様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２０ 日現在

機関番号：３４４０６
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１３２１

研究課題名（和文）多様な歌唱入力に対応した楽曲検索システムの開発

研究課題名（英文）Query-by-Singing music information Retrieval system supporting various singing style

研究代表者
鈴木　基之（Suzuki, Motoyuki）

大阪工業大学・情報科学部・教授

研究者番号：３０２８２０１５

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では歌唱音声を入力とした楽曲検索システムの構築を目的とし，様々な要素技術の開発を行った。まずは歌唱音声を高精度に認識する方法を開発した。歌唱は音符と音韻が対応していることから，音符の区切り時刻を利用した認識法を開発し，精度を向上させた。
次に誤りを含む歌詞に頑健な検索方法を開発した。認識誤りに対しては単語系列ではなく音素系列を利用し，人間の記憶誤りに対しては，誤り傾向を検索スコアに反映させることで検索精度の向上を達成した。
最後にメロディと歌詞それぞれから得られた検索結果を組み合わせる方法の検討を行った。検索位置をあわせることで精度向上を目指したが，期待されるほどの成果は得られなかった。

研究成果の学術的意義や社会的意義
歌唱音声認識の精度向上に大きな貢献をした。歌唱音声の認識が難しい事は従来から知られていたが，音響モデルや言語モデルを適応させる程度しか対処法が提案されていなかった。本研究では音符の区切り時刻を利用する，という新たな発想を取り入れ，認識精度を大きく向上させることができた。更に歌唱音声において任意の位置に無音区間が挿入される可能性があること，それが認識性能を劣化させる大きな原因であった事を初めて明らかにした。
また，歌詞を用いた楽曲検索において，認識誤りだけではなく，人間の記憶誤りにも注目し，適切に対処を行うことで検索精度を向上させたことも大きな貢献である。

研究成果の概要（英文）：In this research, we had developed various elemental technologies with the aim of constructing a music retrieval system using singing voice as input.First, we had developed a method to recognize singing voice with high accuracy. Since a note generally corresponds to a mora in singing voice, we developed a recognition method that uses note boundary information to improve the accuracy.
Next, we developed a robust retrieval method for lyrics containing errors. For recognition errors, we used phoneme sequences instead of word sequences, and for human memory errors, we improved the retrieval accuracy by reflecting error tendencies in the retrieval score.
Finally, a method for combining the retrieval results obtained from both melody and lyrics was studied. We tried to improve the accuracy by matching the retrieval positions, but did not achieve the expected results.

研究分野：音声情報処理

キーワード：楽曲検索　歌唱音声認識　歌詞誤りに頑健な検索　歌詞の記憶誤り

１．研究開始当初の背景

Web からの楽曲の購入等の場面で楽曲を検索する場合，そのほとんどは曲名や歌手名といったメタ情報をキーとして検索が行われている。しかしこの方法では，「曲は知っているけど，曲名や歌手名はあやふや」といった楽曲の検索は難しい。またそもそも楽曲を検索する場面においては，ユーザは楽曲を思い浮べていると思われるため，メタ情報による検索ではなく，楽曲自体を検索キーとして入力できた方が，より自然なインターフェースであると言える。

楽曲自体を検索キーとして使用する場合，ユーザが楽曲の一部を歌唱して入力することになる。このようなシステムは Query-by-Humming/Singing と呼ばれ，2000 年代前半に世界的にさかんに研究が行われた。これらのシステムのほとんどは歌唱データから音楽的特徴量（音の高さや長さ）を抽出し，それを検索キーとして用いている。

一方で，歌唱の多くでは歌詞が用いられているにもかかわらず，従来の検索システムでは歌詞の情報は用いられていない。その主な理由は，歌詞情報を正しく抽出できない事にある。歌詞情報は歌唱音声を音声認識する事で得られるが，歌唱音声は通常発話と異なり，独特の発声法やリズムにあわせた発話長となるため，音声認識精度が非常に悪い。音響モデルを歌唱音声に適応させることである程度の精度向上を達成している研究もあるが，その精度はいまだ十分ではない。また，歌唱入力の際に歌詞の記憶が曖昧であることも多く，誤った歌詞で歌唱する可能性がある。こうした曖昧性を含む歌詞情報から，どう頑健に検索すればよいか，という問題を解決した研究は存在しない。

そこで本研究では，上記のような問題点を解決することで歌詞情報を積極的に検索に利用し，歌唱誤り等も含めて多様な歌唱入力に対応した高精度楽曲検索システムの構築を目指す。

２．研究の目的

本研究の目的は，前述した歌唱音声の認識に関する問題点を解決し，最終的に多様な歌唱入力に対応した楽曲検索システムを構築することである。システムの全体像を図 1 に示す。このシステムでは，検索対象となる楽曲データから，事前に音符の高さや長さの系列情報（メロディ情報）のデータベースと，歌詞データベースの 2 つを準備しておく。ここで歌詞データベースには，各曲についての正しい歌詞データだけではなく，誤りやすい歌詞の情報（研究課題 2）も登録しておく。歌唱音声が入力されると，そこからメロディ情報と歌詞情報（研究課題 1）を抽出し，それらをその信頼度を考慮した上で適切に統合（研究課題 3）して曲を検索する。

３．研究の方法

(1) 歌唱音声の高精度音声認識法の開発（研究課題 1）

歌唱音声では通常音声と比較して発話長（音符の長さに対応）が大きく異なり，それが誤認識の主な原因となっている。そこで本研究では，メロディ情報の抽出時に得られた音符の区切り時刻情報を用いて認識精度を向上させる方法を開発する。具体的には，音符の区切り時刻らしさを各時刻ごとに求め，その値を認識時に考慮することで，1 音符は（おおよそ）歌詞のひらがな 1 文字と対応する，という知見を認識システムに導入する。また歌唱音声では，任意の位置で息継ぎを行う可能性があるため，たとえ単語の途中であっても無音区間が生じる可能性がある。そこで，任意の位置に無音区間を想定した認識を行い，精度の向上を目指す。

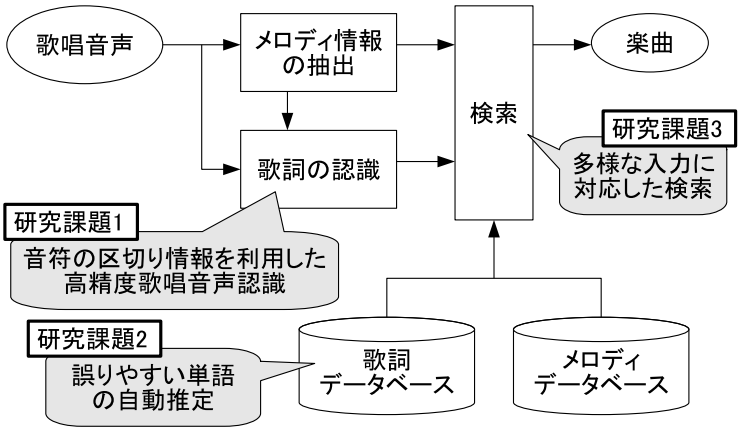

図 1 開発するシステムの全体構成

(2)誤りを含んだ歌詞にも頑健な楽曲検索法の開発（研究課題 2，3）

曖昧な記憶での歌唱において，歌詞はどのように誤る可能性があるのか，人間が行いやすい誤りに注目した検索方法を開発する。具体的には，発音が似ている単語や意味が似ている単語を事前に抽出し，それらの単語間距離を他の単語との距離よりも小さく設定することで，誤った単語が入力された時にも検索できるよう対処する。

また，通常は検索対象楽曲の歌詞のみから言語モデルを構築するため，そこに現れない単語が認識結果として出力されることはない。そのため，歌詞とは異なる単語で歌唱されると，その単語ではなく，言語モデルに登録されている単語の中から，比較的発音が似ている単語が出力されることとなる。この単語は意味的にも発音的にも歌唱された単語とはあまり類似していないため，検索にとって悪影響を及ぼす。そこで，一般の日本語から構築された言語モデルを用い，入力に忠実な（誤った単語を含む）単語系列を得た上で，そこから高精度に検索を行う方法を開発する。

(3)メロディ情報と歌詞情報を用いた検索システムの構築

歌詞情報から得た検索結果と，メロディ情報から得られた検索結果を統合し，最終的な検索結果を表示するシステムを構築する。その際，それぞれの検索位置が矛盾しないように検索結果を統合することで，偶然類似した部分があることでスコアが高くなってしまった候補を排除する。

４．研究成果

(1)歌唱音声の高精度音声認識法の開発

歌唱音声からの歌詞の認識が一般の音声認識と比べて精度が悪くなる原因のひとつは，単語の湧き出し誤りの増加である。一般に歌詞は楽譜にあわせて歌唱されるため，対応する音符の音長によっては，不自然に長く発声される。そのため，この部分で単語の湧き出し誤りが頻発することとなる。

この問題に対処するため，音符の区切り時刻らしさの情報を認識時に利用することとした。「区切り時刻らしさ（尤度)」を計算し，この値を音声認識に用いる特徴量ベクトルの次元を拡張して加える。対応する音響モデルには，すべての音韻の後ろに特殊な HMM を挿入した。「区切り時刻らしさ」の値は 0 から 1 の範囲の値となるように正規化されているため，特殊な HMM の出力確率分布の平均値には 1 を設定し，区切り時刻に近い（高い値の）時に対応するようにした。一方，通常の音素に対応する HMM には 0 を設定した。なおどちらも分散の値はある程度大きい値としている。27 名の歌唱による歌唱音声 198 データを用いて認識実験を行ったところ，認識率は 89.9%となった。一方，区切り時刻を一切使用しない場合は 85.7%であり，およそ 4 ポイント認識精度を向上させることができた。

こうして得られた認識結果を詳細に分析したところ，歌唱中に含まれる無音区間付近において誤認識が数多く観測されることがわかった。通常の音声においては，一般に単語や文節，フレーズといった言語的な「固まり」は一気に発声し，途中で息継ぎを行うことはない。そのため，音声認識システムにおいてもこうした「固まり」の途中で無音区間が挿入される事は想定していない。しかし歌唱音声においては，「固まり」の途中に休符が対応していたり，長い音符を最後まで伸ばさずに途中で息継ぎをしたり，といった現象が多く見られるため，想定していない無音区間が他の単語として認識され，結果として認識率の低下を招いていた。

そこで任意の音韻の後に無音区間が含まれてもよいように音声認識システムの改良を行なった。具体的には，すべての音韻の後ろに挿入している特殊な HMM に，無音区間に対応する音響モデル（sp モデル）を付加した。しかし単純に挿入してしまうと，すべての音韻の後ろで無音区間が存在しなければならなくなる。そこで sp モデル自体をスキップする遷移も挿入し，適切な方が自動で選択されるようにした。このようにして歌唱音声の認識実験を行なったところ，認識精度を 93.2%まで改善させる事に成功した。

(2)誤りを含んだ歌詞にも頑健な楽曲検索法の開発

一般に人間が歌詞を誤る時は任意の単語に誤るのではなく，ある程度誤り方に傾向があると思われる。そこで，ありがちな誤りとして，発音が似ている単語への誤りと意味が似ている単語への誤りの 2 つを仮定し，それぞれの誤りを許容した検索を行うシステムを開発した。具体的には，それぞれの誤りで出現すると思われる単語とは近い距離を定義し，その距離尺度を利用して検索を行うこととした。

発音が近い単語との距離は，それぞれの単語を音素記号列で表記し，その類似度を用いて単語間距離を定義した。しかし，同じ音素ひとつの誤りであっても，例えば「仙台」と「専売」のように発音が似ている組み合わせもあれば，「仙台」と「千枚」のように，あまり似ていない組み合わせもある。そこで音素を調音方式でグループにわけ，グループ内の音素同士は同じ音素と定義した上で音素間距離を DP によって求めた。なお，母音はすべて違う音素として扱った。

一方，意味が近い単語との距離は，word2vec を用いて数値化した。word2vec は同じような

文脈に表れる単語同士は似た意味を持つことを仮定し，様々な単語をベクトル化したものである。そこで，word2vec で得られたベクトル間のコサイン類似度を計算することで意味的な近さを計算した。

これらふたつの距離の重みつき和で単語間の距離を定義し，その値を用いて検索を行った。Web 上に実際に投稿されていた誤りを含む歌詞 66 曲分を用い，様々な長さに切ることで，擬似的に誤りを含む歌詞入力データを作成した。重みを変化させて検索を行ったところ，意味に対する重みを 0.8 に設定した時に検索精度が 74%と最も高くなった。一方，誤り傾向を考慮しない検索システムは 64%であり，10 ポイント向上させることができた。また検索結果として出力する曲数も平均で 1.11 個と，同率 1 位で複数出力されることが比較的少なく（誤り傾向を考慮しないシステムは 7.89 個），検索システムとして望ましい動作をしていることがわかった。

このように誤った単語が歌唱に含まれていた場合，歌唱音声認識の精度が大きく低下してしまう可能性がある。従来の歌唱音声認識では認識精度を向上させるため，検索用のデータベース（以下 DB）に合わせて音響モデル，発音辞書，言語モデルを作成している。この場合，DB に存在する単語しか認識しないため，誤った単語が歌唱に含まれると，その単語周辺で誤認識が多発することとなる。

そこで，このような歌詞誤りを含む歌唱をそのまま認識するため，歌唱音声認識システムにおいて発音辞書，言語モデルを大語彙に対応させることを検討した。この場合，入力歌唱に対して候補となる単語が増えることから，誤りを含まない歌唱に対しては誤認識が増え，認識精度が下がることが考えられる。そこで，まず言語モデル等を大語彙に対応させた時の音声認識精度を検証し，それがどれほど検索に影響するのかを検証した。その結果，歌唱誤りを含まない音声に対しては，認識精度が 50%まで低下し，検索性能も 91%から 80%まで落ちることがわかった。

誤認識の傾向を見ると，発音は似ている別の単語へと誤認識していることが多いことがわかった。そこで，こうした誤認識がおきても高精度に楽曲検索を行うため，事前に誤認識傾向を分析し，誤認識しやすい単語同士は距離が近い，として楽曲検索を行う方法と，単語ではなく，音素系列に変換した上で検索を行う方法のふたつについて，検索精度を比較した。その結果，認識結果を音素系列に変換して検索を行う方法は 93%となり，大語彙モデルを用いたとしても従来の言語モデルでの検索性能を上回ることがわかった

(3)メロディ情報と歌詞情報を用いた検索システムの構築

入力された歌唱音声から，メロディ情報と歌詞情報を抽出し，それぞれを用いて検索を行った後に結果を統合する方法を開発した。この方法を用いることで，歌唱音声中の歌詞やメロディに誤りが含まれていたとしても，両者が同時に起こらなければもう一方の検索結果と統合することで修正され，結果的により頑健な検索が行われる事が期待される。

この時，独立して行われた検索結果に対して統合が行われるため，検索スコアがそれぞれデータベースの曲中の「どこの位置」から計算されたのかが考慮されていない。例えばあるデータベース中の曲に対し，歌詞はその先頭付近に類似し，メロディは中盤付近に類似していたとしても，統合スコアはそれら両者の加算となるため，本来より過度に高いものとなってしまう。

そこで，それぞれの検索において「曲中の位置」の情報に注目し，位置に矛盾のないようにスコアを統合する方法を提案した。事前にデータベース中のすべての曲に対し，歌詞中の各単語がメロディのどの位置に対応しているか，といった対応表を作成しておき，検索時にはそれぞれ対応する位置にあるスコア同士を重み付きで加算する。その後様々な位置での統合スコアの最大値を，その曲に対するスコアとして定義した。なおこの方法においては，検索アルゴリズムの性質上，入力歌唱の最終単語の位置を対応するメロディの位置とあわせるだけであり，先頭単語の位置は必ずしも対応するメロディ位置とあっているとは限らない事に注意が必要である。

提案方法の有効性を示すため，歌唱音声を入力とした楽曲検索実験を行った。歌唱音声は男性 19 名，女性 8 名が童謡をアカペラで歌唱している 198 データを利用した。この時，実際に楽曲検索システムを利用する際には様々な長さで歌唱される可能性があることから，歌唱データを 1 小節ごとに分割し，それらを新ためていくつか連結することで，様々な長さの歌唱音声を作成した。検索システムには童謡 45 曲の歌詞情報とメロディ情報をデータベースに登録した。

検索実験の結果を見ると，歌唱音声の長さが 1 小節の時は位置あわせ有りが 64.5%に対して位置あわせ無しが 68.8%，2 小節の場合は 80.0%に対して 85.6%と，いずれも位置あわせを行わない方がよいという結果となった。なお，3 小節以上の長さがある歌唱に対しては，どちらもほぼ同等の結果となった。

こうしたデータについて詳細に分析を行ったところ，位置あわせの方法に問題があることがわかった。メロディ検索においては，楽譜における 4 分音符の位置ごとに検索スコアが計算される。一方歌詞検索においては，各単語ごとに検索スコアが出力される。この両者を統合する際に位置あわせを行うため，ある単語が楽譜において 4 分音符の位置と対応していない場合は，対応するメロディ検索スコアが存在しないために統合できず，結果として検索スコアは無視され

ることとなる。こうした事から検索精度が低下している事がわかった。一方で長さが 3 小節以上の場合は含まれる歌詞が長いことから，歌詞だけでほぼ正しい曲を検索できてしまうため，どの方法でも高い検索精度となっていた。

そこでメロディ検索の方法を修正し，より短い音符の位置ごとに検索スコアを計算するようアルゴリズムを修正した。計算量は増えるが，その分精密に位置合わせが実現でき，検索精度の向上が期待できる。計算位置を 8 分音符ごとに変更して楽曲検索実験を行なったところ，多少検索精度を向上させることができたが，それでも位置あわせ無しの精度を超えることはできなかった。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 鈴木 基之，杉田 裕亮	４．巻 61
２．論文標題 音符区切り情報を用いた高精度歌唱音声認識	５．発行年 2020年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 798-806
掲載論文のDOI（デジタルオブジェクト識別子） 10.20729/00204230	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Suzuki Motoyuki、Tomita Sho、Morita Tomoki	４．巻 -
２．論文標題 Lyrics Recognition from Singing Voice Focused on Correspondence Between Voice and Notes	５．発行年 2019年
３．雑誌名 Proc. INTERSPEECH 2019	６．最初と最後の頁 3238-3241
掲載論文のDOI（デジタルオブジェクト識別子） 10.21437/Interspeech.2019-1318	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 鈴木 基之，竹中 智美	４．巻 2019-MUS-123
２．論文標題 主旋律に注目したクラシック音楽の自動擬音語変換	５．発行年 2019年
３．雑誌名 情報処理学会研究報告 音楽情報科学	６．最初と最後の頁 1-5
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 鈴木基之，富田翔	４．巻 2018-MUS-119
２．論文標題 音符区切り位置の推定誤りに頑健な高精度歌唱音声認識	５．発行年 2018年
３．雑誌名 情報処理学会研究報告	６．最初と最後の頁 1-4
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

1．発表者名 Suzuki Motoyuki
2．発表標題 Lyrics Recognition from Singing Voice Focused on Correspondence Between Voice and Notes
3．学会等名 INTERSPEECH 2019（国際学会）
4．発表年 2019年

1．発表者名 鈴木 基之
2．発表標題 主旋律に注目したクラシック音楽の自動擬音語変換
3．学会等名 情報処理学会 音楽情報科学研究会
4．発表年 2019年

1．発表者名 鈴木基之
2．発表標題 音符区切り位置の推定誤りに頑健な高精度歌唱音声認識
3．学会等名 情報処理学科 音楽情報科学研究会
4．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	杉田　裕亮 (Sugita Yusuke)		

6．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	富田　翔 (Tomita Sho)		
研究協力者	竹中　智美 (Takenaka Tomomi)		
研究協力者	森田　朋希 (Morita Tomoki)		
研究協力者	小野　桂太郎 (Ono Keitaro)		
研究協力者	竹之下　一真 (Takenoshita Kazuma)		
研究協力者	石橋　萌香 (Ishibashi Moeka)		
研究協力者	廣田　誠二 (Hirota Seiji)		
研究協力者	高橋　悠介 (Takahashi Yusuke)		

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	木村　秋人 (Kimura Akito)		
研究協力者	栗山　聖也 (Kuriyama Seiya)		

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関