Vol.2018-MUS-121 No.17
2018/11/21

グループアイドルソングを対象とした歌唱者ダイアライゼーション手法の基礎的検討

須田 仁志[1,a)]　深山 覚[2,b)]　中野 倫靖[2,c)]　齋藤 大輔[1,d)]　後藤 真孝[2,e)]

概要：本稿では，複数人が歌唱している楽曲に対して誰がいつ歌っているかを推定する歌唱者ダイアライゼーションの基礎的な検討を行う．とくに本稿ではグループアイドルソングのような複数の歌唱者が交互に歌ったり同時に歌ったりする楽曲を対象とする．本稿では伴奏音を除去した歌声を用いてアイドルソングのデータセットを構築した．またこれらの歌声に対して，歌唱者の音響モデルを未知とした手法と既知とした手法の 2 手法を用いて歌唱者ダイアライゼーションを行った．歌唱者の音響モデルを未知とした手法には，会話音声に対する話者ダイアライゼーションで広く用いられている修正ベイズ情報量規準を用いた手法を利用した．また音響モデルを既知とした手法では，i-vector を用いた話者認識を利用して短時間での歌唱者認識を繰り返し行うことで推定した．推定結果から，歌唱者の音響モデルの有無により大きな性能の差があること，また音響モデルが既知であっても短時間での歌唱者認識だけでなく適切な後処理によって推定誤りを減らせることが確認できた．

1. はじめに

ボーカルのある楽曲を考えたとき，歌唱している人数が 1 人か複数人かで大きく区別することができる．とくに複数人で歌唱している楽曲は，合唱の形態だけでなく歌唱者が交互に歌う形態を採用することがある．こうした演出は，グループアイドルソングやアニメソング，日本や世界のバンド音楽，複数人のアーティストがコラボレーションして作られた楽曲などに見ることができる．邦楽においてはこうした時間方向での歌い分け手法をとくにパート割りと呼ぶ場合があり，本稿でもこのように表現する．

パート割りの大きな役割として，歌う箇所が歌唱者によって異なることで，音楽を用いて注目する歌唱者を切り替える演出が可能になる点がある．これによってグループアイドルの各アイドルやアニメの各キャラクターなどを自然に印象づけることができる．一方で，誰がどの部分で歌っているかは，よくその歌唱者を知っている聞き手でも難しいことがある．また，そうしたパート割りの分析を行った場合は歌詞と対応付けることが多く，音声に対して直接メタデータなどの付与を行うことは少ない．どの歌唱者がいつ歌っているかを特定できれば，単にその結果を示すことだけでなく，同時に歌っている人数からの盛り上がり推定や，歌唱者の声質に応じた音楽演出が可能になる．また歌声の音響モデルを歌唱者ごとに適応することが可能になり，より高性能な歌詞認識や伴奏音抽出への応用が期待できる．本研究では，こうしたパート割りのある楽曲に対して歌声と歌唱者との対応をとることを考える．

パート割り推定に類似の問題として会話音声から「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があり，会議・ニュース・電話音声などを主な対象として研究がなされている．話者ダイアライゼーションからのアナロジーにより「誰がいつ歌っているか」を推定する技術を歌唱者ダイアライゼーションと呼ぶことができる．歌唱者ダイアライゼーションは，民族音楽を対象として行った研究の報告がなされている [1]．しかし，背景雑音や楽器音の存在，歌唱者の切り替わりが多いことなどが原因で推定の誤りが非常に多い点，また信頼性の高い正解ラベルを与えることが難しい点が指摘されている．また会話音声と歌声では話者（もしくは歌唱者）の重なり方や音素の継続長などの音響的特徴が異なるため，話者ダイアライゼーションの手法をそのまま適用することが適切であるとは限らない．

[1] 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST), Umezono, Tsukuba, Ibaraki 305-8568, Japan
a) hitoshi@gavo.t.u-tokyo.ac.jp
b) s.fukayama@aist.go.jp
c) t.nakano@aist.go.jp
d) dsk_saito@gavo.t.u-tokyo.ac.jp
e) m.goto@aist.go.jp

そこで本稿では，複数の歌唱者がそれぞれソロで同一の楽曲を歌唱した歌声が存在する楽曲群を利用し，これらから伴奏音を除去した歌声を用いて歌唱者ダイアライゼーションに対する基礎的な検討を行う．これにより，伴奏音除去を含む前処理による影響を小さくできるほか，正解ラベルを適切に与えることが可能になる．話者ダイアライゼーションには，話者の音響モデルを未知とした手法が広く用いられている．本稿ではこの手法のほかに，歌唱者の音響モデルを既知として短時間フレームでの歌唱者認識を繰り返すことでパート割りを推定する手法を用い，それぞれの性能を確認した．音響モデルが未知の場合では，歌唱者のほかに音素による影響を大きく受けることが確認され，音響モデルが既知の場合と比較して性能が大きく低下することが示された．また音響モデルが既知の場合では，瞬時に歌唱者が入れ替わるような不適切な結果が多く見られるものの，適切な後処理によってある程度の推定が可能であることが確認された．

2. 歌唱者の音響モデルを未知とする手法

本節では各歌唱者の音響モデルを未知とした場合の手法を述べる．これは話者ダイアライゼーションに広く用いられている手法と同様である．この手法は，前処理，セグメンテーション，クラスタリング，後処理の 4 手順に大別できる [2].

2.1 前処理

一般的な話者ダイアライゼーションの前処理では，発話区間の検出を行い，無音やノイズのみの区間を計算対象から除外する処理を行う．また音声強調やリバーブ除去などの処理を行う．とくに複数のマイクが利用可能な場合は到達時刻の差（time-delay-of-arrival; TDOA）を利用して事前処理を行うことでダイアライゼーション精度を向上させる手法が提案されている [3].

本稿では歌声なしの音源を利用してあらかじめ伴奏音を除去した音源を用いることを前提に，前処理として歌声区間の検出のみを行う．ミックス処理が施されている音源では歌唱者ごとに異なるパンが振られている場合がありこれを用いて歌唱者を分離することも考えられるが，本稿ではモノラル音声として扱いそのような条件に依存しない手法を検討する．

2.2 セグメンテーション

各セグメントが単一の話者による発話になるようにセグメンテーションを行う．各時刻の音響特徴量がすべて同じ話者による音響特徴量と仮定して分割していく手法（トップダウン・分割型クラスタリング）と，逆にすべて異なる話者による音響特徴量と仮定して結合していく手法（ボトムアップ・凝集型クラスタリング）に大きく分けられ，どちらも広く用いられている．どちらの手法においても何らかの規準で分割前と分割後もしくは凝集前と凝集後の評価を行い，分割・凝集位置や分割・凝集を止める点を決定する．評価規準には修正ベイズ情報量規準（modified Bayesian information criterion; mBIC），一般化尤度比（Generalized Likelihood Ratio; GLR）規準，情報量変化（Information Change Rate; ICR）規準などが広く用いられている [4–6]. 本節では mBIC を用いたトップダウン型の手法について述べる．

修正ベイズ情報量 mBIC はモデル M に対して次のように定義される．

$$\mathrm{mBIC}(M) = \log L(\mathcal{X}, M) - \frac{\lambda}{2}\#(M)\log N \qquad (1)$$

ここで $\log L(\mathcal{X}, M)$ は観測系列 $\mathcal{X}$ に対するモデル M の対数尤度，$\#(M)$ はパラメタ数である．λ は重み定数であり，実験的に決定する．mBIC にもとづくトップダウン方式のセグメンテーションでは，各歌唱者による音響特徴量が確率分布に従って生成されていると仮定する．本稿ではこの確率分布を単一のガウス分布と仮定した．そして，あるセグメントに対して全体を 1 つの歌唱者による発声とする仮説 H_0 と，そのセグメントが前半と後半で 2 つの歌唱者によって構成されているとする仮説 H_1 を用意し，そのベイズ情報量の高い仮説がより適当であるとする．具体的には観測音響特徴量系列を $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ とし，

$$H_0 : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (2)$$

$$H_1 : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_1} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1),\quad \boldsymbol{x}_{N_1+1}, \ldots, \boldsymbol{x}_N \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \qquad (3)$$

の 2 つの仮説を用意する．ここで $\boldsymbol{\mu}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ および $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ はそれぞれ各正規分布の平均ベクトルおよび分散共分散行列である．具体的なベイズ情報量は，特徴量の次元数を m として，H_0 では

$$\mathrm{mBIC} = -\frac{N}{2}\log(2\pi)^m|\boldsymbol{\Sigma}| - \frac{\lambda}{2}\frac{m(m+3)}{2}\log N \qquad (4)$$

となる．H_1 についても同様に計算すれば，H_0 と H_1 のベイズ情報量の差 ΔmBIC は

$$\Delta\mathrm{mBIC} = \frac{1}{2}[N\log|\boldsymbol{\Sigma}| - N_1\log|\boldsymbol{\Sigma}_1| - N_2\log|\boldsymbol{\Sigma}_2|] - \frac{\lambda}{2}\frac{m(m+3)}{2}\log N \qquad (5)$$

となり，この値が正であれば仮説 H_1 が適切であると推測することができる．ここでは $N_2 = N - N_1$ としている．なお分割点 N_1 は最も尤度が高くなるような点を選べばよい．この作業を BIC が増加しなくなるまで繰り返すことで，音声をセグメンテーションすることができる．

2.3 クラスタリング

セグメントに対して，同一話者の発話が同じクラスタに属するようにクラスタリングを行う．セグメンテーションと同様に mBIC を利用する手法のほかに，セグメントに対する i-vector を抽出してこれを比較する手法も用いられる．十分な精度のセグメンテーションが行えない場合があるため，セグメンテーションとクラスタリングを同時に行ったり，加えて話者の発話長モデルを利用したりする手法なども提案されている [7, 8]．本稿ではクラスタリングにおいても mBIC を利用している．

2.4 後処理

上の手順で得られるダイアライゼーション結果を修正するための処理を行う．歌っている歌唱者と状態を対応付けた隠れマルコフモデル（hidden Markov model; HMM）を構成し，上の結果を初期値としてビタビ探索を行うなどの手法を用いる．本稿では状態の出力となる音響モデルを 8 混合 GMM でモデル化しビタビ探索を行うことで，最終的なダイアライゼーション結果を得ている．

2.5 予想される手法の問題点

mBIC によりセグメンテーションやクラスタリングを行う場合，セグメント内の音響特徴量を統計モデルで表すことを前提としているため，短時間での音響特徴量の統計的な性質がその結果に影響する．すなわち，この手法は歌唱者の声質にもとづいてセグメンテーションやクラスタリングが行われることを期待しているが，継続長の長い音素によってこれらの結果に影響を及ぼす．これは長時間同じ音響特徴量が観測されることで，その時間の特徴量をモデル化したときの尤度が高くなりやすくなるためである．会話音声ではフィラーや言いよどみになどにより起きうるが，歌声の場合は音価によって音素の継続長が決まるためこうした影響を受けやすくなると考えられる．後段に HMM を用いた修正処理も行うが，初期状態の音響モデルは前段のセグメンテーション・クラスタリングの結果にもとづき構築されるため，これらの処理の性能が最終的な性能に大きく影響する．こうした要因により，歌唱者ごとの音響モデルが存在しない条件では，十分なダイアライゼーション性能を得ることが難しいと考えられる．

3. 歌唱者ごとの音響モデルを既知とする手法

本稿では話者ダイアライゼーションで広く用いられる手法のほかに，音響モデルを既知としたダイアライゼーション手法を適用する．歌唱者が既知であれば，i-vector を用いた話者識別 [9] を短い間隔で行うことで，各時刻での歌唱者が推定できる．また本稿では，ごく短時間で歌唱者が次々と入れ替わるような推定結果を抑制するための後処理として，各時刻において周辺の時刻で推定された歌唱者から最多の歌唱者を選ぶ方法で平滑化を行う．すなわち，時刻 t において推定されたラベルを $y(t)$，短時間歌唱者認識の間隔を τ として，次式にもとづき平滑化後のラベル $\hat{y}(t)$ を得る．

$$\hat{y}(t) = \mathrm{maj}\{y(t-n\tau), y(t-(n-1)\tau), \ldots, y(t+n\tau)\} \tag{6}$$

ここで maj は最多のラベルを選ぶ関数である．本稿では $\tau = 100\,\mathrm{ms}$，$n = 15$ とした．

なお本稿で扱う歌声には歌唱者全員で歌う部分が存在するが，i-vector による話者認識は通常一人による発話を前提として認識を行うため手法をそのまま適用することができない．そこで本稿では，同時歌唱の歌声を新たな 1 人の歌唱者による歌声としてモデル化することで，音源分離などを行わずに歌唱者の推定を行う．

4. 実験条件

本稿では，グループアイドルソングのデータセットとしてゲーム『アイドルマスター』で使われている楽曲を用いた．これはゲーム内にパート割りの演出がある楽曲が多数存在し，その楽曲を各歌唱者がソロで歌った歌声および歌声なしの音源が一部市販されているためである．伴奏音の除去には歌声りっぷ[*1]を用いた．歌声にはリバーブなどのエフェクトがかかっている場合が多く，また歌声りっぷによる伴奏音の除去時に再合成が行われているため，理想的なドライボーカルの歌声ではないことに留意する．サンプリング周波数はすべて 16 kHz とし，音響特徴量には 12 次元のメル周波数ケプストラム係数（mel-frequency cepstral coefficients; MFCC）およびそのデルタ特徴量を用いた．

ダイアライゼーションの対象音声は各歌唱者のソロの歌声をミックスして作成した．楽曲には『思い出をありがとう』を用い，歌唱者は A（水瀬伊織），B（星井美希），C（如月千早）の 3 名とした．本稿では問題を単純化し基礎的な検討を行うため 3 人中 2 人のみが歌っている状況は作らないものとした．すなわち 3 人全員が同時に歌っている・いずれかの歌唱者がソロで歌っている・誰も歌っていないの 5 通りの状態に限られるとした．このうち誰も歌っていない状態についてはパワーにもとづく歌声区間検出により事前に認識の対象から除外している．また 3 人が同時に歌っている部分については，3 人とも同じ音高のユニゾンとした．

声の重なり合いについては，会話音声に対して用いられる重なり合う前後の音声から推定する手法や音源を分離して認識する手法を適用することは難しい．そのため，歌唱者の音響モデルが未知の手法においては，3 人それぞれのソロ部分とユニゾン部分の 4 つにクラスタリングされることを期待する．また音響モデルが既知の手法においては，

[*1] http://www.vector.co.jp/soft/win95/art/se127635.html

3 人のソロのモデルのほかにユニゾンのモデルを用いて 4 クラス識別を行う．

ダイアライゼーション結果の客観評価には diarization error rate（DER）を用いた [10]．DER は 1) 誤った歌唱者でラベリングされた時間（error），2) 誤も歌っていない区間に歌唱者がラベリングされた時間（false alarm），3) 歌っている歌唱者がいる区間に誰も歌っていないとラベリングされた時間（miss）の 3 つを足し合わせた時間の，音声の総時間に対する割合である．

5. 実験 1: 歌唱者の音響モデルが未知の条件

本節および次節ではパート割りの存在する楽曲に対して実際にダイアライゼーション手法を適用することで，それぞれの手法の性能を議論する．本節では歌唱者それぞれの音響モデルを未知とした条件でのダイアライゼーション手法，すなわち話者ダイアライゼーションで広く用いられている手法について検討する．

5.1 mBIC にもとづくセグメンテーション

事前実験として，短い歌声に対してセグメンテーションを行った．ここではダイアライゼーションの対象音声とは異なる，2 人の歌唱者が前半と後半でそれぞれ歌い分ける 6.6 秒の歌声を用いた．

モデルのパラメタ数の重み定数 λ を 0 から 10 まで選び，歌声に対してセグメント数を 1 から 7 まで仮定したときのそれぞれにおける最大の修正ベイズ情報量を計算した結果を**図 1** に示す．この値が最も大きいセグメント数が各 λ において最適なセグメント結果であると推定される．$\lambda = 0$ の場合はベイズ情報量が単調増加したが，これはモデルの尤度のみを評価しているためである．また λ が大きいほどモデルのパラメタ数に対する重みが大きくなるため，セグメント数が多くなるほどベイズ情報量が小さくなった．

この歌声に対して λ を 2 と 3 に選びセグメンテーションを行った結果を**図 2** に示す．ともに歌声のはじめとおわりの短い無音部分がセグメントに分割されている．どちらの λ においても，歌唱者の切り替わり点ではない場所でセグメントの切断が発生している．またどちらも 5.2 秒付近から 6.4 秒付近までに 1 つのセグメントが割り当てられているが，この部分は 1 つの音素/u/にあたる部分である．mBIC にもとづくセグメンテーションでは音響モデルを仮定せず 1 つのガウス分布でモデル化するため，このように歌声に見られる継続長の長い音素の影響を受けやすいと考えられる．

5.2 音響モデルを未知としたダイアライゼーション

mBIC にもとづくセグメンテーションおよびクラスタリングと HMM を用いたビタビ探索を行うことで，歌唱者の音響モデルを未知としてダイアライゼーションを行った．ここではダイアライゼーションツールキットである LIUM [11] を用いた．この結果を**図 3** に示す．1 つめのクラスタと歌唱者 A の歌唱部分がよく合致しているが，3 つめと 4 つめの両方のクラスタがユニゾン部に割り当てられているなど誤りも多く見られる．DER は 34.8%となった．

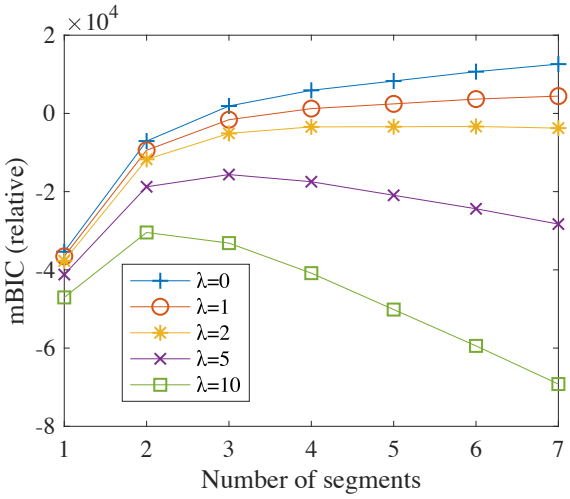


図 1 セグメント数を 1 から 7 まで仮定したときの修正ベイズ情報量．λ は mBIC におけるモデルのパラメタ数に対するハイパーパラメタ．修正ベイズ情報量の値は定数項を無視して計算した相対値．

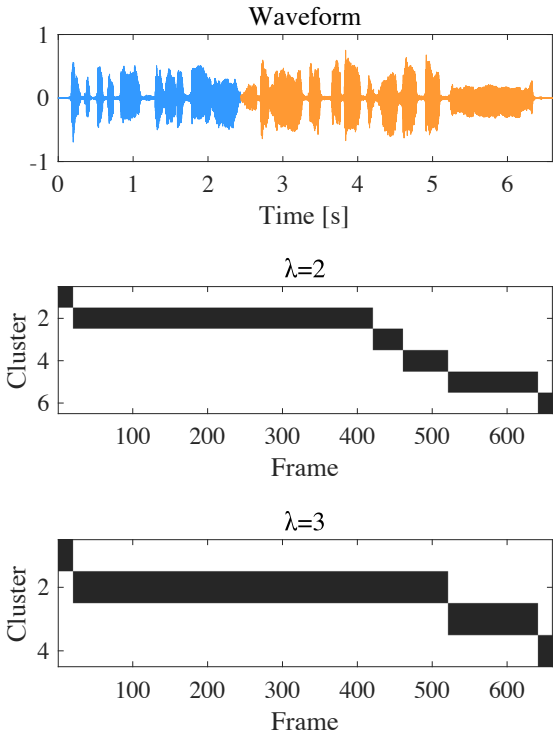


図 2 2 人の歌唱者による歌声に対して mBIC にもとづくセグメンテーションを行った結果．波形が青色の部分と橙色の部分がそれぞれの歌唱者のパートであることを示す．

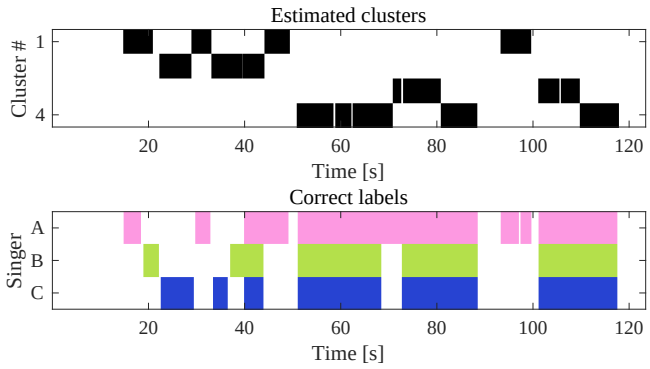


図 3 LIUM を利用した，歌唱者の音響モデルを用いない手法での歌唱者ダイアライゼーションの結果．上段が推定結果，下段が正解ラベルを示す．歌唱者の認識を行っていないため，どのクラスタとどの歌唱者が対応付けられているかは推定していない．

6. 実験 2: 歌唱者の音響モデルが既知の条件

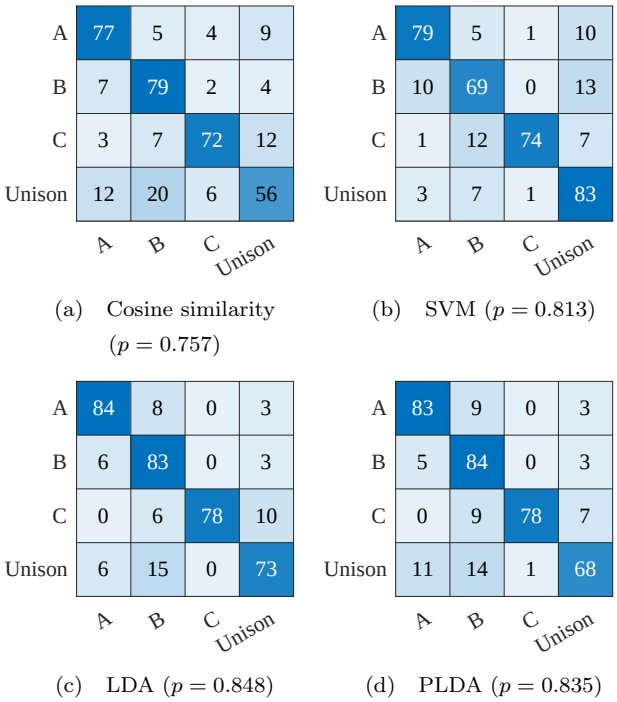
本節では歌唱者それぞれの音響モデルを既知とした条件で，短時間歌唱者認識を利用したダイアライゼーション手法の性能について検討する．歌唱者認識は 1 秒間の音声から i-vector を抽出して行った．話者空間モデルである universal background model（UBM）には 2048 混合の混合ガウスモデルを用い，学習には多数話者音声データベース APPBLA における女性話者のうち 1185 名の発声した ATR 音素バランス文 20580 発話（およそ 30 時間分）を用いた．i-vector の次元数は 100 次元とし，判別器の学習には認識対象の楽曲を含まない 15 曲のソロおよびユニゾンの歌声を用いた．

6.1 短時間歌唱者認識

事前実験として，各歌唱者のソロと 3 人のユニゾンの歌声それぞれに対して，余弦類似度（cosine similarity），サポートベクトルマシン（SVM），線形判別分析（LDA），probabilistic LDA（PLDA）の 4 種の判別法を用いて短時間歌唱者認識を行った．これにより得られた混同行列を**図 4** に示す．SVM を除いては同様の傾向が見られ，歌唱者 A と B の歌声をユニゾンと判定する誤りが多く見られた（混同行列の左下部分）．これは 2 歌唱者の声質が近く，またユニゾンの音声が 3 人の歌声の平均的な特徴量になるため，その結果ユニゾンの音声が 2 歌唱者の特徴量に近くなったと考えられる．推定性能を表す正解率は，余弦類似度では 0.757，SVM では 0.813，LDA では 0.848，PLDA では 0.835 となり，LDA が最も優れていた．ここで，正解率とは全フレームのうち正解のラベルを与えたフレームの割合を示す．

6.2 短時間歌唱者認識による歌唱者ダイアライゼーション

5.2 節と同じ歌声に対してダイアライゼーションを行った．判別法には前節の結果から LDA を用いた．ダイアライゼーションの推定結果を**図 5** に示す．平滑化前の結果では短時間で歌唱者が入れ替わるような不適切な箇所が多く見られるが，平滑化を行うことで改善している．平滑化後の結果では 93 秒付近から 100 秒付近に大きな誤りが見られる．この部分は本楽曲においてセリフにあたる部分であり，歌声でなくセリフであることが音響特徴量系列の統計量を評価する i-vector に大きく影響を与えたと考えられる．DER は平滑化前が 28.2%，平滑化後は 11.9%となり，ともに 5.2 節の結果より低い誤り率が得られた．



(a) Cosine similarity ($p = 0.757$)　(b) SVM ($p = 0.813$)

(c) LDA ($p = 0.848$)　(d) PLDA ($p = 0.835$)

図 4 4 つの判別手法を用いて短時間歌唱者認識を行い得られた混同行列．各行が推定ラベルに，各列が正解ラベルに対応する．たとえば最も左下の値は歌唱者 A の歌声をユニゾンと推定したフレーム数を表す．Unison は 3 人の歌唱者が同時に歌った歌声を示す．p は推定の性能を示す正解率．

7. 考察

歌唱者の音響モデルを未知とした実験では，歌唱者 B と C の 2 歌唱者の識別ができないなど，DER が高く誤りの多い結果となった．音響モデルが未知の場合は，与えられた短時間の音声断片から歌唱者の音響モデルを仮定して認識を行うため，緻密な音響モデルを用いることができず，また音素の影響を受けやすくなると考えられる．

歌唱者の音響モデルを用いて歌唱者認識にもとづいてダイアライゼーションを行った場合には，平滑化をせずとも歌唱者未知での手法より DER が低くなり，平滑化を行うことでさらに正解に近い結果を得た．音響モデルの有無による差は事前知識による合理的な差であると考えられるが，音響モデル未知の手法に大きな改善の余地があること

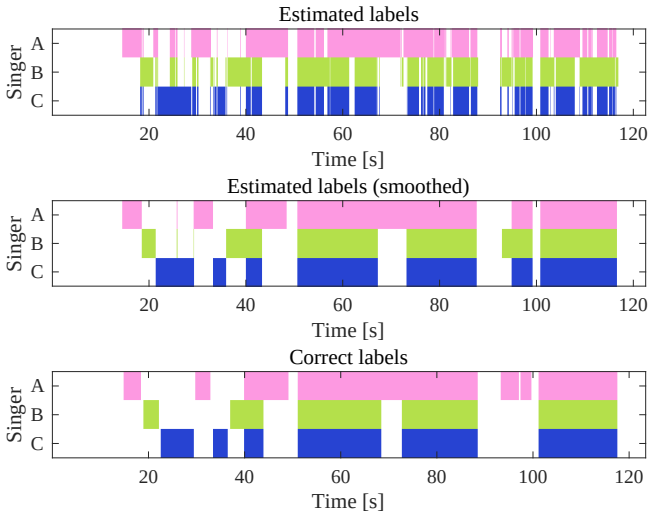


図 5 短時間歌唱者認識にもとづくダイアライゼーションの結果．上段が平滑化前の推定ラベル，中段が平滑化後の推定ラベル，下段が正解ラベルを示す．

が認められた．また音響モデル既知の手法での推定性能は平滑化によるところが大きく，歌唱者認識手法そのものと平滑化手法の両手法に対してさらなる検討が必要である．

8. おわりに

本稿では「誰がいつ歌っているか」を推定する歌唱者ダイアライゼーション技術に関する基礎的な検討を行った．既存の音源を用いる場合には伴奏音の除去や正解ラベルの付与などの問題が挙げられる．そこで伴奏音を除去し正解ラベルを確実に与えるために，特定のゲームに用いられているグループアイドルソングを活用したデータを整備した．またこのデータに対して，話者ダイアライゼーションに広く用いられている手法および歌唱者認識を利用した手法の 2 手法を適用しダイアライゼーションを行った．歌唱者の音響モデルを用いない手法では推定の誤りが多く見られた．また音響モデルを用いて歌唱者認識を行った場合でも，歌唱者認識自体の誤りが無視できず，歌唱区間の長さを考慮した処理が必要であることを確認した．

本稿では重なり合った 3 人の歌声を，ソロの歌声とは異なる歌唱者による歌声として認識させる手法を用いた．しかし 3 人中 2 人のみが歌っている場合を考慮しておらず，これを考慮する場合同様の手法では組合せ爆発を起こしモデル数が大きく増えてしまう．重なり合った声の音響モデルはソロの音響モデルの平均的なモデルになることを考慮すると，同じ数の話者認識よりも問題が難しくなり，推定精度が大きく下がることが考えられる．事前に同時歌唱者数のみを推定しておくなどの手法により，モデルの候補数を減らす工夫が求められる．また本稿では検証実験に 1 曲のみを用いたが，楽曲や歌唱者の組み合わせを様々に選ぶことでより詳細な手法の検討を行いたい．

謝辞 本研究の一部は JST ACCEL（JPMJAC1602）の支援を受けた．

参考文献

[1] Thlithi, M., Barras, C., Pinquier, J. and Pellegrini, T.: Singer Diarization: Application to Ethnomusicological Recordings, *5th International Workshop on Folk Music Analysis*, pp. 124–125 (2015).
[2] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G. and Vinyals, O.: Speaker Diarization: A Review of Recent Research, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 2, pp. 356–370 (2012).
[3] Anguera, X., Wooters, C., Peskin, B. and Aguiló, M.: Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System, *Machine Learning for Multimodal Interaction*, pp. 402–414 (2005).
[4] Chen, S. S. and Gopalakrishnan, P. S.: Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion, *DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132 (1998).
[5] Gish, H., Siu, M.-H. and Rohlicek, R.: Segregation of Speakers for Speech Recognition and Speaker Identification, *International Conference on Acoustics, Speech, and Signal Processing*, pp. 873–876 (1991).
[6] Vijayasenan, D., Valente, F. and Bourlard, H.: An Information Theoretic Approach to Speaker Diarization of Meeting Data, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 7, pp. 1382–1393 (2009).
[7] Ajmera, J. and Wooters, C.: A Robust Speaker Clustering Algorithm, *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 411–416 (2003).
[8] Kotti, M., Benetos, E. and Kotropoulos, C.: Computationally Efficient and Robust BIC-Based Speaker Segmentation, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 5, pp. 920–933 (2008).
[9] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B.: Speaker Verification Using Adapted Gaussian Mixture Models, *Digital Signal Processing*, Vol. 10, pp. 19–41 (2000).
[10] Tranter, S. E. and Reynolds, D. A.: An Overview of Automatic Speaker Diarization Systems, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp. 1557–1565 (2006).
[11] Rouvier, M., Gay, P., Khoury, E., Merlin, T. and Meignier, S.: An Open-source State-of-the-art Toolbox for Broadcast News Diarization, *INTERSPEECH* (2013).