鼻歌検索のための音楽特徴量の抽出と評価

張 夢堯† 成 凱†

†九州産業大学 〒813-8503 福岡県福岡市東区 2-3-1

E-mail: †chengk@is.kyusan-u.ac.jp

あらまし 鼻歌検索とは，ユーザのハミングを検索キーとして，そのメロディを含む楽曲やその部分の検索を行うものであり，ハミング検索とも呼ばれる．鼻歌検索は，音楽の特徴に基づいているため，音楽音響信号の特徴を数値化した音楽特徴量を抽出する必要がある．楽曲特徴量が様々あるが，どれが有効であるかについてまだ体系的に評価されていない．本研究では，音楽データのメロディ検索に利用できる楽曲特徴量を調査し，評価を行う．そのうえ，楽曲特徴量抽出の高速化を図る．

キーワード 音楽検索，鼻歌検索，特徴量抽出

1. はじめに

音楽検索では，曲名，歌手名，歌詞等の情報がなく路上で何気なく聞いた曲を検索したいとき，鼻歌検索が有効である．鼻歌検索とは，ユーザのハミングを検索キーとして，そのメロディを含む楽曲やその部分の検索を行うものであり，ハミング検索とも呼ばれる．鼻歌検索は，音楽データベースの楽曲およびハミングから数々の音楽特徴量を抽出し，その特徴量を用いて類似検索を行う．鼻歌検索は，音楽の特徴に基づいており，音楽音響信号の特徴を数値化した音楽特徴量を抽出する必要がある．音楽特徴量は音楽信号の一つの側面，あるいは幾つかの側面から捉える特徴を数値化したものであるため,異なる音楽特徴量は，音楽の異なる特性を表現することができる．音楽特徴量の選択は鼻歌検索の性能に直接的な影響を与える．鼻歌検索のための音楽特徴量について，多くの研究が行われてきた[1][2][3][14]. それらの音楽特徴量の中で，どのような特徴量を利用することが有効であるかはまだ体系的に解明されていない．

本研究では，音楽情報検索に関連する音楽特徴量を調査し，それぞれの特徴量の鼻歌検索の有効性を検証する．さらに，鼻歌検索によく使われる特徴量の抽出を実施し，音楽特徴量の検索精度，抽出速度を体系的に評価する．検索の精度と効率を向上させるとともに鼻歌検索技術の発展に貢献する．

2. 音楽データ

2.1. 音楽の三要素

音楽は音から生まれ，メロディ，ハーモニー，リズムという三要素が欠かせない[1]．メロディはピッチの上昇下降によって構成され，音楽の流れを作り出す役目があり，音楽から感じられる様々な情緒はメロディが基になっている．音楽のメロディの多くは，西洋音楽の「C，$C^{\#}$，D，$E^{\flat}$，E，F，$F^{\#}$，G，$A^{\flat}$，A，$B^{\flat}$，B」のピッチで構成されている．このように 1 オクターブの間で 12 音に均等に分けられて，メロディに使うことができるピッチのセットが音階である．この 12 音階の隣接音同士は等間隔で半音刻みになっていて，演奏の上ではこの 12 音階を扱う．

ハーモニーは，ある音に別のピッチの音を同時に重ねることによって形成されるもので，構成されるハーモニーによってメロディに様々なボリュームを与えることができ，各音間のピッチの関係とその展開がハーモニーを形成する．リズムは，楽曲に統一感を与えるために繰り返しによって生み出される時間パターンのまとまりで，強弱のアクセントの違いが各リズム固有の時間パターン形成する．リズムによってその曲の感じがほとんど決まる．

音楽の三要素が組み合わさって音楽になる．メロディが横軸に展開し，ハーモニーが縦軸で響き，そこにリズムでアクセントを付けると音楽ができる．

2.2. 基本周波数とピッチ

音楽の形成は，楽器，人間の声帯といった固体に動撃や力を与えると固体全体あるいはその一部を振動させることである．基本周波数とは，信号を正弦波の合成で表したときの最も低い周波数成分の周波数を意味する．人間の耳の聞こえ方は個人差が大きく，周波数によっても聞こえ方が異なるが，一般的には 20Hz～20,000Hz が可聴範囲とされる．音声分野では，有声音の基本周波数の別称としてピッチという用語が広く用いられている．一方，聴覚分野では，基本周波数は物理量であるが，ピッチは心理量，つまり主観的な属性である[12]. したがって，ピッチは音の大きさ，スペクトル分布，存続時間，時間変動など，様々な要因の影響を受ける．

2.3. 音楽データの形式

音楽データを大きく分けると，オーディオデータと MIDI データという二つの種類がある[1].

オーディオデータは，一般的に，音波をデジタル信

号に変換したデータである．音の波形はアナログ信号である.アナログ信号の大きな特徴に「連続性を持つ」ことがあげられる．デジタル信号はアナログ信号をある時間ごとに読み取った値の羅列として表される．オーディオデータのフォーマットには，圧縮方法によって，主に非圧縮データ，可逆圧縮データ，非可逆圧縮データの格納に用いられるものがある.非圧縮形式は，録音対象が複雑な音楽でも全くの静寂であっても，単位時間当たりに同じ量のビットを記録する．

多数のオーディオに基づいた鼻歌検索では非圧縮データの「WAV」を使っておる．そのほか,「AIFF」も非圧縮データである．非可逆圧縮は，一般には元データを復元することができない．音響心理学等様々な技法を使用し，可聴域にない音や，ある音でマスクされて聴き取り難い音を省いて圧縮するため，同じ音源のファイルより数分の一のサイズになるが，体感的な音質はそれなりに保たれる．このフォーマットが最も利用している形式である．その中の「MP3」は 1 番認知度の高いファイル形式である．その後継として，より高音質を実現させるために生まれた「AAC」がよく使用される．他にも「WMA」,「Vorbis」なども非可逆圧縮方式である．可逆圧縮は元データと同一のデータを保持したままサイズを削減することができる．再生時は解凍され元の非圧縮形式に戻ることができるため，音質面ではオリジナルのデータと変わらないということになる.この形式で最も一般的なのは「FLAC」, Apple では「ALAC」という独自のフォーマットを採用している．

一方，MIDI は musical instrument digital interface の略で，音の高さや発音のタイミングなどの演奏情報を格納したデータ形式である．例えば，ピアニストがピアノの鍵盤を指で押さえたり,ペダルを踏んだりする．この「動作」が「演奏情報」で，つまり「どのように弾いたのか」を表す．このような情報を電子的に表したものが「MIDI」である．ピアノでト「C」の音を強く弾いた場合,「60」の鍵盤を「120」の強さで弾いた，といった具合に,MIDI では演奏情報がすべて数字で表される．この演奏情報は，MIDI 対応の楽器間で送受信したり，SMF (Standard MIDI File) 形式のファイルとして保存しておいて，楽器上だけでなくコンピュータ上で再生したりもできる．

コンピュータに格納された MIDI データをバイナリ形式で見ると,MIDI データはヘッダ部とデータ部によって構成され，ヘッダ部には MIDI のデータフォーマット，チャンネル数，テンポ値などが記録され，データ部には音の情報(時刻, 音高, 長さ)やコントロール情報(ピッチベントやサスティーンペダルなど)が格納されている.MIDI データは電子楽譜に近い情報をもつとも言える．それゆえ，オーディオデータに比べると，データサーズは格段に小さくなる．MIDI の情報には，MIDI チャンネルという 1 から 16 の番号が割り当てられている．この MIDI チャンネルを使って，1 本の MIDI ケーブルで同時に 16 パート用の情報を送る仕組みになっている．これは「16 種類の楽器を同時に鳴らせる」ということを表す．

ファイルとして保存された MIDI データは，オーディオデータよりサーズが小さいだけでなく，編集できるという特性が持っている．つまり，演奏を間違えた箇所を修正したり，テンポを変えたり，移調したりなどが容易に行なえる．このため, MIDI を活用することで，音楽制作や楽器の練習が効果的に行なえる．

3. 鼻歌検索に使われる音楽特徴量

音楽情報処理には，音楽音響信号から低次の特徴量抽出から，音高，音長，和音などの音楽単位の特徴量がある．例えば，音程，音長差，基本周波数などである．これらの基礎的な特徴量から，音楽の三要素であるメロディ，リズム，ハーモニーに基づいたもっと高次の特徴量を抽出することが不可欠である[9]．音楽情報処理において，よく使われる特徴量は，平均平方２乗エネルギー(Root Means Square Energy)，ゼロ交差率(Zero Crossing Rate)，短時間フーリエ変換（Short Time Fourier Transform: STFT），Constant-Q 変換，ログメルスペクトログラム(Log-melspectrogram:MSPEC)，メル頻度ケプストラム係数（Mel Frequency Cepstrum Coefficients:MFCC），Beats Per Minute(BPM)，Chroma Energy Normalized(CENS)などがある．これらの特徴量を基に，音楽の内容の意味付け，音楽データの解析などが行われ，鼻歌検索のための技術を実現していく必要がある．

3.1. メル尺度

メル尺度は音高の知覚尺度である．メル (mel) という名称は,「メロディ (melody)」に由来し，音高の比較に基づく尺度である[9]．音響学には，音高は一定時間に振動が何回あるかで決まる．1 秒間に周期が何回あるかを「周波数」と呼び，Hz（ヘルツ）という単位で表す．人間の聴覚には，周波数の低い音に対して敏感だが，周波数の高い音に対して鈍感という特性があるため，人間は線形スケールで周波数を知覚しない．例えば, 人間は 100Hz と 200Hz の音の違いを簡単に見分けることができる．ただし，同じように，1000Hz と 1100Hz の違いを見分けることが難しい．

3.2. クロマ特徴量

クロマ特徴量は，音高−時間表現およびピッチクラス−時間表現である[11]．メル尺度に基づいた特徴とは異なり，クロマ特徴は直接に音楽の十二音階を利用す

ることで，音楽の本質をより良く表現することができるため，鼻歌検索などの音楽情報検索によく応用されている．

クロマベクトルを求めるには，ある音階の振幅強度を求めるためには，その音階の周波数付近の振幅強度の平均値を算出する必要がある．音響信号から周波数成分を抽出するために，フーリエ変換を用いることが多い．フーリエ変換による周波数解析では，直交周波数分割多重方式をも用いて行う．直交周波数分割多重方式とは，デジタル変調方式の一つで，隣り合う周波数の搬送波同士の位相を互いに直交させて周波数帯域の一部を重なり合わせ，高密度な周波数分割を行う手法である．この直交周波数の間隔は，解析する区間の時間長を T とすると，1/T(Hz)となるため，周波数分解能を高めるためには，ある程度の時間フレーム長が必要となる．特に低音域では，微小の周波数の変化で音程が変わるため，隣接する音程の周波数の差よりも細かい周波数分解能でのフーリエ変換が必要になる．

3.3. 短時間フーリエ変換 (STFT)

信号の性質に関しては，定常信号と非定常信号に分類される．音楽信号などの非定常信号は，複数の正弦波状の成分を含む周波数構造を持ち，それは時間と共に変化する．確定定常信号は，離散的周波数成分から構成されている．これはさらに基本周波数とその整数倍の周波数からなる周期信号と，2 つ以上の独立した基本周波数とその高調波成分が混合した疑周期信号とに区別される場合がある．

短時間フーリエ変換(STFT)は，音声波形に対して窓関数をずらしながらかけ，窓関数で切り取られた区間それぞれをフーリエ変換することである[20]．窓関数は，ある有限区間以外で 0 となる関数である．ある関数や信号に窓関数が掛け合わせられると，区間外は 0 になり，有限区間内だけが残るので，無限回の計算が不要になり数値解析が容易になる．窓関数は他の関数にかけて利用する．窓関数を使うと，ある関数のある区間をそのまま，もしくは加工して切り出すことができる．音声信号の分析では，音声区間全体でのフーリエ変換を行うのでなく，ある区間で時間窓をかけて周波数分析を行う．関数に窓関数を掛けると，関数のある区間が切り出されるため，窓関数を掛けた関数をフーリエ変換すると,有限期間のフーリエ変換ができる．すフーリエ変換を行えば，周波数成分がわかるが，切り出した信号がその区間で波の形になることはない．そこで，窓関数をつかって両端をつなげ，周期関数とみなして分析を行う[17]. そして, 時間窓をシフトしながら繰り返し分析を行って，スペクトルの時間的な変化を見える．

窓関数は何種類がある．これは，複数のメインローブ，サイドローブが合成された場合に，周波数特性の解析観点での性質が変わるためである．窓関数には，メインローブが狭い（周波数分解能が良い）とサイドローブが低い（ダイナミックレンジが広い）という二つの特長が要求される．しかし, この二つはトレード・オフの関係にあり, 両立させるには限界がある． そのため，ある状況では最適だった窓関数が，別の状況ではそうではないということも起こる．音声信号の分析では，ハン窓，ハミング窓がよく使われる

STFT は統計的特性が時間により変化する非定常信号の分析に使われる信号処理手法の一つであり，非定常信号の周波数成分の時間変化を捉えるために，短時間毎に信号を切り出しフーリエ変換したものである．切り出し時間窓とフーリエ変換の長さを別に設定することにより必要な周波数分解能を保ったまま時間分解能を良くする工夫がなされている．実質的には，STFT で時間により動くウィンドウを使って分析する信号のいくつかのフレームを抽出する．時間ウィンドウがとても狭い場合，抽出された各フレームは，フーリエ変換が使えるように定常に表示される．時間軸に沿ってウィンドウが移動して行くにつれて，周波数と時間の相違の関係が見えてきている．

3.4. ログメルスペクトルグラム (MSPEC)

通常の信号波形は多数の異なった正弦波が合成されたものと見ることができる．このとき周波数の関数として，各正弦波の周波数成分の振幅，または振幅および位相を複素数によって表したものを周波数スペクトラムという．スペクトラムは周波数，信号成分の強さという二つの要素を表す二次元信号である．スペクトログラムとは，スペクトラムをシーケンスに並べたものであり，音声をはじめとする様々なシーケンスデータの分析に使われる可視化手法のことである[21]. このスペクトログラムは時間，周波数，信号成分の強さという三つの要素で構成された三次元グラフである. スペクトログラムの形式は様々なバリエーションがある．最も一般的な形式では，横軸が時間を表し，縦軸が周波数を表す．そして，各点の明るさや色である時点のある周波数での強さを表す．

ログメルスペクトログラムとは，周波数領域ではなくメル尺度で音声のスペクトログラムである．メルスケールスペクトログラムはスペクトログラムとメルスケール変換の組み合わせであり，音声信号にメルフィルタバンクを適用した特徴量として使われる．周波数がメル尺度に変換された STFT のこととも言える．メルスペクトルグラムの抽出は以下のプロセスがある．

3.5. メル周波数ケプストラム係数 (MFCC)

メル周波数ケプストラム係数(MFCC)は人間の音高知覚を考慮し，オーディオデータに基づく鼻歌検索に

よく使われる特徴量である[16]．メル尺度に基づくケプストラムの係数ということである．

ケプストラムとは，スペクトラムのアナグラムによる導出語であり，音声波形をフーリエ変換して得たパワースペクトルについて，その値の対数をとり，さらに逆フーリエ変換した結果を指す．

音声信号は，声帯の振動や摩擦による乱流などの音源信号に，声道，口腔，鼻腔の形状などによって決まる調音フィルターが畳み込まれたものである．周波数領域では，調音フィルターの振幅伝達特性が音源信号のパワースペクトルに比べて滑らかに変化する関数であるという性質を用いて両者を分離することである．

ケプストラムの変数の次元は時間と同じになるが，これにはケフレンシーという言葉を用いることが多い．これは周波数のアナグラムである．信号$x(n)$のフーリエ変換を$X(e^{jw})$とすると，ケプストラム$c(m)$は

$$c(m) = \frac{1}{2\pi}\int_{\pi}^{2\pi} log\left|X(e^{jw})\right|e^{jwn}dw$$

音楽信号処理の場合，ケプストラムで低次成分において個人差の大きいピッチ成分を除去して，音韻の特定にとって重要な声道の音声特性(すなわち，口腔の形)のみを抽出できる．高次成分において声道の特性を除去して，ピッチ成分を抽出できる．

メル周波数ケプストラム係数(MFCC)は，ケプストラムそのものではなく，ケプストラムと同じケフレンシー領域に定義される低次元のスペクトル情報である．人間の聴覚上重要な周波数成分が引き伸ばされて，ケプストラム全体における割合が増える，メル周波数ケプストラムの特徴量の次元数が減り，計算の負荷が減ることができるといった利点があるため，鼻歌検索でよく使われている．しかし，雑音のスペクトルが特定の帯域に集中している場合，ケプストラムのすべての係数に影響を及ぼす特徴もある

3.6. Constant-Q 変換 (CQT)

音楽では，すべての音符が数オクターブの十二音階で構成されているため，十二音階は，ピアノの 1 オクターブの 12 の半音に対応する．これらの半音間の周波数比は$2^{1/12}$である．明らかに，同じピッチの 2 オクターブの場合，高オクターブは低オクターブの 2 倍の周波数になる．たとえば，西洋音楽の基本周波数（F0）は，次のように定義できる．

$$F_k = 440Hz \times 2^{\frac{k}{12}}, k\epsilon[-50,40]$$

したがって，音楽では，音は指数関数的に分布するが，フーリエ変換によって得られたオーディオスペクトルは線形に分布し，二つの周波数は一対一で対応させることはできない．そのため，数多くの音楽分析では Constant-Q 変換(CQT)を用いている[15]. 音楽分析では，Constant-Q 変換は音源分離，音楽信号分析，およびオーディオエフェクトなど，幾つかの方面に役立っている．

3.7. Chroma Energy Normalized (CENS)

STFT と CQT によるクロマ特徴の量は，多くの類似した曲の同じ特徴を明らかにしているが，特性にはまだ多くの違いが存在している．したがって，さらに量子化と平滑化を適用して，局所的なリズム，テンポの変化による局所的な変動の影響をさらに減らす手法―CENS(Chroma Energy Normalized) が提案されている[4]．CENS を文字通りで日本語に翻訳すると，クロマエネルギ正規化統計である．クロマバンド内のエネルギー分布に関する短時間の統計を考慮することにより，さらに高度な抽象化を追加すると，スケーラブルで堅牢なオーディオ機能のファミリーを構成する CENS 特徴量が得られる．CENS は，大きな窓関数で統計を実行して，速度，明瞭さ，音楽の装飾（ビブラートやアルペジオコードなど）の局所的な偏差を滑らかにすることであり，オーディオマッチングや類似性などのタスクに最適である[5]．

4. 評価実験

各音楽特徴量を用いて鼻歌検索することで，特徴量による検索の精度を別々に評価するための実験方法と実験結果について述べる．

4.1. 実験データ

音楽データベースには，ネット上と MIREX2020(音楽情報検索コンテスト)上でダウンロードしたポピュラー音楽，クラシック音楽，純音楽を含む 200 曲を用いた．MIREX2020 は第 16 回音楽情報検索評価コミュニケーションで公開されたデータセットであり，その中の一部のクラシック音楽，純音楽を採用している．ネット上の部分は，NetEase Cloud Music 上の一部のポピュラー音楽を採用いている. フォーマットは WAV にする．

GarageBand でレコーディングした 26 のハミングを入力クエリとしている. フォーマットは WAV にする．

評価するために，部分音楽片の長さ(ウィンドウ長)を統一する必要がある．本実験では，すべての楽曲データとハミングデータのうちの 10 秒間分を用いた．オーディオチャンネルは統一的に「モノ」としている．

4.2. 特徴量抽出

データを特徴量に変換する作業は特徴量抽出と呼ばれる．本実験では，オーディオデータを短時間フーリエ変換(STFT)，ログメルスペクトルグラム(MSPEC)，メル周波数ケプストラム係数 (MFCC)，Constant-Q 変換(CQT)，Chroma Energy Normalized (CENS)の 5 つの特徴量を抽出した．その中で，MSPEC

とMFCCは基本周波数をメル尺度に変換して用いた特徴量であるため，クロマに対応していない．STFT, CQT, CENS は，メル尺度に変換するプロセスはない特徴量であるため，その３つの特徴量をクロマ特徴量として抽出することである．各特徴量行列式の shape は(各特徴量の次元数,時間フレームの長さ)である．時間フレームの長さはオーディオデータの標本点数をフレーム周期に相当する標本点数で割った数の次の整数(例えば 9.23→10)である．時間フレームの長さの算出は次の式による．

$$\text{時間フレーム長} = \left\lceil \frac{sampling\ rate \times time}{hop\_length} \right\rceil$$

sampling rate は毎秒のサンプリング数を表す．音楽業界の標準としては 44100 としておる．time は本実験で利用されたデータの部分音楽片の長さとした 10s にしている．hop_length は連続するフレーム間のサンプル数を表し，普通にフーリエ変換での窓長の 1/4 としている．フーリエ変換において，窓長は 2 の冪乗(512，1024，2048)とすることが多い．本実験では，窓長は 2048 として，hop_length は 512 となる．したがって，時間フレーム長の計算結果は 862 である．

(1) ログメルスペクトルグラム(MSPEC)

MSPEC の抽出では，librosa.feature.melspectrogram を用いた．音の周波数[Hz]をメル尺度に変換する際に，特徴量の次元数を落とし，低周波成分ほど分解能を高く，高周波成分になるほど分解能を低くする．メル尺度の既定階数(次元数)n_melは 128 であり，各フレームの周波数成分が 128 個のメル周波数に対して抽出されていることを意味している．MSPEC 特徴量のデータは(128,862)の行列式である．

(2) メル周波数ケプストラム係数(MFCC)

MFCC の抽出では，librosa.feature.mfcc を用いた．MFCC の次元数n_mfccは 12~24 がよく使われる．n_mfccを大きい値にすると，メル周波数スペクトル包絡のより細かい成分まで考慮することができる．一般的に，MFCC の次元数は 12 または 20 である．本実験では，n_mfcc =20 としており，MFCC 特徴量のデータは(20,862)の行列式である．

(3) 短時間フーリエ変換(STFT)

STFT の抽出では，librosa.feature.chroma_stft を用いた．先に STFT を計算して，音楽の十二音階に基づいたクロマベクトルに変換したため，次元数は 12 となっておる．クロマの STFT 特徴量のデータは(12,862)の行列式である．

(4) Constant-Q 変換(CQT)

CQT の抽出では，librosa.feature.chroma_cqt を用いた．CQT は低周波数で分解能が良い特性を持っている．音楽以外の要素の干渉を排除することができるため，STFT より明らかになっていると見える．クロマの CQT 特徴量のデータは(12,862)の行列式である．

(5) Chroma Energy Normalized (CENS)

CENS の抽出では，Librosa.feature.chroma_cens を用いた．CENS はより長い窓長で統計し，平滑化することで，テンポ，アーティキュレーション，およびトリルやアルペジオなどの音符グループの実行における局所的な偏差を滑らかにすることである．STFT と CQT より安定性を持っていると見える．CENS 特徴量のデータは(12,862)の行列式である．

4.3. 類似度計算

ハミングを入力際には，歌の把握がたりないため，テンポをオリジナル曲により早くなったり，遅くなったりすることが多い．したがって，もっと柔軟性を持っている DTW 距離を選用している[18].

図 1 は「風の住む町」ー中村雅俊という純音楽のオリジナル曲とそのレコーディングしたハミングの一部のシーケンスデータの DTW 計算の視覚化である．紫色の線はオリジナル曲を表し，オレンジ色はハミングを表す．ハミングしたデータを見ると，一番目のピークはオリジナル曲より早くなって，二番目のピークはオリジナル曲より遅くなったことが見える．DTW 距離を計算する際に，同じ形になっている部分が対応するようなパスになっている．このように位相がずれているだけ実際は似ているデータを直感的な類似度を表すことができる．

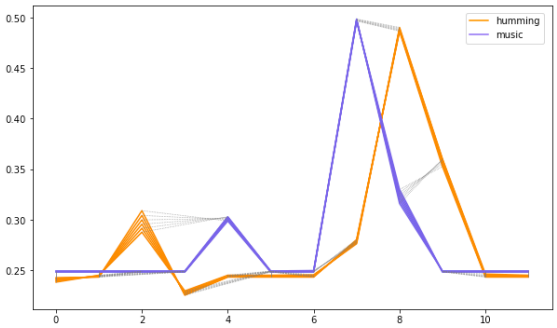

図 1 DTW 距離の計算例

個別評価

個別評価では，正解率，MRR 値，平均抽出速度という 3 つの評価指標で鼻歌検索において各特徴量の検索精度と抽出速度を評価する．

正解率 (Accuracy)は評価用データに含まれるすべてのサンプルのうち，正解を当てることができたサンプルの割合を示す指標である．本実験では，Top1, Top5, Top10 を別々に評価した．Top1 はクリエと最も類似した楽曲として正解の楽曲が選ばれたもの，Top5 はクエリと類似する上位 5 曲に正解楽曲が存在したもの，Top10 は上位 10 曲に正解楽曲が存在したクエリの割

合を表す.

MRR (Mean Reciprocal Rank) は複数のクエリとそれに対応するランキングが与えられているときに，そのランキングを与える検索アルゴリズムの性能を評価する指標である．Reciprocal Rank は，順位の逆数のことを表す．クエリ集合を Q とすると，MRR は次の式で書ける．

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

$rank_i$はランクのi番目は正解曲であると表す．MRR 値の逆数値は，ランクの調和平均に対応している．

平均抽出速度．鼻歌検索をする際に，特徴量の抽出速度は検索効率と直接関係している．各特徴量の抽出速度は有意差が存在しているため，特徴量の平均抽出速度を評価指標の一つとしている．平均抽出速度は抽出数と抽出時間の比である．

4.3.2. 総合評価

総合評価は，本実験で評価したすべての特徴量を組み合わせて評価することである．各特徴量により算出した類似度を「Min-Max normalization」の方法で正規化し，正規化した結果を加算した結果を総合類似度とする．データベースのすべての楽曲(A で表す)の中の楽曲 Q とハミング入力 H の類似度を$sim(Q,H)$とすると，正規化値$sim_{norm}(Q,H)$の計算式は以下のように与えられる．各特徴の検索精度の差を考慮しており，重みつけが必要になっている．本実験では，各特徴量の MRR 値とすべての特徴量 MRR の和の比の値を重みにする．各正特徴量による類似度の正規化値の重み付け平均を総合類似度$sim_{comp}(Q,H)$とする．

4.4. 実験結果

4.4.1. 正解率による評価結果

表 1 と図 2 は，各特徴量によるランクした結果を統計し，Top1，Top5，Top10 の結果と棒グラフで可視化することである．

表 1 正解率の評価結果

	Top1	Top5	Top10
STFT(クロマ)	53.84%	61.53%	69.23%
MSPEC(メル	23.08%	34.62%	42.31%
MFCC(メル尺	38.46%	42.31%	50.00%
CQT(クロマ)	50.00%	57.69%	73.08%
CENS(クロマ)	65.38%	76.92%	84.62%
総合評価	80.77%	88.46%	92.31%

図 2 より，鼻歌検索において，クロマ特徴量はメル尺度に基づいた特徴量に比べてより高い正解率を持っていることが示している．MFCC は MSPEC により離散コサイン変換のプロセスを追加したため，Top1 の結果は MSPEC により明らかに上がっており，Top5 と Top10 も小幅に上がっている．クロマの CQT を用いた検索の精度は STFT と比べると，Top10 の正解率が高いが，Top1 と Top5 は STFT より少し低いことが示している．CENS は量子化と平滑化のプロセスが追加しているため，検索の正解率が一番高いと示している．

総合評価により検索の正解率は個別の特徴量に比べると，Top1，Top5，Top10 は全面的に向上していることが示している．

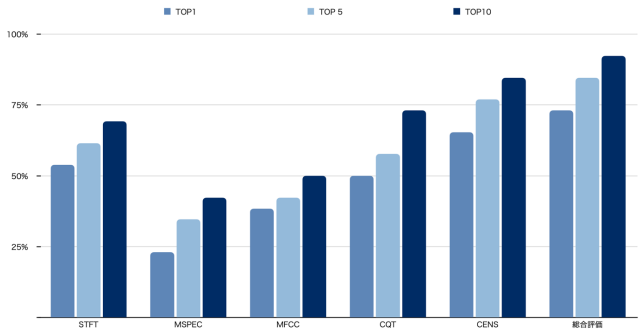

図 2 正解率の評価結果

4.4.2. MRR 値による評価結果

図 3 は各特徴量による検索したランクを基づいて計算した MRR 値を棒グラフで表示することである．

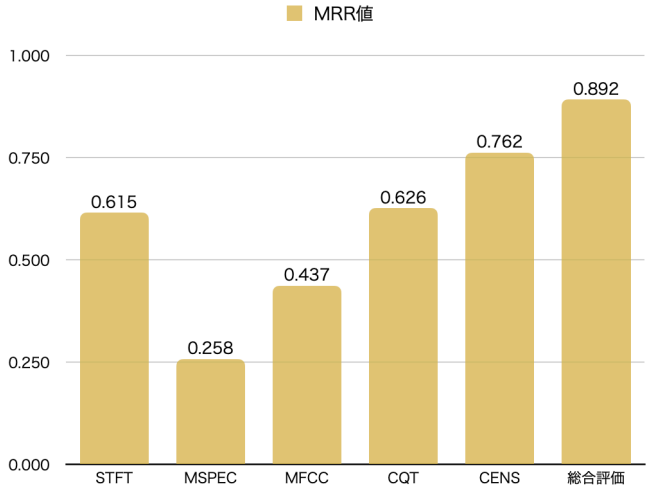

図 3 MRR 値

図 3 より，クロマ特徴量による鼻歌検索の MRR 値はメル尺度に基づいた特徴量よりはるかに大きいことを示している．したがって，クロマ特徴量はメル尺度に基づいた特徴量に比べてより高い精度を持っていることが分かる．メル尺度に基づいた特徴量を見ると，正解率では MFCC は MSPEC よりそれほど高くないが，

MRR 値から見ると MFCC の MRR 値は MSPEC よりはるかに大きいことを示している．この現象は MFCC が離散コサイン変換で，係数間の関係を減らすことで，ハミングの不正確性を大幅に改善し，鼻歌検索の精度を MSPEC に比べて大きく向上させていることが分かる．クロマ特徴量を見ると，クロマの STFT の MRR 値は CQT と大差ない．これは，CQT は STFT の冗長性を改善し，より音楽情報処理に適しているが，鼻歌検索では特に向上させていないことを示している．CENS の MRR 値は一番高い．これは，鼻歌検索において，CENS がほかの特徴量に比べるとより良い性能を持っていることを示している．

総合評価により検索は五つの特徴量の組み合わせであるため,正解曲のランク位は極端な値が少ない．そこで,総合評価の MRR 値は個別の特徴量より高い，検索の精度が個別の特徴量より向上されたと示している．

4.4.3. 平均抽出速度による評価結果

図 4 は各特徴量の平均抽出速度を棒グラフで表示することである．

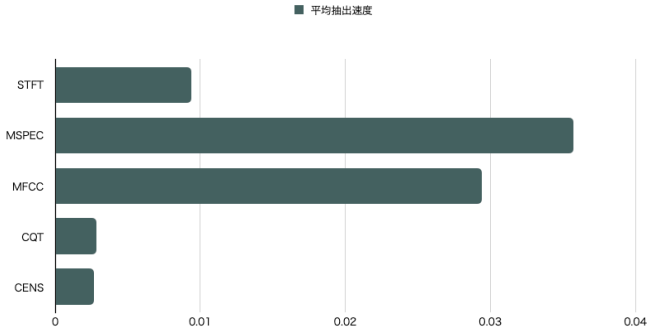

図 4 各特徴量の平均抽出速度

図 4 より，メル尺度に基づいた特徴量は，周波数をメル尺度によってメル周波数に変換すれば良いため，クロマ特徴量の抽出速度よりはるかに速いことを示している．クロマの STFT は本実験では比較的高い精度を示しているが，CQT と CENS に比べると抽出速度がはるかに速いことが分かる．計算コストが高いため，クロマの CQT と CENS の抽出速度は比較的遅いことを示している．

4.5. 考察

実験結果により，クロマ特徴量は鼻歌検索に適すであることが分かる．これはクロマ特徴量が音楽データの周波数とエネルギーを音楽の十二音階に格納することで，音楽の本質をより良く表現することができるためである．メル尺度に基づいた特徴量は，人間の聴覚特性を考慮しており，より早い時期に鼻歌検索によく利用されたが，音楽の形式とメロディが複雑になるにつれて，検索の性能がクロマ特徴量に比べて良くなくなってきた．クロマ特徴量の中で，CENS は音楽とハミングの特性をより安定的に表現することができるため，鼻歌検索に適していることが分かる．

評価実験をする際に，そのほかにも BPM，平均平方２乗エネルギー(RMSE)，ゼロ交差率(ZCR)といった幾つかの音楽情報処理によく使われる特徴量を抽出した結果，それらの特徴量は音楽の特性の一部を反映することができるが，鼻歌検索には適していない．

考察としてそれらの特徴量は鼻歌検索に適しない原因は以下だと考えられている．BPM は一分間の拍数を表し，音楽の三要素の中のリズムを表現できる特徴量であるが，低次元特徴量である．したがって，二つの楽曲のリズムは同じ可能性があるため，二つの楽曲から抽出した PBM は同じ数値になることが存在している．音楽の特性を十分的に表現できなく，特に鼻歌検索のような不正確性がある音楽情報検索には作用は微々たるものである．平均平方２乗エネルギ(RMSE)は，音声信号のエネルギーを良く表現できるが，音声信号の波形を見ると振幅が小さいが，実際は穏やかではない場合がある．したがって，RMSE を使うとソロパートなど，合奏に比べ音量が負けやすいところは数値が小さくでてしまうため，鼻歌検索に適しない．ゼロ交差率(ZCR)は，音声信号を時間領域の波形で見たときに，値が正負の入れ替わりを表現する特徴量であり，不安定さの表す手法とされている．しかし，高音は元々波の間隔が狭いため，高音が多い楽曲では適切に表現できない．

音楽データ量は増え続ける膨大な今は，従来の早い年に提案された簡単な特徴量は不十分である．そこで，これらの特徴量を音楽の三要素に関する特徴量と融合し，もっと次元数の高い特徴量を得て，鼻歌検索に適用することが求めていると考えられる．

5. 終わりに

音楽情報検索の一つとして，鼻歌検索は音楽データベースの楽曲およびハミングから数々の音楽特徴量を抽出し，その特徴量を用いて類似検索を行う．音楽特徴量は音楽信号の一つの側面，あるいは幾つかの側面から捉える特徴を数値化したものであるため，異なる音楽特徴量は，音楽の異なる特性を表現することができる．しがたって，音楽特徴量の選択は鼻歌検索システムの性能に直接的な影響を与える．

鼻歌検索のための音楽特徴量については，1990 年代半ばから数多くの研究が進んでいる．簡単な特徴量から，もっと次元数の高い特徴量を計算することで，音楽の特性をより良く反映し，鼻歌検索技術を向上させることを求められている一方，どのような特徴量が有効であるかを体系的な評価も少ない．

本研究では，鼻歌検索によく使われる音楽特徴量を調査した上で，評価実験を行った．結果，クロマベクトルに基づいた特徴量はメル尺度に基づいた特徴量より検索の精度が高いことを示している．その中で，CENS より検索の精度は一番高い，200 曲の小規模データベースに対して Top1 は 65.38%，Top5 は 76.92%，Top10 は 84.62%の検索精度を示している．さらに，5 つの特徴量を組み合わせて総合評価の結果は，Top1，Top5，Top10 が別々73.08%，88.46%，92.31%に達成した．抽出速度において，メル尺度に基づいた特徴量が速いことを示している．検索精度と抽出速度の両方から見ると，クロマベクトルに基づいた STFT のコスパが一番良いことを示している．

鼻歌検索によく使われている音楽特徴量は音楽特性をうまく反映することができるが，人の音感と歌唱力によって，調子外れに歌ったり，調子外れでなくでもオリジナルよる調子があがったり，下がったりする場面が多い．調子の変わりは必ずしもちょうどのオクターブとは限らない．この問題に対して，様々な研究もあるが，大規模データベースに対する鼻歌検索の技術は未熟で検索の精度も向上することを求められている．

今後の課題として，ハミングをオリジナルに還元するために，ハミングの中の個々の音符間の相対的関係も考慮に入れて，この相対的関係を注目し，全体的なパターンを推定できる特徴量を抽出することを検討すべきだと考えられる．

謝辞　本研究は令和 3 年度 KSU 基盤研究費による支援を頂いており，ここで謹んで感謝の意を表する．また，発表会で座長やコメンテータの方から有益なコメントをいただき、今後の参考になった．

参　考　文　献

[1] Huang, Y. P., Lai, S. L., Chang, T. W., Horng, M. S. (2012, June). Query-by-Humming/Singing of MIDI and Audio Files by Fuzzy Inference System. In 2012 Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing pp. 41-46. IEEE.

[2] McNab, R. J., Smith, L. A., Witten, I. H., Henderson, C. L., & Cunningham, S. J. (1996, April). Towards the digital music library: Tune retrieval from acoustic input. In Proceedings of the first ACM international conference on Digital libraries pp. 11-18.

[3] McNab, R. J., Smith, L. A., & Witten, I. H. (1995). Signal processing for melody transcription.

[4] Meinard Müller, Frank Kurth, and Michael Clausen: Chroma-Based Statistical Audio Features for Audio Matching. Proceedings of the Workshop on Applications of Signal Processing (WASPAA), USA, 275–278, 2005.

[5] Meinard Müller and Sebastian Ewert: Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. Proceedings of the International Conference on Music Information Retrieval (ISMIR), Miami, Florida, USA, pp. 215–220, 2011.

[6] D. Byrd and T. Crawford. Problems of music information retrieval in the real word. In Information Processing and Management, pp 249–272, 2001.

[7] J. Shen, J. Shepherd, and A. H. Ngu. Integrating hetegrogeneous features for efficient content-based music retrieval. In Proc. 13th CIKM Conference, 2004

[8] 小島千か. 音楽鑑賞の指導と評価に関する実践的研究―西洋音楽における音楽の諸要素と視覚的イメージの関連に着目して. 音楽教育実践ジャーナル 5.2 (2008): 142-149.

[9] 奥乃博，北原鉄朗，吉井和佳. 楽曲の特徴量抽出と検索技術. 電気学会誌 127.7 (2007): 417-420.

[10] 小川樹，森勢将雅. メルケプストラムを加工した音声の音質を計測する知覚モデルの開発と評価. 電子情報通信学会論文誌 D 103.4 (2020): 205-214.

[11] 石田颯人，木村昌臣. 度数表記と Chord2Vec を利用した楽曲類似度指標の提案. 研究報告音声言語情報処理 (SLP) 2019.21 (2019): 1-5.

[12] 日本音響学会編，音響学入門ペディア，コロナ社，2017 年 3 月

[13] 川村 新，黒崎 正行，音声&画像の圧縮/伸長/加工技術，CQ 出版，013 年 4 月

[14] 小杉尚子，櫻井保志，山室雅司，串間和彦. (2004). 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, 45(1), 333-345.

[15] 小舘亮之. オーディオ信号の低ビットレート分析/合成符号化. 情報処理学会研究報告オーディオビジュアル複合情報処理 (AVM) 1996.17 (1995-AVM-012) (1996): 49-56.

[16] 辻俊明，佐藤航陽，境野翔. (2021). 力覚情報のメル周波数ケプストラム係数に基づく接触動作の認識. 日本ロボット学会誌, 39(2), 173-176.

[17] 小野順貴. (2016). 短時間フーリエ変換の基礎と応用. 日本音響学会誌, 72(12), 764-769.

[18] 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 (TOD) 45.SIG07 (TOD22) (2004): 163-178.

[19] WAV ファイルや MIDI ファイルについて - ミュージック CD
https://www.megasoft.co.jp/mcdd/gogo/music_a_gogo_d09.html

[20] ゼロからわかる STFT
https://fast-d.hmcom.co.jp/blog/stft-from-zero/

[21] Python で音の STFT 計算を自作！スペクトログラム表示する方法
https://watlab-blog.com/2019/05/19/python-spectrogram/