概 要

歌声の音響信号からその声質を反映したベクトル表現を得ることで，対応する歌手の識別や似た歌声を持つ歌手の検索が可能となる．こうした特徴量表現の獲得を目的として，本研究は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を，歌声に特化した形で導入する．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とする．これは，学習データ内の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，提案手法では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対応する歌手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．実際に本研究では，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した．これは提案手法の,「声質は似ていないが歌い方を似ている歌手を探す」あるいは「歌い方は似ていないが声質は似ている歌手を探す」といった新たな形式の歌声検索という応用可能性を示唆するものである.（794 字）