# 深層学習分類モデルを用いた J-POP における発声法の推移の分析

関西学院大学大学院 理工学研究科 人間システム工学専攻 2 年 47021804 菅野翔平

## 1. はじめに

カラオケなどアマチュアが歌を歌う機会において，最近のポピュラー音楽が歌いにくくなっていると指摘する声を聞くことが少なくない．ここ数年間の男性ボーカルのポップス楽曲では高音域で歌唱することが多くなったと言われ，歌唱が難しくなった主要因の一つとして考えられる．発声法の視点からは，地声・裏声の他にミックスボイスと呼ばれる歌唱法がプロの中で使用されるようになった．Yee らの論文[1]によるとミックスボイスは，高音域を喉の負担を抑えて発声することが可能であり，地声や裏声とは区別して考える必要があると述べられている．一般的にもミックスボイスが認知されつつあるが，この発声法に対して調査した文献は未だ少ないというのが現状であり，詳細な分析が求められている．卒業研究では，現在のアーティストの発声法が徐々にミックスボイスに変化していると仮説を立て，メロディの音高上昇の現状と発声法の変化について調査を進めてきたが，手作業によるアノテーション手法では楽曲数を確保できず，膨大なデータを調査することは困難という問題点があった．

そこで本稿では，問題に対して既存楽曲からデータセットを独自で作成，様々な発声法を自動判別するシステムを作成することでアプローチする．そのシステムを活用し対象楽曲を拡張することで，部分的に行っていた発声法の推移の状況について網羅的に調査を行う．また最終的な展望として，歌唱練習支援システム等の応用に繋げることを目標に本研究を進める．

## 2. 各発声法の位置付け

本研究で作成するシステムに求められる要件として，仮説に挙げたミックスボイスの他に，地声，裏声，裏声を使わず地声を張り上げて発声するプルの 4 クラスを判別できるようにする．関連研究をもとに本稿では，声帯が伸びずに閉じた状態で発声された声を地声，声帯が伸びて開いた状態で発声された声を裏声，声帯を完全に閉めず喉の負担が少ない状態で発声された声をミックスボイス，喉の負担が大きく張り上げる地声をプルと定義し進める．分類システムの作成にあたって，発声法を判別する深層学習モデルの作成と，そのモデルの学習に必要なデータセット制作の大きく二つの工程からアプローチする．

## 3. システムの実装

### 3.1 データセットの作成

分類モデルを作成するにあたり，本研究に適した学習に必要となるデータセットが存在しないため，ESC-50 データセットを参考に J-POP のヒットソング全 312 曲を使用しデータセットを作成する．なお対象楽曲は性別による音域，声質の差を無くすため男性ボーカルのみに限定し，フルサイズの音声をデータとして使用する．音源には一般にボーカル音以外に伴奏音が含まれているため，facebook research が公開している音源分離ライブラリ Demucs[2]を使用しボーカル音のみを抽出する．次に，発声法のラベルづけを行う長さとして最適だと考えられる 1 秒の長さに音声をカットし，それらの音声クリップから log-mel spectrogram, Chromagram, F0 の 3 種類の音響特徴量をマルチチャンネル画像のように連結し，深層学習の入力データとして利用する．

音声データをクリップに分割した結果，合計で 69668 の音声クリップを収集した．ボーカル抽出した音声データの中には，コーラスやユニゾンなど歌声が重なっている音声も存在するため，ノイズが多い音声を省いた後，4 種類の発声法に加えて，コーラスのラベルを追加した合計 5 種類でアノテーションを行う．この音声クリップは前処理後のボーカルの抽出音のみで構成されているため，筆者の耳でもアノテーションが可能である．アノテーションの結果，地声，コーラス，裏声，ミックスボイス，プルの順に［26417, 25482, 297, 785, 1251］の有声クリップを得た．

### 3.2 分類モデルの検討

本研究では，山本ら[3]を参考に CNN ベースのモデル 3 種類を実装し，それぞれの精度の比較検討を行う．今回使用するモデルとして，PyTorch で提供されている CNN ベースのモデル，VGG, ResNet, ConvNet を比較する．

4.1.2 節で述べたように，データセットには発声法の分類の妨げとなるコーラスの音声が問題となるが，本稿ではまずコーラスと非コーラス(地声，裏声，ミックス，プル)の 2 クラス分類を行うモデルと，地声，裏声，ミックス，プルを判別する 4 クラス分類を行うモデルを作成することで対処する．

作成したデータセットにはクラスの分布に不均衡があるため，Stratified-5Fold Cross Vocalization 手法でモデルの精度を比較する．損失関数に多クラス分類で多用される交差エントロピー誤差，最適化関数に Adam，学習率を 1e-4 とし，事前学習は行わず出力のクラス数に合うように入出力のパラメータを変更し学習を行う．評価指標として，正解率，適合率，再現率，F 値から各モデルの精度比較を行う．

学習の結果，2 クラス分類では，全ての項目で最も数値の高い VGG19 モデルを，4 クラス分類では，全ての項目において最も数値の高い VGG16 モデルを採用し，年代ごとの発声法の推移について網羅的に調査する．

**表 1** 2 クラス分類モデルの精度比較

| | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| VGG16 | 0.921 | 0.922 | 0.919 | 0.920 |
| VGG19 | 0.922 | 0.922 | 0.921 | 0.922 |
| ResNet34 | 0.908 | 0.909 | 0.907 | 0.908 |
| ResNet152 | 0.908 | 0.909 | 0.906 | 0.907 |
| ConvNet tiny | 0.864 | 0.868 | 0.861 | 0.863 |
| ConvNet base | 0.869 | 0.871 | 0.867 | 0.868 |

**表 2** 4 クラス分類モデルの精度比較

| | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| VGG16 | 0.948 | 0.758 | 0.687 | 0.715 |
| VGG19 | 0.946 | 0.717 | 0.679 | 0.695 |
| ResNet34 | 0.942 | 0.702 | 0.652 | 0.673 |
| ResNet152 | 0.940 | 0.701 | 0.646 | 0.670 |
| ConvNet tiny | 0.937 | 0.657 | 0.591 | 0.615 |
| ConvNet base | 0.934 | 0.649 | 0.593 | 0.616 |

# 4. メロディ音高と発声法の推移の調査

## 4.1 調査のアプローチ

前章で述べたモデルを使用し，J-POP の楽曲に対して発声法の推移の調査，またメロディ音高の推移について分析する．年代ごとの楽曲の傾向を分析するためには，その年を代表するヒットチャートを収集する必要がある．そこで本研究では 1971 年から 2007 年までの CD 売り上げランキング，2008 年から 2022 年までの Billboard JAPAN Hot 100 Year End ランキング上位曲を用いる．二つの指標から男性ボーカルの楽曲を 1 年ごとに 30 曲選曲し，52 年間で合計 1560 曲の楽曲に対して，メロディの音高上昇の状況，また分類システムを利用し発声法の推移の調査を行うことで，冒頭に述べた仮説の検証を行う．

## 4.2 メロディ音高の調査

本節では，対象楽曲に対してピッチ推定手法を使用し，年代ごとのメロディ音高の状況について調査する．メロディのピッチを取得するために，本研究では従来の基本周波数の推定手法を上回る CREPE[4]を利用する．入力する音声データはデータセット作成手順と同様に，Demucs での音源分離，無音部分の除去等の前処理を行う．分析対象合計 1560 曲に対してピッチ推定を行い，1 曲ずつ平均値を算出する．30 曲の平均値を各年の値として算出し，メロディ音高の推移の状況をまとめたグラフを図 1 に示す．横軸を年代，縦軸を周波数(Hz)で表している．

結果として，1 次近似に正の傾きが見られ，メロディの平均値が上昇していることが明らかになった．1971 年代は 250Hz(C4)程度の高さであるが，2020 年代では，290Hz(D4)程度の周波数を示しており，約全音分のピッチが上昇していると言える．また 2000 年代にも 2020 年代と同様に上昇しており，この状況には B'z や GLAY などのアーティストの流行が大きく影響していると考えられる．

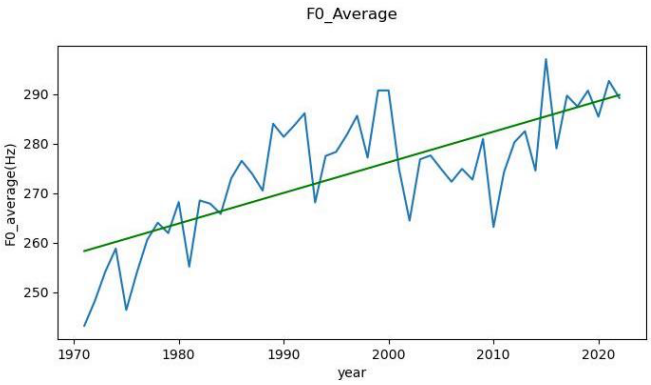


図 1：年間 30 曲のメロディ音高の平均値

## 4.3 発声法の推移

4.1 節で収集した対象楽曲に分類モデルを活用し，各発声法の推移について調査を行う．まず 2 クラス分類を行いコーラスと予想されるクリップを除いた後，4 クラス分類を行い，地声，裏声，ミックス，プルのクリップ数を集計する．年ごとに各発声法のクリップ数を集計し割合をまとめたグラフを図 2 に，コーラス以外の 4 クラスの発声法を使用された曲数でカウントしたグラフを図 3 に示す．

図 2 によると，コーラスは 1970 年代の 30%前後から徐々に割合が上がり，70%近くの割合を示している 2013 年を境に少しずつ減少していることがわかる．反対に非コーラスの音声が減少していることから，音を重ねる編曲が増えていると捉えられる．次に図 3 によると，ミックスボイス，プルは増加傾向にあり，2000 年代前後に一度上昇，2010 年前後に減少した後，2020 年代で 10 年にかけて再び上昇している．反対に，裏声に関しては減少傾向にあり，2000 年代にピークを迎えていることがわかる．

ミックスボイス，プルの推移に関して，2000 年代の増加には前節と同様，B'z や GLAY などのロックバンドが流行したこと，2010 年代の減少には嵐や関ジャニ∞などのアイドルソングの流行が大きく影響していると考察できる．

発声法が変化したためメロディが高くなったのか，メロディが高くなったため，歌えるような発声法に変化したのか，二つの因果関係を証明することは難しいが，双方に相関関係があることはこれらのデータから言える．

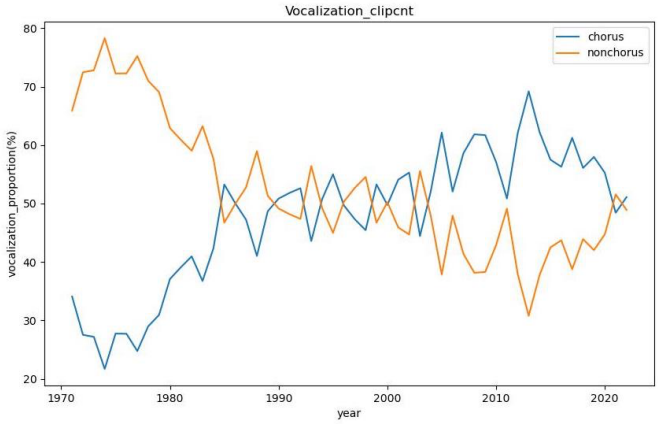


図 2：コーラスと非コーラスの推移（クリップ数カウント）

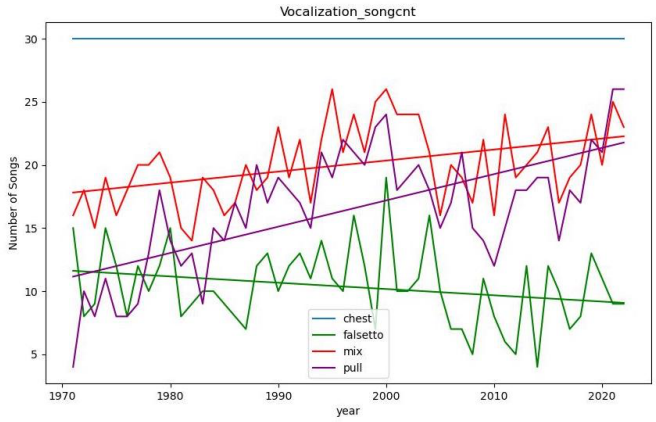


図 3：発声法の推移（曲数カウント）

# 5. おわりに

本稿では，発声法を分類するシステムの提案を行い，J-POP のヒットチャートを対象に発声法の推移とメロディ音高の調査を行った．結果として，年々男性ボーカルのピッチが上昇していること，またコーラスの音声が増えていることや，ミックスボイス，プルなどの多様な発声法の使用率が上がっていることを明らかにした．

今後の予定として論文誌への投稿や，本モデルを活用した歌唱練習支援システムの作成などの応用に繋げたいと考えている．

## 参考文献

[1] Lee, Y., Oya, M., Kaburagi, T. et al.: Differences Among Mixed, Chest, and Falsetto Registers A Multiparametric Study, *J. of Voice*, pp.1-19(2021).
[2] Défossez, A., Usunier, N., Bottou, L. et al.: Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed, *Proc. Audio and Speech*, pp.1-12(2019).
[3]山本雄也，Nam, J，寺澤洋子ほか：歌唱テクニックの識別における hand-crafted 特徴量と深層学習抽出特徴量の比較，情報処理学会研究報告，Vol.2021-MUS-130 No.30(2021).
[4] Kim, J.W., Salamon, J., Li, P. et al.: Crepe :A convolutional representation for pitch estimation, *IEEE Int'l Conf. on Acoustics, Speech and Signal Proc.*, (ICASSP), pp.161-165(2018).