6T-05

歌唱者ダイアライゼーションに向けた歌唱者識別手法の比較

田中麻衣 † 北原鉄朗 †
† 日本大学文理学部情報科学科

1. はじめに

複数人で歌唱している楽曲は，歌唱者の歌うフレーズが交互に入れ替わるような形式をとることがある．こうした時間軸での歌い分けを示した言葉としてパート割りという言葉がある．消費者にとって，パート割りの情報は，応援する歌唱者がどこを歌っているかを表す重要な情報となる．

本研究の目的は，パート割り，つまり，ある楽曲のどこからどこまでを誰が歌っているかを自動で推定することである．この処理は,「歌唱者ダイアライゼーション」と呼ばれている．関連する研究には，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ライゼーション[1]，アーティストの同定において背景音の除去を導入したもの[2] などがある．しかし，パート割りを識別する研究は少なく，歌声分析によるパート割り識別の基礎的検討を行った須田らの研究[3] 以外にはあまり行われていない．

課題となるのは，与えられる音源が背景音を含んでいるために，歌唱者の音響的特徴を忠実に抽出できない点である．須田らの研究[3] ではカラオケ音源を使用した音源分離手法が用いていたが，この手法ではカラオケ音源がない楽曲に適用できない．近年，音源分離技術の発展が目覚ましく，聴感上の歪みがかなり少ない状態で歌唱音源を抽出できる．この技術を活用すれば，ある程度の歌唱者ダイアライゼーションはできる可能性がある．

そこで本稿では，カラオケ音源を用いない音源分離手法で歌唱音源を抽出し，歌唱者識別を行い結果を比較する．識別手法には事前学習を用いない手法と，事前学習を用いる手法を用い，それぞれ精度を比較した．

2. 手法

本稿ではニューラルネットワークによる事前学習により作成された音源分離モデル demucs[5] を用いる．これによりカラオケ音源を用いずに背景音を分離する．demucs[5] により分離された音源を用いて，事前学習を必要としない手法と事前学習を必要とする手法で歌唱者の識別を行う．複数の話者が同時に歌唱する場合も考慮して話者ラベルを設定する．たとえば，2 人組のグループを対象とする場合「話者 A」「話者 B」「話者 A と話者 B が同時に歌唱」「歌唱なし」の 4 つのラベルを設定する．

2.1 事前学習を必要としない手法

事前学習を必要としない手法として，次の手法を試行する．

- ダイアライゼーションツールキット「LIUM」[4] を用いて歌唱者ごとのセグメンテーションを行った後，セグメントに対して k-means 法でクラスタリングを行う．
- LIUM を用いずに，0.5s, 1.0s, または 2.0s ごとに音響信号を切断し，得られたセグメントに対して k-means 法でクラスタリングを行う．特徴量には 20 次元の MFCC に対して時間軸方向に平均を取ったものを用る．

2.2 事前学習を用いた手法

教師付き学習によって話者識別モデルを学習し，0.5s, 1.0s, または 2.0s ごとのセグメントに対して話者識別を行う．入力は各セグメントから抽出される特徴量 MFCC の時系列，出力は話者ラベルを one-hot ベクトルとして表現したものである．また，学習のモデルとして，LSTM (long short-term memory) と CNN (convolutional neural network) を用いる．エピック数はいずれも 100 とした．

2.2.1 LSTM

2 層の LSTM を用いる．1 層目の LSTM では隠れ層のノード数を 16 とし，活性化関数に tanh 関数を利用した．2 層目の LSTM では隠れ層のノード数を 8 とし，1 層目と同様に活性化関数に tanh 関数を利用した．出力層の活性化関数には softmax 関数を利用した．

2.2.2 CNN

本稿では，2 層の畳み込み層（フィルタサイズ：3 × 3）と最大値プーリング層（サイズ：2 × 2）の組合せを複数繰り返し，全結合層を 2 層経て one-hot ベクトルを出力する．活性化関数には ReLU 関数（出力層のみ softmax）を用い，ドロップアウトおよびバッチ正規化も導入した．

3. 実験

3.1 実験条件

本稿ではグループソングのデータセットとして男性ボーカル 2 人 (北川悠仁，岩沢厚治) で構成されている「ゆず」の楽曲を用い，各楽曲に対して「ボーカル無し」「北川が歌唱」「岩沢が歌唱」「二人が歌唱」の 4 クラスへの分類を試みた（ただし，クラスタリングについてはラベルは付与されない）. パート割の正解データは，第 1 著者が実際に聴いて 10ms 単位で作成した．

3.2 事前学習を必要としない手法

「ゆず」の楽曲全 25 曲に対して 2.1 節で述べた 2 つの手法でクラスタリングの実験を行った. クラスタリングの評価には adjusted Rand index (ARI)[6] を使用した．

3.3 事前学習を用いた手法

事前学習を用いた手法では，加齢による歌声の変化の影響を確認するため，(1) 25 曲のうち発売年が古いもの 13 曲を学習データとして残り 12 曲をテストデータにする場合（年代順割当），(2) ランダムに 12 曲を学習でデータ，13 曲をテストデータにする場合（ランダム割当）の両方で実験を行った．0.5s, 1s または 2s ごとに分割した各セグメントに対し

Comparison of singer identification methods towards singer diarization
by Mai Tanaka and Tetsuro Kitahara (Nihon University)

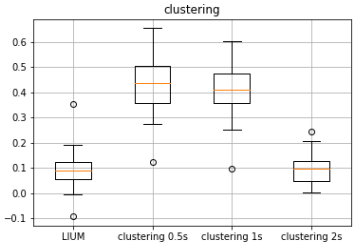

図 1 クラスタリング結果の評価（ARI）

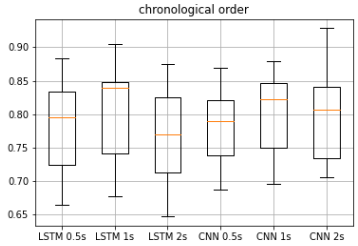

図 2 年代順割当の場合の正解率

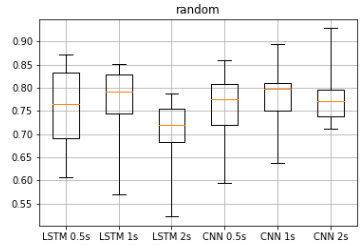

図 3 ランダム割当の場合の正解率

て 10ms ごとに抽出した 20 次元の MFCC の時系列を入力とし，10ms ごとの正解ラベルに対してセグメントごとに求めた最頻値を出力とした．評価には正解率を用いた．

3.4 実 験 結 果

3.4.1 事前学習を必要としない手法

クラスタリングの結果を**図 1** に示す．0.5s で分割した際の ARI が最も高く，セグメントが短いほど ARI が高いことがわかる．特に「人生芸無」では 0.5s で 0.656 であるのに対し，2s では 0.0019 であった．これは，セグメントが長いため, セグメント内でソロの歌唱パートと 2 人の同時歌唱パートが混在することになり，本来ソロパートとして識別されるべきパートが同時歌唱パートとして識別されてしまったからと考えられる．一方，0.5s，1s で ARI が最低だった「未練歌」（それぞれ 0.123，0.0984）は，2s の方が ARI が高かった（0.1383）これは,「未練歌」は同じ歌唱者が歌うパートが長いためであると考えられる．

LIUM を用いたものは，LIUM を用いずに 2s でセグメント分割した場合と同程度の ARI であった．LIUM が出力するセグメントのほとんどが 2s 以上であり，なかには 10s 以上のものもあったからと考えられる．LIUM を用いたもので比較的よい結果を残した楽曲として「人生芸無」が挙げられる．LIUM によるセグメントは 2〜4s のものが多く,「人生芸無」は 2〜3s で歌唱者が入れ替わっていることが，その要因と考えられる．

3.4.2 事前学習を用いた手法

年代順割当の場合の結果を**図 2** に示す．LSTM を用いた手法では，1s で分割した際の正解率の中央値が 0.8395 となっていた．特に「天国」では 0.9048 をとっていた．しかし 2s では正解率が下がる結果となった．CNN を用いた手法では 1s，2s で分割した際の正解率の中央値が 0.8 以上となっていた．特に 2s では「天国」で 0.9285 をとっていた.「天国」は全モデルで 0.83 以上の正解率であった．これは「天国」のパート割りが入れ替わりが少なく，同じ歌唱者による歌唱が長く続く単純な構造になっているからだと考えられる．

ランダム割当の場合の結果を**図 3** に示す．LSTM では 1s で分割した際の正解率の中央値が 0.7194 となっていた．一方，CNN ではセグメントが長くなるほど正解率が上がる傾向にある．特に「天国」「保土ヶ谷バイパス」「桜木町」では 2s で 0.9286，0.9123，0.9197 であった.「天国」「保土ヶ谷バイパス」は歌唱者の切り替わりが少ない，片方の歌唱者のソロパートがないなどパート割りが単純であるという共通点がある．他にも「風吹く町」「陽はまた昇る」「未練歌」等パート割りが単純な楽曲は存在したが，他の楽曲よりも高音域を歌っていることにより普段高音域を歌っている別の歌唱者に認識されたり，コーラスや楽器音が分離音に残っており，無歌唱の区間が歌唱者のパートして識別されるなどの誤識別により正解率が下がっていた.「桜木町」は，他 2 曲と比較するとパート割りが単純ではないが，音源分離によって背景音がほぼ除去されていたために，誤識別が少なかったと考えられる．

テストデータの比較をしたとき，CNN ではランダム割当の場合が優れており，2s が良い結果を残していた．LSTM ではテストデータによる差はなかったが，どちらも 1s でよい結果を残していた．

4. お わ り に

本稿では，複数歌唱者による歌唱の「パート割」の自動推定を目指し，事前学習を用いる手法と用いない手法の両方で歌唱者識別を試みた．どちらの手法でも音源を 0.5〜2s のセグメントに分割して識別を行うため，歌唱者交代の頻度などによって精度が上がるセグメント長が異なることが分かった．今後は，識別結果の平滑化などを行って精度を上げるとともに，推定されたパート割を見ながら楽曲鑑賞するアプリの開発も目指していく．

謝辞 本研究は，科研費 22H03711，21H03572 の支援を受けた．

参 考 文 献

1) Hanifa, R., M., Isa, K., and Mohamad, S., “A review on speaker recognition: Technology and challenges”, Computers & Electrical Engineering, Vol.90, No.107005, March 2021.
2) Sharma, B., Das, R. K., Li, H.“On the Importance of Audio-source Separation for Singer Identification in Polyphonic Music”, INTERSPEECH 2019，2019.
3) 須田仁志， 斎藤大輔, “歌声分析によるグループアイドルソングのパート割り構造認識に関する基礎的検討”，東京大学大学院工学系研究科電気系工学専攻 修士論文，2019.
4) “LIUM SpkDiarization”, https://projets-lium.univ-lemans.fr/spkdiarization/, 閲覧日: 2023-1-11.
5) D efossez, A., “Hybrid Spectrogram and Waveform Source Separation”, ISMIR 2021 MDX Workshop, p. 11, 2021.
6) Chacón, J. E., “A close-up comparison of the misclassification error distance and the adjusted Rand index for external clustering evaluation.” The British journal of mathematical and statistical psychology，2019.