愛知工科大学紀要 第 18 巻 Bulletin of Aichi University of Technology(2020 年度)pp.15- 20(2021.3)

研究論文

Convolutional Neural Network を用いた楽曲からのアーティスト識別および類似アーティストの推定

堀 弘志* 實廣 貴敏*

(2020 年 10 月 1 日受理)

Identification of Artists and Estimation of Similar Artists from Music Using Convolutional Neural Network

Hiroshi HORI* Takatoshi JITSUHIRO*

(Received October 1, 2020)

Abstract

This paper proposes the algorithm for the identification of music artists by the convolutional neural network (CNN) using the long-term average of Mel-frequency cepstral coefficients as acoustic feature parameters. Experimental results show that 83% accuracy rate was obtained for the identification of the five idol groups. Furthermore, we propose the estimation method of similar artists by the CNN using the same feature parameters. When we compared the degree of similarity for the estimated similar artists between one subject and the trained CNN, we found that they tended to be roughly similar.

キーワード: 楽曲, 畳み込みニューラルネットワーク, アーティスト識別, 類似アーティスト推定, 長時間平均
Keywords: music, convolutional neural network, artist identification, estimation of similar artists, long-term average

1. はじめに

近年，音楽情報処理分野の発展により，歌声やリズムに関する研究活動が世界的に活発に取り組まれ，学術的な観点からだけではなく，産業応用的な観点からも注目を集めている．このような研究は，歌声合成，歌詞認識，歌声検索など多岐にわたっている．

また，音楽を聴く媒体と環境が増えたことにより音楽検索の需要は高まってきている．例えば，鼻歌検索や歌詞検索などを使い気になった曲を検索したり，音楽再生アプリや動画サイトなどでオススメとして出たりなど，普段生活している中で利用する頻度が多い．しかし，これまでの鼻歌検索や歌詞検索などでは，推薦曲がユーザーの好みと異なったものが出たりする．明確にどういった基準で推薦される曲が決まっているのかは，ブラックボックスとなっていて分かってはいない．しかし，インターネットの質問サイトなどでは，「あなたへのおすすめはどういった基準で選ばれているのか」や「おすすめはどういった仕組みになっている」など，多くの人がオススメで自分の思っていたものが推薦されていない．例えば，動画サイトや音楽再生アプリなどのオススメは，他の人が聴いたときに一緒に聴かれている曲がオススメになっていたり，音楽のデータの情報から同じ会社や同じジャンルのものがオススメとして推薦されていたりする．必ずしも曲調など音響的な特徴から類似のものを探しているわけではない．音楽サイトのオススメなどでは，ジャンルが違ったとしていても曲調などが似ているアーティストを調べたい人は多く，実際にそういった基準で検索するニーズは高い．また，パターン認識技術において，深層学習(Deep Learning)が効果的であることが明らかになり，音楽情報処理分野においても，Deep Learning Network (DNN)を用いた研究が盛んである．例えば，音楽データから Convolutional Neural Network (CNN) [1]を用いて音楽ゲーム譜面自動生成する検討[2]や，DNN による音楽ジャンル分類[3]が行われている．また，楽曲間で歌手の歌声(声色)がどれだけ似ているかを，歌声の類似度として定量的に求める手法は範囲が広く，歌声の類似度に基づいた楽曲の自動分類[4]や，ある曲の歌手の歌声とよく似た歌声を持つ楽曲を検索するシステム[5]など，様々な

* 愛知工科大学工学部情報メディア学科，〒443-0047 愛知県蒲郡市西迫町馬乗 50-2
Department of Media Informatics, Aichi University of Technology, 50-2 Manori, Nishihasama-cho, Gamagori-shi, Aichi 443-0047, Japan

研究事例がある. ただし, これらはいずれも歌手固定で行われている.

本研究では, 曲の音響特徴量からCNNを用いてアーティストの類似度を出し, オススメのアーティストを推薦する. まず, CNN を推奨に使うアーティストの楽曲で学習し, それらアーティストの曲からアーティストを識別出来るようにする. 類似度はその CNN によって得られるアーティストの確率を用い, 未知のアーティストの曲が入力されたとき, 最も確率が高いアーティストが最も類似度の高いアーティストとみなす. 本研究では, 特徴量として楽曲を短時間フレーム分析によりメル周波数ケプストラム特徴量を得た後, その長時間平均を計算したものを用いる. これを CNN への入力とする.

本論文の構成を述べる. 2 節では, CNN によるアーティスト識別方法について, 3 節では, アーティスト識別の評価実験について述べる. 4 節では, 2, 3 節で学習したCNNを用いた類似アーティストの推定方法を説明し, その結果を被験者との比較について述べ, 5 節にて本論文をまとめる.

2. Convolutional Neural Network (CNN)を用いた楽曲からのアーティスト識別

CNN と特徴量分析についての説明を行う. 近年, パターン認識をはじめ, 多くの分野で機械学習技術は成功している. 今回は, 色々な機械学習の技術がある中で, CNN を用いる. また, 特徴量分析の特徴量は, メル周波数ケプストラム特徴量を用いる. 以降では, 2.1 節において CNN について説明し, 2.2 節において特徴量分析, 2.3 節において識別手法について説明する.

2.1. Convolutional Neural Network (CNN)について

Figure **1** は, 一般的な CNN の構造を示す. 通常のニュートラルネットワークと違い, 畳み込み層とプーリング層を持っているのが特徴である. 全結合層以降では DNN が使用されている. これによって線形分離可能ではないデータを識別が可能となる.

CNN は画像認識分野において最も顕著な成功をおさめている. 脳の視覚野の構造における知見を元に, ニューロン間の結合を局所に限定し, 層間の結合を疎にしていることを特徴としている. 画像の局所的な特徴量を担う畳み込み層と, 局所ごとに特徴をまとめあげるプーリング層を繰り返した構造の後, 全結合層によって画像判別を行っている. 畳み込み層では, 畳み込み操作を行うフィルタを掛けて特徴マップを出力する. プーリング層は, 最大値または平均値を特徴として抽出することで, データの次元削減を行い, 演算量を下げる. 全結合層は画像判別の層であり, 1 次元出力ができる形態に変換する. 物体の認識において必要不可欠である入力の平行移動に対する不変性を段階的に加え, 入力の解像度を少しずつ落としながら異なるスケールで隣接する特徴の共起をとり, 識別に有効な情報を選択的に上層へ渡していくネットワークとなっている.

これらの特徴から画像以外の分野でも多く利用されるようになった. 本研究では, 入力は楽音であるが, 限定された数のアーティストを識別, さらには未知の入力に対して類似のアーティストを出力したい. そこで, CNN の機構によれば, ニューラルネットワークにより, 入力された多次元の特徴量から, 識別に必要な特徴を抽出, 小さい次元の出力を行うことが可能と考えられる.

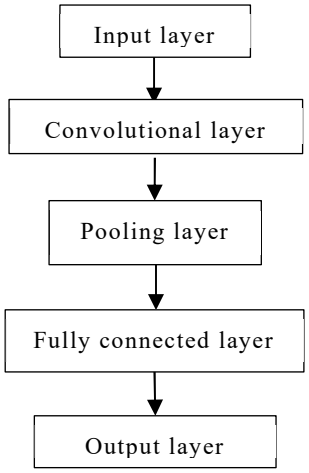

Figure 1 The general structure of a CNN.

2.2. 音声特徴量

音声や楽音の特徴量分析には, 一般に 20～40 ms 程度の音声区間で信号を取り出し, 時間窓をかけて周波数分析を行う. この処理を 5～20 ms 程度でずらしながら, 音声全体を分析していく. これが分析フレームになる. その音声の長さに応じて, 分析フレームが増えることになる. 本研究では, 各フレームの分析に, 人の聴覚特性を考慮した周波数分析である, メル周波数ケプストラム(Mel-Frequency Cepstral Coefficient, MFCC)を用いる.

このままでは, 楽音により得られる分析フレームの数が変動し, ニューラルネットワークへの入力は難しい. 本研究では, あらかじめ学習されたアーティストを入力楽曲で識別, さらに, 未知の楽曲入力から最も近いアーティストを選択するのみである. そこで, まず, 各フレームの特徴量を正規化するため, 標準化変換を行う.

次に，分析フレーム全体の長時間平均を求める．これにより 1 フレーム分の特徴量ベクトルになる．この長時間平均を用いる方法は，音声による話者認識でかつてよく利用された方法であり，今回は楽曲全体のおおまかな特徴が長時間平均に現れることになる．

2.3. 識別手法

CNN を用いてアーティスト識別を行う．Figure **2** に CNN の学習手順を示す．楽曲の音響データをアーティスト名と対にして，CNN 学習への入力とする．CNN の入力層へは，まず特徴量分析器において，メル周波数ケプストラム MFCC を求め，さらに，標準化変換し，長時間平均 MFCC を求め，入力する．

Figure **3** に CNN によるアーティスト名の推定手順を示す．アーティスト名を推定したい楽曲の音響データをまず特徴量分析器に入力する．MFCC を求め，標準化変換し，それらの長時間平均 MFCC を求める．それを CNN に入力し，推定されたアーティスト名を得る．

このように CNN を学習し，アーティストの推定，すなわち，アーティスト名の識別を行う．実験では，まず上記の手順を用いて，CNN の楽曲音響データからのアーティスト名識別能力を評価する．

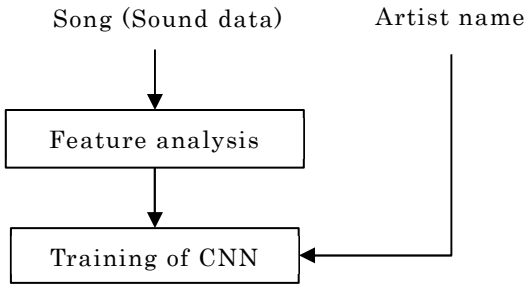

Figure 2 Training of the CNN that identifies artist names from acoustic data of songs.

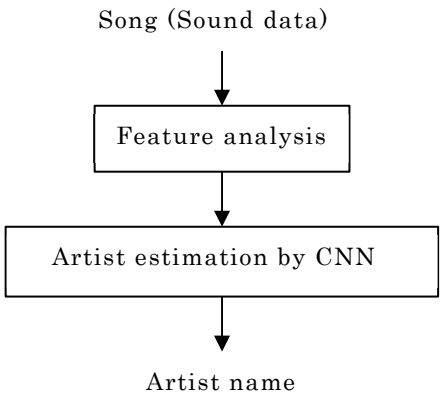

Figure 3 the estimation of artist names by the CNN.

The

3. CNN によるアーティスト識別の実験

3.1. 実験内容

本実験では，アーティスト名識別の精度を検証する．対象として，5 組のアーティスト(でんぱ組.inc，CY8ER，ももいろクローバーZ，AKB48，乃木坂 46)を用いた．アイドルと呼ばれるアーティストたちであるが，近年のアイドルブームによりたくさんのアイドル・グループが存在し，曲数が豊富であるため，実験するには適している．また，その成り立ちからも似ているアイドルも多く，アイドル検索のニーズがある．グループのアイドルを選択している理由は，昔と比べ，ソロのアイドルよりグループのアイドルの方が人気があり，ニーズがあると考えられる．Table **1** に第 1 著者が考える選択したアーティスト 5 組の特徴をまとめる．

Table 1 The features of the five selected artists.

でんぱ組.inc	6 人グループのアーティスト．それぞれが入り乱れながら歌う．激しい曲が多い．
CY8ER	5 人グループのアーティスト．それぞれがパートごとに分かれて歌う．ゆっくりとした曲が多い．
ももいろクローバーZ	4 人グループのアーティスト．パートごとに歌うのと全員で歌う数がほぼ同じ．激しい曲が多い．
AKB48	約 20 名で歌うアーティスト．パートごとに分かれて歌うことが少なく，ほとんどが複数人で歌われている．様々なジャンルの歌を歌う．
乃木坂 46	AKB48 と同じ会社のグループで歌い方や曲調がよく似ている．AKB48 との比較として採用．

3.2. 実験条件

実験条件を示す．学習データはでんぱ組.inc，CY8ER，ももいろクローバーZ，AKB48，乃木坂 46 の 5 組から計 16 曲ずつ計 80 曲を用いた，評価データは，上記と同じアーティスト 4 曲ずつ計 20 曲を用いた．これらはすべて違うデータであり，販売されている楽曲のまま，音声と演奏を分離するようなことは行わず，曲の長さも特別には統一していないものを用いた．

楽曲の品質は 44.1 kHz サンプリング周波数，16 ビットであった．音声分析にはフレーム長を 20 ms，フレーム周期を 15 ms とした．音声特徴量は 20 次元 MFCC とし，長時間平均 20 次元 MFCC を CNN の入力とした．CNN に関しては，入力層を 20 ユニット，出力層をアーティスト 5 組に合わせて 5 ユニットとした．中

間層はユニット数を 256 に固定したものの，畳み込み層，および，プーリング層を一層ずつ増やして精度を評価した．

3.3. 実験結果

Figure 4 に隠れ層を 3～10 層に変えた時の正解率を示す．精度が最も良いのは隠れ層が 6 層の時で学習回数が 100 回の時，正解率は 83%であった．また，隠れ層が 5 層の時，正解率は 60%で最も精度が低かった．さらに，8 層以上では 65%で変化がなかった．

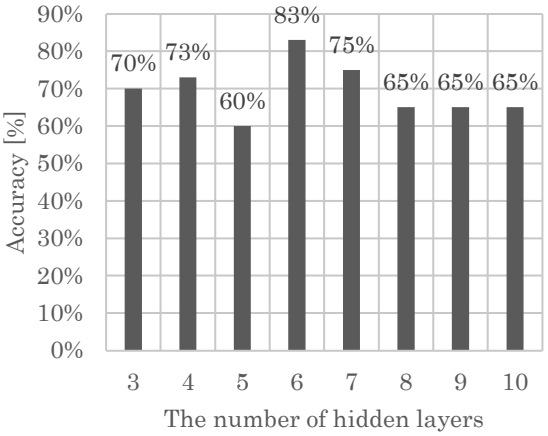

Figure 4 Accuracy rates for the number of hidden layers.

4. CNN による類似アーティストの推定

4.1. 類似アーティスト推定の手法

3 節での実験により，限られたアーティスト数ではあるが，楽曲の音響データから CNN により，比較的高い確率で識別が可能であることが分かった．本研究の本来の目的である，類似アーティストの推定を検討する．Figure 5 に提案する類似アーティストの推定方法を示す．3 節で行ったような，推奨したいアーティストの楽曲を CNN で学習しておく．その CNN へ未知アーティストの楽曲を入力し，その時に得られる学習済みアーティストに対する確率のうち，最も確率が高いものを推奨する類似アーティストであると判定するとする．

ここで用いる音声特徴量は MFCC の長時間平均である．楽曲全体の平均であるので，時間的変動を表現することはできないが，楽曲全体で現れる周波数特性の代表的なものを表現できる．おおまかな特徴として見ることができるので，類似アーティストを検索するには適していると考えられる．

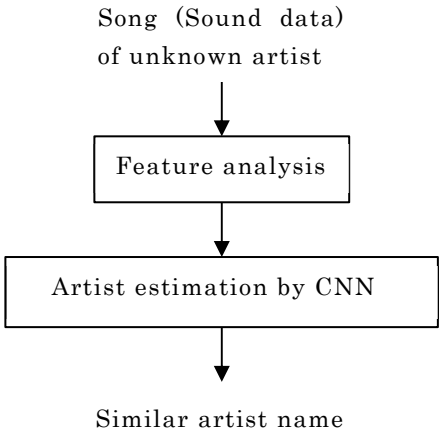

Figure 5 The estimation method of a similar artist from a song of an unknown artist.

4.2. 実験

4.2.1. 実験内容

3 節での実験で学習された CNN を用い，そこで学習されたアーティストとは別である未知のアーティストの楽曲を入力とし，そのときに得られる学習済みアーティストそれぞれの確率の大きさ順に推奨される類似アーティストとする．この推奨された類似アーティストが，人間により似ていると推定されるアーティストと同様な傾向があるのか，全くないのかを検証する必要がある．そこで，ここで用いたアーティストの楽曲に詳しい被験者に楽曲を聞いてもらいながら，その差異を検討する．

全くアイドルを知らない被験者に対する評価も検討したいところだが，あまり簡単ではないと考えている．1 曲ごとの類似度を評価するのであれば，1 曲ずつ比較して聞いてもらい，評価することが可能である．しかし，本研究で考えている推奨したい類似アーティストは，ユーザが好むアーティストに対して，楽曲の集合として見た場合に，よく似ていると思えるアーティストである．そのように類似アーティストを見つけるには，推奨するアーティストと入力される未知のアーティストの双方の楽曲に通じている必要がある．したがって，それらのアーティストをよく知らない被験者では，あらかじめ多くの複数の楽曲を聞く必要がある．今回は，十分な時間がなかったため，実験のしやすい，実験対象の楽曲に日ごろから慣れ親しんでいる被験者で行った．

4.2.2. 実験条件

実験条件を示す．学習データはでんぱ組.inc，CY8ER，ももいろクローバーZ，AKB48，乃木坂 46 の 5 組から 16 曲ずつ計 80 曲を用いた．評価データは，上

記と同じアーティスト 4 曲ずつ計 20 曲を用いた. これらはすべて違うデータであり, 音声と演奏を別にすることなく, そのまま用い, 曲の長さもそのまま統一していないものを用いた. 予測データとしては, 上記のアーティストの曲で学習と評価に使用していない曲を, 各アーティスト 3 曲ずつ計 15 曲, 未知のアーティスト 4 組から 1 曲ずつ計 4 曲を用いて, それぞれの曲に対しての類似アーティスト推定を行う. 未知のアーティストは日向坂 46, BiSH, SKE48, E-girls を用いる. これらアーティストに詳しくない方にも多少理解してもらえるよう, 被験者の主観であるが, 音響的な類似アーティストをあげておく. 日向坂 46 と SKE48 は, 同様に同じ会社のグループである AKB48 と乃木坂 46 と特徴が似ている点が多い. BiSH はももいろクローバーZ と似ている点が多い. E-girls はでんぱ組.inc と似ている点が多い. CNN は隠れ層 6 で固定し, 学習は 3 節と同様に, 各アーティスト 16 曲で合計 80 曲を用いた.

CNN の推定結果がどの程度, 人の感覚に近いか, 被験者と比較を行った. ここでは, 評価対象のアイドルに詳しい 20 代男子学生 1 名に, 既知のアーティストの曲を聴いてもらったのちに, 未知のアーティストを聞いてもらい, どのアーティストに近いかを選択してもらった. そのパーセンテージを類似度スコアとして CNN の推定結果と比較した.

4.2.3. 実験結果

6だけ太字？

Figure **6**, 7, 8, 9 にそれぞれ, 未知のアーティストである, BiSH, E-girls, 日向坂 46, SKE48 に対する, 学習済みアーティストのでんぱ組.inc, CY8ER, ももいろクローバーZ, AKB48, 乃木坂 46, それぞれの類似度スコアをレーダーチャートにより示す. 類似度スコアは CNN により出力される確率をパーセンテージで表したものである.

CNN による類似度は学習済みアーティストに対して, あまり大きな差が出ていない場合が多い. 20%前後が多く, 5 組の学習済みアーティストであることを考えると, 未知アーティストの楽曲に対して, 明確には識別できていないと考えられる. 類似度を推定するには, 音声特徴量からより工夫が必要と考えられる.

類似度のトップスコアのものだけ見ると, CNN によると, BiSH は CY8ER に, E-girls はでんぱ組.inc に, 日向坂 46 は CY8ER に, SKE48 は AKB48 に類似しているという結果になった. これだけでは人間との比較は難しいが, レーダーチャートとして, 学習済みアーティスト全体との比較をすると, BiSH 以外は類似度の傾向が比較的近いといえる. ただし, 被験者はこれらの楽曲に慣れ親しんだ人とはいえ, 1 名のみの主観であるので, より多くの被験者による比較が必要である.

Figure 6 Similar scores for Bish.

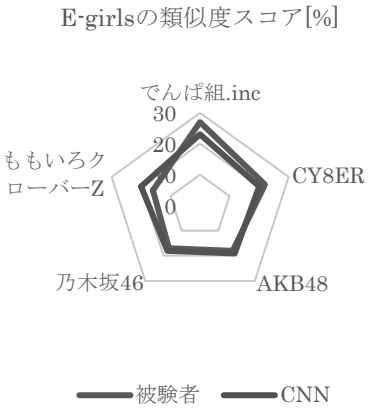

Figure 7 Similar scores for E-girls.

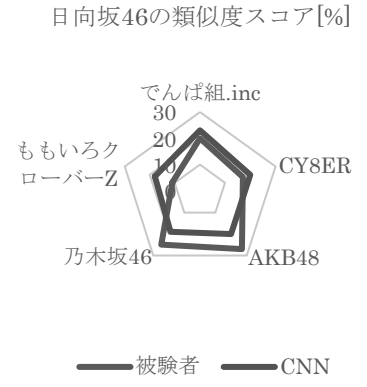

Figure 8 Similar scores for Hinatazaka46.

Figure 9 Similar scores for SKE48

5. おわりに

本研究では，楽曲の音響特徴量から CNN を用いてアーティストの識別を行い，その学習された CNN を用い，未知アーティスト楽曲に対する確率から，類似するアーティストを推奨する方法の検証を行った．さらに，被験者 1 名の主観で未知のアーティストがどの学習済みアーティストに類似しているかをスコア化したものと比較して，人間と CNN との類似度スコアの比較を行った．

アーティストの識別では，楽曲を短時間フレーム分析により，メル周波数ケプストラム特徴量を得た後，その長時間平均を求めて用いた．隠れ層の Convolutional 層と Pooling 層を一層ずつ増やし CNN の構造を変えたもので評価を行った．隠れ層が全 6 層の時，正解率は 83%で最も高い精度が得られた．さらに，隠れ層が 8 層以降では正解率が 65%で変化がなかった．音響特徴量に長時間平均を用いたということで，時間的変動は考慮できていないが，楽曲に含まれる主な周波数成分を抽出でき，それを元に識別することで比較的高い精度を得られることが分かった．

類似アーティストの推定方法では，4 組の未知アーティストの楽曲に対して，5 組の学習済みアーティストから類似のアーティストを選択する実験を実施した．それらの類似度スコアと，これらのアーティストの楽曲に詳しい被験者 1 名が評価する結果を比較した．CNN による類似度スコアは全般的には 5 組とも比較的近い値を出力しており，必ずしも明確に識別できているわけではないのがわかった．ただし，類似度スコアの 5 組の学習済みアーティストに対する偏りは，これらアーティストに詳しい被験者 1 名の主観によるスコアによる偏りに近いところも多く，メリハリは少ないものの，大きな傾向はつかめていると考えられる．

今後の課題を述べる．今回は「アイドル」と呼ばれるアーティストにのみ注目し，評価を行ったが，枠組みとしては，楽曲のジャンルによらないので，より広いジャンルでの評価を行う必要がある．今回は，一般的な CNN を用いて実験を行ったが，他のニューラルネットワーク，例えば，Long short-term memory (LSTM)や，Deep Neural Network (DNN)に関する各種技術(Dropout など)を使用することで識別精度の向上を検討することがあげられる．また，類似アーティストの推定では，評価自体がまだ不十分であり，被験者を増やして実施する必要がある．あるいは，専門的に詳しくない被験者でも評価しやすい手法を検討することも考えられる．類似度スコアに関しては，単純に CNN の出力する確率を用いたが，より適したスコア，より適した音響特徴量を検討していく必要がある．

参考文献

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. of the IEEE, pp. 2278–2324 (1998).

[2] 柴崎大地，酒井充，丸山博，“ニューラルネットワークを用いた音楽ゲーム譜面自動生成の検討”，情報処理学会第 80 回全国大会講演論文集 2018(1), pp. 165-166, 2018 年

[3] 赤羽慎，藤田侑介，王龍標，甲斐充彦，“セグメントレベル特徴量を用いた楽曲のジャンル分類”，情報処理学会研究報，vol. 2014-MUS-103, no. 22, 2014 年

[4] W.-H. Tsai, and H.-M. Wang, D. Rodgers, S.-S. Cheng and H.-M. Yu, “Blind clustering of popular music recording based on singer voice characteristics,” Proc. ISMIR 2003, pp. 167-173 (2003).

[5] H. Fujiyama and M. Got , “A music information retrieval system based on voice timbre,” Proc. ISMIR 2007, pp. 467-470 (2007).