Vol.2022-MUS-135 No.9
2022/9/15

歌声のための自己教師あり対照学習による特徴量表現の獲得手法

矢倉 大夢[1,2,a)]　渡邉 研斗[2,b)]　後藤 真孝[2,c)]

概要：本研究では，歌声に特化した自己教師あり対照学習による特徴量表現の獲得手法を提案した．画像ドメインを中心に発展してきた自己教師あり対照学習は，教師データなしでロバストな特徴量表現の獲得を可能にしてきた．これは，あるサンプルの特徴量表現とそのサンプルを自動変換したものの特徴量表現が近づくようにニューラルネットワークを学習することで実現される．提案手法では歌声の性質を踏まえ，ピッチシフトとタイムストレッチの 2 つを用いてサンプルを変換し，学習を行う．ただし，一般的な自己教師あり対照学習とは異なり，あるサンプルの特徴量表現とそのサンプルをピッチシフトやタイムストレッチしたものの特徴量表現を識別するようにニューラルネットワークを学習する．これにより，声質や歌唱表現の違いに敏感な特徴量表現の獲得を可能にする．本研究ではその効果を，500 人の歌声サンプルから歌手ラベルを識別するタスクによって検証を行った．その結果，上記のようにピッチシフト・タイムストレッチを適用して獲得された特徴量表現を識別器の入力とすることで，これらの変換を用いずに獲得された特徴量表現を入力とした場合に比べ，識別精度が 9.12%向上することが確認された．さらに提案手法は，変換の適用方法を変更することにより，声質や歌唱表現のいずれかのみに敏感な特徴量表現を獲得するよう拡張することができる．実際，そうした特徴量表現によって歌のジャンル，歌手の性別，発声技法を捉えられることが確認できており，これは提案手法のさらなる応用可能性を示唆するものである．

1. はじめに

特徴量（高次元空間のベクトル）表現の獲得は，音楽推薦など様々な応用に紐づく音楽情報処理の要素技術の 1 つである [1,2]．しかし，歌声は音楽において重要である [3] にも関わらず，歌声に特化した特徴量表現の獲得手法の研究，特に深層学習を用いた手法の研究は多くない．これは，深層学習では大規模なデータセットが必要になる一方で，そのようなデータセットを準備し，適切なアノテーションを付与することのコストが障壁になっているためである．

そこで本研究では，**図 1** のように自己教師あり対照学習を導入することで，多声音楽から分離された歌声から特徴量表現を獲得する手法を提案し，その有用性を確認した [4]．自己教師あり対照学習では，データセットに含まれるそれぞれのサンプルについて，その特徴量表現と，そのサンプルを自動変換したものの特徴量表現とが近くなるように，そして異なるサンプルの特徴量表現とは遠くなるように，ニューラルネットワークを学習する．ここでの「自動変換」として，例えば画像ドメインではノイズの付加，色の歪み，反転などがよく用いられており，これにより，こうした変換に対してロバストな（影響を受けにくい）特徴量表現がアノテーション情報なしに獲得できる [5]．この学習方法は画像・映像ドメインを中心に発展してきており，様々なタスクでの精度向上に貢献してきた [5,6]．

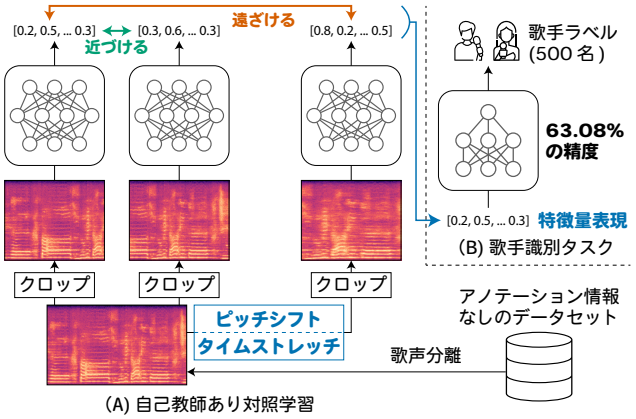


図 1: （A）自己教師あり対照学習を工夫しながら導入することで，歌声に特化した特徴量表現を獲得する．（B）獲得された特徴量表現は各歌手の違いを捉えるものとなっており，単純な識別器でも高い精度で歌手の識別が可能となる．

提案手法は，この自動変換に歌声の性質を考慮して「ピッチシフト」と「タイムストレッチ」の 2 つの変換を導入することで，歌声に特化した特徴量表現の獲得を可能にする．ただし一般的な自己教師あり対照学習とは異なり，元のサンプルの特徴量表現と変換後のサンプルの特徴量表現が遠ざかるようにネットワークを学習する．これにより，これ

1 筑波大学
2 産業技術総合研究所
a) hiromu.yakura@ipsj.or.jp
b) kento.watanabe@aist.go.jp
c) m.goto@aist.go.jp

ら自動変換に対して敏感な（影響を受けやすい）特徴量表現を獲得することが期待できる．

さらにピッチシフトについては，時間領域でのリサンプリングの後にタイムストレッチを適用することによって音高を変更するというナイーブな方法で自動変換を行う．これにより，変換対象サンプルの音高の変化に加えて，敢えてフォルマントも変化させる．フォルマントが変わると声質も変わるので，このナイーブなピッチシフトに敏感な特徴量表現は「声質の違い」を捉えられるものになると考えられる．同様にタイムストレッチについても音高が変化してしまうナイーブな方法で自動変換を行う．これにより変換の対象となるサンプルにおけるビブラートの速さ（細かさ）や音高の切り替わりにおける基本周波数（F0）の勾配など，歌唱表現の要素を変化させることになる [7]．そのため，このナイーブなタイムストレッチに敏感な特徴量表現は「歌唱表現の違い」を捉えられるものになると考えられる．

また提案手法は，声質と歌唱表現の両方を踏まえた歌手固有の特徴量表現を獲得できるだけでなく，そのいずれかのみに特化した特徴量表現を獲得するよう拡張することもできる．例えば，ピッチシフトされたサンプルの特徴量表現を元のサンプルの特徴量表現と近づけながら，タイムストレッチされたサンプルの特徴量表現を遠ざけるようにネットワークを学習すれば，歌唱表現の違いには敏感だが声質の違いには左右されないような特徴量表現を獲得できる．これにより，提案手法の応用可能性を大きく拡がる．例えば，声質は異なるが歌唱表現は類似している歌声を検索するということは，メル周波数ケプストラム係数（MFCC）のような従来の特徴量では実現が困難だが，提案手法を用いればアノテーション情報なしに実現できる．

本研究ではまず，提案手法における自己教師あり対照学習の有効性の検証を，獲得した特徴量表現を歌手識別のタスクに応用することによって行った．その結果，500 名の歌手を対象に 63.08%の精度を達成し，そのうち 9.12%はピッチシフトとタイムストレッチの導入に起因するものであった．次に，声質と歌唱表現のいずれかのみに敏感となるようにネットワークを学習し，獲得した特徴量表現の性質を検討した．その結果，アノテーション情報を用いずとも，声質または歌唱表現のどちらかのみが類似する歌声の検索が実現できることが示唆された．

2. 背景

2.1 自己教師あり対照学習

1 節で述べたように，自己教師あり対照学習はアノテーション情報のないデータセットからの特徴量（高次元空間のベクトル）表現の獲得を可能にする強力な手法である [8,9]．**図 2** A に示すように，自己教師あり対照学習は正例および負例と呼ばれる概念を導入する．正例はデータ

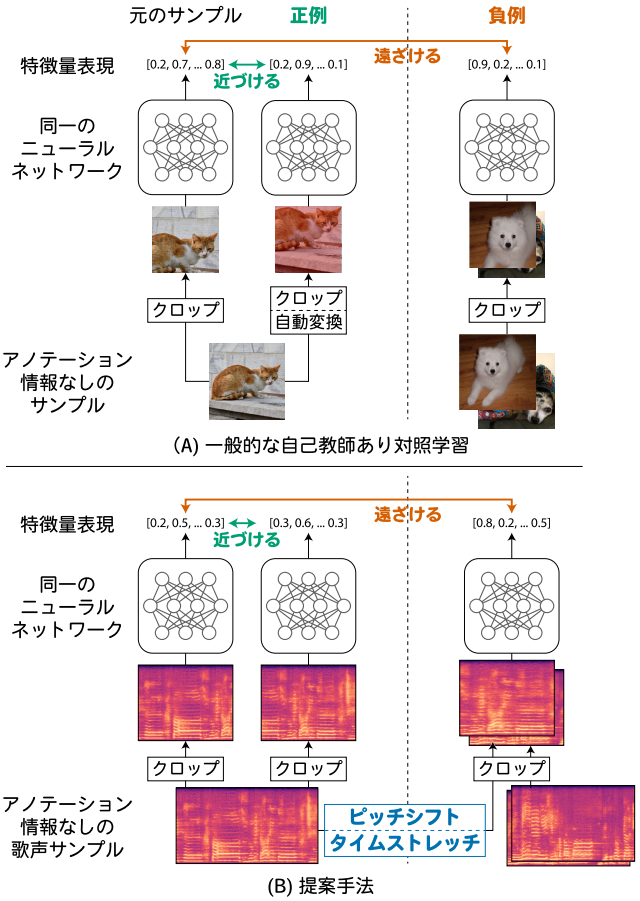


図 2: （A）一般的な自己教師あり対照学習はデータセット中のサンプルに自動変換を適用して正例を得ることで，ロバストな特徴量表現を獲得する．（B）提案手法は，ピッチシフトとタイムストレッチを適用して負例を得ることで，細かな違いに敏感な特徴量表現を獲得する．

セット内のあるサンプルに何らかの自動変換を適用することで得られるものを指し，負例はそのサンプル以外のデータセット内の他のサンプルを指す．そして，当該サンプルの特徴量表現が正例の特徴量表現と近く，かつ，負例の特徴量表現と遠くなるようにネットワークを学習する．これにより，データセット中のサンプル同士をロバストに区別する特徴量表現を獲得でき，未知のサンプルを含めたサンプル間の類似度計算も可能になる．実際，こうして獲得された特徴量表現が，画像ドメインにおける画像分類や物体認識，動画ドメインにおける行動認識など様々なタスクにおいて性能向上に寄与することが知られている [5,6]．

ここで，データセット $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ を用いて，ニューラルネットワーク $f_\theta(\cdot)$ を学習することで，そうした特徴量表現の獲得を行うものとする．このとき自己教師あり対照学習では，データセットのそれぞれのサンプル $\boldsymbol{x}_i$ に対し，K^+ 個の正例 $X_i^+ = \left\{\boldsymbol{x}_{i,1}^+, \boldsymbol{x}_{i,2}^+, \cdots, \boldsymbol{x}_{i,K^+}^+\right\}$ および K^- 個の負例 $X_i^- = \left\{\boldsymbol{x}_{i,1}^-, \boldsymbol{x}_{i,2}^-, \cdots, \boldsymbol{x}_{i,K^-}^-\right\}$ を用意する．そして，以下の損失関数を最小化するようにパラメータ θ を最適化することによって，ネットワークの学習を行う．

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{K^+}\log\frac{\hat{\mathcal{L}}_\theta\left(\boldsymbol{x}_i, \boldsymbol{x}_{i,j}^+\right)}{\hat{\mathcal{L}}_\theta\left(\boldsymbol{x}_i, \boldsymbol{x}_{i,j}^+\right) + \sum_{k=1}^{K^-}\hat{\mathcal{L}}_\theta\left(\boldsymbol{x}_i, \boldsymbol{x}_{i,k}^-\right)}$$

$$\hat{\mathcal{L}}_\theta(\boldsymbol{a}, \boldsymbol{b}) = \exp\left(\frac{f_\theta(\boldsymbol{a})\cdot f_\theta(\boldsymbol{b})}{\tau}\right) \quad (1)$$

なお，τ は獲得される特徴量表現の凝集性を調整するハイパーパラメータである [9]．また前述の通り，正例 X_i^+ は $\boldsymbol{x}_i$ に様々な自動変換を適用することによって得られるのに対し，負例 X_i^- には $K^- = N-1$ として $X \setminus \{\boldsymbol{x}_i\}$ が充てられることが多い．

この定式化をベースに，学習のさらなる効率化を図る手法が多く提案されている [8,9]．例えば，SimCLR [8] は一定数の正例を X_i^+ として事前に用意しておく代わりに，学習中に自動変換を適用して正例を生成するという手法を採っている．その自動変換に係るパラメータを動的に変更することで正例の多様性を高め，よりロバストな特徴量表現の獲得を可能にしている．また，MoCo [9] はモメンタムエンコーダを導入することで，学習時の計算効率とメモリ効率の向上を図っている．

2.2 歌声の特徴量表現の獲得

このようにアノテーション情報なしに特徴量表現が獲得できるという利点があるにも関わらず，音楽情報処理における自己教師あり対照学習の活用はまだ多くない [10–12]．例えば CLMR [12] は，SimCLR [8] をベースにガウスノイズの付加や周波数フィルタの適用といった自動変換を正例の生成に導入することで，楽曲へのタグ付けタスクにおいて state-of-the-art の性能を達成したと報告している．音響信号の特徴量を獲得するという点ではこれら手法を歌声へと適用し，歌手識別タスクなどに応用することは可能だが，歌声に特化した手法ではないために声質や歌唱表現のいずれかのみを考慮するといったことはできない．

もちろん，自己教師あり対照学習を用いない手法の提案は多くあり，特に歌手識別を目的にした研究は広く行われてきた [13–20]．これらは，MFCC や線形予測符号（LPC）といった古典的な特徴量表現を用いたもの [13–16]，ビブラートなどを考慮するよう独自の特徴量表現を設計したもの [17,18]，また深層学習を用いたもの [19,20] に大別することができる．しかし，あくまで歌手識別を目的としており，自己教師あり対照学習を用いた場合のように柔軟な類似度計算を行うことはできない．

検索などへの応用も踏まえて類似度計算を可能にした手法としては，MFCC や LPC に混合ガウス分布や潜在ディリクレ配分法を組み合わせた手法がある [21,22]．ただし，MFCC や LPC は主に声質を反映するため，歌唱表現までを考慮に入れることが難しいという制限がある．他にも深層学習，特に対照学習を取り入れた手法も存在する [23,24] が，自己教師ありではなく歌手情報のアノテーションを必要とするという点に違いがある．本研究は自己教師ありで，歌声の性質を踏まえた自動変換の導入によって，声質または歌唱表現のどちらかのみに着目した類似度計算も可能にした点が提案手法の重要な新規性の一つとなっている．

3. 自己教師あり対照学習による歌声の特徴量表現の獲得

3.1 提案手法

提案手法は MoCo [9] を拡張し，ピッチシフトとタイムストレッチを導入することによって歌声に特化した特徴量（高次元空間のベクトル）表現の獲得を可能にした．MoCo は一般的な自己教師あり対照学習（図 2 A）と同様に，データセット内のサンプルの特徴量表現と，そのサンプルからの自動変換によって得られた正例の特徴量表現との内積が大きくなるようにネットワークを学習する．一方，提案手法では導入した 2 つの自動変換によって（正例ではなく）負例を生成し，その特徴量表現と元のサンプルの特徴量表現との内積が小さくなるように学習する（図 2 B）．これにより，ピッチシフトとタイムストレッチによって生まれる変化に敏感な特徴量表現を獲得することができる．

なお自動変換によって負例を得る手法は，Tao ら [25] が動画を対象に既に提案しており，動画のフレームを時間軸方向にシャッフルして得られた動画を負例として用いたところ，人物行動認識タスクにおいて高い性能を達成する特徴量表現を獲得できたと報告している．そこで本研究でも予備実験として，歌声を時間軸方向にシャッフルして得られたものを負例として用い，獲得された特徴量表現を検証してみた．しかし，時間軸方向のシャッフルは歌声の音響信号に不自然なアーティファクト（非連続的な変化）を生じさせ，結果としてそうしたアーティファクトにのみ敏感な特徴量表現となることが判明した．つまり，歌声の特徴を捉えた特徴量表現を獲得するには，自動変換によって得られる負例が，他の歌手による自然な歌声としてデータセットに含まれうるようなものである必要がある．

そこで本研究で新規に導入したのが，ピッチシフトとタイムストレッチの 2 つの変換である．これらの変換を負例の生成に用いて学習されたニューラルネットワークは，それぞれ声質と歌唱表現を区別するような特徴量表現を出力するようになる．例えば，声質はスペクトル包絡とフォルマントに依存する [26] ために，時間領域でのリサンプリングに基づくナイーブなピッチシフト[*1]を適用された歌声は，これらの成分が大きく変化してしまい，異なる声質の歌声として知覚されるようなものになる．そのため，ピッチシフトした歌声を区別するようにネットワークを学習させる（すなわち，ピッチシフトした歌声を負例として扱う）と，声質の変化に敏感な特徴量表現が獲得できる．

[*1] ここで TD-PSOLA [27] のようなフォルマントを維持するピッチシフトを適用すると，声質に敏感な特徴量表現は獲得できない．

タイムストレッチは，ビブラートの速さ（細かさ）や音高の切り替わりにおける F0 の勾配など，微小時間内に発生する歌唱表現（アーティキュレーション）を変化させてしまう．一方で，同じ歌手によるビブラートの速さは曲のテンポによらずほぼ一定であることが知られており [28]，その歌手の歌唱表現を際立たせる重要な要素となっている [29]．また，音高の切り替わりにおける F0 の勾配も個人の歌唱スタイルを反映するもので，同じく歌唱表現を捉える手がかりとなる [30, 31]．つまり，音高が変化してしまうナイーブなタイムストレッチによってこれらを変化させることは，その歌声に含まれる歌唱表現の個人性を損なわせることにつながる．そのため，タイムストレッチした歌声を区別するようにネットワークを学習させる（すなわち，タイムストレッチした歌声を負例として扱う）と，歌唱表現の変化に敏感な特徴量表現が獲得できる．

ここで図 2 B に立ち戻ると，ネットワークの学習のためには正例も用意する必要がある．これは，データセット内のサンプルを一定の秒数にクロップしてからネットワークに入力するという実装にすることによって解決できる．つまり，あるサンプルに含まれる歌声は同一の歌手によるものだと考えられるため，そのサンプルの別の位置のクロップを正例として，それらから得られる特徴量表現との内積が大きい（つまり，類似度が高い）ものになるようネットワークを学習すればよい．これには，データセットに含まれる歌声の長さが統一されていなかった場合に，可変長の入力を受け付けるような構造のネットワークを用いずとも，固定長の入力を受け付ける比較的単純な構造のネットワークを用いることができるという利点もある．さらにクロップする位置を動的に変更するようにすれば，SimCLR [8] と同様に正例の多様性を高めることにもつながる．

ここまでを 2.1 節の表記に従ってまとめると，提案手法は導入した自動変換を用いて負例 X_i^- を拡張していると言える．つまり，一般的な自己教師あり対照学習ではデータセット中の他のサンプルのみからなる X_i^- を，ピッチシフトあるいはタイムストレッチしたバージョンの $\boldsymbol{x}_i$ によって補完していることになる．一方，正例 X_i^+ には $\boldsymbol{x}_i$ と同じ歌声からクロップされた $\tilde{\boldsymbol{x}}_i$ が充てられる．以上により声質と歌唱表現の変化に敏感で，各歌手の違いを捉えられるような特徴量表現を獲得することができる．

3.2 データ

提案手法を実現するためには，多くのサンプルを含む歌声のデータセットを用意する必要がある．特に自己教師あり対照学習は，ネットワークの学習に用いるサンプル数が多くなるほど，様々なタスクでの性能向上を示すことが知られている [32]．そこで，Million Song Dataset [33] に対応する 30 秒の楽曲サンプルからなるデータセットを構築した既存研究 [12, 34] からヒントを得て，新たな大規模データセットを構築した．そしてすべての 30 秒の楽曲サンプルに Spleeter [35] を適用して歌声を分離し，さらに歌唱区間検出 [36] を用いて検出された 0.5 秒以上の連続する無音区間をカットするという前処理も行った．

なお，このデータセットには楽曲ごとのアーティスト名の情報も含まれているが，提案手法の自己教師あり対照学習ではアーティスト名やその他のメタデータを除いた音響信号のみを用いるようにした．ただし，このアーティスト名の情報を用いて，事前に 4 節で用いる検証用データセットを分割することは行った．具体的には，データセット中に 50 曲以上の登録があるアーティスト 500 名をランダムに抽出し，それらアーティストの楽曲を取り除いたものを学習用データセットとした．結果として，328,418 曲からなる学習用データセットを自己教師あり対照学習に用いた．

3.3 実装

こうして構築したデータセットを用いて，3.1 節で述べた提案手法に従ってニューラルネットワークを学習した．ここで，学習の対象となるネットワークには，既存研究 [20] における教師あり学習での歌手識別にも用いられていた CRNN [37] を選んだ．これを，5 秒にクロップされた歌声を入力として，256 次元の特徴量表現ベクトルを出力するよう PyTorch を用いて実装した．[*2]

また，学習中に負例を得るために適用する自動変換としては「音高を 3 半音上げる」「音高を 3 半音下げる」「速度を 1.70 倍にする」「速度を 0.65 倍にする」という 4 種類を用意した．これらのパラメータの決定には，3.1 節で述べた観点が影響している．つまり，変化度合いが大きすぎると自動変換によって生まれる不自然なアーティファクトを捉えるような特徴量表現となってしまう一方で，変化度合いが小さすぎると対照学習のための有効な負例にならない可能性がある．そうした点から変換後のサンプルの自然さと，元のサンプルとの区別が十分につくかという観点とのバランスを取って，上記の通りに決定した．

4. 歌手識別タスクでの評価

提案手法によって獲得された特徴量表現の有効性を確認すべく，本研究ではまず歌手識別タスクでの検証を行った．ここでは，特徴量表現を入力としてアーティスト名のラベルを出力するように識別器を学習し，その精度を既存の特徴量表現を使用して学習した場合と比較した．

4.1 データ

検証には，3.2 節にて自己教師あり対照学習に用いる学習用データセットから取り除いておいた 500 名のアーティストの歌声を用いた．そして，各アーティストごとに 50

[*2] ソースコードを `https://github.com/hiromu/contrastive-singing-voices` にて公開している．

曲をランダムに抽出し，合計 25,000 曲分のサンプルからなる検証用データセットを構築した．そしてアーティストごとに `train`：`validation`：`test` = 40 曲：5 曲：5 曲の分割を設定し，最終的な精度評価には `test` に割り当てられた合計 2,500 曲分のサンプルを用いた．

4.2 手順

前述の通り，獲得された特徴量表現と対応するアーティスト名のラベルによって識別器を教師あり学習することで歌手識別を行ったが，ここで用いた識別器としては非常に「浅い」（表現力の弱い）ものを選んだ．これは，SimCLR [8] や MoCo [9] でも取り入れられている検証アプローチである．ここで，そのような表現力の制約された識別器であっても応用タスクで高い精度を出せるということが確認できれば，その入力として用いられた特徴量表現はサンプルの特徴をよく捉えたものであると言える．本研究では CLMR [12] の実験設計に従って，獲得した特徴量表現とアーティスト名のラベルの対応関係を 512 次元の隠れ層を持つ 3 層パーセプトロンに学習させた．

比較対象としては，楽曲へのタグ付けタスクにおいて state-of-the-art の性能を達成した（2.2 節参照）CLMR [12] で獲得された特徴量表現を用いた．妥当な比較条件となるように，3.2 節と同じ 328,418 曲分の学習用歌声データセットを使用して CLMR の自己教師あり対照学習を行い，そうして得られたネットワークによって 4.1 節の 25,000 曲分の検証用データセットから特徴量表現を求めた．そして，それら特徴量表現を入力として同じ 3 層パーセプトロンを学習し，精度評価を行った．

またベースラインとして，2.2 節で述べた既存研究にて歌手識別に用いられてきた古典的な特徴量表現も用意した．具体的には，Wang 及び Tzanetakis [23] がベースラインとしていた，クロマベクトル（12 次元），MFCC（20 次元），スペクトル重心・ロールオフ・スペクトル（3 次元）の平均と標準偏差からなる 70 次元の特徴量表現を用いた．そして前述の条件と同様に 3 層パーセプトロンを学習し，精度評価を行った．

4.3 結果

精度評価の結果を**表 1** に示す．提案手法に基づき，ピッチシフト及びタイムストレッチによって生成された負例を用いながら CRNN の学習を行った場合に，最も高い精度（Top-1 精度：63.08%）を得られることが確認できた．なお既存研究では，Million Song Dataset [33] に含まれる 500 名のアーティストの歌手識別タスクに対して，深層学習ベースの手法によって 39.3%の精度が得られたという報告 [24] がある．学習に用いたデータセットが異なるのに加え，学習手法についても「教師あり」と「自己教師あり」という違いがあるために結果の直接比較は難しい[*3]が，提案手法は比較的高い精度を達成していると言える．

表 1: 500 名からなる歌手識別タスクの実験結果.

特徴量表現	Top-1 精度	Top-5 精度
MFCC 等からなるベースライン	0.20%	1.00%
CLMR [12]	47.96%	73.64%
提案手法（負例の自動生成なし）	53.96%	76.60%
提案手法（タイムストレッチを利用）	61.00%	81.16%
提案手法（ピッチシフトを利用）	61.28%	81.72%
提案手法（上記 2 つを利用）	**63.08%**	**82.16%**

また CLMR [12] との比較を通して，提案手法の利点をより明らかにすることができる．具体的には，提案手法で用いたネットワークのパラメータ数（489 k）は CLMR（2.9 M）に比べて遥かに小さいにも関わらず，提案手法はピッチシフトやタイムストレッチを用いない場合においても同等の精度を達成することが確認された．これは，一般的な自己教師あり対照学習と同様に，CLMR が（タイムストレッチは用いていないものの）ピッチシフトを正例を得るための自動変換として用いていることに起因すると考えられる．つまり提案手法とは逆に，ピッチシフトで生まれる声質の変化に対してロバストな（声質の変化を区別しない）特徴量表現になっていると示唆される．

加えて，提案手法はピッチシフトとタイムストレッチをその学習に組み込むことによって，さらなる精度向上を実現している．特にその両方を負例の生成に用いることによって，表 1 では最も高い精度（Top1 精度：63.08%）を達成している．これは，この 2 つの自動変換によって得られた特徴量表現が声質や歌唱表現の違いに敏感なものになるという仮説に合致するもので，その特徴量表現が CLMR を超える精度を達成する上で大きく貢献した．

一方で，ベースラインとして用意した特徴量表現で 3 層パーセプトロンを学習した結果は，ランダムな出力を行う識別器とほぼ変わらない精度となった．これは，ベースラインの特徴量表現が artist20 [15, 16] のような 20 名程度の歌手識別タスクでは有効だったものの，より多数の歌手を含むデータセットではその有効性が損なわれたという Wang 及び Tzanetakis の報告 [23] に合致するものである．また，パーセプトロンの出力を具体的に確認したところ，一部の歌手にその推論結果が集中しており，結果としてランダムな出力に近い精度になってしまったということが分

[*3] Lee 及び Nam らの手法 [24] はアノテーション情報を用いた教師あり学習である上に，提案手法が用いるような歌声分離を適用して構築されたデータセットではなく，DAMP [38] のようにクリーンな歌声のデータセットを必要とすることから，本研究では比較条件に含めなかった．ただし，参考値として Lee 及び Nam [24] の公開している DAMP [38] での学習済みネットワークで得た特徴量表現を用いて，同じ 3 層パーセプトロンでの精度評価を行った．その結果，Top-1 精度が 47.9%，Top-5 精度が 71.2%となり，依然として提案手法の方が高い精度となることが確かめられた．

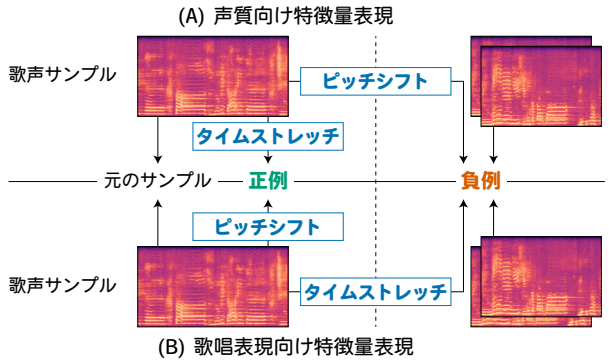


図 3: （A）タイムストレッチされたサンプルを正例に充てることで，声質の違いを捉える特徴量表現が得られる．（B）逆にピッチシフトされたサンプルを正例に充てることで，歌唱表現の違いを捉える特徴量表現が得られる．

かった．ここからも，提案手法によって獲得された特徴量表現は同じパーセプトロンでも 500 名のアーティストの識別を可能にする情報を含んでいたという点で，提案手法の有効性を支持できる．

5. 声質または歌唱表現に特化した特徴量表現

4.3 節ではピッチシフトとタイムストレッチを負例の生成に用いたが，1 節で述べたようにこれらを正例の生成に用いることで異なる性質を持つ特徴量表現が獲得できる．つまり，$\boldsymbol{x}_i$ にピッチシフトまたはタイムストレッチのいずれかを適用した得られたものを，負例 X_i^- ではなく正例 X_i^+ に充てて自己教師あり対照学習を行う．これによってアノテーション情報なしに，声質または歌唱表現のどちらかのみに敏感な特徴量表現を得られると期待できる．

本研究では，実際にこれらの自動変換の適用方法を切り替えてニューラルネットワークの学習を行い，獲得された特徴量表現の性質を検証した．具体的には，**図 3** A のようにタイムストレッチされたサンプルを正例に，ピッチシフトされたサンプルを負例に充てる実装と，図 3 B のようにピッチシフトされたサンプルを正例に，タイムストレッチされたサンプルを負例に充てる実装を用意した．そして 3.3 節 と同様に 3.2 節のデータセットで CRNN を学習し，データセット内のそれぞれのサンプルの特徴量表現を得た．以下，図 3 A の実装で獲得した特徴量表現を「声質向け」の特徴量表現と呼び，図 3 B の実装で獲得した特徴量表現を「歌唱表現向け」の特徴量表現と呼ぶ．

5.1 楽曲ジャンル

我々はまず，音楽学において楽曲ジャンルと歌唱スタイルとの関連性が指摘されていること [39] を踏まえ，特徴量表現と楽曲ジャンルの関連性を検討した．ここでは，データセット内の各サンプルのメタデータに含まれていたジャンル情報 (3.2 節参照) を使用し，ジャンルによって含まれる歌声の多様性がどう変化するのかを調査した．より具体的には，ジャンルごとに分散 (対応するジャンルに含まれる特徴量表現の平均からの L_2 距離の二乗平均）を計算し，その値によって**表 2** に主要なジャンルを並べた．

表 2: 2 種類の特徴量表現のジャンル内での分散の比較．

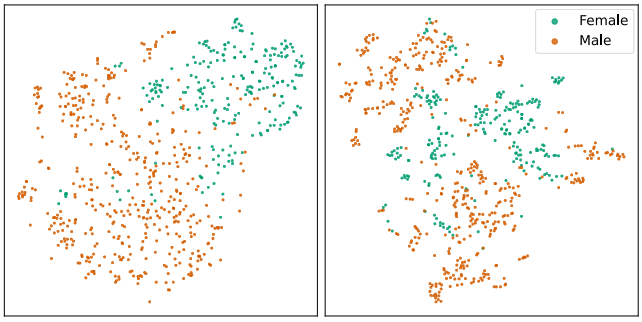
声質向けの特徴量表現	歌唱表現向けの特徴量表現	分散
Alternative	J-Pop	大
Rock	Rock	↑
Pop	Pop	
J-Pop	Alternative	
Folk	Anime	
Country	Hard Rock	
Blues	Folk	
Anime	Blues	
Metal	Reggae	
Hard Rock	Metal	
Reggae	Country	
Hip-Hop/Rap	Hip-Hop/Rap	小

表 2 では，独特の歌唱スタイルを持つことで知られる *Hip-Hop/Rap* はどちらの特徴量表現でも分散が最も小さくなっている．逆に，幅広いスタイルの曲を含むことの多い *Alternative*，*Rock*，*Pop*，*J-Pop* は上位にランクインしており，多彩な声質や歌唱表現を持つことを反映したものと考えられる．ここで，Malaway [39] はヒップホップやラップにおいて，アーティキュレーションや韻律よりもライムに歌手の個人性が現れると指摘しており，表 2 の結果はその指摘と整合していると言える．

また，*Country* は声質向けの特徴量表現と歌唱表現向けの特徴量表現で，大きく異なる傾向を示した．Malaway [39] はカントリー音楽に関して，Feld ら [40] の人類学的見地からの分析によると歌唱スタイルが統制されやすいことを指摘しており，これは歌唱表現向けの特徴量表現の分散が小さくなっていることと一致する．逆に，*Anime* は歌唱表現向けの特徴量表現の分散が大きいのに対し，声質向けの特徴量表現の分散が小さいという逆の傾向を示した．これは，このジャンルではアニメキャラクターの声優によって歌われた楽曲が多く含まれており，そうした場合に演技上の工夫として異なる歌唱表現を使い分けることは容易でも，異なる声質を使い分けながら歌うのは身体構造上の制約から相対的により難しいことが影響している可能性がある．

5.2 歌手の性別

別の観点として，特徴量表現と歌手の性別の関連性も検討した．これは，同じ性別の歌声は異なる性別の歌声よりも相対的に近い声質を持つ（つまり，性別の違いは声質向けの特徴量表現では区別しやすく，歌唱表現向けの特徴量表現では区別しにくい）と想定できるためである．ただし 3.2 節のデータセット内のメタデータに歌手の性別情報は含まれていなかったため，表 2 にリストされている 12 のジャンルごとに 5 名ずつのアーティストをランダムに抽出



(A) 声質向け特徴量表現　(B) 歌唱表現向け特徴量表現

図 4: t-SNE による特徴量表現の可視化．声質向けの特徴量表現をプロットした（A）は，性別間の区別が歌唱表現向けの特徴量表現をプロットした（B）よりも明確である．

し，手動で性別情報を付加した．そして**図 4** に示すように，t-SNE [41] を使用して 2 次元空間にマッピングすることにより，特徴量表現と歌手の性別の関係を可視化した．

その結果，声質向けの特徴量表現（図 4 A）では，歌唱表現向けの特徴量表現（図 4 B）と比べて、各性別がより明確に分離して分布していた．この結果と 5.1 節の結果はいずれも，声質向けと歌唱表現向けの特徴量表現が，それぞれ歌声の異なる性質を捉えていることを支持している．

5.3 VocalSet での検証

さらに，VocalSet [42] を用いて声質向けおよび歌唱表現向けの特徴量表現の性質を定量的に確認した．このデータセットには，20 名の歌手によるビブラートやトリルなどの発声技法ごとの歌声のサンプルが含まれており，声質向けの特徴量表現は歌手ごとの違いを，歌唱表現向けの特徴量表現は発声技法ごとの違いを捉えられると期待できる．そこでこれらのサンプルに対し，声質向けおよび歌唱表現向けの特徴量表現を得るように学習した前述のニューラルネットワークを適用し，対応する特徴量表現を得た．そして，それらの特徴量表現が類似しているサンプルがどういったものとなっているかを確認することで，特徴量表現の持つ性質を検証した．

より具体的には，データセット内の各サンプルをクエリとしたときに，他のサンプル全てを特徴量表現の類似度（コサイン類似度）によってランク付けするとどのような順序になるかを計算した．そして，上位にランク付けされた歌声がクエリと共通の歌手あるいは発声技法のものとなっているかどうかを，Mean Reciprocal Rank（*MRR*）[*4] [43] と Top-k 精度（*Prec@k*）を計算することによって調べた．ここでは，20 人の歌手による 9 種類の発声技法[*5]の歌声を使用した．そのため，ランダムにランク付けしたときの *Prec@k* は，歌手と発声技法についてそれぞれ 0.05 と 0.11 となる．

表 3: クエリとした歌声サンプルと同じ歌手または同じ発声技法のサンプルが，獲得した特徴量表現によって似ているとみなされる度合い（数値が大きいほど似ている）．

特徴量表現	歌手			発声技法		
	MRR	Prec@1	Prec@5	MRR	Prec@1	Prec@5
声質向け	0.826	0.747	0.601	0.756	0.640	0.525
歌唱表現向け	0.788	0.699	0.511	0.816	0.730	0.564

実験結果を**表 3** に示す．我々の予想の通り，同じ歌手のサンプルを検索する場合には，声質向けの特徴量表現の方が歌唱表現向けの特徴量表現よりも優れていることが確認できた．逆に，同じ発声技法のサンプルを検索する場合には，歌唱表現向けの特徴量表現の方が優れていることも確認できた．加えて Urbano ら [44] のガイドラインに従って Student の t 検定を行ったところ，*MRR*，*Prec@1* および *Prec@5* のいずれも，歌手と発声技法の両方について $p < 0.001$ という結果となった．つまり，声質向けと歌唱表現向けの特徴量表現では，優劣が有意に異なる実験結果が得られていたことが確認できた．

これらの結果から，ピッチシフトとタイムストレッチのいずれかによって正例を生成し，もう一方で負例を生成するという提案手法によって，声質や歌唱表現のどちらかについての違いを区別するような特徴量表現を得られたことがわかる．特に，ここで使用されている特徴量表現を得るための学習において，VocalSet のサンプルを（アノテーションを含め）一切使用していないことは特筆に値する．獲得された特徴量表現が，学習に用いたデータセットとは異なるデータセットの歌声の特徴を捉えられていたことは，提案手法の汎用性の高さを示唆している．

6. まとめ

歌声に特化した自己教師あり対照学習の新たな手法を提案した本研究の貢献を，以下にまとめる．

- ピッチシフトまたはタイムストレッチされたサンプルを区別するようニューラルネットワークを学習するという新たな手法により，アノテーションなしで声質や歌唱表現の違いに敏感な特徴量表現の獲得を可能にした．
- 500 名の歌手識別タスクで 63.08%という高い精度を達成し，獲得した特徴量表現が歌手識別タスクの精度の向上に役立つことを確認した．
- 提案手法の対照学習で正例と負例を変更する工夫により，声質または歌唱表現のいずれか一方について類似している歌声を検索することも可能なことを示した．

本研究によって得られた歌声の特徴量表現は，単なる歌手識別に留まらない幅広い応用の可能性を持っており，今後様々なタスクに応用していく予定である．

[*4] クエリと同じ歌手または発声技法のサンプルが 1 位にランク付けされた場合に，MRR は 1 となる．
[*5] VocalSet [42] に含まれる 10 種類の発声技法のうち，*spoken* は歌声でない発話を収録したサンプルとなっているため除外した．

謝辞 本研究は JST（JPMJAX200R, JPMJCR20D4）および JSPS（JP21J20353, JP21H04917）からの支援を受けた.

参考文献

[1] Ramírez, J. and Flores, M. J.: Machine learning for music genre: Multifaceted review and experimentation with Audioset, *J. Intell. Inf. Syst.*, Vol. 55, No. 3, pp. 469–499 (2020).
[2] Schedl, M.: Deep learning in music recommendation systems, *Front. Appl. Math. Stat.*, Vol. 5 (2019).
[3] Demetriou, A. M. et al.: Vocals in music matter: the relevance of vocals in the minds of listeners, *ISMIR*, pp. 514–520 (2018).
[4] Yakura, H. et al.: Self-Supervised Contrastive Learning for Singing Voices, *IEEE/ACM Trans. Audio, Speech, Language Process.*, Vol. 30, pp. 1614–1623 (2022).
[5] Jaiswal, A. et al.: A survey on contrastive self-supervised learning, *Technologies*, Vol. 9, No. 1 (2021).
[6] Jing, L. and Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 43, No. 11, pp. 4037–4058 (2021).
[7] Umbert, M. et al.: Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges, *IEEE Signal Process. Mag.*, Vol. 32, No. 6, pp. 55–73 (2015).
[8] Chen, T. et al.: A simple framework for contrastive learning of visual representations, *ICML*, pp. 1597–1607 (2020).
[9] He, K. et al.: Momentum contrast for unsupervised visual representation learning, *IEEE/CVF CVPR*, pp. 9726–9735 (2020).
[10] Wu, H. et al.: Multi-task self-supervised pre-training for music classification, *IEEE ICASSP*, pp. 556–560 (2021).
[11] Carr, A. N. et al.: Self-supervised learning of audio representations from permutations with differentiable ranking, *IEEE Signal Process. Lett.*, Vol. 28, pp. 708–712 (2021).
[12] Spijkervet, J. and Burgoyne, J. A.: Contrastive learning of musical representations, *ISMIR*, pp. 673–681 (2021).
[13] Zhang, T.: Automatic singer identification, *IEEE ICME*, pp. 33–36 (2003).
[14] Fujihara, H. et al.: Singer identification based on accompaniment sound reduction and reliable frame selection, *ISMIR*, pp. 329–336 (2005).
[15] Zhang, X. et al.: A novel singer identification method using GMM-UBM, *CSMT*, pp. 3–14 (2019).
[16] Murthy, Y. V. S. et al.: Singer identification for Indian singers using convolutional neural networks, *Int. J. Speech Technol.*, Vol. 24, No. 3, pp. 781–796 (2021).
[17] Nwe, T. L. and Li, H.: On fusion of timbre-motivated features for singing voice detection and singer identification, *IEEE ICASSP*, pp. 2225–2228 (2008).
[18] Loni, D. Y. and Subbaraman, S.: Timbre-vibrato model for singer identification, *ICTIS*, pp. 279–292 (2019).
[19] Nasrullah, Z. and Zhao, Y.: Music artist classification with convolutional recurrent neural networks, *IJCNN*, pp. 1–8 (2019).
[20] Hsieh, T. et al.: Addressing the confounds of accompaniments in singer identification, *IEEE ICASSP*, pp. 1–5 (2020).
[21] Fujihara, H. and Goto, M.: A music information retrieval system based on singing voice timbre, *ISMIR*, pp. 467–470 (2007).
[22] Nakano, T. et al.: Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity, *IEEE ICASSP*, pp. 5202–5206 (2014).
[23] Wang, C. and Tzanetakis, G.: Singing style investigation by residual siamese convolutional neural networks, *IEEE ICASSP*, pp. 116–120 (2018).
[24] Lee, K. and Nam, J.: Learning a joint embedding space of monophonic and mixed music signals for singing voice, *ISMIR*, pp. 295–302 (2019).
[25] Tao, L. et al.: Self-supervised video representation learning using inter-intra contrastive framework, *ACM Multimedia*, pp. 2193–2201 (2020).
[26] Sundberg, J.: *The science of the singing voice*, Northern Illinois University Press, DeKalb, IL (1987).
[27] Moulines, E. and Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Commun.*, Vol. 9, No. 5-6, pp. 453–467 (1990).
[28] Seashore, C. E.: *The vibrato*, Studies in the Psychology of Music, University of Iowa, Iowa City, IA (1932).
[29] Humphrey, E. J. et al.: An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music, *IEEE Signal Process. Mag.*, Vol. 36, No. 1, pp. 82–94 (2019).
[30] Ohishi, Y. et al.: Parameter estimation method of F0 control model for singing voices, *Interspeech*, pp. 139–142 (2008).
[31] Kako, T. et al.: Automatic identification for singing style based on sung melodic contour characterized in phase plane, *ISMIR*, pp. 393–398 (2009).
[32] Goyal, P. et al.: Scaling and benchmarking self-supervised visual representation learning, *IEEE/CVF ICCV*, pp. 6390–6399 (2019).
[33] Bertin-Mahieux, T. et al.: The million song dataset, *ISMIR*, pp. 591–596 (2011).
[34] Pons, J. et al.: End-to-end learning for music audio tagging at scale, *ISMIR*, pp. 637–644 (2018).
[35] Hennequin, R. et al.: Spleeter: A fast and efficient music source separation tool with pre-trained models, *J. Open Source Softw.*, Vol. 5, No. 50, p. 2154 (2020).
[36] Kum, S. and Nam, J.: Joint Detection and Classification of Singing Voice Melody Using Convolutional Recurrent Neural Networks, *Appl. Sci.*, Vol. 9, No. 7 (2019).
[37] Choi, K. et al.: Convolutional recurrent neural networks for music classification, *IEEE ICASSP*, pp. 2392–2396 (2017).
[38] Smith, J. C.: Correlation analyses of encoded music performance, PhD Thesis, Stanford University (2013).
[39] Malawey, V.: *A blaze of light in every word: Analyzing the popular singing voice*, Oxford University Press, Oxford, UK (2020).
[40] Feld, S. et al.: Vocal anthropology: From the music of language to the language of song, *A Companion to Linguistic Anthropology*, John Wiley & Sons, Ltd, chapter 14, pp. 321–345 (2005).
[41] van der Maaten, L. and Hinton, G.: Visualizing data using t-sNE, *J. Mach. Learn. Res.*, Vol. 9, No. 86, pp. 2579–2605 (2008).
[42] Wilkins, J. et al.: VocalSet: A singing voice dataset, *ISMIR*, pp. 468–474 (2018).
[43] Craswell, N.: Mean reciprocal rank, *Encyclopedia of Database Systems*, Springer, p. 1703 (2009).
[44] Urbano, J. et al.: Statistical significance testing in information retrieval: An empirical analysis of type I, type II and type III errors, *ACM SIGIR*, pp. 505–514 (2019).