収音を仮定した音声信号処理方式ではほとんど正しく検索されないことが分かる．

一方，“tel_hum_tel_SC” と “mic_hum_mic_SC” の違いは，音声の収音方式とその音声信号の処理方式の組合せの違いで，前者は電話回線経由で収音したハミングを電話回線経由専用の信号処理方式で処理した場合の検索精度で，後者はマイクで収音したハミングをマイク収音専用の信号処理方式で処理した場合の検索精度である．両者の検索精度はほとんど差がないと評価できるので，この結果から電話を用いたハミング検索システムが，マイク収音の検索システムとほぼ同じ検索精度を実現したことが確認できた．なお，“tel_hum_tel_SC” の方が “mic_hum_mic_SC” より若干検索精度が高いのは，本論文で使用した複数の倍音の高調波構造を手がかりにした音声信号処理方式が非常に有効だったということを示している．

次に,「男声と女声」に関する実験結果について述べる（図 8 参照）．173 個のハミングデータのうち，男声によるものは 134 個，女声によるものは 39 個であった．図 8 における “tel_hum_tel_SC” は，図 7 と同様で，男声と女声を合わせた検索精度で，“male” は男声，“female” は女声による検索精度である．この結果から，男声/女声による検索精度の違いはほとんどないことが分かる．女声の方が若干精度が高いのは，先に述べたとおり，ハミングデータにおいて基本周波数成分が存在する場合が多いからではないかと考えられる．以上より，本論文で提案した基本周波数推定手法は，基本周波数成分の存在の有無が分からない音声信号に対しても有効であることが分かる．

5.2 特徴のある部分による検索結果のみを用いた最終検索結果の作成の効果

本評価実験では，4.3 節の x の値を 0.8 とした．この条件の下で，本論文で使用した 173 個のハミングのうち，$x \geq 0.8$ に該当したのは 75 件であった．この 75 件のうち，本方式を用いることで検索順位が上がったものは 25 件，順位の下がったものは 5 件だった．残りのハミングは順位が変わらなかった．順位が下がったもののうち，26 位以下になったものは 2 件で，1 位から 2 位になったものが 1 件，2 位から 13 位になったものが 1 件，18 位から 23 位になったものが

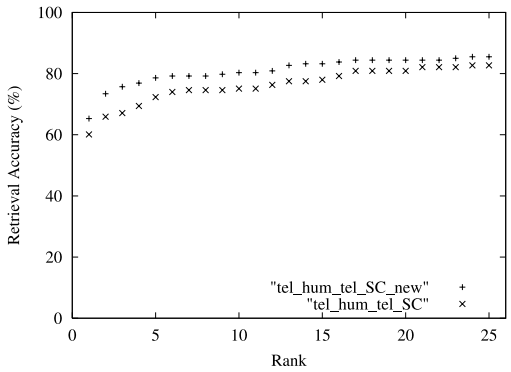

図 9 特徴的な検索キーのみを用いて最終検索結果を作成した場合の検索精度
Fig. 9 Retrieval accuracy without result based on featureless subsequence of humming.

表 2 ピッタリ再生機能の有無に基づく再生開始音符の差
Table 2 Differences of the number of notes to start playback for the cases with/without "just-play function".

		ピッタリ再生機能あり			
		0	1	2	3
機ピ	0〜4	67	6	0	0
能ッ	5〜9	28	7	0	0
なタ	10〜14	12	3	0	0
しリ	15〜19	9	2	2	2
再	20〜24	7	0	0	0
生	25 以上	3	0	0	0
	合計	126	18	2	2

1 件だった．これに対し，順位が上がったものの中で 1 位に検索されるようになったものは 10 件，また 26 位以下から 25 位以内に入ったものは 7 件だった．

図 9 に検索精度を示す．"tel_hum_tel_SC_new" は，図 7 における "tel_hum_tel_SC" に対して，本方式を適用した場合の検索結果を示している．図 9 から，本方式を用いることで各順位において約 5%程度検索精度が向上していることが分かる．これらの結果から，本方式の有効性を確認することができた．

5.3 ピッタリ再生機能の効果

173 件のハミングデータに対して，5.2 節の検索精度評価において 25 位以内に検索された 148 件のハミングについて，ピッタリ再生機能を用いた場合と用いなかった場合の再生開始音符の差の比較結果を**表 2** に示す．表 2 は，MIDI データと楽譜で確認した，ユーザが実際にハミングを開始した音符と，再生された音符との間の音符数を示している．音符数が 0 というのは，ユーザがハミングした音符から再生されたということである．この表では，たとえばピッタリ再生機能を使用しないで再生すると，ハミングした音符から 5 から 9 個ずれた音符から再生されるが，ピッタリ再生機能を使用すると，ユーザがハミングした音符から再生されるハミングは 28 件だったということが分かる．ピッタリ再生機能を使用しない方が，使用するよりもハミングを開始した音符をより正確に特定できたのは 2 件で，それ以外はすべて同等またはより正確に特定することができた．この 2 件が，ピッタリ再生機能を用いると正しい発声開始音符から再生されなかった原因は，1 つは，ハミングの最初の声が小さすぎて音声信号処理の際に最初の発声を検出できなかったからで，もう 1 つは，ハミングの中に発声すべき正しい長さよりも長く発声している部分があったからであると考えられる．以上より，本論文で提案したピッタリ再生機能を用いると，ほぼ正確にユーザがハミングした部分を特定することができるということが確認できた．

5.4 電話を用いたハミング検索における検索結果の提示方法に関する議論

本論文では 1 位に検索された曲を再生することでユーザに対する検索結果の提供とした．そのためにハミングしたと思われる部分をできるだけ正確に特定することが必要で，ピッタリ再生機能を提案した．またこの機能を用いることで，ほぼ正確にハミングした部分を特定することができることを定量的に確認した．実際に実験システムを使ってみると，この機能は，電話のように聞こえてくる音声の品質が良くない状況において非常に有効であることも分かったが，やはりハミング検索システムの検索結果の提示という観点では，上位 5 位くらいまでの曲名も提示できた方が望ましいと思われる．携帯電話などはディスプレイも装備しているので，ハミングを収録した後，検索結果はメールで通知するという方法も考えられる．また，音声合成システムなどと連携させて検索結果の曲名を読み上げるという方法も考えられる．今後は検索結果の提供方法を充実させる方向で，システムの改善を進めたい．

ピッタリ再生機能は，ユーザがハミングした部分を特定するための機能であり，楽曲の特定の部分の再生以外にも有効な使い方がある．楽曲の特定の部分を瞬時にポインティングすることの必要性はまだあまり認識されていないが，今後，楽曲に歌詞データを含め様々なメタデータが時系列的に付与されることが想定されるので，重要な技術になると考えられる．音楽は時系列データなので，従来は音楽データの特定の部分をアクセスする場合，楽譜を使ったり，当該部分にたどり着くまで流し聞きをしたりするなどしなければならなかった．しかし，ポインティングしたい部分をハミングするだけで，楽曲の所望の部分に高速に正確にアクセスできれば，部分的に聞き取れなかった歌詞の言葉

を調べたり，その部分で使用されている楽器を調べたりすることが簡単にできるようになると考えられる．

6. まとめ

本論文では，電話を使ったハミング検索システムについて述べた．

電話を通して収音されたハミングは，電話回線の帯域制限により，基本周波数情報や基音の高調波成分を十分に取得できないので，発声音高を特定するのが困難である．これに対して，本論文では限られた周波数帯域でもピークを持つ周波数チャネルどうしの関係から基本周波数を正確に推定する手法，およびその情報を用いて各発声の音高を特定する手法を提案し，その有効性を検索精度を基に定量的に確認した．

また，電話機はインタフェース機器としては自由度が低く，検索結果として類似度つきの楽曲リストのような複雑な情報を出力することは難しい．これに対し，本論文では検索結果として 1 位に検索された楽曲を，ハミングされた部分を正確に特定して再生する手法を提案し，ピッタリ再生機能として実現した．この機能を用いるとほぼ正確にハミングした部分を特定できることを定量的に確認した．今後はこれを基に，よりサービスに適したハミング検索結果の提示方法を検討していく予定である．

さらに，今まで蓄積した検索結果の調査から，入力されたハミングの一部が複数の楽曲と近い距離を持つ場合は正解曲を限定できず，結果的に検索精度を低下させる原因の 1 つになっていたことが分かった．これに対し，本論文ではハミングのそのような部分による検索結果を用いずに，最終検索結果を作成する手法を提案し，検索精度が全体的に約 5%向上することを確認した．

これらの提案手法の効果は電話を使ったハミング検索環境を実際に構築することで定量的に確認した．

今後の課題として，特にハミングの一部が複数の楽曲と近い距離を持つ場合に検索精度が低下した事実をふまえ，音楽データの音楽的性質と特徴空間内でのデータの分布に関する検討や，さらに音楽的性質がマッチングへ及ぼす影響に関する検討などが考えられる．

参考文献

1) 小杉尚子，小島 明，片岡良治，串間和彦：大規模音楽データベースのハミング検索システム，情報処理学会論文誌，Vol.43, No.2, pp.287–298 (2002).
2) 小杉尚子，櫻井保志，山室雅司，串間和彦：SoundCompass：ハミングによる音楽検索システム，情報処理学会論文誌，Vol.45, No.1, pp.333–345 (2004).
3) 小杉尚子，櫻井保志，森本正志：ハミング検索のための音楽データ自動時間正規化手法，情報処理学会論文誌：データベース，Vol.45, No.SIG7 (TOD22), pp.163–178 (2004).
4) 朝日新聞：歌ってみたかったあの歌タッタタッタタ — 一発選曲カラオケ店登場，2000.12.7 夕刊.
5) http://www.shazam.com/uk/do/home: JUST HIT 2580 ON YOUR MOBILE.
6) http://shazam.jp. SHAZAM を使って音楽認識しよう！
7) http://www.gracenote.com/gn_products/mobileMusic.html: Gracenote Mobile MucisID^{SM}
8) Kosugi, N., Sakurai, Y. and Morimoto, M.: SoundCompass: A Practical Query-by-Humming System — Normalization of Scalable and Shiftable Time-Series Data and Effective Subsequence Generation, *Proc. ACM SIGMOD International Conference on Management od Data*, pp.881–886 (2004).
9) 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，電子情報通信学会論文誌，Vol.J77-D-II, No.8, pp.1543–1551 (1994).
10) 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，信学技報，pp.25–32，電子情報通信学会 (1998).
11) 柳瀬隆史，高須淳宏，安達 淳：メロディからの特徴抽出による曲検索システム，情報処理学会研究報告，Vol.99, No.39, pp.1–6 (1999).
12) 橋口博樹，西村拓一，岡 隆一，赤坂貴志：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 PRMU200-107〜118, Vol.100, No.443, pp.79–86 (2000).
13) Kosugi, N., Nishihara, Y., Kon'ya, S., Yamamuro, M. and Kushima, K.: Let's Search for Songs by Humming!, *Proc. ACM Multimedia 99* (*Part 2*), p.194 (1999).
14) Sonoda, T., Goto, M. and Muraoka, Y.: A WWW-based Melody Retrieval System, *Proc. ICMC 1998*, pp.349–352 (1998).
15) Sonoda, T. and Muraoka, Y.: A WWW-based Melody-Retrieval System — An Indexing Method for A Large Melody Database, *Proc. ICMC 2000*, pp.170–173 (2000).
16) http://www.utagoe.com/search/index.html: utagoe
17) Nishimura, T., Hashiguchi, H., Takita, J., Zhang, J.X., Goto, M. and Oka, R.: Music Signal Spotting retrieval by a Humming Query Using Start Frame Feature Dependent Continuous Dynamic Programming, *2nd Interna-*

tional Conference on Music Information Retrieval (2001).
18) 橋口博樹，西村拓一，矢部博明，岡 隆一，赤坂貴志：鼻歌による音楽検索と歌詞音声検索の統合処理の検討，音楽情報科学研究会研究報告 39-9，pp.57–62, 情報処理学会 (2001).
19) 西村拓一，岡 隆一，滝田順子，後藤真孝：類似メロディー区間検出による音楽時系列検索の高速化，音楽情報科学研究会研究報告 39-10，pp.63–70, 情報処理学会 (2001).
20) 小杉尚子，西原祐一，紺谷精一，山室雅司，串間和彦：ハミングを用いた音楽検索システム，情報処理学会研究報告 99-DBS-119，pp.49–54, 情報処理学会 (1999).
21) Kosugi, N., Nishihara, Y., Sakata, T., Yamamuro, M. and Kushima, K.: A Practical Query-By-Humming System for a Large Music Database, *Proc. 8th ACM International Conference on Multimedia*, pp.333–342 (2000).
22) Kosugi, N., Nishimura, G., Teramoto, J., Mii, K., Onizuka, M., Konya, S., Kojima, A., Kataoka, R., Honishi, T. and Kushima, K.: Content-based Retrieval Applications on a Common Database Management System, *Proc. 9th ACM International Conference on Multimedia*, pp.599–600 (2001).
23) Kosugi, N., Nagata, H. and Nakanishi, T.: Query-by-Humming on Internet, *Proce. 14th DEXA 2003*, pp.589–600 (2003).
24) 日本経済新聞：メロディを口ずさんで選曲，2001.1.19.
25) ASCII: 音声認識ソフト大集合！ http://review.ascii24.com/db/review/soft/voice/2001/03/08/623839-006.html
26) 朝日新聞：音声認識で一発選曲 — 進化するカラオケ，2001.11.26 科学面「技あり」. http://www.asahi.com/science/waza/011126.html
27) Faloutsos, C., Ranganathan, M. and Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Database, *Proc. ACM SIGMOD, International Conference on management of Data*, pp.419–429 (1994).
28) 古井貞熙：ディジタル音声処理，東海大学出版会 (1992).
29) Goldin, D.Q. and Kanellakis, P.C.: On Similarity Queries for Time-Series Data: Constraint Specification and Implementation, *Proc. 1st International Conference on the Principles and Practice of Constraint Programming*, pp.137–153 (1995).
30) 新井 純ほか：最新音楽用語事典，リットーミュージック (1998).
31) 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.2212–2223 (2001).
32) UlrichMichels/角倉一朗：図解音楽事典，白水社 (1998).
33) 後藤真孝：音楽音響信号を対象としたメロディーとベースの音高推定，電子情報通信学会論文誌 D-II，Vol.J84-D-II, No.1, pp.12–22 (2001).
34) 鈴木直美：[明解] インターネット時代の標準ファイルフォーマット事典，インプレス (1998).

(平成 16 年 3 月 20 日受付)
(平成 16 年 6 月 29 日採録)

(担当編集委員 今井 正和)

小杉 尚子

1993 年慶應義塾大学理工学部電気工学科卒業．1995 年同大学院理工学研究科計算機科学専攻修士課程修了．同年日本電信電話（株）入社．以来，リアルタイム OS の研究を経て，現在は音楽検索システムの研究開発に従事．

櫻井 保志（正会員）

1991 年同志社大学工学部電気工学科卒業．同年日本電信電話株式会社入社．1996 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．1999 年同大学院博士後期課程修了．工学博士．現在，NTT サイバースペース研究所に所属．2004 年よりカーネギーメロン大学客員研究員（2005 年までを予定）．索引技術，情報検索に関する研究開発に従事．

森本 正志（正会員）

1986 年京都大学工学部情報工学科卒業．1988 年同大学院工学研究科情報工学専攻修士課程修了．同年日本電信電話（株）入社．1996 年米国スタンフォード大学客員研究員．2002 年京都大学情報学研究科知能情報学専攻博士後期課程在籍．現在，NTT サイバースペース研究所において，画像・音楽・映像などのメディア・ハンドリング技術の研究開発に従事．電子情報通信学会，映像情報メディア学会各会員．