論文　**Original Paper**

ランダムフォレスト法を用いた歌声が聞き手に与える印象の単語推定に関する研究

田　中　晶　之[*1]・中　村　嘉　志[*2]

A Study on Estimating Impression of Singing Voice using Random Forest Method

Masayuki Tanaka[*1] and Yoshiyuki Nakamura[*2]

Abstract: 歌声の印象の要因を突き止めて推定するためには歌声を聴いた際に感じる印象を分析することが必要である。本研究では，高評価された楽曲に付されている印象を表す言葉に着目し，歌声とその歌い方を単語の側面から分類する。分類した楽曲から特徴を量として抽出し，その特徴量を用いて歌声に対して聴き手が感じる印象を単語推定する手法を提案する。音響特徴量の抽出にはメル周波数スペクトラム係数（MFCC）を用い，機械学習のランダムフォレスト法を利用した分類器を構築し，印象を表す単語推定を行った。その結果，提案手法による一致率は全体としては高くなかったが，特定の単語推定においては８割以上の一致率が示された。この結果から，提案手法を活用すれば，同じような印象を与える楽曲検索に応用することが可能である。

Key words: 歌声情報処理，単語推定，ランダムフォレスト法，メル周波数ケプストラム係数，音響特徴量

1. はじめに

楽曲が聴き手に与える印象は，歌唱者の歌い方によって大きく変わる。ここで言う楽曲とはポピュラー音楽のことであり，現代までに多くの楽曲が作られている。その中で，聴き手が楽曲を聴き感動して涙を流す，といった現象が起きることがある。この現象は特にプロのアーティストのライブでみられることが多い。その場合，聴き手側のアーティストへの思い入れや楽器の音，会場の雰囲気なども関係していると考えられる。このように楽曲の印象というのは主観的なものであるがゆえに様々な要因が絡み合っている。

一方，プロではない人のアカペラの場合でも同様のことが起こり得ることが知られている［1］。聴き手へ与えられる情報はほぼ歌声とその歌い方のみであることから，楽曲が聴き手に与える印象は歌声と歌い方に要因が占められると考えられる。このとき，音楽理論を基にした歌唱採点システムの評価基準用に，音程やテンポの正確性や，こぶし・しゃくり・ビブラートといった歌唱テクニックではない要因が聞き手に伝わって印象を左右していると考えられる。なぜなら，これらの音楽理論に基づいた歌唱は，近年では音声合成技術によって正確に再現が可能であるが，そうして生成された歌唱は工学技術的な良さはあっても，ここで述べている印象とは異なるものだからである。

印象というのは人間の情動的な反応の末に定着するものである。したがって，印象の要因を分解して突きとめることは容易ではない。ここで，印象を表した言葉に着目してみる。現代ではアマチュアでも手軽に動画を投稿でき多くの人に歌唱を聴いてもらうことができる。特に「歌ってみた」動画と呼ばれる，様々な楽曲をカバーしてアマチュアが歌唱した動画投稿がされており，その中で再生数が多い歌唱には多くの聴き手のコメントが綴られている。すなわち，歌声の印象が言葉（以降，印象語）として表出されている。

歌声の印象の要因を突き止めることは容易ではないが，同じような印象が持たれる歌唱動画には同じような印象語を含むコメントが付される。これらの印象語に着目すれば，人間の情動の末に表出された言葉から歌唱動画を分類することができる。印象に至った要因を探ることは困難であっても，印象語に基づく分類によって歌唱に共通した特徴を抽出することは可能である。

そこで本研究では，高評価された楽曲を印象語の側面から分類し，それらに含まれる歌声と歌い方の特徴を分析する。分析した特徴を量として抽出すれば，機械学習によって楽曲が与える印象を推定することが可能となる。推定結果を客観的指標として利用することにより，

[*1]国士舘大学大学院工学研究科
[*2]国士舘大学理工学部／大学院工学研究科

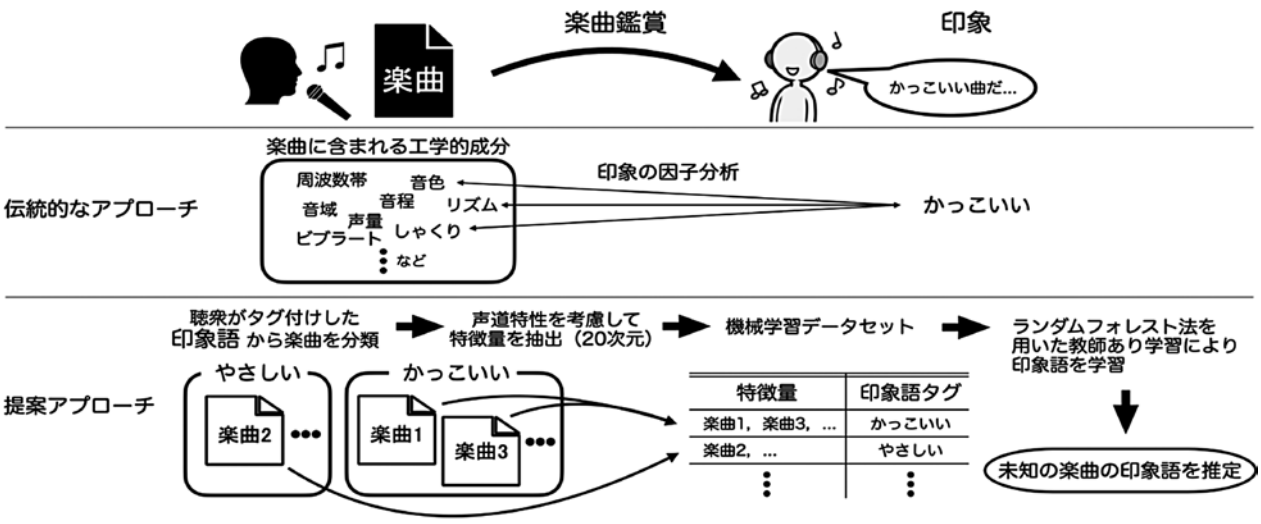

図1　本研究のアプローチ

歌声や歌い方によってどんな印象が持たれるかを事前に知ることができる。また，歌唱の仕方の改善や表現力向上に応用することも可能である。

本稿では，楽曲から歌声を抽出し，その音響特徴量を分析することで，印象を言葉として推定する手法と結果について述べる。特に，歌声を対象とする観点から演奏などの歌声以外を極力排除する目的で声道特性を用いて特徴量を抽出し，推定結果への影響を可能な限り減少させることを試みる。また，音響特徴量の分析と印象語の推定に機械学習モデルのひとつであるランダムフォレスト法［2］を用いて少ないデータ数でも推定した印象語が歌声にコメントとして付された印象語と一致するような分類器の構築を目指す。研究アプローチのイメージを**図1**に示した。

以降，本稿の構成は次のとおりである。第2章では，歌声の印象を推定する関連研究と，本研究で用いる音響特徴量，機械学習モデルのランダムフォレスト法について述べる。第3章では，本研究の方針と歌声の音響特徴量を分析し，印象を推定する手法とそれの実装について詳述する。第4章では，実験の結果を示して考察をする。第5章はまとめである。

2. 関連研究

2.1　先行研究

これまでにも歌声の印象を評価・推定する研究はなされている。印象推定において，金厲らは，歌声の印象評価度を構築に基づく多様な印象の自動推定手法を提案した［3］。これは，ポピュラー音楽におけるアマチュア女性歌唱者の歌声を対象として，歌声の多様な印象を適切に評価可能な評価尺度を構築し，印象得点と音響特徴量を用いた重回帰分析を行っている。実験には，評価者には未知のメロディ・歌詞であるオリジナルのメロディを作成・利用した。その結果，47種の印象評価語と歌声の印象評価に関わる3因子（迫力性，丁寧さ，明るさ）を歌声の音響特徴量と対応付けた。しかし，この研究の場合，男性における歌声の印象は考慮していない。本研究の目的である，楽曲データから抽出した歌声の印象を推定するためには，歌声以外も含まれているデータを用いるほか，女性の歌声だけでなく，男性の歌声も対象に入れて分析する必要がある。

また，歌声の評価において，山根らは，音声合成システムの音源データ検索のための声質評価推定を行った［4］。主観的な声質評価値に着目し，それを用いた音源データ（歌手の歌声）の検索方法について検討し，歌声合成用の音源データに対して，声質評価値を自動推定する方法を提案した。声質特徴量の抽出には統合確率密度に基づく手法を応用している。混合正規分布モデルを用いてモデル化することで音韻などの周波数特性の影響を極力受けない，声質を表す特徴量の抽出を実現している。また，評価値推定の回帰モデルには，カーネル回帰分析を用いている。それにより，年齢・性別に関する評価値推定で高い精度を得られることを示した。この研究に対し本研究では，年齢，性別に関係なく印象の推定をすることを目指す。

2.2　印象語

楽曲サイトや動画投稿サイトでは，一般人の楽曲に対する印象を表すコメントがされている。本稿では，先行研究［3］を参考に，歌声の印象を表す言葉を印象語と呼んでいる。この印象語は楽曲全体に対して付けられているものも多いが，歌声に対して付けられているものも十分にある。このことから，歌声に付けられた印象語のコメントを集め，一番コメントが多かった印象語を歌声の音響特徴量と結びつけることで，分類する際のラベルとして学習に用いる。**表1**に，本研究で用いる印象語の一覧を示す。これらの印象語は楽曲に寄せられたコメントを

表1　印象語一覧

透明感のある	力強い
かっこいい	悲しい
かわいい	やわらかい
やさしい	

分析し，特に多かったものを選んでまとめたものである。

2.3　音響特徴量

楽曲に関する研究では，音響特徴量の分析が多く見られる。しかし，既存の楽曲から歌声を抽出して分析を行った研究は著者らの知る限り少ない。藤原らの研究［5］では，楽曲に含まれる種々の伴奏音が混在している音楽音響信号から，伴奏音を抑制した音響信号を抽出した。その抽出した音響信号から歌手名を同定する際の特徴量としてメル周波数ケプストラム係数（Mel-Frequency Cepstral Coefficients，以降，MFCC）を用いている。そこで本研究では，楽曲から直接抽出した歌声の音響特徴量として，人の周波数知覚特性を考慮した音響特徴量であるMFCCを用いる。

MFCCはメル周波数軸上で計算されるケプストラム係数である［5］。ケプストラム分析とは，スペクトラルの包絡と微細構造，つまり，声道特性（声帯から対外へ出るまでの空間での包絡な共振振動）と声帯振動を分離する手法である。ケプストラム係数は，ケプストラムに対し，対数パワースペクトルを離散コサイン変換することで計算される。スペクトル包絡はケプストラムの低次の係数に表現され，微細構造は高次の係数に表現される。メル周波数とは，人間の聴覚特性に適合した対数周波数軸のことである。MFCCの計算においては，まず，音声信号に対してメルフィルタバンクを行う。次に，対数を取り離散コサイン変換を行う。最後に，低次の係数を抽出する。本研究では，20次元のMFCCを用い，それに印象語を結びつけたものを聴き手が受け取る歌声の音響特徴量として用いる。

2.4　ランダムフォレスト法

本研究で扱う音響特徴量は，20次元のMFCCと印象語のみで構成されている。これを機械学習用データセットと見なしたときに，1つの印象語に対して20個のパラメータが存在することとなる。この20個のパラメータ数は学習用としては少ない。そのため，データセットのパラメータ数が少なくても分析可能なランダムフォレスト法を本研究では用いる。

ランダムフォレスト法は機械学習手法の一つであり［6］，複数の決定木を用いたアンサンブル学習である。図2に，ランダムフォレスト法のイメージを示した。

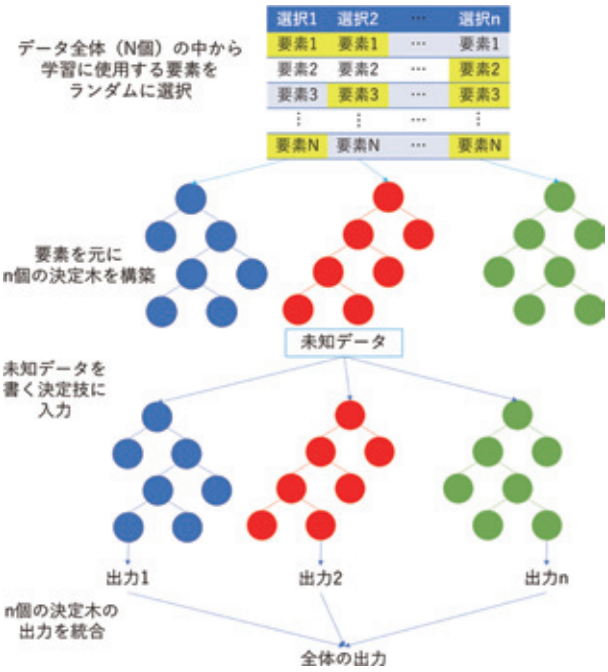

図2　ランダムフォレスト法

ランダムフォレスト法には2種類のランダム性が取り入れられている。このランダム性によって，個々の決定木が弱識別器として相補的に働き，高い予測性能と汎化性能を得ることができる。

2種類のランダム性は個々の決定木を構築する際の学習データのサンプリングと決定木のノードでの分割関数の学習に導入されている。前者について，ある決定木の学習データを選択する際，要素の中からランダムに，決定木ごとに選択する要素の重複を許してランダムに抽出する。後者について，ノードにおける分割関数のパラメータが取り得る値の集合の中から，値の選択をランダムに行う。これらを各決定木の根ノードから終点の葉ノードまで処理を再帰的に繰り返すことで学習を行う。

ランダムフォレスト法の評価は，構築した全ての決定木から得られた予測結果を統合する。この際の処理には一切ランダム性はない。統合には主に相加平均や相乗平均が用いられる。

3.　歌声の印象推定

3.1　研究対象

本論文では先行研究［3］と同様に，「歌声の印象」を「歌声を聴いた際にその歌声に対して生じる主観的な感覚」と定義する。特に，その感覚が言葉として表出した印象語を歌声の印象とする。歌声の印象は歌唱者の技術に大きく影響されるが，本研究ではアマチュアがプロの歌唱を参考にアレンジを加えて歌唱することを想定する。アマチュアの歌声の印象を推定するためには，まずプロの歌声で正しく印象語の推定を行えることを示す必要がある。このことから，プロが歌う楽曲のデータから

抽出した歌声に含まれる声道特性（MFCC）を特徴量としてとらえ，そこから生じる歌声の印象を分析する。

本研究では，アマチュアがプロの歌を聴いてそれを参考に歌うことを想定し，分析対象楽曲の条件を以下のように定める。まず，楽曲については日本人に馴染みのある「日本のポピュラー音楽」を対象とする。

次に，歌唱者は「楽曲ファイルに含まれるアーティスト」を対象とする。最後に，歌声の評価者は「音楽に関する専門知識を持たない一般人」を想定する。評価者の十分な人数の確保が難しいため，評価は楽曲配信サイトや動画配信サイトの一般人のコメントを分析したもの基準とする。ただし，コメントの中にアーティストの歌声について一度も言及されていない楽曲は本研究では対象外とした。

3.2　実装

本研究での，分類器の作成から入力・出力までの処理の手順を以下に示す。

1．楽曲へのコメントから，歌声に対する印象語が複数付されている楽曲を探す
2．付されている印象語の数を基に楽曲を**表1**に従って分類する
3．楽曲ファイルから不要な周波数成分を取り除く
4．3から歌声のない時間を取り除き，歌詞のフレーズの時間信号ごとに切り分ける
5．4で切り分けた歌声のデータに分類した印象語をラベルとして付ける
6．歌声のデータから特徴量（MFCC）を抽出し，リスト化する
7．6のリストを基に，ランダムフォレスト法で学習させ，分類器を作成する
8．作成した分類器に未知の歌声のデータを入力し，推定結果を出力するこのうち，1～7を事前準備，8を実験の項目で詳述する。

3.3　事前準備

事前準備の流れを**図3**に示す。まず，研究に用いる楽曲を集め，印象語の分析をした。楽曲を集める際，よく歌われる楽曲の方が多くの印象語をコメントとして付されていると考え，カラオケランキング［7］の上位曲を中心に収集した。また，一曲ごとに付されている印象語の数をまとめた。

次に，歌声を印象語で分類した。まとめた印象語の中から歌声に対して一番多く付されていた印象語をその歌声の印象とした。それを43曲分行い，歌声を**表1**の印象語にしたがって分類した。その際，一曲に対し複数人が歌っている楽曲については，ソロで歌っている部分を歌声ごとに分類した。また，**表1**にはない印象語が一番多かった歌声については先行研究［3］を参考に**表1**の中で一番近い印象語に分類した。

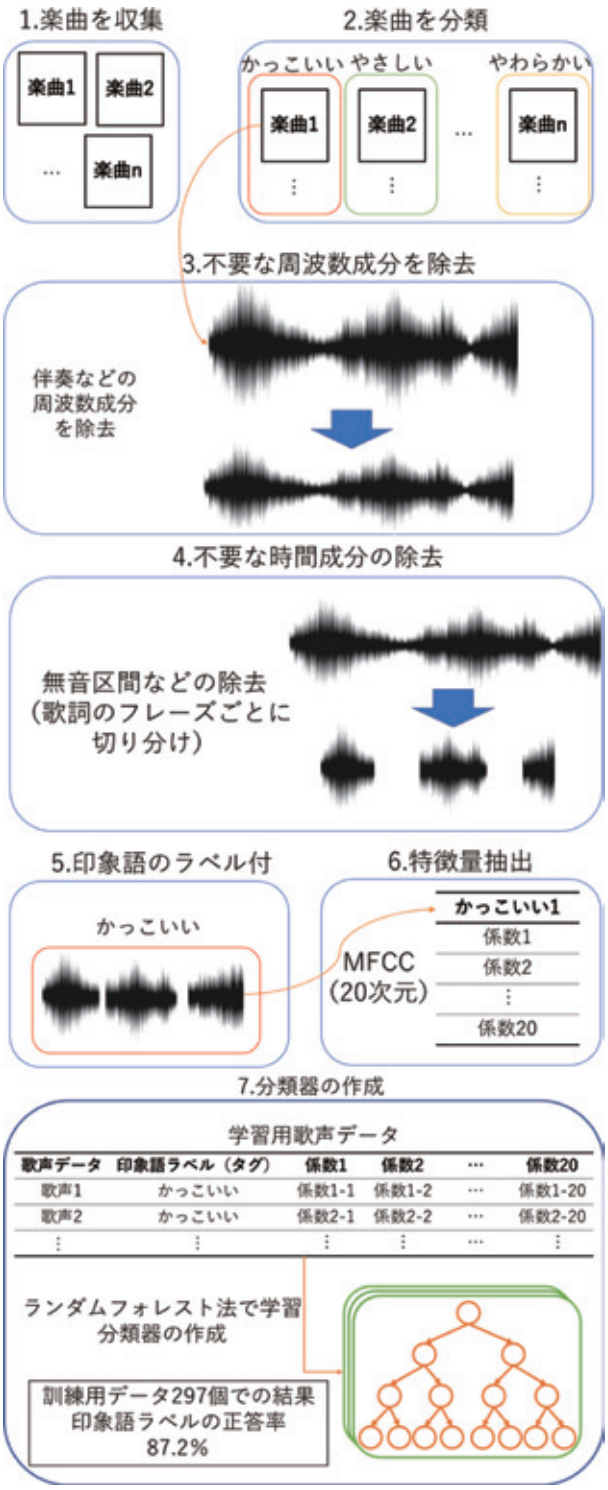

図3　事前準備

その次に，集めた楽曲ファイルを用意し，不要な周波数成分を除去した。最初，楽曲はmp3ファイルであったため，機械学習で扱いやすいwavファイルに変換した。変換した後，ボーカル抽出ツール［8］を用いて伴奏や楽器の周波数成分を取り除き，ほぼ歌声のみのデータにした。

表2　層化5分割交差検証結果

回数	一致率(%)	
	MFCC	F0
1回目	61.8	47.5
2回目	48.9	94.2
3回目	91.0	92.4
4回目	96.0	92.8
5回目	39.6	46.0
平均	74.7	74.6

その後，歌声のみのデータを歌詞のフレーズごとの時間信号に切り分けた。周波数成分を取り除いただけでは歌声のない時間が存在するので，手作業でその時間を取り除いた。また，学習するデータ数を増やすため，歌詞のフレーズごとの時間信号に切り分けた。この切り分けたデータを本研究では歌声データと呼び，全部で1,485個になった。切り分けた歌声データに対して，事前に分類した印象語のラベル付けを行った。

最後に，1,485個の歌声データそれぞれから，特徴量を抽出し，リスト化した。Pythonライブラリのlibrosa[9]を用いて20次元のMFCCを特徴量として抽出した。その後，抽出したMFCCをデータセット化し，csvファイルに出力した。

作成したcsvファイルのデータセットを基にランダムフォレスト法によって印象語の分類器を作成した。まずデータセット1,485個を8：2の割合で学習用データ1,188個と訓練用データ297個に分けた。

次に学習用データ1,188個を用いてランダムフォレスト法によって特徴量を学習させ，印象語を推定する分類器を構築した。作成した分類器に訓練用データ297個一致率を確認したところ，訓練データでは87.2%の一致率であった。

また，汎用性の評価のために層化分割交差検証を行った。今回はデータセットを5分割し，回数ごとの一致率と平均を求めた。また，比較のために，人が感じる音の高さと密接に関係する音響特徴量として用いられる基本周波数F0でも同様の前処理を行い，ランダムフォレスト法による分類と層化分割交差検証を行った。それぞれの結果を**表2**に示す。これらの結果から，この分類器を実験に使用できると判断した。

3.4　印象語推定実験

実験用の楽曲は**表1**の7種の印象語に分類できるものをそれぞれ1曲ずつ，新たに用意した。これらの楽曲にも，事前準備の1～6と同様の手順を用い，20次元のMFCCを抽出したデータセットのリストを作成した。

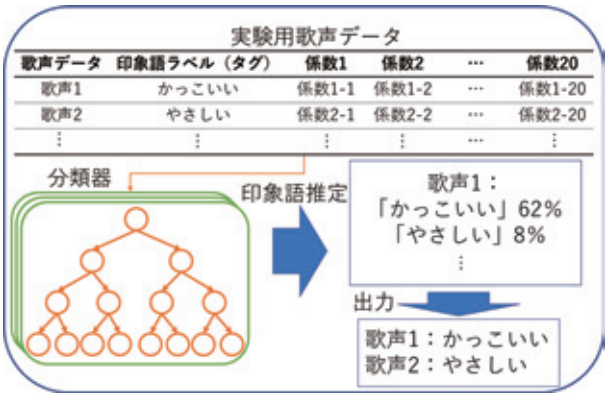

図4　印象語推定

このリストのデータ数は各印象語のラベルをつけたデータが30個ずつ，計210個である。これを実験用データとする。そのテストデータに対してランダムフォレスト法のクラス分類を行い，各楽曲データの印象語推定の出力を調べた。出力は，各印象語の推定割合を比較し，一番割合が高い印象語を推定結果として出力する。印象語推定実験のイメージを**図4**に示した。

4.　結果と考察

分類器に実験用歌声データ210個を入力した際の印象語推定結果の一致率をまとめた。全体の推定結果を**表3**に，各印象語の推定結果を**表4**に示す。なお，**表3**の印象語のうち，「透明感のある」「かっこいい」「かわいい」の3曲は女性，「やさしい」「力強い」「かなしい」「やわらかい」の4曲は男性の楽曲を用いた。

実験の結果，**表3**より，7曲の歌声データのうち3曲は「透明感のある」印象語，残りの4曲は「かっこいい」印象語を推定した。また，**表4**を見ると，全体の一致率は18.1%であった。そのうち「透明感のある」印象語は一致率が85.7%と高くなった一方で，他の印象語の推定結果は低くなった。これら2つの表から，この分類器の特徴として，女性の歌声は「透明感のある」印象と推定しやすく，男性の歌声は「かっこいい」印象と推定しやすいことがわかる。また，**表3**より，男性，女性ともに「かわいい」，「悲しい」，「やわらかい」印象とは推定されにくいことも特徴として挙げられる。

前者の理由として，それぞれの印象語の歌声として用いたアーティストが性別で偏っていたことが考えられる。「透明感のある」印象語に関しては女性ほとんどで，「かっこいい」印象語に関しては全てが男性であった。「透明感のある」印象語の曲に関しては，分類器作成の際に用いた楽曲アーティストは男性も含め全体的に声域が高かったことも挙げられる。

後者の理由として，実験に用いた楽曲のテンポが全体的に速かったことが挙げられる。「悲しい」印象語の学習に用いた楽曲はテンポが比較的ゆっくりである曲が多

表3 実験用歌声データの印象語推定の割合

歌声データラベル	印象語							推定結果
	透明感のある	かっこいい	かわいい	やさしい	力強い	かなしい	やわらかい	
透明感のある	87%	0%	0%	3%	10%	0%	0%	透明感のある
かっこいい	30%	7%	10%	23%	23%	3%	3%	透明感のある
かわいい	50%	10%	0%	23%	17%	0%	0%	透明感のある
やさしい	0%	93%	0%	7%	0%	0%	0%	かっこいい
力強い	0%	73%	0%	7%	17%	3%	0%	かっこいい
かなしい	7%	93%	0%	0%	0%	0%	0%	かっこいい
やわらかい	0%	30%	0%	23%	7%	17%	23%	かっこいい

表4 印象語推定結果

印象語ラベル	一致率（%）
透明感のある	85.7
かっこいい	6.67
かわいい	0.00
やさしい	6.45
力強い	6.67
悲しい	0.00
やわらかい	24.1
全体	18.1

く，前者の理由もあり推定ができなかったと考えられる。

以上のことから，この分類器は「透明感のある」印象を推定することに長けていると考えられる。理由として，声域が高い，すなわち声の周波数が高い歌声を「透明感のある」印象として推定していることが挙げられる。この分類器を使うことで，歌声の周波数の高さから，「透明感のある印象」を持たせる歌声を出せているかを確認する1つの指標にできる。

今後の分類器の改善点として，各印象語の学習に用いるアーティストの男女比を同じにすること，違うテンポでも同じ印象語に分類できるようにすること，1つの曲に対して，複数の印象語のラベルを付けて学習，分類を行うことの3点が挙げられる。また，本研究では，学習，実験用の歌声データを共にCDや楽曲配信サイトから入手したが，今後は学習用データに動画配信・投稿サイトの楽曲を用いる，テスト用歌声データにアマチュアのものを用いることを検討する。

5. おわりに

本研究では，歌声から聴き手が感じる印象を印象語として推定することを目指して，まず楽曲から歌声の特徴量としてMFCCを抽出した。その後，抽出したMFCCに対してランダムフォレスト法を用いることで印象語を推定，分類する実験を行った。その結果，男声や女声，曲のテンポによって歌声の印象語の推定に影響があることが示された。

以上から，楽曲から抽出した歌声の印象推定が可能であることが示された。ただし，全体的な推定結果の一致率が低いため改善する必要がある。推定する印象語の種類や学習に用いる楽曲の歌声については議論の余地があるため，今後の課題とする。

参考文献

［1］さとうそら：【歌うま高校生】会場中が涙した感動の歌声【花 アカペラver.】〈https://www.youtube.com/watch?v=OECo8KN95Q&ab_channel=%E3%81%95%E3%81%A8%E3%81%86%E3%81%9D%E3%82%89〉（参照 2021-09-11）.

［2］Breiman, L.: Random Forests. Machine Learning, Vol.45, No.1, pp.5-32（2001）.

［3］金礪愛，中野倫靖，後藤真孝，菊池英明：歌声の印象評価尺度の構築に基づく多様な印象の自動推定手法，情報処理学会論文誌，Vol.57，No.5，pp.1375-1388（2016）.

［4］山根壮一，小林和弘，戸田智基，他5名：歌声合成システムの音源データ検索のための声質評価推定法，情報処理学会研究報告，Vol.2015-MUS-108，No.6，pp.1-6（2015）.

［5］藤原弘将，北原鉄郎，後藤真孝，他3名：伴奏オン抑制と高信頼度フレーム選択に基づく楽曲の歌手名同定手法，情報処理学会論文誌Vol.47，No.6，pp.1831-1843（2006）.

［6］波部斉：ランダムフォレスト，情報処理学会研究報告，Vol.2012-CVIM-182，No.31，pp.1-8，（2012）.

［7］2019年カラ鉄年間カラオケランキングTOP10,000：カラオケの鉄人，入手先〈https://www.karatetsu.com/ranking/total/2019〉（参照2021-09-05）.

［8］VocalCanceler2：まらしいらぼ，入手先〈https://mahoroba.logicalarts.jp/vocalcanceler2/purchase.php〉（参照2020-10-25）.

［9］Mcfee, B., Raffel, C., Liang, D., et al.: librosa–Audio and Music Signal Analysis in Python, Proc. Proceedings of the 14th Python in Science Conference（SciPy 2015）, pp.18-24（2015）.