話者認識技術を用いた歌唱者のダイアリゼーションの有効性評価

森 楓里 (指導教員：粕川正充)

1　はじめに

ボーカルのある楽曲において、歌唱している人数が複数の場合があり，これはグループやアーティストによるコラボレーション楽曲に見られる．このような楽曲を時間方向で歌い手ごとに分けることを邦楽ではパート割りと呼ばれ、歌い手聴き手双方において重要な情報となってくる．現在，歌詞や楽譜などの情報からパート割りの分析を行われることが多く，歌声の音声データに直接アプローチをかける手法はあまり存在しない．音声データから歌唱者の推定ができれば，結果の表示だけでなく，歌唱者の声質に応じた演出やより高性能な歌詞認識などへの応用が期待できる．

本研究では既存の手法である会話音声から「誰がいつ話しているか」を推定する話者ダイアリゼーション技術を用いて,「誰がいつ歌っているか」を推定する歌唱者のダイアリゼーションを行った．具体的には歌唱者 A と B の 2 人でのデュエットソングの場合,「A のソロパート」,「B のソロパート」,「A と B の 2 人でのデュエットパート」に分け，その結果を時間軸方向で歌い手ごとに結果をプロットする．

歌唱者ダイアリゼーションは民族音楽に対して行った研究 [1] が発表されている．しかしながら，この研究では評価に用いた楽曲の伴奏音やノイズの影響で推定誤りが多くなったことが問題点として挙げられている．

そこで本研究では，LaSAFT[2] を用いて複数の歌い手による楽曲から伴奏音を除去した後，完全教師あり話者ダイアリゼーション技術である UIS-RNN[3] を用いて歌唱者の識別を行った．大規模の会話学習データから学習済みの UIS-RNN 話者認識モデルを歌唱データの歌唱者認識に使用することで，話者ダイアリゼーション技術の歌唱者ダイアリゼーションにおける有効性の評価を行った．

2　研究概要

話者ダイアリゼーションは通常以下の手順である．

1. 前処理 無音区間を取り除き，発話区間を抽出する．

2. 特徴量の抽出 発話区間をいくつかのセグメントとして区切り，各セグメントから特徴量を抽出する．

3. クラスタリング 特徴量を用いて各セグメントに対して話者 ID を割り当て，話者数を推定する．

Ex. 再計算 追加で話者数などに制約がある場合に，再度 2. に戻ってセグメンテーションを行い，計算し直す．

2.1　前処理

複数の歌い手による楽曲において伴奏が除かれた音源を使用するため，伴奏除去を行った．本研究では歌声抽出において一番ドライボーカルに近かった LaSAFT を用いた．この手法による歌声の抽出は伴奏及びノイズ除去が完全ではないため理想的なドライボーカルの歌声ではないことに留意しつつ処理を行った．

得られた音源から，発話区間 (歌声部分) を音声区間検出技術 (VAD) を用いて抽出した．

2.2　特徴量の抽出

発話区間をいくつかのセグメントとして区切り，各セグメントから特徴量を抽出した．特徴量として話者認証ディープニューラルネットワークモデル d-vector[4] の隠れ層の出力を使用した．

具体的には UIS-RNN において使用した会話学習データの d-vector 抽出法 [5] を参考に歌唱データの特徴量を抽出した．

2.3　クラスタリング

セグメントに対して，歌唱者が同一の部分が同じクラスタに属するようにクラスタリングを行った．

話者ダイアリゼーションにおけるクラスタリングに関しては K-Means[6], スペクトラルクラスタリング [6] などの教師無しで行う方法が主流となっている．一方，UIS-RNN とは話者ダイアリゼーションを行うため完全教師ありのニューラルネットワークであり，音声データを訓練データとして学習させるものである．

本研究では抽出した特徴量を元に，大規模な会話データにより学習された UIS-RNN と，比較対象として教師無しで行えるスペクトラルクラスタリングの 2 つをクラスタリング手法として比較した．

2.3.1　スペクトラルクラスタリング

スペクトラルクラスタリングは機械学習のうちの教師なし学習に分類される．データからノード (点) を生成し，そのノードとデータの類似度に応じたグラフを生成し，クラスタに分類するアルゴリズムである．

2.3.2　UIS-RNN

話者ごとの特徴量を抽出した d-vector を入力データとして学習を行った．図 1 のように話者ごとに RNN の隠れ層の状態を定義し，状態を飛び飛びで保持する形でモデル化した. これによって初期条件として話者の人数を設定する必要をなくしている．話者データとして d-vector を入力すると，現在のデータから次のデータにかけて話者が入れ替わるのかどうかを予測した．話者が変わっていると判別した場合には，違う話者のネットワークに変更し，再計算される．ここまでを繰り返すことによって話者のクラスタリングを行った．

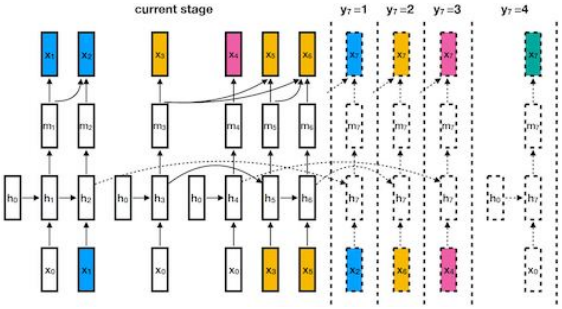

図 1: UIS-RNN の概要図

3　実行例

実行例として複数人で歌唱している市販楽曲 4 曲に対して実験を行った．音声データのサンプリング周波数は 44.1KHz，量子化ビット数は 16bit である．

また，本研究では問題を単純化し基礎的な検討を行うため，複数人で歌唱している楽曲のうち「各々がソロで歌っている部分」と「全員で歌っている部分のみ」を抽出し，歌唱者のダイアリゼーションを行った．

ダイアリゼーションの推定結果の例として楽曲 1 を UIS-RNN を用いてクラスタリングしたものを図 2 に示す．図 3 は，楽曲 1 の正解データを自分自身でラベル付けし，示したものである．横軸が時間で縦軸がクラスタリングされたの歌唱者 ID となっており，具体的には下から「女性と男性のデュエットパート」,「男性のソロパート」,「女性のソロパート」となっている．

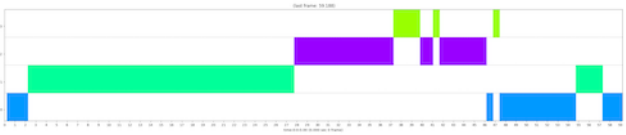

図 2: 楽曲 1 の UIS-RNN による実行例

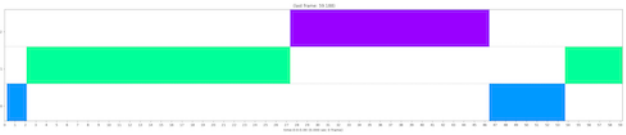

図 3: 楽曲 1 の正解データ

ダイアリゼーションの評価指標には DER (Diarization Error Rate) を用いた．

$$DER = \frac{FS + MS + SE}{T}$$

FS（False alarm Speech）... 発話者なしの区間で発話と誤判定した時間，MS（Missed Speech）... 発話者ありの区間で発話なしと誤判定した時間，SE （Speaker Error）... 話者を誤った時間，T ... 総発話時間を示す．

表 1: 計測結果 (DER)

手法	楽曲 1	楽曲 2	楽曲 3	楽曲 4
UIS-RNN	9.5%	16.7%	40.2%	20.8%
スペクトラル	34.7%	32.1%	43.1%	2.4%

民族音楽に対して行った研究 [1] では DER が 43.1% であったのに対して本研究の DER が改善されているのは用いた楽曲の伴奏音を除去し，ドライボーカルに近づけることができたことが要因である．また，クラスタリング手法で DER を比較するとスペクトラルクラスタリングに比べて UIS-RNN の方が DER の数値が低くなっている．しかしながら，UIS-RNN の手法においては楽曲間で DER に差があり，これは異なる歌唱者でも類似した声で同じようなテンポの曲調を歌唱した際に間違えて同一歌唱者であるとクラスタリングされることや，特にデュエットパートで同じ歌唱者でも各々の声の大きさが逆転した際に別々にクラスタリングされてしまうことが原因にあげられる．

4　まとめ

本研究では，話者ダイアリゼーション技術を複数の歌謡者による楽曲に用いて,「誰がいつ歌っているか」を推定する歌謡者のダイアリゼーションを行いその有効性を検証した．クラスタリングにおいて，会話データを用いて事前学習された UIS-RNN と，教師無しで行えるスペクトラルクラスタリングの二手法で行った．その結果，事前学習を行った話者認識モデルである UIS-RNN をクラスタリングに用いた方が推定の誤りが少なかった．しかしながら，男女や声質に差のある歌唱者達によるデュエットにおいてはダイアリゼーション推定が上手に行えたが，声質や歌い方の似ている歌唱者が同じフレーズを歌唱した場合や同じ歌唱者でも声の大きさや曲調が変わった場合に推定の誤りが多く見られ，楽曲の特徴によって結果に差が見られた．これは会話のデータセットを事前学習に用いたため，抑揚 (音の上下，大小) や，裏声などの話し声には無い歌声独特の特徴に対応しきれなかったことが原因である．

4.0.1　今後の課題

事前学習に歌声を用いることで，これらの特徴に対応した歌唱者認識モデルを構築したい．会話の訓練データセットはオープンソースとして公開されているものが多いものの，現状として歌声の訓練データセットで大規模に公開されているものがあまりないため，その訓練データセット (ソロの歌声) を楽曲から抽出して用意する必要がある．また，特徴量の抽出においては d-vector 以外にも GMM(gaussian mixture model) スーパーベクトルに因子分析の手法を用いることで MFCC から特徴量を抽出する i-vector や，近年開発された x-vector という i-vector を発展させた特徴量も存在する．これらで特徴量を抽出し，特徴量抽出法に差をつけての有効性の評価の実験したい．

参考文献

[1] Thlithi et al., ”Singer diarization: application to ethnomusicological recordings.” (2015)

[2] Woosung Choi et al., “ LaSAFT: Latent Source Attentive Frequency Transformation for Conditioned Source Separation., ” (2020)

[3] Aonan Zhang et al., “ Fully Supervised Speaker Diarization “ (2019)

[4] E. Variani et al., “ Deep neural networks for small footprint text-dependent speaker verification “ (2014)

[5] W. Xie et al., “ Utterancelevel aggregation for speaker recognition in the wild ” (2019)

[6] Quan Wang et al.,“Speaker diarization with lstm ” (2018)