修士論文

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

平成 31 年 1 月 31 日

指導教員　　齋藤 大輔　講師

東京大学大学院工学系研究科

電気系工学専攻　37-176447

須田 仁志

本論文は東京大学大学院工学系研究科に修士号授与の要件として提出した修士論文である.

内容梗概

本論文では，複数人が歌唱する楽曲を対象として，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を扱う．歌唱者ダイアライゼーションを扱う上で，伴奏音が含まれた実音声を用いて十分に検討を行うことは難しい．そこで本論文では，歌声が抽出可能で歌唱者ダイアライゼーションの検討に適した日本のグループアイドルソングの楽曲群に着目し，これらの楽曲を用いて歌唱者ダイアライゼーションに関する基礎的な検討を行う．グループアイドルソングでは，パート割りと呼ばれる時間的な歌い分け演出を採用することがある．本研究で扱う歌唱者ダイアライゼーションは，いわばグループアイドルソングに対するパート割り構造の認識にあたる．

本論文では，いくつかの手法を用いて歌唱者ダイアライゼーションを試みる．第 1 の手法は，会話音声などに用いられてきた標準的なダイアライゼーション手法で，歌唱者の人数や各歌唱者に関する知識を必要としない．第 2 の手法は，歌唱者の認識を短時間ごとに行う手法で，歌唱者の知識を必要とするが歌唱者の認識さえ行うことができれば性能に期待ができる．実験により，手法の違いにより楽曲や歌唱者間でのダイアライゼーション性能の傾向が異なることが明らかになった．さらに，第 3 の手法として 2 手法を組み合わせた手法を用いて実験を行い，現状の歌唱者ダイアライゼーション手法における限界と課題点を確認した．

パート割りの認識に関連し，さらに本論文では声の重なりのみを検出するための手法を検討する．音声から同時に発声している人数を推定する手法はいくつか検討がなされてきた．しかし，本研究で扱うグループアイドルソングでは，3 人以上の女性歌手が同じ音高で発声する歌声を扱わなければならず，既存の音響特徴量から推定する手法では難しいことが明らかになった．本論文では，スペクトルやケプストラムにもとづく従来の特徴量だけでなく，波形や自己相関関数などの信号により近い特徴量を用い，同時歌唱の認識に対して直接的なアプローチを検討する．1 人か 2 人かを判別する実験により，声の基本周波数の違いに着眼し，従来用いられてきた音響特徴量と比較して長時間の音声から得た特徴量を用いることで同時歌唱の認識が行える可能性を示した．

目次

第 1 章

序論

1.1 本研究の背景: 音楽に対する情報処理

音楽に対する情報処理は，音声に対するそれの後を追うようにして発展してきた．こうした技術によって，インターネットを経由して楽曲をダウンロードし，プレイリストを用いて再生順序を自由に並び替えたり，イコライザを用いて好みの音に近づけたりすることが今や可能になっている．しかし，これらの情報処理技術は広範な音声に対して適用されるものであり，音楽的な内容にもとづいたものではない．

音楽に特化した情報処理を用いることで，音楽鑑賞をさらに豊かにする種々の技術やアプリケーションが開発されている．たとえば MIXTRAX[*1] とよばれる音楽再生アプリケーションでは，読み込んだ楽曲を自動で解析することで，サビを抽出・連結して盛り上がりを維持したまま楽曲を遷移したり，選曲した楽曲に似た曲調やテンポの楽曲を優先して再生したりする機能を利用可能にしており，新たな音楽の聴き方を提案している．Songle[*2] とよばれる Web サービスでは，YouTube やニコニコ動画などの動画投稿サービスに投稿された楽曲を分析し，サビへの頭出しを可能にしたり，声色の似た歌手の楽曲を検索したりすることを可能にしている [1]．Songle では，ビート，コード，メロディなどの音楽的な情報を可視化することも可能で，楽曲に対する理解を深めることができる．また，Songle と連携した Songle Sync[*3] とよばれるサービスでは，Songle によって解析された音楽情報を元に多様な機器を制御することで，ロボットや光の演出を演奏している楽曲と連動させることができる [2]．このような技術は，単なる信号処理技術ではなく，音楽に特化した情報を抽出する音楽情報処理技術によって支えられている．

1.2 本研究の目的: 歌唱者ダイアライゼーション

音楽情報処理は幅広い音楽に対して研究がなされており，前述のようにボーカルのある楽曲に対してもメロディーやサビ検出が可能になっている．しかし，ボーカルのある楽曲を扱う技術の多くは歌唱者が 1 人であることを仮定している．一方，ポピュラーソングやアニメソングを振り返れば複数人が歌唱している楽曲を多く見ることができ，音楽情報処理においてこうした楽曲を無視することはもはやできない．

複数人が歌唱している楽曲のうち，A メロや B メロを各歌唱者がソロで歌いサビで全員で歌うなどといった時間的な歌い分けを採用することがある．こうした歌い分けの構造は，音高によって分離する方法と区別してパート割りと呼ばれることがあり，モーニング娘。や私立恵比寿中学などのグループアイドルの楽曲，『ハレ晴レユカイ』や『Cagayake! GIRLS』などのアニメソングをはじめとして日本国内でも多くの楽曲に見ることができる．パート割りは，自然と着目する歌手を遷移させることができるなど，他の楽曲構造と同様に作

[*1] http://www.mixtrax-global.com
[*2] https://songle.jp/
[*3] http://api.songle.jp/sync

り手の意図が反映される演出であり，楽曲を分析する上で重要な情報となる．パート割りを自動で分析することができれば，それを可視化して示すことだけでなく，歌唱している人数からサビを推定したり，Songle Sync などで実現される楽曲に応じた演出をパート割りに応じて変化させたりする応用が期待できる．

複数人の歌唱における歌唱者の認識については，特定の歌唱者がその楽曲に含まれているかを認識する target singer detection（TSD）や，各時刻に対してそれを行う target singer tracking（TST）として検討がなされている [3,4]．こうした研究は多くてもデュエットの音声を仮定しており，また楽曲内での時間的構造を明らかにしているわけではない．パート割り認識の目的は，数人の歌唱者からなる楽曲に対して，誰がいつ歌っているかの楽曲構造を明らかにすることであり，現状の TST では不十分である．

パート割り認識に近い問題として，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と呼ばれる研究課題がある．話者ダイアライゼーションは，電話やニュースなど幅広い会話音声を対象として研究がなされてきた [5]．話者ダイアライゼーションの連想から，パート割り認識は歌唱者ダイアライゼーションと表現することができる．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われた研究が既に報告されている [6]．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．歌唱者ダイアライゼーションはあくまで研究課題として提案されているに過ぎず，これに関して基礎的な検討が十分に行われているとは言えない．

本研究は，歌唱者ダイアライゼーションという大きな問題を分割し，基礎的な検討を行うことを目的とする．本研究では，特に日本のグループアイドルソングに着目して歌唱者ダイアライゼーションを議論する．グループアイドルソングなどの楽曲は歌唱者の声質が近く識別が難しい上，ユニゾンと呼ばれる同じ音高で歌われる部分によって分離が困難である場合も多く，音声信号処理の観点からも十分に議論の余地がある課題である．本研究では，グループアイドルを模したゲーム中の，同じ楽曲を複数の歌唱者がそれぞれソロで歌唱した楽曲群に着目する．そしてこの楽曲群を用いて，歌唱者や楽曲による性能の違いを含め，歌唱者ダイアライゼーションに関して詳細に分析する．

1.3 本論文の構成

本論文は，10 の章と 1 つの付録からなる．

第 2 章から第 4 章では，本論文で用いる技術について解説する．第 2 章では，ダイアライゼーションの基礎となる話者認識と，話者認識のための音声分析について述べる．第 3 章では，話者ダイアライゼーションの手法として広く用いられてきた手法について述べる．第 4 章では，歌声に対する情報処理技術について述べる．

第 5 章から第 8 章では，歌唱者ダイアライゼーションを 3 手法を用いて行った実験について述べる．第 5 章では，すべての実験に共通したデータセットの詳細や評価手法について述べる．第 6 章から第 8 章では，適用した 3 手法それぞれについて，用いた手法の詳細および実験の結果とそれに対する考察を述べる．

第 9 章では，パート割りに認識に求められる同時に歌唱している人数の推定について述べる．この章ではダイアライゼーションを直接扱わないが，同時に歌唱している音声を扱う本研究において重要な考察となる．

第 10 章では本論文を結び，今後の展望や課題を述べる．

また，付録では，本論文で行った実験で用いるために整備したデータセットの詳細について述べる．

論文の本論となる各章の構成を図 1 に示す．

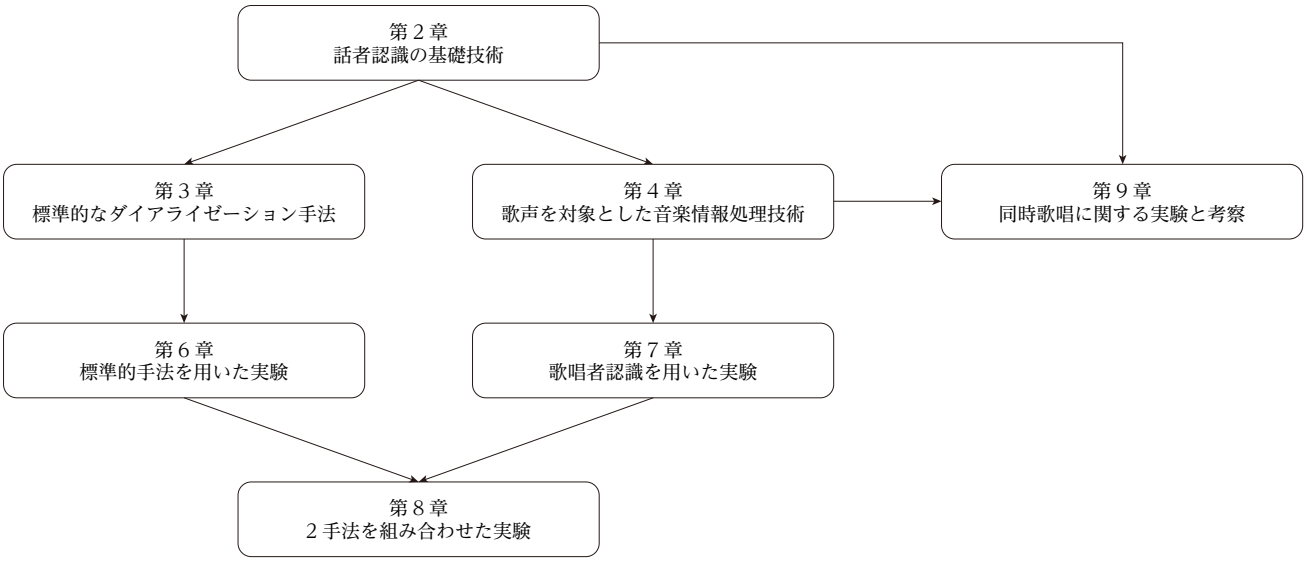


図 1　本論文における各章の関係を表したダイアグラム.

第 2 章

話者認識の基礎技術

本章では，ダイアライゼーションの基礎となる基本的な話者認識技術について述べる．話者認識は，音声から抽出した音響特徴量系列を，学習済みの音響モデルと比較・照合することで行われる．多くの手法では，音響特徴量系列をそのまま用いず，音響特徴量系列から話者を表現するベクトルを得ることで，話者の緻密な音響モデルを構築し高い性能の識別を実現する．このような話者認識の枠組みを図 2 に示す．

2.1　音響特徴量

話者認識のみならずあらゆる音声分析においては，音声を信号波形のまま用いず，音響特徴量の系列を抽出して利用する．どのような音響特徴量を用いるかは利用する目的によって異なり，話者認識においては話者情報をよく表す音響特徴量を得る必要がある．音響特徴量の多くは，周波数解析を行って得られたスペクトルから抽出する．このような音声分析過程の概略を図 3 に示す．

2.1.1　短時間フーリエ変換とスペクトログラム

入力となる音声は時々刻々と変化する非定常な信号であり，各時刻においてその瞬間の音響的特徴を得られることが望ましい．各時刻から特徴量を抽出するには，注目する時刻周辺の音声のみを取り出して周波数解析すればよい．短時間で音声を切り出すことで，注目した時刻周辺に関して時変性を無視して分析することができる．

各時刻周辺を取り出す操作は，音声に対して矩形関数をかけ合わせることに値する．すなわち，時刻 t での音響的特徴に関心がある場合には，音声 $s(\tau)$ のうち t の周辺の時間幅 T に含まれる信号 $s_t(\tau)$ を次式のように得ればよい．

$$s_t(\tau) = \begin{cases} s(t+\tau) & \left(-\frac{T}{2} \le \tau \le \frac{T}{2}\right) \\ 0 & (\text{otherwise}) \end{cases} \tag{1}$$

これは矩形関数 $h_{\text{rect}}(t)$ を用いて次のように表式できる．

$$s_t(\tau) = s(t+\tau) h_{\text{rect}}\left(\frac{\tau}{T} + \frac{1}{2}\right), \quad h_{\text{rect}}(\tau) = \begin{cases} 1 & (0 \le \tau \le 1) \\ 0 & (\text{otherwise}) \end{cases} \tag{2}$$

$s_t(\tau)$ をフーリエ変換することで，時刻 t におけるスペクトル $S_t(f)$ を得ることができる．

$$S_t(f) = \mathcal{F}(s_t(\tau))(f) \tag{3}$$

しかし，矩形関数をかけ合わせた信号をフーリエ変換すると，音声のスペクトルに矩形関数のフーリエ変換が畳み込まれることになる．矩形関数のフーリエ変換は標本化関数であるから，これはある周波数のスペクト

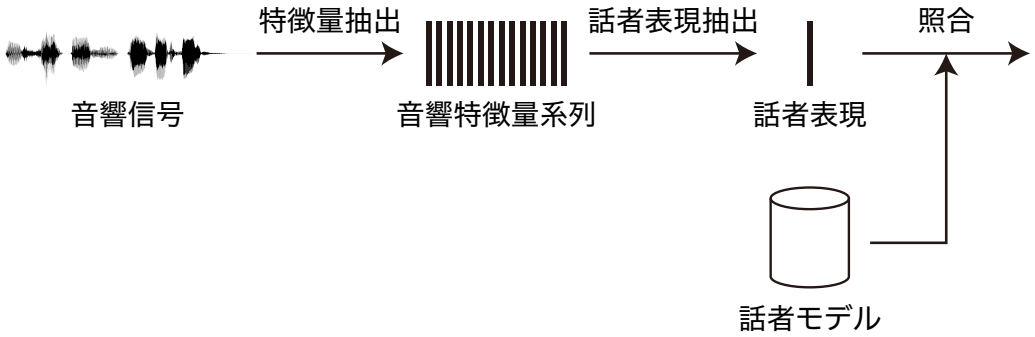


図 2　話者認識システムの概要．音声から音響特徴量の系列や話者表現を抽出し，学習済みのモデルと比較することで，その話者らしさを音声から得る．

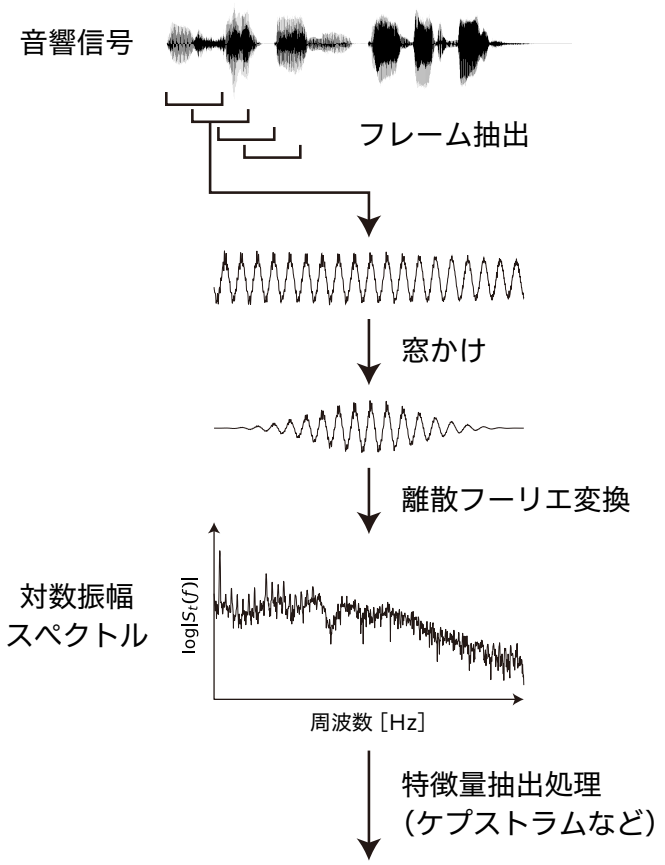


図 3　音響特徴量抽出のための音声分析の概要．短時間の音声に対する分析を，一定の時間間隔で繰り返し行うことで，時変な特徴量の分析を実現する．

ルが前後の周波数に影響を及ぼすことを意味する．この現象をスペクトル漏れと呼び，本来信号が存在しないはずの周波数にあたかも存在するかのように観測されてしまうため，音声を分析する上で望ましくない．

音声をある時間幅で切り出す場合，スペクトル漏れは決して避けることができない．しかし，矩形関数の代わりに別の関数を用いることで，スペクトル漏れを低減することが可能である．ここで用いる関数を窓関数とよび，窓関数を音響信号とかけ合わせて短時間の信号を得ることを窓かけと呼ぶ．一般の窓関数 $h(\tau)$ を用

いることで，時刻 t 付近の窓かけされた信号は次式のように得ることができる．

$$s_t(\tau) = s(\tau + t)h\left(\frac{\tau}{T} + \frac{1}{2}\right) \tag{4}$$

この操作を短い時間間隔で行い各時刻でフーリエ変換することでスペクトログラム $Y(t,f) = |S_t(f)|^2$ を得ることができる．このような手順でスペクトログラムを得る手法を短時間フーリエ変換と呼び，もっとも基本的な音声分析処理である．

窓関数には多くの種類が提案されており，そのうちの多くは有限台で対称な山型の関数である．本論文で用いるメル周波数ケプストラム係数を求める際には，一般的に次式で表されるハミング窓を用いる．

$$h(\tau) = \begin{cases} 0.54 - 0.46\cos 2\pi\tau & (0 \le \tau \le 1) \\ 0 & (\text{otherwise}) \end{cases} \tag{5}$$

また式 (2) のように矩形関数を窓関数として用いる場合には，この窓関数を矩形窓と呼ぶ．

音声分析においては，短時間フーリエ変換を行うための信号の素片をフレームと呼ぶ．音響信号を分析する際には，フレームをどのような長さで切り出し，どの程度の周期でフレームを得るかを決定する必要がある．このとき，フレームの長さ T をフレーム長，フレームを切り出す周期をフレーム周期もしくはフレームシフトと呼ぶ．

2.1.2 ソースフィルタモデル

スペクトルのどのような成分に話者情報が含まれるかは，声の生成モデルを導入することで推し量ることができる．

声は，声帯による励振によって空気が振動して音源が発生し，声道という共鳴管によって音色付けされることで声として観測される．この生成過程は，声帯振動によって発生した音源波と，声道の形状により決まる周波数特性を持った声道フィルタによってモデル化できる．音源波を $g(t)$，声道フィルタのインパルス応答を $v(t)$，声の音声信号を $x(t)$ とすると，声の生成モデルは次のように表式できる．

$$x(t) = g(t) \otimes v(t) \tag{6}$$

ここで $\otimes$ は畳み込みを表す．$g(t)$，$v(t)$，$x(t)$ のフーリエ変換をそれぞれ $G(f)$，$V(f)$，$X(f)$ とすれば，式 (6) の両辺をフーリエ変換することで，声の生成モデルは次式のように表せる．

$$X(f) = G(f)V(f) \tag{7}$$

この生成モデルは音源（ソース）と声道フィルタによる線形システムであることから，ソースフィルタモデルと呼ばれる．音声を分析する際には，多くの場合このモデルを仮定する．

式 (7) で示した生成モデルは，声帯振動による有声音のみを仮定したモデルである．無声音を考慮する場合には，音源波をホワイトノイズとしてモデル化することが多い．また，声は空気を通して伝わるため，空気中の放射による影響を受ける．無声音と放射特性を考慮することで，ソースフィルタモデルは次式のように表せる．

$$X(f) = R(f)V(f)\begin{cases} G(f) & (\text{有声音のとき}) \\ 1 & (\text{無声音のとき}) \end{cases} \tag{8}$$

ここで，$R(f)$ は周波数領域での放射特性を表す．音声合成を行う場合には有声無声を考慮する必要があるため，このようなモデルを仮定して合成を行うことが多い．一方，話者認識や音声認識においては有声無声の

区別を行わずとも十分に認識を行えることが多いため，このような場合分けモデルを採用せず分析することもしばしばある．

2.1.3 スペクトル包絡とケプストラム

有声部の音源波は三角波などで近似され，声の高さを表す基本周波数以外の情報を持たない．これ対し，話者や母音の情報は声道の特徴に反映される．そのため，認識に必要な話者や母音の情報を得るためには声道フィルタを推定すればよい．

声道フィルタは声道の共鳴特性によって決定され，共鳴が生じるいくつかの周波数にピークを持つ特性を持つ．これらの周波数はフォルマント周波数と呼ばれ，声道フィルタを特徴づけるパラメータである．最も低いフォルマント周波数は第 1 フォルマント周波数と呼ばれ，声道長すなわち声道全体の長さに依存する．第 1 フォルマント周波数は基本周波数よりも高いことが多く，また声道フィルタはなめらかな特性を持つため，声道特性はスペクトル領域で大域的な特徴として現れる．すなわち，スペクトルの概形が声道フィルタを示すものとして解釈され，これをスペクトル包絡と呼ぶ．話者認識においては，話者情報がスペクトル包絡に現れることを利用して，話者情報を保持する特徴量を得る．

スペクトル包絡を抽出する最も基本的な手法として，ケプストラムを利用する手法がある．式 (7) の両辺に対して絶対値の対数を計算すると，次式で示すようにソースフィルタモデルは和の形で表される．

$$\log|X(f)| = \log|G(f)| + \log|V(f)| \tag{9}$$

ケプストラムとは，対数スペクトルを逆フーリエ変換したものであり，$\log|X(f)|$ に対応するケプストラムは $\mathcal{F}^{-1}\log|X(f)|(t)$ となる．ケプストラムにおける時間軸はケフレンシーと呼ばれる．声道フィルタがスペクトルの概形に現れるとすれば，ケプストラム領域では低いケフレンシーの領域に現れる．対数スペクトル領域で式 (9) のように音源波と声道フィルタが和の形に分離できるのと同様，ケプストラム領域でも音源波と声道フィルタは和によって表現される．そのため，ケプストラムのうちケフレンシーの高い成分をゼロとして低い領域のみを抽出し，再度フーリエ変換することで，声道フィルタの特性のみを持つ対数スペクトル $\log|V(f)|$ を得ることができる．このように，ケプストラム領域で一部の帯域のみを抽出する処理をリフタリングと呼ぶ．

音声からスペクトルを抽出し，ケプストラムを利用してスペクトル包絡を得た例を図 4 に示す．音源波は，スペクトル領域では基本周波数および倍音成分による調波構造と呼ばれる細かな櫛状の構造を作り，これはケプストラム領域では高いケフレンシーに現れる．これに対しスペクトル包絡は低いケフレンシーに現れるため，これを取り出してスペクトル領域に戻すことでスペクトル包絡を得ることができる．

基本周波数が極度に高い音声からスペクトル包絡を抽出する場合，調波構造によるスペクトルが疎になり，スペクトル包絡が明確にならない問題がある．第 1 フォルマント周波数は，声道長が比較的短くフォルマント周波数が高くなりやすい女性でも母音によっては 300 Hz 程度まで低くなる [7]．一方，基本周波数はこの周波数よりも高くなる場合があり，とくに歌声ではメロディーに依存して高くなりやすい．音声認識や話者認識でスペクトル包絡を特徴量として用いる場合，スペクトル包絡の曖昧さが直接音響特徴量の曖昧さとなるため，認識の性能に影響しうる．本研究で用いた歌声において見られるこのような例を図 5 に示す．

2.1.4 メル周波数ケプストラム係数（MFCC）

音声認識では，音響特徴量としてメル周波数ケプストラム係数（mel-frequency cepstral coefficients; MFCC）が広く用いられる．人の聴覚では，内耳中の蝸牛にある基底膜が音声の周波数に応じて異なる振動をするこ

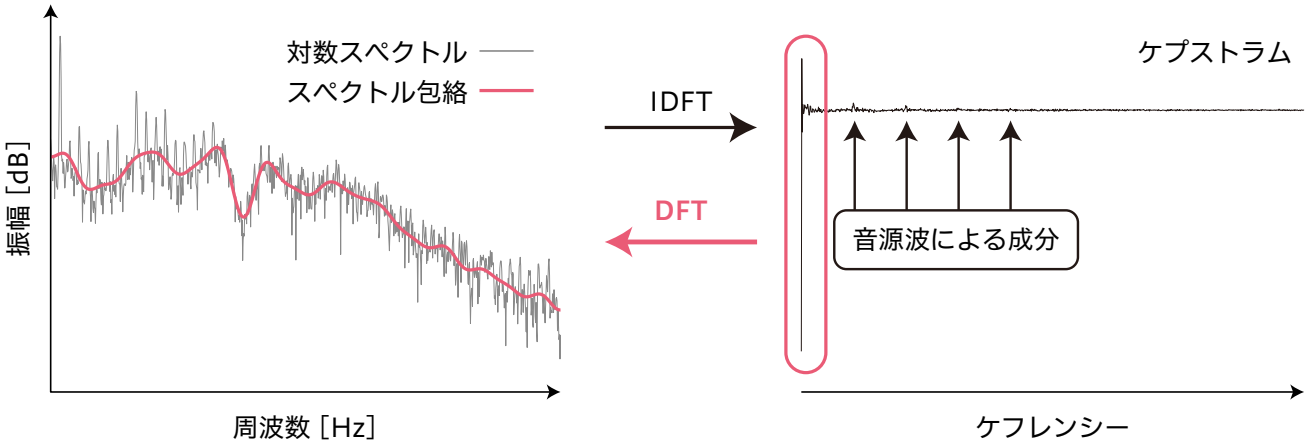
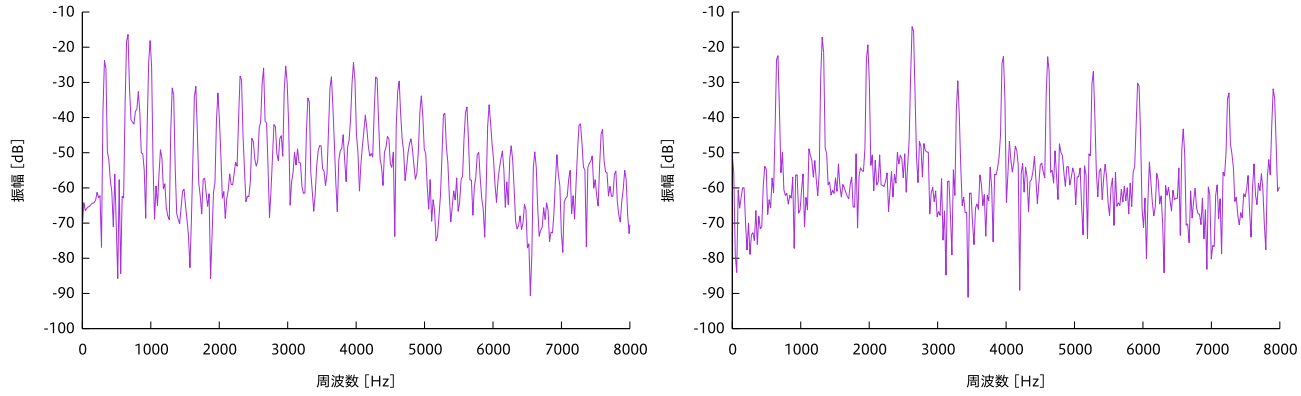


図 4　スペクトルからケプストラムを得てリフタリングを行いスペクトル包絡を求めた例．高いケフレンシーに現れる音源波の成分をリフタリングによって除去することで，スペクトル包絡のみを選択して得ることができる．



(a) 基本周波数 330 Hz でのスペクトル．

(b) 基本周波数 660 Hz でのスペクトル．

図 5　高さの異なる同一母音「エ」のスペクトルの比較．ともに歌唱者 E の歌う楽曲 G から抽出した．図 5a や母音の情報にもとづけば第 1 フォルマント周波数は 800 Hz 付近であると推定されるが，基本周波数が高いスペクトルではスペクトル包絡が曖昧になり第 1 フォルマント周波数も不明瞭になっている．楽曲および歌唱者の詳細については第 5 章および付録 A を参照されたい．

とで周波数分析を行っている．これは，異なる中心周波数と帯域幅を持つフィルタを多数用意し，それぞれのフィルタを通過した音響信号のパワーを得ていると解釈できる．MFCC はこのような人の聴覚における基底膜振動を模倣するフィルタバンクを用いて抽出されるため，聴覚の特性に近い音響特徴量とみなされる．

基底膜を模倣したフィルタバンクには，メル周波数軸上での三角状のフィルタを用いる．メル周波数とは，非線形な聴覚上の音の高さを周波数と関連付けた量で，周波数と次の関係式を持つ．

$$f_{\mathrm{mel}}\,[\mathrm{mel}] = 2595 \log_{10}\left(1 + \frac{f\,[\mathrm{Hz}]}{700}\right) \tag{10}$$

フィルタバンクの中心周波数は，メル軸上で 0 Hz から等間隔に設定する．サンプリング周波数 16 kHz の場合は，0 kHz から 8 kHz までメル軸上で等間隔に 22 の周波数を選び，0 kHz と 8 kHz を除いた 20 の周波数を中心としたフィルタバンクを構成することが多い．

振幅スペクトルに対してフィルタバンクを適用したのち，対数をとり離散コサイン変換を行う．これによって，ケプストラムと同様の軸をもつ音響特徴量を得ることができる．得られた特徴量のうち，低次のものを MFCC として用いる．サンプリング周波数 16 kHz の音声では，20 次元のうち低次の 12 次元を特徴量とする場合が多い．

MFCC はスペクトルに対して直接フィルタバンクを適用するため，声帯振動や放射の特性がそのまま現れる．声帯振動の特性はおよそ −12 dB/oct，放射特性はおよそ 6 dB/oct であるとされているため，MFCC を計算する際には事前に 6 dB/oct の 1 次の有限インパルス応答（finite impulse response; FIR）フィルタを適用することで 2 つの特性を補償する．1 次 FIR フィルタは次式のように実装される．

$$y(n) = x(n) - cx(n-1) \tag{11}$$

ここで $x(n)$ と $y(n)$ はそれぞれ入力と出力の離散信号，c はフィルタの係数である．6 dB/oct のフィルタを実現するには $c = 0.97$ とする．

2.1.5 Δ 特徴量と ΔΔ 特徴量

ケプストラムや MFCC などの音響特徴量は，着目した時刻からのみ計算された静的な特徴量である．一方，音響特徴量がどのように遷移したかなどを表す動的な情報を利用することで，認識や合成の性能を向上できる．このような動的成分を特徴量として扱えるよう，静的特徴量の 1 階時間微分である Δ 特徴量や，2 階微分である ΔΔ 特徴量などが多く導入される．ある時刻 t での音響特徴量 $\boldsymbol{x}_t$ に対する Δ 特徴量および ΔΔ 特徴量は，それぞれ次式のように計算されることが多い．

$$\Delta \boldsymbol{x}_t = \frac{\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}}{2} \tag{12}$$

$$\Delta\Delta \boldsymbol{x}_t = \frac{\boldsymbol{x}_{t+1} - 2\boldsymbol{x}_t + \boldsymbol{x}_{t-1}}{4} \tag{13}$$

これらを静的特徴量と結合して $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta \boldsymbol{x}_t^\top, \Delta\Delta \boldsymbol{x}_t^\top\right]^\top$ のように構成することで，動的成分を含めた音響特徴量を用いることができる．ここで $^\top$ はベクトルの転置を表す．$\boldsymbol{X}_t$ のように静的特徴量と動的特徴量を結合した特徴量を，とくに静的動的特徴量と呼ぶ．

2.2 音響特徴量から話者を推定する手法

話者認識は，MFCC のような音響特徴量の系列を入力として，学習済みの話者の音響モデルと比較し，その話者との類似度を得る技術である．話者認識のうち，話者照合は特定の話者による発話かどうかを判別する技術，話者識別は学習済み話者からどの話者による発話かを推定する技術であり，本論文では主に話者識別を扱う．また本節では，単一の話者による発話を対象とした話者識別を扱う．

古典的な話者識別法として，参照用の音響特徴量を保持しておき，与えられた音響特徴量ベクトルと参照ベクトルの距離から推定する手法がある [8]．しかし，音響特徴量は発話内容によって大きく変動するため，音響特徴量の平均ベクトルを用いた話者のモデル化は，発話内容に対するロバスト性や話者数が多い場合の識別性の観点から不十分である．

そこで，話者ごとの音響特徴量系列を確率分布としてモデル化し，入力の音響特徴量を学習済みの確率分布と比較して評価する手法が提案された．音響特徴量を確率分布でモデル化することができれば，入力の音響特徴量に対してそれがモデル話者による発話かどうかを確率として評価することができ，すべてのモデル話者の中から最も確率の高いモデルを選ぶことで話者識別が可能になる [9]．

さらに緻密な話者モデルを構築するため，多数の話者から universal background model（UBM）を構築し，これを話者ごとに適応する手法が提案されている [10, 11]．独立に話者モデルを構築する手法と比較して，話者ごとの音響モデル構築に必要な発話数が少なく認識の性能が優れる．本節では UBM の話者適応を用いた話者認識手法について述べる．

2.2.1 混合ガウスモデル（GMM）

混合ガウスモデル（Gaussian mixture model; GMM）とは，D 次元のベクトル $\boldsymbol{x}$ の確率密度関数を次式のようなガウス分布の重み付き和でモデル化するものである．

$$p(\boldsymbol{x} \mid \lambda) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \Sigma_m) \tag{14}$$

$$= \sum_{m=1}^{M} w_m \frac{1}{\sqrt{(2\pi)^D |\Sigma_m|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_m)^\top \Sigma_m^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_m)\right] \tag{15}$$

ここで，M は GMM の混合数，m は混合のインデックス，w_m，$\boldsymbol{\mu}_m$，Σ_m はそれぞれ m 番目の混合の重み，平均ベクトル，分散共分散行列である．w_m は正の重みであって，$w_m > 0$ かつ $\sum_m w_m = 1$ である．また，$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ は平均ベクトル $\boldsymbol{\mu}$，分散共分散行列 Σ をもつ多変量正規分布である．λ は混合数および各混合の重み，平均ベクトル，分散共分散行列を合わせたパラメータセットを表し，次式のように表現できる．

$$\lambda = \{M, w_1, \ldots, w_M, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M, \Sigma_1, \ldots, \Sigma_M\} \tag{16}$$

D 次元のベクトル系列 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$ を適切にモデル化するガウス混合モデルのパラメータは EM アルゴリズムによって推定できる．ここでいう適切なパラメータとは，ベクトル系列全体の尤度が最大となる，すなわち対数尤度の和が最大となるようなパラメータを指す．

MFCC のようなケプストラム領域の音響特徴量は，次元間の相関が小さいとされている．そのため，音響特徴量系列を GMM でモデル化する場合，分散共分散行列のうち次元間相関を表す非対角成分を 0 とした対角共分散を用いることが多い．

2.2.2 Universal Background Model（UBM）と話者適応

UBM は，多数の話者による発話から抽出した音響特徴量を学習した音響モデルである．多数の話者から学習することで，人の声の背景知識をモデル化することができる．また，単一話者の音響モデルは UBM を適応することで得ることができる．UBM の適応には，最大事後確率（maximum a posteriori; MAP）推定による適応（MAP 適応）が用いられる．UBM を GMM でモデル化することで，UBM の学習や適応を簡便に定式化することができる [10]．

UBM が次式の GMM で表されるものとする．

$$p(\boldsymbol{x} \mid \lambda_{\mathrm{UBM}}) = \sum_{m=1}^{M} w_m p_m(\boldsymbol{x}) \tag{17}$$

$$= \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \Sigma_m) \tag{18}$$

また，適応したい話者による発話から得られた音響特徴量系列を $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ とする．ここで，GMM

の各混合に対する特徴量 $\boldsymbol{x}_t$ の事後確率は次のようになる．

$$\gamma_m(\boldsymbol{x}_t) = \frac{w_m p_m(\boldsymbol{x}_t)}{\sum_{m'=1}^{M} w_{m'} p_{m'}(\boldsymbol{x}_t)} \tag{19}$$

また，次式で表されるパラメータ n_m を導入する．n_m は混合 m の occupation count と呼ぶ．

$$n_m = \sum_{t=1}^{T} \gamma_m(\boldsymbol{x}_t) \tag{20}$$

$\gamma_m(\boldsymbol{x}_t)$ および n_m を用いれば，GMM の EM アルゴリズムと同様にモデル全体の事後確率を最大化することで，話者 s に対して適応した GMM のパラメータを次のように求めることができる．

$$p(\boldsymbol{x} \mid \lambda_s) = \sum_{m=1}^{M} \hat{w}_m \mathcal{N}\left(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_m, \hat{\Sigma}_m\right) \tag{21}$$

$$\hat{w}_m = \left[\alpha_m \frac{n_m}{T} + (1 - \alpha_m) w_m\right] \beta \tag{22}$$

$$\hat{\boldsymbol{\mu}}_m = \alpha_m E_m[\boldsymbol{x}] + (1 - \alpha_m)\boldsymbol{\mu}_m \tag{23}$$

$$= \alpha_m \frac{1}{n_m} \sum_{t=1}^{T} \gamma_m(\boldsymbol{w}_t)\boldsymbol{x}_t + (1 - \alpha_m)\boldsymbol{\mu}_m \tag{24}$$

$$\hat{\Sigma}_m = \alpha_m E_m\left[\boldsymbol{x}\boldsymbol{x}^\top\right] + (1 - \alpha_m)\left(\Sigma_m + \boldsymbol{\mu}_m\boldsymbol{\mu}_m^\top\right) - \hat{\boldsymbol{\mu}}_m\hat{\boldsymbol{\mu}}_m^\top \tag{25}$$

$$= \alpha_m \frac{1}{n_m} \sum_{t=1}^{T} \gamma_m(\boldsymbol{x}_t)\boldsymbol{x}_t\boldsymbol{x}_t^\top + (1 - \alpha_m)\left(\Sigma_m + \boldsymbol{\mu}_m\boldsymbol{\mu}_m^\top\right) - \hat{\boldsymbol{\mu}}_m\hat{\boldsymbol{\mu}}_m^\top \tag{26}$$

ここで β は $\sum_m \hat{w}_m = 1$ となるよう選ぶ．また，α_m は m 番目の混合の適応係数で，次式によって求める．

$$\alpha_m = \frac{n_m}{n_m + r} \tag{27}$$

r は定数であり，2048 混合の場合は 16 を選ぶことが多い．

2.2.3 スーパベクトル

UBM と各話者の GMM が学習できれば，次式のように尤度比を求めることで話者 s に対するスコアを得ることができる [10]．

$$\Lambda(X \mid s) = \log p(X \mid \lambda_s) - \log p(X \mid \lambda_{\mathrm{UBM}}) \tag{28}$$

しかし，話者ごとの GMM は異なるとはいえど大きな差はないため，尤度比から話者を認識する手法では多数の話者から識別を行うことが難しい．また，各話者について GMM のパラメータを保持する必要がある上，特徴量系列に対する尤度をすべての話者に対して計算する必要があるため，計算機上で取り扱いにくい．

そこで，話者適応を平均ベクトルのみに関して行い，平均ベクトルを結合したスーパベクトルから話者表現を得る手法が提案された [12, 13]．スーパベクトルのベクトル長は混合数と特徴量の次元数の積となるが，スーパベクトルを低次元の空間に写像することで扱いやすく識別しやすい話者表現を得ることができる．

2.2.4 i-vector

スーパベクトルから抽出される識別性能の高い話者表現として，i-vector が提案されている [14]．i-vector を用いた話者認識法では，スーパベクトルが次式のように発話に依存する因子と依存しない因子で表現でき

ると仮定する．

$$\boldsymbol{M} \approx \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{29}$$

式 (29) の第 2 項がスーパベクトルのうち発話に依存する要素となる．ここで，$\boldsymbol{M}$ がスーパベクトル，$\boldsymbol{m}$ が発話に非依存なベクトル，$\boldsymbol{T}$ は全変動行列，$\boldsymbol{w}$ が全因子とよばれ，$\boldsymbol{w}$ そのものが i-vector である．$\boldsymbol{m}$ および $\boldsymbol{T}$ は，$\boldsymbol{w}$ が平均 $\boldsymbol{0}$ で分散共分散行列が単位行列の正規分布に従うように選ぶ．式 (29) をもとにすれば，次式もまた成立する．

$$p(\boldsymbol{M} \mid \boldsymbol{w}) = \mathcal{N}(\boldsymbol{M}; \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}, \Sigma) \tag{30}$$

ここで，Σ は対角共分散行列である．$\boldsymbol{T}$ および Σ は EM アルゴリズムによって推定でき，また i-vector は $\boldsymbol{T}$ と Σ から推定することができる [15,16]．また，多数の発話から話者の i-vector をモデル化する際には，発話 i-vector の平均をとることで話者 i-vector を得ることができる．

2.2.5 i-vector による話者認識

発話 i-vector を得ることができれば，その i-vector を学習済みの話者 i-vector と比較する必要がある．

古典的な手法では，話者 i-vector と発話 i-vector のコサイン類似度により，入力 i-vector に対するその話者らしさを与える．ここで，発話 i-vector $\boldsymbol{w}$ と話者 s の i-vector $\boldsymbol{w}_s$ のコサイン類似度を次式で定義する．

$$\cos(\boldsymbol{w}, \boldsymbol{w}_s) = \frac{\boldsymbol{w} \cdot \boldsymbol{w}_s}{\|\boldsymbol{w}\| \|\boldsymbol{w}_s\|} \tag{31}$$

ここでは $\|\boldsymbol{w}\|$ は $\boldsymbol{w}$ の L2 ノルムを表す．

また，線形判別分析（linear discriminant analysis; LDA）やサポートベクトルマシン（support vector machine; SVM）などの分類手法を用いることもできる．i-vector をそのままに線形判別することは難しいため，どちらの手法においてもコサインカーネルなどを用いたカーネルトリックが有効である．

さらに，確率的線形判別分析（probabilistic linear discriminant analysis; PLDA）を用いる手法も提案されている [17]．PLDA を用いた判別法では，i-vector を話者内変動と話者間変動に分離し，対象とする話者による発話かどうかを尤度比によって評価する．PLDA を用いる手法は話者内での変動を考慮して評価するため，他の判別法より比較的性能が高いことが知られている．

2.3 i-vector 以外の話者表現

i-vector ではなく，音響特徴量を入力としてニューラルネットワークを用いて話者表現を得る手法がいくつか提案されている．動的な情報を与えるため，着目しているフレームの前後数フレームの音響特徴量を結合したベクトルを入力とすることが多い．i-vector とは異なり，ニューラルネットワークの構成や入力の音響特徴量を自由に選ぶことができる．

2.3.1 x-vector

x-vector は, ニューラルネットワークを用いて得られる話者の埋め込み表現の 1 つである [18,19]．x-vector の学習には，音響特徴量を入力，話者の確率を出力とするニューラルネットワークの識別器を構成する．この識別器は最終層の活性化関数を softmax とすることで話者に対する確率を得るが，その 1 つ前の層から得たベクトルを x-vector と呼ぶ．x-vector を得るニューラルネットワークの構成の概略を図 6 に示す．音響特徴

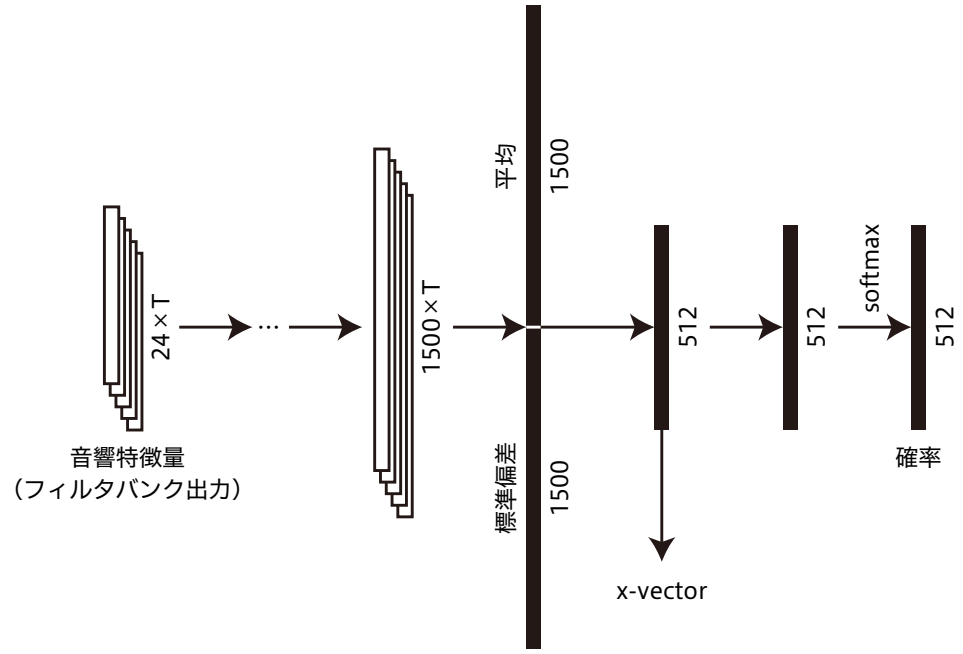


図 6　x-vector を抽出するためのニューラルネットワークの構成．識別器として構成したニューラルネットワークの中間層から話者表現を抽出する．T は音声全体のフレーム数を示す．

量には，MFCC を計算する際に用いるフィルタバンクの出力を利用する．i-vector と比較して性能が高く，異ドメイン音声に対する頑強性があると指摘されている [20]．中間層から得られた特徴量を用いる手法は様々な分野で広く適用されているが，それらの手法と同様，目的とする話者表現として正当かどうかについては疑問の余地がある．

2.3.2　Triplet Loss

話者表現を学習する上で，同じ話者の話者表現は近く，異なる話者の話者表現は遠くなることが要請される．この要請を損失関数として反映した triplet loss を用いて話者表現を学習する手法が提案されている [21]．損失関数の選び方により話者表現の分布を選べること，また短時間の音声から話者表現を抽出することが可能である点で優れている．2 つの音響特徴量に対して，話者表現の空間内での距離を求めることができるため，会話音声における話者の切り替わり点の推定に適用することができる [22]．

第 3 章

ダイアライゼーションの標準的な手法

本章では，会話音声などに対して用いられる標準的なダイアライゼーション手法について述べる．多くの手法では話者の人数や話者数などを事前に与えず，与えられた音声のみから話者の数とそれぞれの話者がいつ発話しているかを推定する．

ダイアライゼーションの文脈では，動画などを用いたマルチモーダルなダイアライゼーションが検討されることがある [23]．歌唱者ダイアライゼーションにおいても，プロモーションビデオやライブ映像などを用いてマルチモーダルなダイアライゼーションを検討することが考えられるが，本論文では CD に収録されている歌声のみを用いるためマルチモーダルなダイアライゼーションについては扱わない．

3.1 ダイアライゼーションの基本原理

ダイアライゼーションの手法は，同一の話者が発話していると思われる区間で分割し，分割されたセグメントのうち同一の話者による発話と思われるセグメントに同一のラベルを与える，という 2 手順を基礎として構成されている [5]．この手法の概念図を図 7 に示す．

3.2 標準的なダイアライゼーション手法

セグメンテーションとクラスタリングの 2 手順のみで十分な性能のダイアライゼーションを行うことは困難である．多くの手法では，この 2 手順に種々の手法を組み合わせることでダイアライゼーションを実現する．

3.2.1 前処理

話者認識を行いやすくするため，発話区間の検出，雑音除去，リバーブ除去などの処理を行う．

発話区間検出（voice activity detection; VAD）では，発話のない無音や雑音のみの区間を除外する．音声認識や話者認識の前処理として用いられるが，発話か非発話かを判別する音声認識技術の 1 つと捉えることができる．発話・非発話の誤りは直接ダイアライゼーションの誤りとなるため，高い性能の VAD は数値的な性能を上げる点で重要である．パワーに対して閾値を設定することで検出する手法が最も簡便であるが，雑音に影響されやすいなどの問題点がある．GMM や隠れマルコフモデル（hidden Markov model; HMM）を用いて統計的に推定する手法が広く用いられているほか，さらに性能の高い VAD を実現するためニューラルネットワークを用いる手法も多く用いられている [24–26]．

雑音除去およびリバーブ除去は，音声認識や話者認識において基本的な前処理であり，多数の手法が提案されている [24]．雑音除去には，時間領域や周波数領域でフィルタを適用する手法が広く知られている．ニューラルネットワークを用いた雑音除去も広く研究がなされており，波形に対して直接雑音除去を行う手法も提案

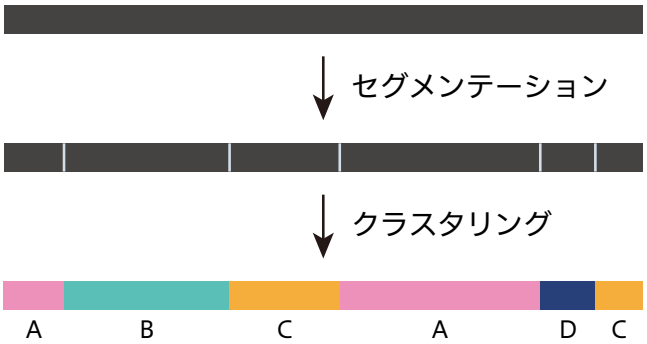


図 7　標準的なダイアライゼーション手法の概念図．音声を同じ話者が話していると思われる区間で分割し，各セグメントに対して話者に対応するラベルを与えることでダイアライゼーションは実現できる．

されている [27, 28]．話者認識や音声認識においては音響特徴量を得ることが目的であるため，雑音やリバーブを除去した音声を合成する必要はなく，スペクトログラムや用いる音響特徴量に対して雑音・リバーブ除去を行うことができればよい．

3.2.2　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．両手法によるセグメンテーションの概念図を図 8 に示す．

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 [29–31]．本節では mBIC を用いたセグメンテーション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．

$$\mathrm{mBIC} = \log L(\mathcal{X}, M) - \frac{\lambda}{2}\#(M)\log N \tag{32}$$

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，$\#(M)$ はモデルのパラメータ数，N は観測系列 $\mathcal{X}$ の長さ $|\mathcal{X}|$ である．λ は重み定数であり，実験的に決定する．λ によってモデルサイズに対する重み付けを変えることで，セグメントの分割のされやすさを指定することができる．なお，モデル選択に用いられる本来のベイズ情報量規準においては $\lambda = 1$ である [32]．

mBIC にもとづくトップダウン型のセグメンテーションでは，各話者による音響特徴量がそれぞれ異なる何らかの確率分布に従って生成されていると仮定する．そして，あるセグメントに対して全体を 1 つの話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 つの話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 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$，確率分布を単一のガウス分布とし，次式で表される 2 つの仮説を用意する．

$$H_0 : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \tag{33}$$

$$H_1 : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_1} \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1),\ \boldsymbol{x}_{N_1+1}, \ldots, \boldsymbol{x}_N \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2) \tag{34}$$

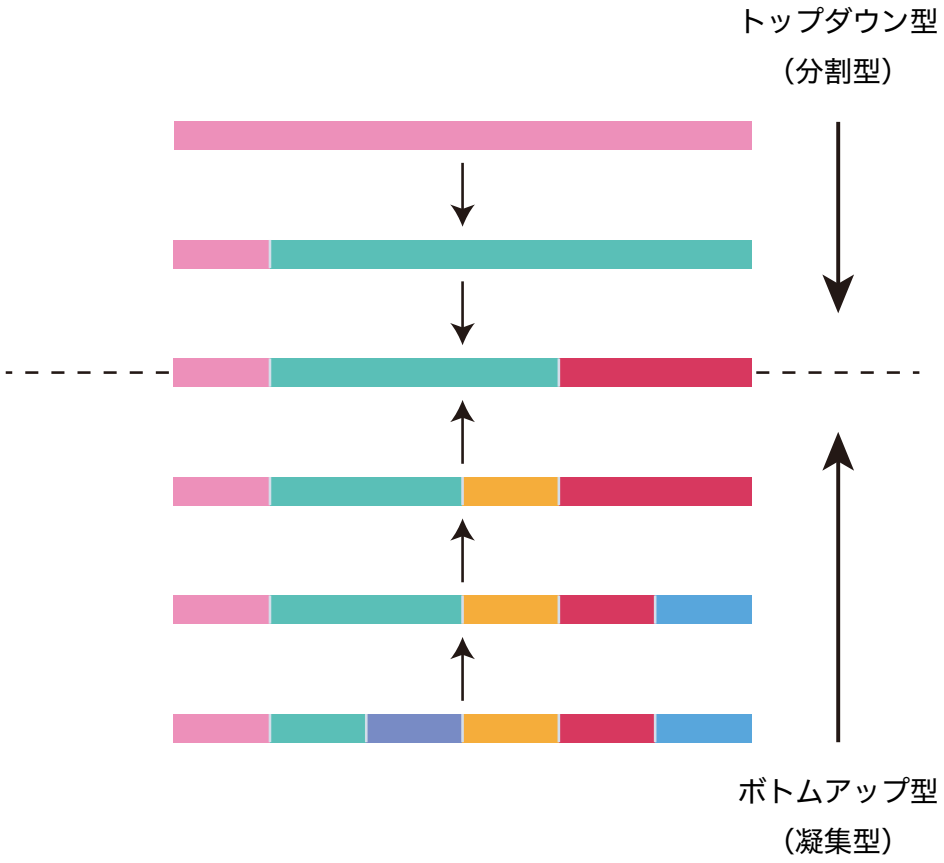


図 8　トップダウン型およびボトムアップ型のセグメンテーションの概念図．トップダウン型では分割点がない状態から出発し，ボトムアップ型では十分に分割した状態から出発する．理想的にはどちらの手法でも同じ結果を得ることができる．

ここで，$\boldsymbol{\mu}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ および $\Sigma, \Sigma_1, \Sigma_2$ 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．具体的な修正ベイズ情報量は仮説 H_0 と H_1 でそれぞれ次式のようになる．

$$\mathrm{mBIC}(H_0) = -\frac{N}{2}\log(2\pi)^m|\Sigma| - \frac{\lambda}{2}\frac{m(m+3)}{2}\log N \tag{35}$$

$$\mathrm{mBIC}(H_1) = -\frac{N_1}{2}\log(2\pi)^m|\Sigma_1| - \frac{N-N_1}{2}\log(2\pi)^m|\Sigma_2| - \frac{2\lambda}{2}\frac{m(m+3)}{2}\log N \tag{36}$$

ただし m は特徴量の次元数である．式 (35)，(36) を用いれば，仮説 H_0 と仮説 H_1 の修正ベイズ情報量の差 ΔmBIC は次式のように計算される．

$$\Delta\mathrm{mBIC} = \mathrm{mBIC}(H_1) - \mathrm{mBIC}(H_0) \tag{37}$$

$$= \frac{1}{2}[N\log|\Sigma| - N_1\log|\Sigma_1| - (N-N_1)\log|\Sigma_2|] - \frac{\lambda}{2}\frac{m(m+3)}{2}\log N \tag{38}$$

情報量が高いほど音響特徴量系列を表現するモデルとして適切であることから，ΔmBIC 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．セグメントに対して分割点を決定し，修正ベイズ情報量が上昇しなくなるまで再帰的に繰り返すことで，音声を同一話者と推定される区間でセグメンテーションすることができる．なお，セグメントの分割点は，分割点によらずモデルのパラメータ数が同じであることから，系列の

対数尤度が最大となる点を選べばよい．式 (34) においては次式によって N_1 を選ぶ．

$$\hat{N_1} = \arg\max_{N_1} -\frac{N_1}{2}\log(2\pi)^m|\Sigma_1| - \frac{N-N_1}{2}\log(2\pi)^m|\Sigma_2| \tag{39}$$

ボトムアップ型のセグメンテーションでは，初期状態の修正ベイズ情報量といずれかの隣り合うセグメントを結合した場合の修正ベイズ情報量を比較して，凝集するかを選択する．初期のセグメント数を $K+1$ とすれば，$K+1$ のセグメントに分割しておく仮説 H_K と，いずれかの隣り合うセグメントを結合したときの仮説 H_{K-1} を用いて，それぞれの仮説に対する修正ベイズ情報量を次式のように計算する．

$$\mathrm{mBIC}(H_K) = \sum_{k=1}^{K+1} -\frac{N_k}{2}\log(2\pi)^m|\Sigma_k| - \frac{(K+1)\lambda}{2}\frac{m(m+3)}{2}\log N \tag{40}$$

$$\begin{aligned}\mathrm{mBIC}(H_{K-1}) = &\max_{1\le k'\le K}\left[\sum_{1\le k<k',k'+1<k\le K+1} -\frac{N_k}{2}\log(2\pi)^m|\Sigma_k| - \frac{N_{k'}+N_{k'+1}}{2}\log(2\pi)^m\left|\Sigma_{k',k'+1}\right|\right] \\ &- \frac{K\lambda}{2}\frac{m(m+3)}{2}\log N \end{aligned} \tag{41}$$

$$\begin{aligned}= \mathrm{mBIC}(H_K) &+ \frac{\lambda}{2}\frac{m(m+3)}{2}\log N \\ &+ \max_{k'}\left[\frac{N_{k'}}{2}\left|\Sigma_{k'}\right| + \frac{N_{k'+1}}{2}\left|\Sigma_{k'+1}\right| - \frac{N_{k'}+N_{k'+1}}{2}\left|\Sigma_{k',k'+1}\right|\right]\log(2\pi)^m \end{aligned} \tag{42}$$

ここで，Σ_k は分割前の k 番目のセグメントに属する音響特徴量系列の分散共分散行列，k' および $k'+1$ は凝集するセグメントのインデックス，$\Sigma_{k',k'+1}$ は k' 番目と $k'+1$ 番目のセグメントを結合したセグメントに属する音響特徴量系列の分散共分散行列である．$\mathrm{mBIC}(H_K)$ より $\mathrm{mBIC}(H_{K-1})$ が大きければ，そのセグメンテーションは凝集するべきであると判断できる．トップダウン型の場合と同様，この操作を mBIC が大きくなる限り繰り返すことで，音声のセグメンテーションを実現することができる．

トップダウン型，ボトムアップ型によらず，最小の系列長を適切に決める必要がある．同一話者の発話している最小の時間に制約を設ける必要がある上，あらゆる長さの系列を仮定すると計算コストが大きくなるためである．また，長さが 1 の系列を仮定すると対数尤度が発散することにも注意する．

セグメンテーションは，候補とする分割位置が多くなるほど計算量が増大する．計算量を減らしつつも詳細なセグメンテーションを実現するため，逆ガウス分布を用いて発話長をモデル化してベイズ情報量の計算回数を減らす手法も提案されている [33]．

3.2.3 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，トップダウン型とボトムダウン型の両手法でクラスタリングを行うことができる．この概念図を図 9 に示す．

クラスタリングでは，2 つのセグメントが同一の話者によるものか異なる話者によるものかを繰り返し評価することで行う．セグメンテーションと同様，mBIC を用い情報量が多いを仮説を選ぶことでクラスタリングを行うことができる [29]．また，セグメンテーションと比較して仮説を適用する音声の長さが長く話者表現を抽出することができるため，話者表現の距離を規準としてクラスタリングを行う手法も用いられている [25, 34, 35]．

クラスタリングでは，隣り合うセグメントを同一の話者による発話としてラベル付けすることができる．し

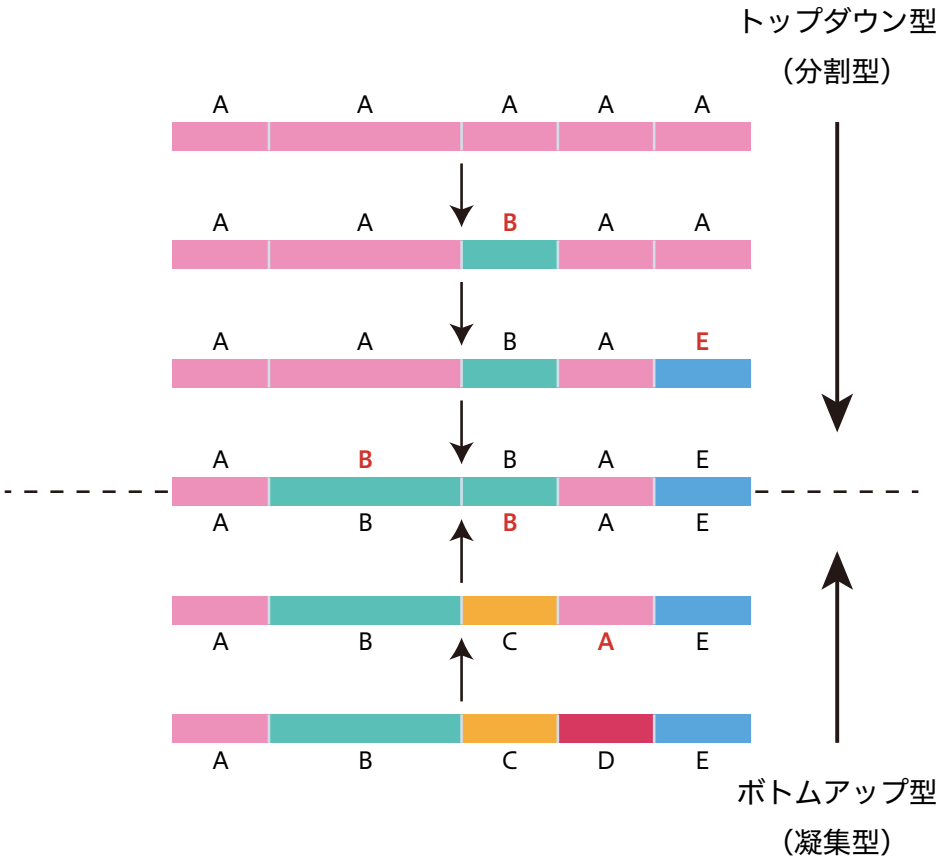


図 9　トップダウン型およびボトムアップ型のクラスタリングの概念図．トップダウン型ではすべてのセグメントが同一の話者である仮説から出発し，ボトムアップ型ではすべてのセグメントが異なる話者である仮説から出発する．セグメンテーションと同様，理想的にはどちらも同じ結果になる．

たがって，実際の話者の切り替わりよりも高い頻度のセグメンテーションが得られても，クラスタリングで不適切な話者の切り替わりを抑制することができる．

トップダウン型とボトムアップ型のどちらを用いても，セグメンテーションが正しい限り，理想的には正しい結果を得ることができる．一方，mBIC などを用いてクラスタリングする場合，構築される音響モデルがトップダウン型とボトムアップ型で異なるため，クラスタリングの結果に違いが生じる [36]．トップダウン型は長時間の音声から音響モデルを構築するため，話者内変動に対して安定であるが話者の識別性に劣る．一方ボトムアップ型は音響モデルが短時間の音声から緻密に作られるため話者識別性に優れるが，発話内容や雑音など話者情報以外から影響を受けやすく不安定である．結果は異なるが性能に大きな差はなく，2 つの手法を組み合わせることでさらに性能が向上する．

3.2.4　後処理

クラスタリングで得られた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．

クラスタリング時に最後に構築されたモデルを用いて，セグメンテーション境界を修正する手法が提案されている [37]．また，単語境界などでごく短時間で発話・非発話区間が切り替わるような場合があり，短時間の非発話区間を抑制するような手法も適用されている．

クラスタリングの性能はセグメンテーションの結果に大きく依存するが，セグメンテーション自体の性能

を十分に確保することは難しい．そのため，セグメンテーションとクラスタリングを同時に行う手法が提案されている．ボトムアップの手法では，HMM の各状態を話者の音響モデルと仮定して十分な数の状態を用意し，各状態に対して話者の発話したセグメントを表すサブ状態を導入して，これを結合していく手法が提案されている [38]．トップダウンの手法では，進化形 HMM（evolutive hidden Markov models; E-HMM）と呼ばれる HMM の学習法を用い，初期の HMM から話者状態を増やしていく手法が提案されている [39]．

3.2.5 マルチチャンネル処理

複数のマイクを用いて収録した音声を利用してダイアライゼーションの性能を向上させる手法がいくつか提案されている．最も単純な手法として，各チャネルの音声を独立にセグメンテーションしその結果を結合することでセグメンテーションの性能を向上させる手法が提案されている [40]．

マルチチャンネルの音声を扱う上では，特定の方向からの音声のみを強調もしくは抑制するビームフォーミングが有効である [24]．ビームフォーミングの最も古典的な手法として，音声がマイクに到達する時間の差（time-delay-of-arrival; TDOA）を利用して，各チャンネルの音声にそれぞれ異なる遅延を与えて加算する遅延和ビームフォーミング（delay-and-sum beamforming）が知られている [41]．この遅延和ビームフォーミングを用いて前処理として雑音を抑制することで，ダイアライゼーションの性能を向上させる手法が提案されている [42]．遅延和ビームフォーミングには時間遅延推定（time delay estimation; TDE）が必要であり，TDE には一般化相互相関関数（generalized cross-correlation; GCC）を用いた GCC-PHAT 法などが知られている [43]．

本研究で扱う歌声を対象としたダイアライゼーションにおいては，CD に収録されている音源がステレオであることから，マルチチャンネルの処理を用いることが可能である．楽曲のミックスにおいては，歌声や楽器ごとに左右の音量を変えるパンニング（panning）と呼ばれる処理を行うことで定位感[*4]を与えている．歌声が左右に振られていることを利用すれば，同時に歌唱している人数を推定して追加情報とする方法が考えられる．またステレオ音声からセンター音声を抽出する手法はいくつか知られており，伴奏音の抑制に用いることも可能である [44]．本論文では CD から抽出した伴奏音のない歌声のみを用い，また同時に歌唱している部分についてもパンニングを考慮しないため，本論文ではマルチチャンネル処理については扱わない．

[*4] 音像が左右前後様々な方向に存在するように感じられること．

第 4 章

ボーカルのある楽曲に対する音楽情報処理の基礎技術

本章では，歌唱者認識をはじめとする歌声に対する分析技術について述べる．歌唱者ダイアライゼーションは歌声を扱う技術の 1 つであり，本章で述べる技術は本研究の基礎となる．

4.1　歌声と話し声の相違点

歌声と話し声では，その音響的な特徴が大きく異なることが知られている [45]．

歌声の基本周波数は話し声よりも大きく変動し，またプレパレーションやビブラートなど歌声に特有の性質が観測される [46]．またスペクトル包絡は，同じ母音でも歌声と話し声で大きく異なる場合があることが指摘されている [47]．そのため，歌声ではなく話し声で学習した音響モデルを歌唱者認識などに利用することは難しく，また音響モデルの構築においても学習に用いる系列長などを十分考慮する必要がある．

4.2　歌唱者認識

歌唱者認識は楽曲を自動で分類するために重要な技術であり，歌唱者認識によって同一歌唱者による楽曲をメタデータなしに探すことができるなど種々の応用が期待される．本論文では，歌声から歌手の特徴を抽出して比較や照合などを行うことを歌唱者認識と呼び，歌唱者認識のうち特に多数の既知の歌手から与えられた楽曲の歌唱者を識別する問題を歌手同定と呼ぶ．

歌唱者認識では話者認識と同様に特徴量として MFCC が用いられることが多い．数人からの歌手同定では，楽曲のうち歌唱している部分を取り出すことで伴奏音の抑制を行わずとも一定の性能で実現できる [48, 49]．また，歌声に対して伴奏音抑制を適用する手法や，特徴量領域で歌声の強調を行う歌手同定法も提案されている [50–52]．さらに，歌唱者の音響モデルを構築せず，楽曲それぞれの特徴量分布から楽曲同士の相互情報量を計算することで，歌手の類似度のみを得る手法も提案されている [53]．

多くの場合，認識対象の楽曲はある程度の長さを持つため，短時間での認識が行えずとも楽曲に対する歌唱者認識の性能を確保することができる．また，歌唱者が 1 人であることを仮定している場合が多く，このような場合には声の調波構造のみを取り出すことで歌声抽出を行うことができる [54]．

2 人の歌唱者が存在する歌声に対しては分離を行わず，尤度比による評価に委ねる手法や，複数の歌唱者が同時に歌っている音声そのものをモデル化する手法が提案されている [3, 4]．

4.3　伴奏音抑制・歌声抽出

伴奏音抑制や歌声抽出は音源分離技術の一種であり，音声認識や話者認識における雑音除去と同様，歌唱者認識や歌詞認識などの歌声処理において重要な技術の 1 つである．伴奏音は雑音と同様に抑制する手法が

多く提案されているが，伴奏音は音楽的な構造を持った非定常的な音声となるため一般的な雑音に比べ抑制は容易ではない．

伴奏音抑制は，スペクトログラムから歌声の特徴が見られる部分を抽出し，ウィーナーフィルタや位相推定などを用いて音声を復元する手法が広く用いられている．抽出する場合，スペクトログラムを直接推定する手法だけでなく，スペクトログラムの各時刻・周波数ビンが歌声か伴奏音を 2 値で表現してマスクを行う手法が用いられる．マスクを適用する手法を拡張し，各時刻・周波数ビンに対するマスクの値に実数値を選ぶ場合には特にソフトマスクと呼ばれ，より高品質な伴奏音の抑制が期待できる．スペクトログラム，マスクのどちらを推定する手法にせよ，音声のスペクトログラムのうち振幅成分のみに着目しており，位相成分を考慮して音声を再合成する手法はほとんど行われていない．

4.3.1 スペクトログラムの特徴を利用する手法

音楽の音響信号では，歌声やピアノの音声のように調波構造を持つものと，ドラムなどのように雑音性が強く調波構造を持たないものに大きく分けられる． そのため，スペクトルの倍音構造をモデル化して抽出することで，スペクトルから選択的に歌声を抽出する手法が考えられる [55].

歌声は楽曲中で変動が多く，楽器音は繰り返し構造が多い．この性質を利用し，ロバスト主成分分析によってスペクトログラムを分解する手法が提案されている [56]．ロバスト主成分分析では，行列 M を次式のように分解する [57].

$$\text{minimize} \quad \|L\|_* + \lambda\|S\|_1 \tag{43}$$

$$\text{subject to} \quad L + S = M \tag{44}$$

ここで，$\|\cdot\|_*$ は行列のトレースノルムを，$\|\cdot\|_1$ は行列の L1 ノルムを表す．トレースノルムは行列の特異値の和，L1 ノルムは行列のすべての要素の和である．λ は正のパラメータで，行列 L の低ランク性と行列 S のスパース性のトレードオフを決定する．M をスペクトログラムとすれば，繰り返し構造がある楽器音のスペクトログラムほど低ランク行列 L で表現され，一方繰り返し構造が少なく調波構造を持つ歌声は疎行列 S で表現されやすくなる．

4.3.2 非負値行列因子分解を用いた手法

スペクトルの特徴をデータから学習する手法として，非負値行列因子分解（non-negative matrix factorization; NMF）を用いる手法が広く知られている．

NMF は，入力の非負値行列を加算モデルで表すことで，入力を実現する因子を分析する手法である．具体的には，非負値行列 $\boldsymbol{Y}$ を，非負値行列 $\boldsymbol{H}$ と $\boldsymbol{U}$ を用いて次式のように近似する．

$$\boldsymbol{Y} \approx \boldsymbol{H}\boldsymbol{U} \tag{45}$$

ここで $\boldsymbol{H}$ は基底，$\boldsymbol{U}$ はアクティベーションと呼ばれる．$\boldsymbol{Y}$ をスペクトログラムと考えれば，ある時刻 t のスペクトル $\boldsymbol{y}_t$ が NMF によって次式で近似できることを意味する．

$$\boldsymbol{y}_t = \sum_{n=1}^{N} u_{n,t}\boldsymbol{h}_n \tag{46}$$

ここで，$\boldsymbol{h}_n$ は $\boldsymbol{H}$ の第 n 列のベクトル，$u_{n,t}$ は $\boldsymbol{U}$ の n 行 t 列の要素を表す．式 (46) は，任意の時刻 t のスペクトル $\boldsymbol{y}_t$ が，N 個のスペクトルテンプレート $\boldsymbol{h}_n$ で表され，各テンプレートの生起の強さが $u_{n,t}$ であること

を表式したものである．NMF では，非負値行列 $\boldsymbol{Y}$ さえ与えられれば，補助関数法によって $\boldsymbol{H}$ と $\boldsymbol{U}$ を推定することができる [58]．

NMF は，分析や分離など音楽音響信号のスペクトログラムを扱う手法として広く用いられてきた [59–61]．歌声の分離についても同様に NMF を用いた手法が提案されている [62]．また，歌声のスペクトルに倍音構造が現れることを利用し，拡張した NMF でスペクトログラムをモデル化することで歌声分離を行う手法も提案されている [63, 64]．

4.3.3 ニューラルネットワークを利用した手法

これまでの手法では，スペクトルの倍音構造に着目して伴奏音抑制や歌声抽出を行う手法が多く用いられてきた．一方，伴奏音にも倍音構造を持つ楽器が多くある上，伴奏音のスペクトルを NMF によって加算モデルで表現するのは困難である．そのため，スペクトログラムを入力としたニューラルネットワークを構成する手法も多く提案されている [65–68]．

第 5 章

実験の構成および条件

本章では，本研究で行った実験の構成と，実験に共通する実験条件について述べる．

5.1 実験の構成

本論文では，第 3 章で述べたダイアライゼーション手法を適用した手法（手法 1）および，第 4 章で述べた歌唱者認識を利用した手法（手法 2）の 2 手法を比較し検討する．また，手法 1 で用いたセグメンテーションと手法 2 で用いた歌唱者認識を組み合わせた手法（手法 3）についても検討する．

本研究では CD などに収録されている楽曲に対して歌唱者ダイアライゼーションを行うことを目標とする．しかし，伴奏音が含まれた歌声を直接扱うことは難しく，基礎的な検討を行うことが困難である．そのため本論文では，歌声が含まれた音声と歌声のない伴奏音のみの音声を用いて作成した，歌声のみの音声を利用する．

5.2 グループアイドルソングのデータセット

本研究で行う歌唱者ダイアライゼーションの基礎的な検討には，歌唱者それぞれの音響モデルが学習でき，また同一の楽曲に対して種々の歌唱者の組み合わせが可能なデータセットが求められる．そこで，複数の楽曲を共通の歌唱者がソロで歌った音声が存在するゲーム内の楽曲に着目し，これらから新たにデータセットを構築した．

楽曲は A から L の 12 曲ある．楽曲 A は手法 1 におけるハイパーパラメータの決定に，楽曲 B，C は手法 2 における i-vector の学習に用いた．楽曲 D 以降の 9 曲は評価に用いた．B，C を除く楽曲には 3 人の歌唱者によるパート割りを行った．楽曲 A，D，E，F，G のパート割りはソロまたは 3 人同時の部分のみから構成され，2 人同時に歌っている区間はないものとした．すなわち，このパート割りにおいては，誰も歌っていない，1 人目の歌唱者のみが歌っている，2 人目の歌唱者のみが歌っている，3 人目の歌唱者のみが歌っている，3 人全員が歌っている，の 5 通りの場合が存在する．ここではこのパート割り方法を「デュオなし」と呼ぶ．一方，楽曲 H，I，J，K のパート割りでは，ソロまたは 3 人同時の部分の他に 2 人で歌う部分のあるパート割りとした．2 人で歌う部分は，1 人目（センター）を除く 2 人が歌っており，センターと残りの 2 人のうちどちらか 1 人のみが歌っている部分はないものとした．すなわち，このパート割りにおいては，デュオなしの場合に加えて 2 人目と 3 人目の歌唱者が歌っている区間が存在し，6 通りの場合が存在する．ここではこのパート割り方法を「デュオあり」と呼ぶ．楽曲 L にはデュオなしとデュオありの両方のパート割りを設け，それぞれ Lf と Lt と呼ぶ．これらのパート割りの構成条件を表 1 に要約する．

歌唱者は A から M の 13 名が存在する．これらの歌唱者すべてが歌った楽曲は存在しないが，楽曲 B，C の 2 曲を用いることでデータセット中に含まれる 2 人同時，3 人同時のあらゆる組み合わせの音響モデルを

表 1　各楽曲のパート割り構成と用途．楽曲のアルファベットは付録 A の表 11 と対応する．

	楽曲名	パート割りの種類	利用目的
A	おはよう!! 朝ご飯	デュオなし	手法 1 におけるハイパーパラメータ決定
B	スタ→トスタ→	パート割り音声を用いず	手法 2 における i-vector の学習
C	MEGARE!		
D	蒼い鳥	デュオなし	手法の評価
E	きゅんっ! ヴァンパイアガール		
F	Honey Heartbeat		
G	隣に…		
H	9:02 pm	デュオあり	
I	エージェント夜を往く		
J	キラメキラリ		
K	THE IDOLM@STER		
Lf	relations	デュオなし	
Lt		デュオあり	

構成することができる．

データセットの構築およびパート割りの詳細については付録 A を参照されたい．

5.3　音声の条件およびダイアライゼーション時の処理

歌声はすべてミックス後に 44.1 kHz から 16 kHz にダウンサンプリングを行った．音響特徴量には 12 次元の MFCC およびそのデルタ特徴量を用いた．パワーにもとづいて歌声区間の検出を行い，誰も歌っていない区間は事前に認識の対象から除外している．

声の重なり合いについては，会話音声に対するダイアライゼーションで用いられるような，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．また，歌唱者認識で利用する i-vector を用いた話者認識手法は，一人の話者が単独で発話した音声を前提としている．そこで，本論文では同時歌唱の音声を 3 人とは別の異なる歌唱者があたかも歌唱したかのようなモデル化を行う．たとえばデュオなしの音声であれば，手法 1 では 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し，手法 2 では 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．デュオありの音声であっても同様に 5 つめのクラスタやモデルを仮定する．このような同時歌唱を 1 つの音響モデルで扱う手法は，デュエットの音声において検討がなされている [4]．

楽曲 B，C を除くパート割りが必要となる楽曲については，すべて 10 通りのパート割りを無作為な組み合わせで作成し，学習および評価データとした．なお，Lf と Lt については同一の歌唱者の組み合わせを用いて 10 通りのパート割りを構成した．

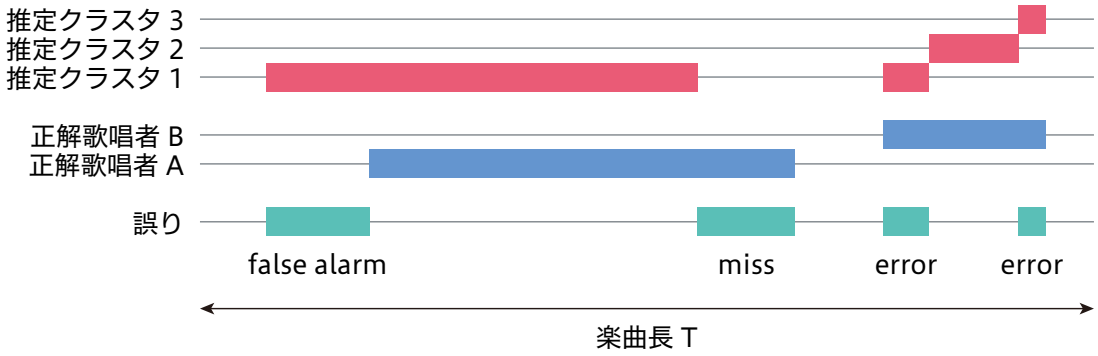


図 10　DER を計算する上で考慮する 3 種類の誤りの模式図．正解歌唱者 A と推定クラスタ 1，正解歌唱者 B と推定クラスタ 2 が対応している．推定クラスタ 3 は対応する正解歌唱者がいないため，すべて誤りの区間となる．ここでは歌唱者のオーバーラップを考慮していない．

5.4　客観評価の指標

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 [69]．DER は S4D [70] を用いて計算した．

DER は，音声の総時間に対する次の 3 つの誤り区間の長さの割合で定義される．

1. 誤った歌唱者でラベリングされた区間（error もしくは confusion）
2. 誤も歌っていない区間に歌唱者がラベリングされた区間（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区間（miss）

この 3 種類の誤りを図 10 に例示する．DER は次式のように，音声全体を S 個のセグメントに分割し各誤りセグメントの長さから計算される．

$$\mathrm{DER} = \frac{\sum_{s=1}^{S} \tau_s \left(\max\left(N_s^{(\mathrm{ref})}, N_s^{(\mathrm{hyp})} \right) - N_s^{(\mathrm{correct})} \right)}{\sum_{s=1}^{S} \tau_s N_s^{(\mathrm{ref})}} \tag{47}$$

ここで，s はセグメントのインデックス，τ_s は対応するセグメントの長さ，$N_s^{(\mathrm{ref})}$，$N_s^{(\mathrm{hyp})}$ は対応するセグメントで推定された同時発話者数と正解の同時発話者数，$N_s^{(\mathrm{correct})}$ は対応するセグメントで正しく推定された話者数を表す．DER の定義では話者がオーバーラップすることを前提としている．しかし，本論文ではオーバーラップ区間を別の話者として扱うためこれを考慮する必要がない．話者のオーバーラップを考慮しない場合，DER は単純に次のように表式できる．

$$\mathrm{DER} = \frac{T_E + T_{FA} + T_M}{T} \tag{48}$$

ここで T は音声全体の長さ，T_E は error 区間の長さ，T_{FA} は false alarm 区間の長さ，T_M は miss 区間の長さを表す．

ダイアライゼーションはあくまでクラスタリング問題であり，各クラスタがどの話者に属するクラスタかを推定しない．一方，DER を得るためには正解ラベルの話者とクラスタの対応が取れていなければならない．そこで，DER が最も低くなるように推定されたクラスタと話者の一対一対応をとる．推定されたクラスタ数が実際の話者数より多い場合には，対応する話者が存在しないクラスタとして扱い，そのクラスタが現れる区間はすべて error もしくは false alarm として計算する．クラスタ数が話者数より少ない場合にも同様に，対

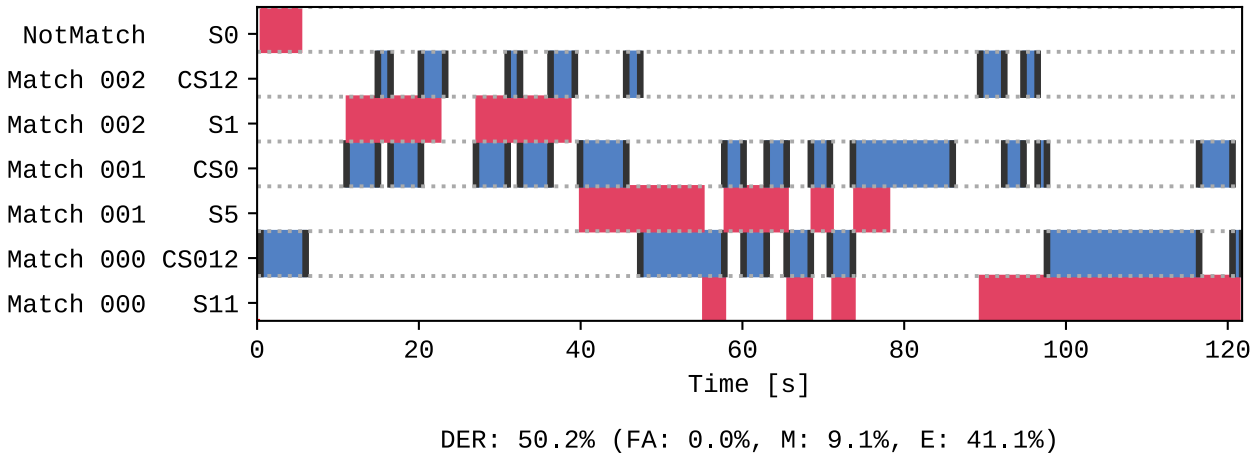


図 11　ダイアライゼーションの結果を示す図の例．これは歌唱者 B，M，C の歌唱した楽曲 J に対する歌唱者ダイアライゼーションの結果である．最下部の FA は false alarm，M は miss，E は error の区間の割合をそれぞれ表す．FA，M，E の和が DER となる．正解クラスタの番号は歌唱者の番号を示す．CS12 であれば，ここでは歌唱者 M，C が歌唱している区間を示す．

応するクラスタが存在しない話者の区間をすべて error もしくは miss として扱う．

DER の計算では，発話区間の曖昧さによって誤差が生じないよう正解ラベルの話者が切り替わる時刻の前後を計算対象に含まないような処理を行う．本論文では，各話者の切り替わり前後 0.5 秒間を DER の計算から除外する区間としている．

5.5　ダイアライゼーション結果の図示

本論文では，ダイアライゼーションを行って得られた結果を図 11 のように示す．青色の帯の行は正解の歌唱部分を示した行で，赤色の帯の行は推定結果を示した行である．青色の帯の左右にある黒の部分は DER の計算から除外された区間である．下から順に，対応する推定歌唱者の存在しない正解歌唱者，推定歌唱者とそれに最も合致した正解歌唱者の組，対応する正解歌唱者の存在しない推定歌唱者が並ぶ．推定歌唱者と正解歌唱者の対応が取れている行は，「Matched n」の形式で対応を示している．図 11 では，すべての正解歌唱者に対して対応する推定歌唱者が存在するため，推定歌唱者と対応が取られていない正解歌唱者は存在しない．

第 6 章

標準的なダイアライゼーション手法を用いた歌唱者ダイアライゼーションの実験

本章では，標準的なダイアライゼーション手法をグループアイドルソングに適用した実験について述べる．

6.1 用いた手法

ダイアライゼーションの実装には，ダイアライゼーションのツールキットである S4D [70] を用いた．3.2 節に示した手法を基礎として，次に述べる手法でダイアライゼーションを行った．

6.1.1 発話区間検出

SIDEKIT [71] の実装により，パワーにもとづいた発話区間検出を行った．ここでは，SNR が 40 dB となるように発話区間検出を行った．

6.1.2 ガウス分布のダイバージェンスにもとづくセグメンテーション

S4D では，mBIC によるボトムアップ型のセグメンテーションを行う．ボトムアップ型のセグメンテーションを行う場合，初期値として音声を十分セグメンテーションした状態が必要になる．最も単純な手法では音声を一定時間ごとに分割した状態を初期値として与えることが考えられるが，音響特徴量にもとづき分割点の候補となる点で分割した状態を初期値とすることで，計算時間の短縮や性能の向上を図ることができる．

mBIC によるセグメンテーションの初期値として，音響特徴量分布のダイバージェンスによって分割したセグメンテーションを与える手法が提案されている [72]．ここで，音響特徴量系列 s_1 および s_2 をそれぞれ対角共分散のガウス分布 $s_1 \sim \mathcal{N}(\boldsymbol{\mu}_1; \Sigma_1)$ および $s_2 \sim \mathcal{N}(\boldsymbol{\mu}_2; \Sigma_2)$ でモデル化し，分布間のダイバージェンスを次式で定義する．

$$D = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \Sigma_1^{-\frac{1}{2}} \Sigma_2^{-\frac{1}{2}} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \tag{49}$$

ただし，対角共分散行列に対する $-\frac{1}{2}$ 乗は次式で定義する．

$$\Sigma^{-\frac{1}{2}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_d} \end{bmatrix} \tag{50}$$

このダイバージェンスは，2 つの分布の平均的な分散をもとにした，分布の平均間のマハラノビス距離にあたる．本論文では，最短のセグメント長が 2.5 秒になるように上述のダイバージェンスを規準にしてセグメンテーションを行った．

6.1.3　mBIC にもとづくセグメンテーション・クラスタリング

初期セグメンテーションをもとに，ボトムアップ型のセグメンテーションおよびクラスタリングを行う．ともに mBIC における定数をハイパーパラメータとして与える必要がある．

6.1.4　HMM を用いた再セグメンテーション

各話者を状態とする HMM を構成してビタビ探索を行う．各状態については，音響特徴量を 8 混合の GMM でモデル化する．GMM の分散共分散行列は対角共分散としている．

6.2　セグメンテーション・クラスタリングにおけるハイパーパラメータの効果

3.2 節で述べたように，mBIC の定数を適切に選ぶことで，モデルのパラメータ数に対する重みが変化し，セグメンテーションやクラスタリングにおける分割・凝集の終了判定の閾値を調整することができる．

mBIC の定数によって，セグメンテーションの結果が変化する例を図 12 に示す．mBIC のパラメータが大きいほどモデルのパラメータ数に対するコストが大きくなるため，セグメント数のより少ない結果が得られる．

また同様に，クラスタリングについての例を図 13 に示す．セグメンテーションの場合と同じように，mBIC のパラメータが大きいほどクラスタ数のより少ない結果が得られる．

6.3　標準的なダイアライゼーション手法を用いた歌唱者ダイアライゼーション

楽曲 A に対して 10 通りのパート割り音声を作成し，この結果をもとに mBIC の定数を決定した．セグメンテーションについては 2，クラスタリングについては 0.95 とした．

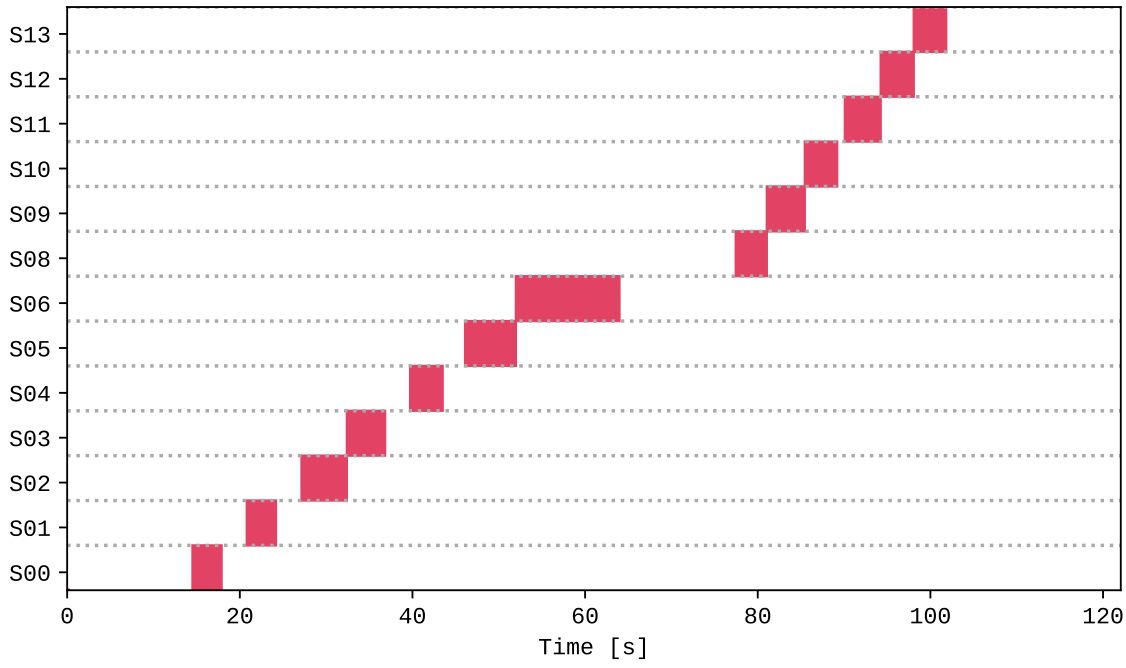
楽曲 D–K，Lf，Lt の 10 曲について，各 10 通りのパート割りを用い，合計 100 曲に対して標準的なダイアライゼーション手法を用いた歌唱者ダイアライゼーションを行った．この結果を表 2 に示す．パラメータの設定に用いた楽曲 A と比較し，評価に用いた他の楽曲では非常に高い DER が示された．デュオの有無に注目すると，デュオのない音声と比較して，デュオのある音声は DER のより低い結果が得られた．

6.4　本節の実験結果の考察

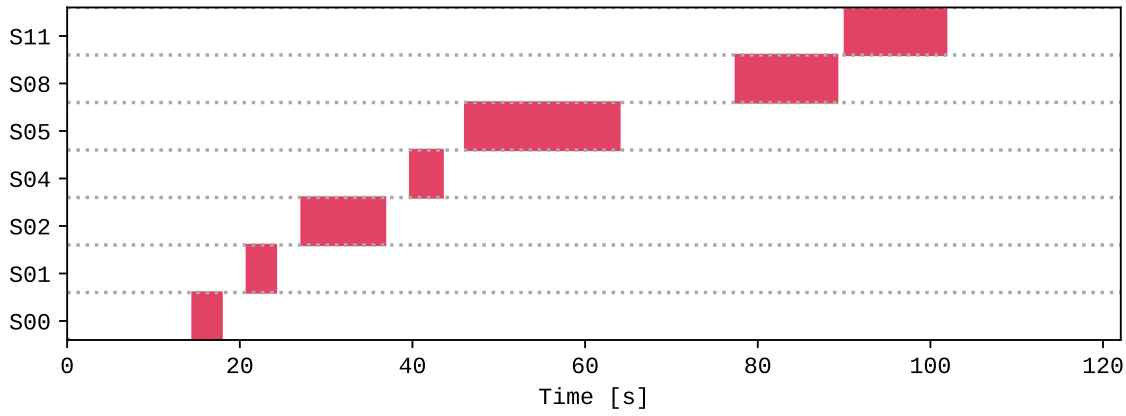
最も DER が高かった楽曲のダイアライゼーションの結果を図 14 に示す．この楽曲のように，正解のクラスタ数よりも推定されたクラスタ数が多い楽曲は 100 曲のうち 83 曲あった．クラスタ数は楽曲によって異なる傾向があり，楽曲 I，J，K を除く楽曲ではクラスタ数が平均して 9 を超えるほど多くなった．デュオパートが存在する楽曲の DER が低かった理由は，クラスタ数が過多な傾向であることから，正解歌唱者に対応するクラスタが増えたためであると考えられる．

また同様に，最も DER が低かった楽曲のダイアライゼーションの結果を図 15 に示す．楽曲 K はパート割りが単純である上，クラスタリングにおいてクラスタ数が増えにくい楽曲だったため DER が低くなったと考えられ，ダイアライゼーション自体の性能は高くないことが示唆される．

ハイパーパラメータの設定に用いた楽曲 A のみ極端に DER が低いことから，楽曲に応じて適切なハイパーパラメータを選ぶ必要があることが示唆される．

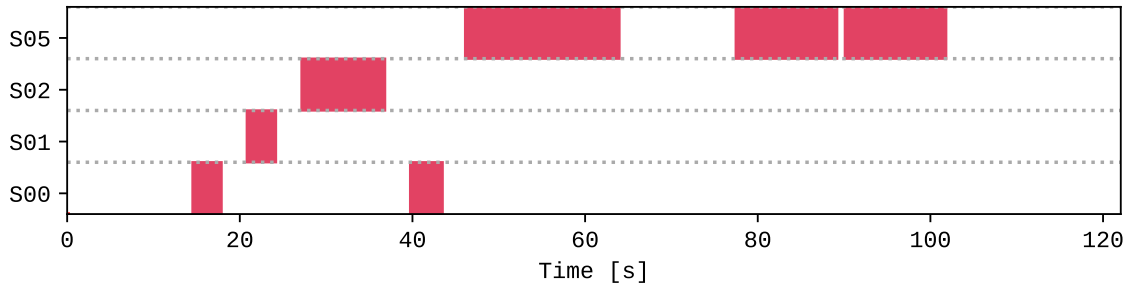


(a) mBIC の定数 λ を 1 とした場合のセグメンテーションの結果.

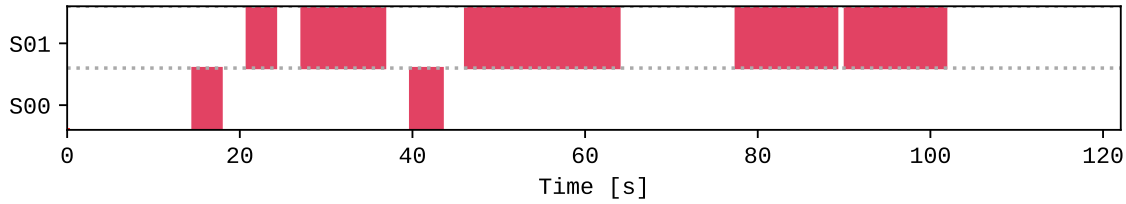


(b) mBIC の定数 λ を 2 とした場合のセグメンテーションの結果.

図 12　mBIC のパラメータによるセグメンテーションの結果における変化の例．ここでは歌唱者 J，H，E の歌唱した楽曲 A を示している．



(a) mBIC の定数 λ を 0.95 とした場合のクラスタリングの結果.



(b) mBIC の定数 λ を 1.5 とした場合のクラスタリングの結果.

図 13　mBIC のパラメータによるクラスタリングの結果における変化の例．ここでは歌唱者 J，H，E の歌唱した楽曲 A を示している．

表 2　標準的なダイアライゼーション手法を利用したダイアライゼーションの実験結果．値は DER [%]．楽曲 A はハイパーパラメータの決定に用いた学習データであるため，全体の値には含まれていない．

楽曲	パート割り	平均	標準偏差
A		35.7	2.0
D		71.0	6.0
E	デュオなし	49.1	1.8
F		67.2	5.5
G		63.6	4.9
Lf		69.5	5.2
H		60.7	3.8
I		65.2	3.2
J	デュオあり	57.0	4.0
K		45.2	11.7
Lt		62.3	5.0
デュオなし		64.1	9.3
デュオあり		58.1	9.3
全体		61.1	9.7

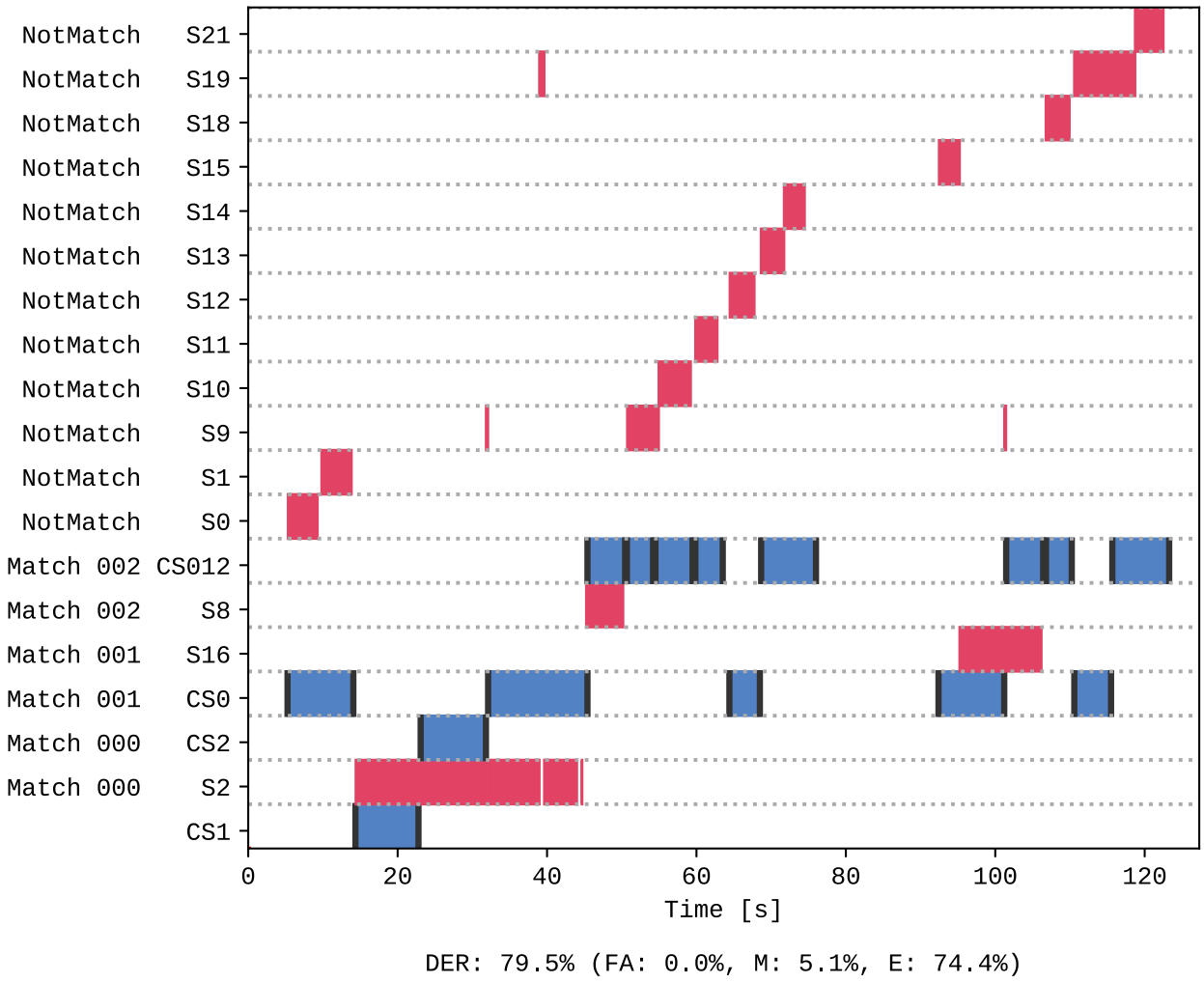


図 14　歌唱者 J，B，C による楽曲 D の歌唱者ダイアライゼーションの結果．本節の実験において最も DER が高かった歌声である．

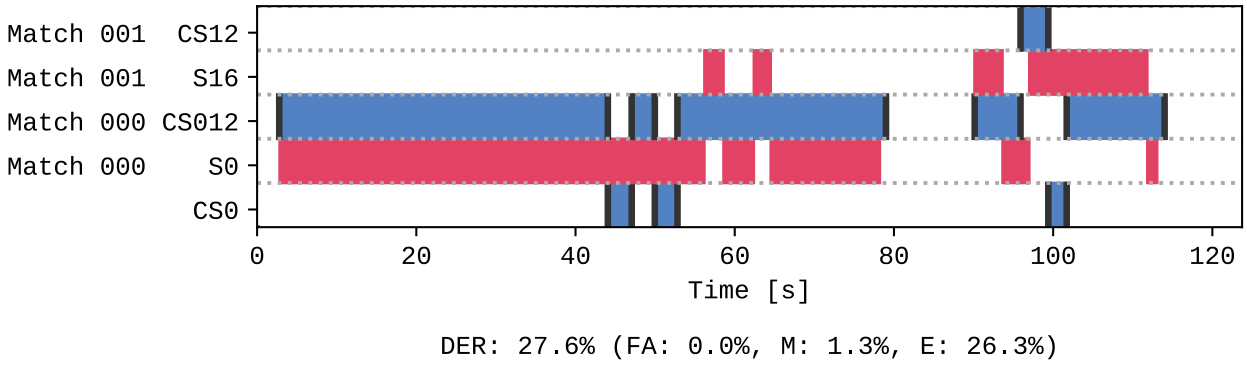


図 15　歌唱者 G，K，C による楽曲 K の歌唱者ダイアライゼーションの結果．本節の実験において最も DER が低かった歌声である．

第 7 章

歌唱者認識を利用した歌唱者ダイアライゼーションの実験

本章では，短時間での歌唱者認識を用いて歌唱者ダイアライゼーションを行った実験の概要について述べる．

7.1 用いた手法

歌唱者の組み合わせとそれぞれの歌唱者の音響モデルが既知である場合，歌唱者ダイアライゼーションを実現する手法として短時間窓で繰り返し歌唱者認識を行う手法が考えられる．本章では，2.2 節で述べた 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行う．この手法の概略を図 16 に示す．

本手法では短時間での歌唱者認識において前後の認識結果を参照しないため，歌唱者がごく短時間で次々と入れ替わるような不適切な推定結果が得られることが考えられる．そこで，歌唱者認識を行った後，前後の時刻での歌唱者認識の結果を利用して推定結果を平滑化する手法を用いる．ここでは，ある時刻 t での歌唱者の認識結果 $y(t)$ を次式のように平滑化して，修正された認識結果 $\hat{y}(t)$ を得る．

$$\hat{y}(t) = \mathrm{mode}\{y(t - n\tau), \dots, y(t), \dots, y(t + n\tau)\} \tag{51}$$

ここで，τ は短時間歌唱者認識を行う間隔，mode は最頻値を選ぶ関数である．このような平滑化手法は，windowed majority vote などと呼ばれる．

フレーム周期は 5 ms，i-vector を抽出するための音声の長さ（チャネル長）は 1 s（200 フレーム），i-vector を抽出する間隔は 100 ms とした．UBM の混合数は 2048 とし，i-vector の識別には PLDA を用いた．UBM および全変動行列の学習，i-vector の抽出および識別には MSR Identity Toolbox [73] を用いた．発話区間検出は信号のパワーにもとづき行った．平滑化では，$n = 10$（平滑化窓 2 秒）と $n = 20$（平滑化窓 4 秒）の 2 通りを適用した．

7.2 短時間での歌唱者認識実験

短時間歌唱者認識そのものの性能を調査するため，構築したデータセットに対して同じ条件で歌唱者認識を行った．ここではすべてソロの歌声に対して認識を行った．この結果を表 3 に示す．全体の正解率は 28.4% となった．歌唱者によって正解率が大きく異なっており，また歌唱者 C，G と誤って推定された例が多く見られた．

また，楽曲を通して最も多かったラベルを選ぶことによって，楽曲ごとに認識を行った結果を表 4 に示す．表 3 の場合と同様，歌唱者によって正解率が大きく異なり，また推定結果が特定の歌唱者に偏る傾向が見られた．

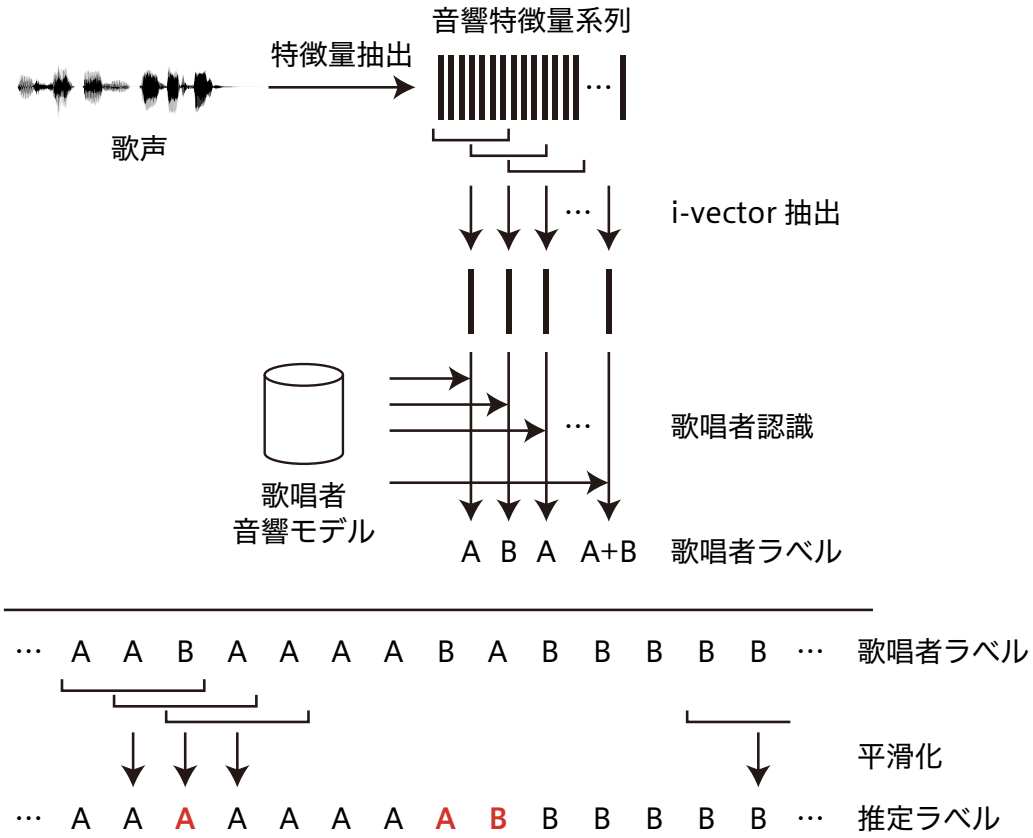


図 16　歌唱者認識を利用した歌唱者ダイアライゼーション手法の概略図.

7.3　短時間での歌唱者認識を利用した歌唱者ダイアライゼーション

楽曲 A，D–K，Lf，Lt の 11 曲について，各 10 通りのパート割りを用い，合計 110 曲に対して歌唱者認識にもとづく歌唱者ダイアライゼーションを行った．楽曲 A については，第 6 章で行った実験ではパラメータ決定に用いているが，本実験では楽曲 B，C を用いて音響モデルを構築したため，楽曲 A についても評価データとしてダイアライゼーションを行っている．この結果を表 5 に示す．平滑化により DER が大きく改善していることが確認された．また，楽曲 Lf および Lt の結果から，デュオありとデュオなしに大きな性能の差は見られなかった．一方，楽曲による性能の差が大きく，楽曲 A，B，L などと比較して楽曲 E，G，I は DER が高い結果となった．特に，楽曲 I は平滑化による大きな改善も見られなかった．

7.4　本節の実験結果の考察

歌唱者認識による歌唱者ダイアライゼーションは，第 6 章で述べた歌唱者の知識がない歌唱者ダイアライゼーションに比べ，平均で 17% 改善した．これは，背景知識の差による合理的な差であると捉えることができる．

7.2 節で述べた歌唱者認識の実験では，フレーム毎の認識率が 28% であり，13 人と少ない歌唱者から歌手同定していることを考えれば，認識自体の性能は低く歌手同定の性能としては不十分である．i-vector による認識では，1 秒程度の短時間の音声を用いて話者を認識することは困難であることが指摘されている [74]．そのため，短時間の音声でも十分に識別性能の得られる話者表現を検討する必要がある．

表 5 によれば，平滑化による DER への貢献が大きいと解釈される．実験の対象とした 110 曲のうち 108 曲

表 3　全歌唱者のソロの歌声に対する，i-vector ごとの歌唱者認識の結果の混同行列．各行が認識対象とした音声の正解歌唱者を，各列が得られた認識結果を表す．太字で示した対角成分が正しく認識できた数を示す．

		推定歌唱者													
		A	B	C	D	E	F	G	H	I	J	K	L	M	正解率
正解歌唱者	A	**132**	21	110	73	13	28	314	116	30	125	117	61	42	11.1%
	B	1	**506**	160	17	0	126	28	34	274	5	4	17	5	43.0%
	C	94	19	**281**	31	2	9	306	68	48	34	83	29	10	27.7%
	D	7	34	35	**31**	2	38	43	5	27	5	27	28	4	10.8%
	E	183	3	87	29	**46**	9	161	338	8	264	133	5	20	3.6%
	F	2	132	89	113	0	**403**	77	39	64	15	25	200	22	34.1%
	G	42	101	273	35	1	37	**620**	24	26	16	24	89	4	48.0%
	H	210	3	64	14	25	3	197	**334**	7	221	76	10	13	28.3%
	I	0	126	153	12	0	41	11	12	**740**	0	3	5	1	67.0%
	J	159	7	113	47	30	14	185	156	12	**237**	156	20	47	20.0%
	K	66	28	102	93	21	35	292	86	21	56	**331**	134	28	25.6%
	L	6	50	104	24	1	29	200	10	25	8	14	**177**	1	27.3%
	M	33	30	135	38	11	30	146	125	44	21	75	49	**16**	2.1%
														平均	28.4%

は平滑化によって DER が改善した．平滑化によってダイアライゼーションの結果が改善した例を図 17 に示す．平滑化前の結果はごく短時間のセグメントが非常に多く，平滑化によりこれが抑制されていることが確認できる．ダイアライゼーションの後処理などで用いられている HMM などを用いた精巧な平滑化手法を用いることで，さらにダイアライゼーションの性能を向上できると考えられる．

本節で述べた実験では，手法 1 と異なり，楽曲 I がひときわ DER が高い．これは，楽曲 I の歌声にハードリミットやディレイなどのエフェクトが強くかかっているため話者表現を適切に抽出できず，歌唱者認識の性能が落ちたためであると考えられる．一方，第 6 章での歌唱者認識を行わない手法では，エフェクトによって音響特徴量が変化しても他の楽曲と同様にダイアライゼーションを行うことができ，歪みによる影響を受けにくいと考えられる．

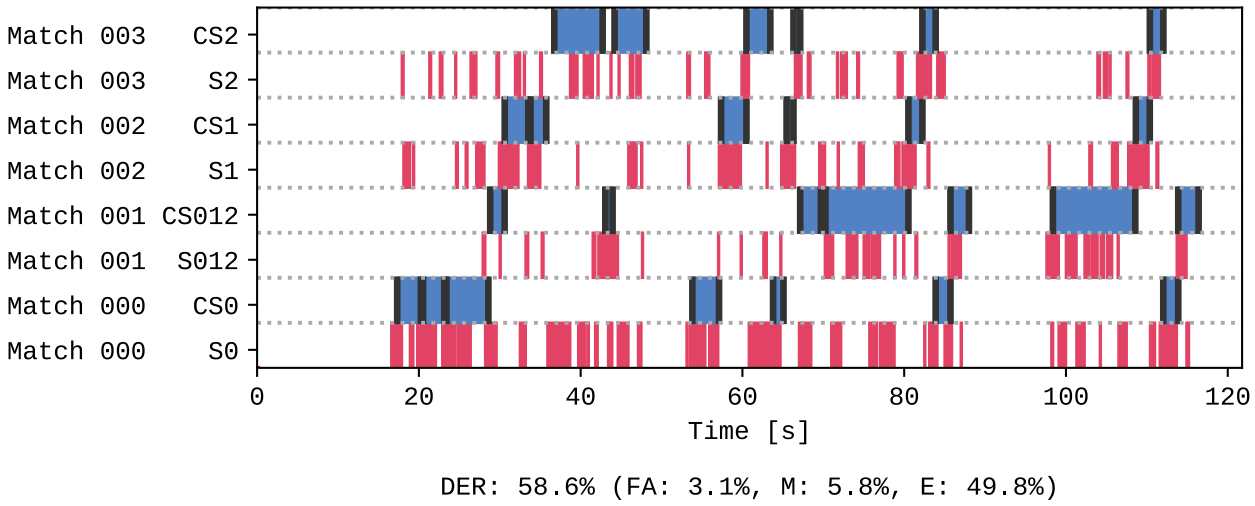
また第 6 章の結果と比較し，本節の実験結果では楽曲内での標準偏差が大きい．これは，歌唱者の組み合わせにより識別の性能が大きく異なることを示唆している．この例として，楽曲 D 中で DER が最も低かった結果と最も高かった結果を図 18 に示す．DER の最も低かった歌唱者 B，H，E の組み合わせでは error が 4.3% と低く，一方 DER の最も高かった歌唱者 J，B，K の組み合わせが error が 59.8% と非常に高い．DER を計算する際には，推定された歌唱者の情報を用いずに推定クラスタと正解クラスタの対応付けを改めて取るため，この対応付けが歌唱者認識の結果と異なる場合がある．歌唱者 J，B，K の組み合わせではこの対応付けが正しく取れておらず，歌唱者認識自体の性能は DER で示されるよりもさらに低いと言える．とくにこの組み合わせでは歌唱者 B のソロパートが 8.4 秒と短いにも関わらず，歌唱者認識の結果では歌唱区間の 98.9 秒間のうち 64.9 秒間で認識されている．この歌唱者の組み合わせにおいては歌唱者 B が誤って推定されやすい傾向があり，これにより大きく誤った歌唱者認識結果が得られたと考えられる．

表 4　全歌唱者のソロの歌声に対する，楽曲ごとの歌唱者認識の結果の混同行列．各行が認識対象とした音声の正解歌唱者を，各列が得られた認識結果を表す．太字で示した対角成分が正しく認識できた数を示す．すべての歌唱者がすべての楽曲を歌唱しているわけではないため，歌唱者によってサンプル数が異なることに注意する．

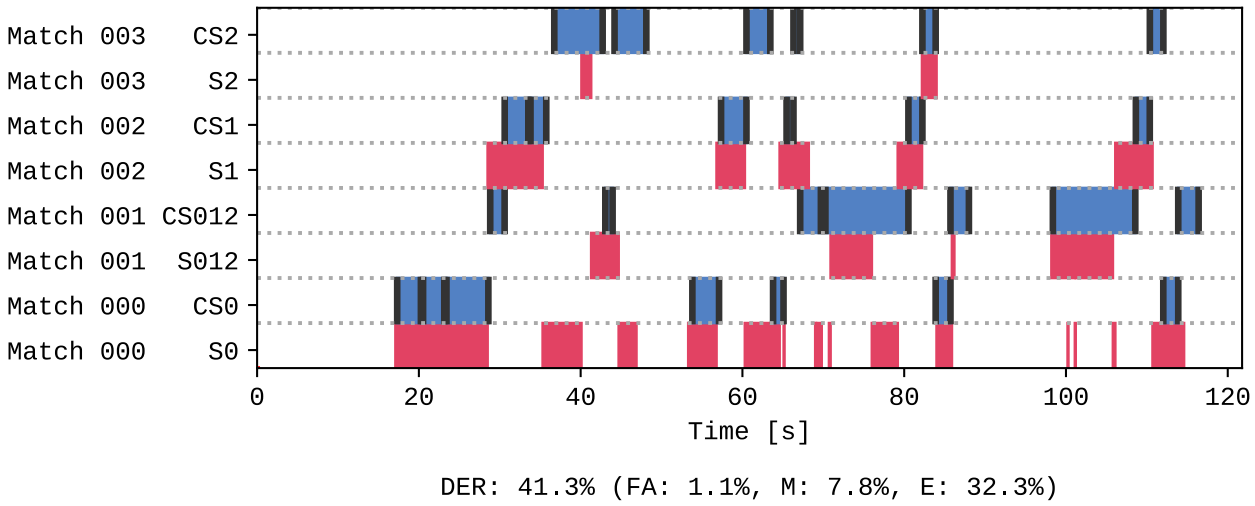
		推定歌唱者													
		A	B	C	D	E	F	G	H	I	J	K	L	M	正解率
正解歌唱者	A	**1**	0	1	0	0	0	7	0	0	3	1	0	0	7.7%
	B	0	**9**	1	0	0	1	0	0	2	0	0	0	0	69.2%
	C	1	0	**5**	0	0	0	5	0	0	0	0	0	0	45.5%
	D	0	0	1	**0**	0	1	1	0	0	0	0	0	0	0%
	E	1	0	0	0	**0**	0	2	6	0	4	1	0	0	0%
	F	0	0	0	0	0	**9**	0	1	0	0	0	3	0	69.2%
	G	0	0	2	0	0	0	**12**	0	0	0	0	0	0	85.7%
	H	5	0	0	0	0	0	1	**6**	0	1	0	0	0	46.2%
	I	0	0	0	0	0	0	0	0	**12**	0	0	0	0	100%
	J	3	0	1	0	0	0	1	2	0	**5**	1	0	0	38.5%
	K	0	0	0	0	0	0	6	1	0	0	**6**	1	0	42.9%
	L	0	0	2	0	0	0	3	0	0	0	0	**2**	0	28.6%
	M	0	0	3	0	0	0	2	1	1	0	1	0	**0**	0%
														平均	45.3%

表 5　歌唱者認識を利用したダイアライゼーションの実験結果．値は DER [%].

楽曲	パート割り	平滑化なし		平滑化 2 秒		平滑化 4 秒	
		平均	標準偏差	平均	標準偏差	平均	標準偏差
A	デュオなし	45.0	8.6	34.1	9.8	29.9	8.5
D		45.7	14.0	38.7	15.1	34.6	8.4
E		58.2	5.8	52.8	8.6	50.2	8.7
F		50.2	6.0	45.3	7.2	42.2	8.5
G		59.1	7.5	54.0	8.7	52.0	10.5
Lf		47.3	9.7	38.0	10.3	32.1	11.6
H	デュオあり	60.3	9.0	54.0	10.2	48.8	10.4
I		63.8	6.3	59.7	7.1	60.2	6.5
J		43.3	8.8	35.1	9.5	34.0	9.7
K		59.4	8.4	51.1	11.4	46.0	13.6
Lt		51.8	7.1	42.2	9.7	37.1	11.2
デュオなし		50.9	10.4	43.8	12.4	40.2	13.5
デュオあり		55.7	10.6	48.4	12.8	45.2	13.8
全体		53.1	10.7	45.9	12.7	42.5	13.8

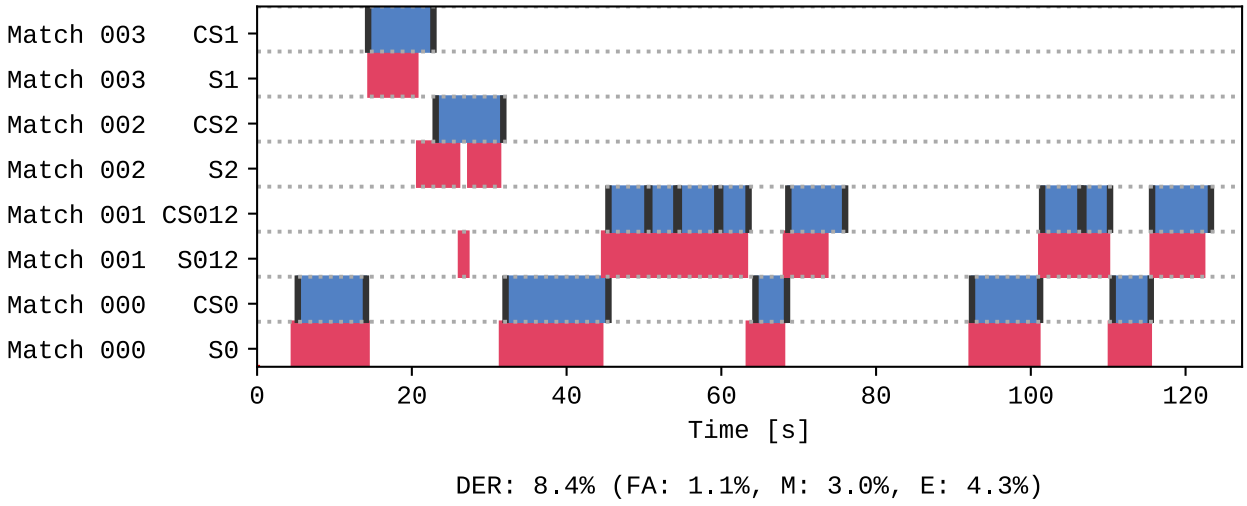


(a) 平滑化前のダイアライゼーションの結果.

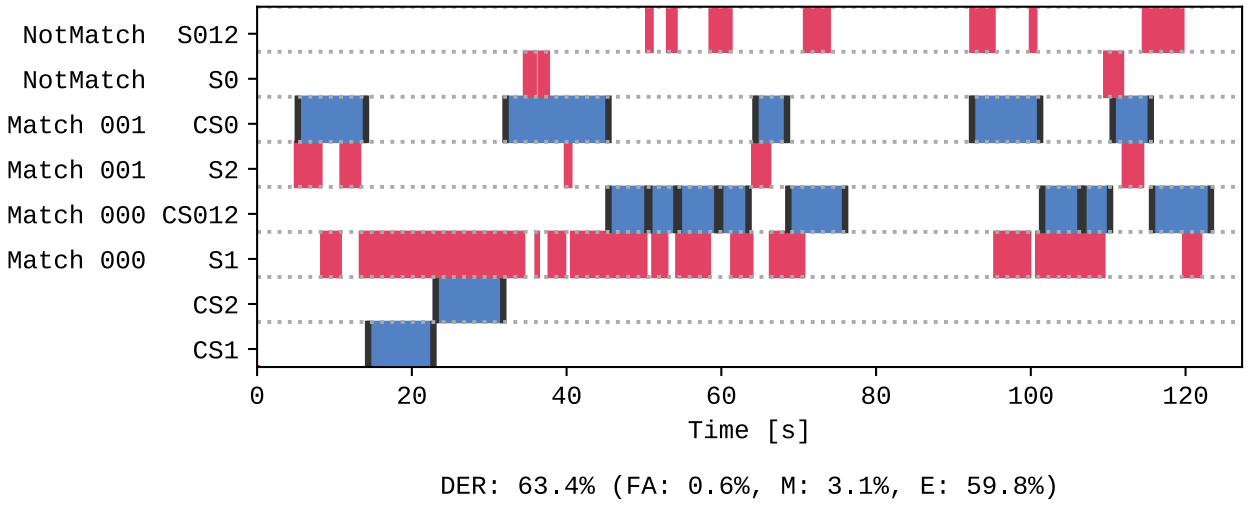


(b) 平滑化後のダイアライゼーションの結果.

図 17　平滑化によってダイアライゼーションの結果が改善した例．図は歌唱者 G，B，D の歌唱した楽曲 E の結果である．



(a) 歌唱者 B，H，E による歌声に対するダイアライゼーションの結果.



(b) 歌唱者 J，B，K による歌声に対するダイアライゼーションの結果.

図 18　同一楽曲であっても歌唱者の組み合わせによってダイアライゼーションの結果が大きく異なる例．ともに楽曲 D の結果である．推定ラベルの数字と正解ラベルの数字は対応している．したがって，図 18b は DER を計算する際の対応付けと実際の歌唱者認識結果に相違がある．

第 8 章

2 手法の組み合わせによる歌唱者ダイアライゼーションの実験

本章では，第 6 章で用いた mBIC によるセグメンテーションと第 7 章で用いた歌唱者認識を組み合わせた手法により歌唱者ダイアライゼーションを行った実験について述べる．

8.1 用いた手法

第 6 章で行った実験では，楽曲によっては mBIC のハイパーパラメータが不適切で，クラスタ数が過多になる問題があった．そこで，本章で述べる実験では標準的なダイアライゼーション手法を用いてセグメンテーションのみ行った．そして，得られた各セグメントに対して第 7 章で構築した歌唱者の音響モデルを用いて歌唱者認識を行った．これによって，クラスタリングにおいて凝集が不十分である問題，および短時間の歌唱者認識で著しくセグメントが多くなる問題に対処できることが考えられる．

8.2 mBIC によるセグメンテーションと歌唱者認識を組み合わせた歌唱者ダイアライゼーション

第 6 章，第 7 章と同様に，楽曲 A，D–K，Lf，Lt の 11 曲について，各 10 通りのパート割りを用い，合計 110 曲に対してダイアライゼーションを行った．セグメンテーションには第 6 章の実験で得られたセグメンテーション結果を用い，歌唱者認識には第 7 章で構築した音響モデルを用いた．楽曲 A についてはセグメンテーションのための mBIC のパラメータの設定に用いているため，評価には用いることができないことに留意する．

この結果を表 6 に示す．平均の DER は第 6 章での実験結果より低い結果が得られた．一方，第 7 章で述べた短時間の歌唱者認識を利用した手法に比べ高い結果となった．

8.3 本節の実験結果の考察

各セグメントに対する歌唱者認識が十分ではなく，認識の誤りによって DER が高くなったと考えられる．多数の結果から信頼されるラベルを選んでいた第 7 章の結果と比較して，セグメントごとに認識を行ったため認識の誤りが反映されやすくなったと推測される．また，歌唱者の音響モデルには第 7 章と同じ 1 秒の音声から抽出した i-vector を用いている一方，評価に用いた i-vector は長さの異なるセグメントから抽出している．学習と認識の音声のチャネル長が異なったため，セグメントに対して適切に評価を行うことができなかったと考えられる．

一方，本節で用いた枠組みではクラスタ数が既知であったため，第 6 章で述べた手法でクラスタリングが十分に行われなかったことによる性能の低下は見られなかった．しかし，これは事前情報による合理的な差

表 6　mBIC によるセグメンテーションと歌唱者認識を組み合わせた手法によるダイアライゼーションの実験結果．値は DER [%]．楽曲 A はハイパーパラメータの決定に用いた学習データであるため，全体の値には含まれていない．

楽曲	パート割り	平均	標準偏差
A	デュオなし	41.3	8.4
D		44.5	11.1
E		52.9	7.4
F		45.6	9.8
G		48.4	9.5
Lf		45.9	10.6
H	デュオあり	51.0	6.9
I		62.1	2.3
J		49.1	4.5
K		27.2	11.3
Lt		44.2	7.2
デュオなし		47.5	9.9
デュオあり		46.7	13.4
全体		47.1	11.7

であり，第 7 章と比較して DER が高いことを考慮すると，本節の手法によってダイアライゼーションの性能が向上したとは言えない．

標準的なダイアライゼーション手法では，セグメンテーションの結果が最終的なダイアライゼーションの性能に大きく影響する．本章で用いた手法でもセグメンテーションによる影響を大きく受けるため，歌唱者認識を用いれば認識できるような短時間での歌唱者の切り替わりを認識できない場合が見られた．この例を図 19 に示す．

8.4　3 手法の比較

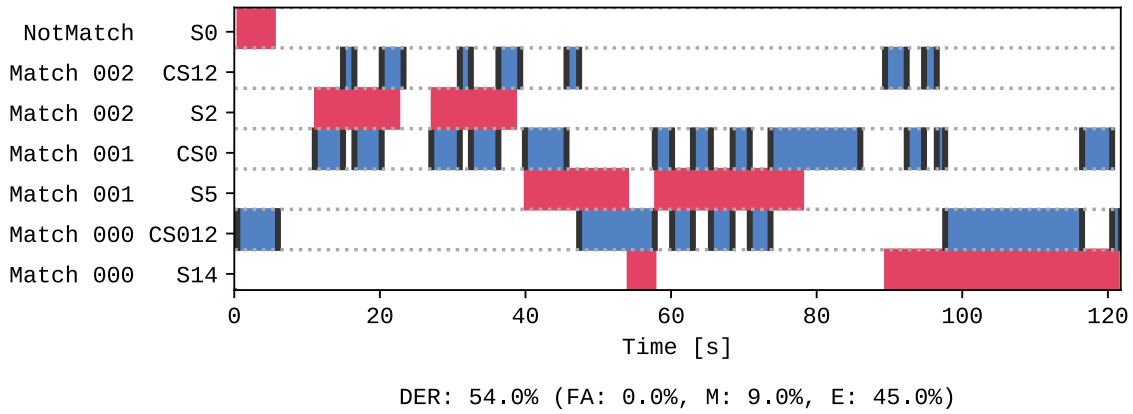
3 手法を用いて行ったダイアライゼーションの結果を表 7 に整理する．

多くの楽曲で手法 2 が最も低い DER を示している．楽曲 E，G についても歌唱者の組み合わせ間でのばらつきを考えれば大きな差とは言えない．一方，楽曲 K については手法 3 が最も DER の低い結果を与えた．手法 3 ではクラスタ数既知の条件でセグメント単位の認識を行っているため，他の楽曲に比べセグメント数が少なく各セグメントが長い楽曲 K では有利であったと考えられる．

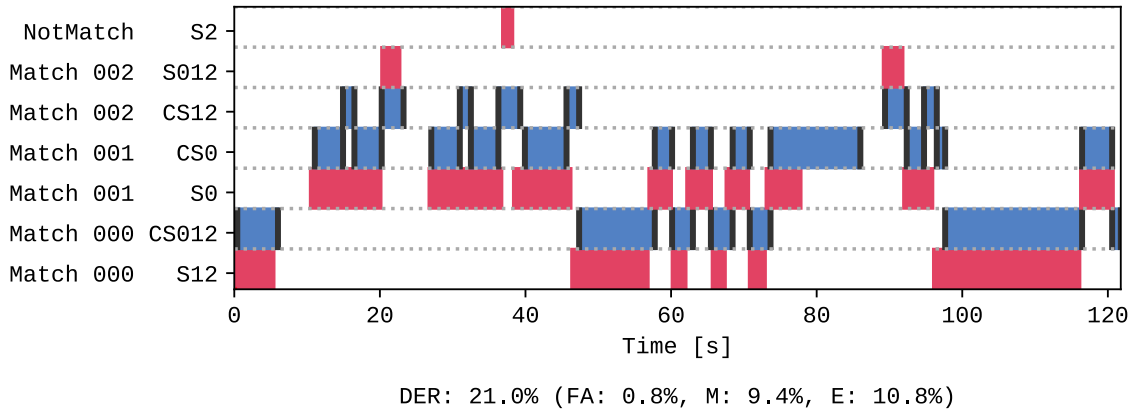
手法 2 と手法 3 では，楽曲間の性能の相違がおよそ同様の傾向を示しており，歌唱者認識の性能が楽曲によって異なることが推測される．一方・手法 1 ではクラスタリングが不十分であることから，どのような楽曲で手法が優れているかを十分に把握することができない．しかし，楽曲 I のような歌唱者認識が難しい音声に対しても手法 1 では他の楽曲と同等の性能が得られることを踏まえれば，手法 2，手法 3 のように既知の音響モデルのみを用いるのではなく，手法 1 のように音響モデルを対象音声から構築する手法を併用することで，さらに高い性能のダイアライゼーションが可能になると推察される．

表 7　3 手法によるダイアライゼーションの実験結果の比較．値は DER [%]．楽曲 A の結果についても示しているが，全体の値はすべて楽曲 A を除いた値を示している．

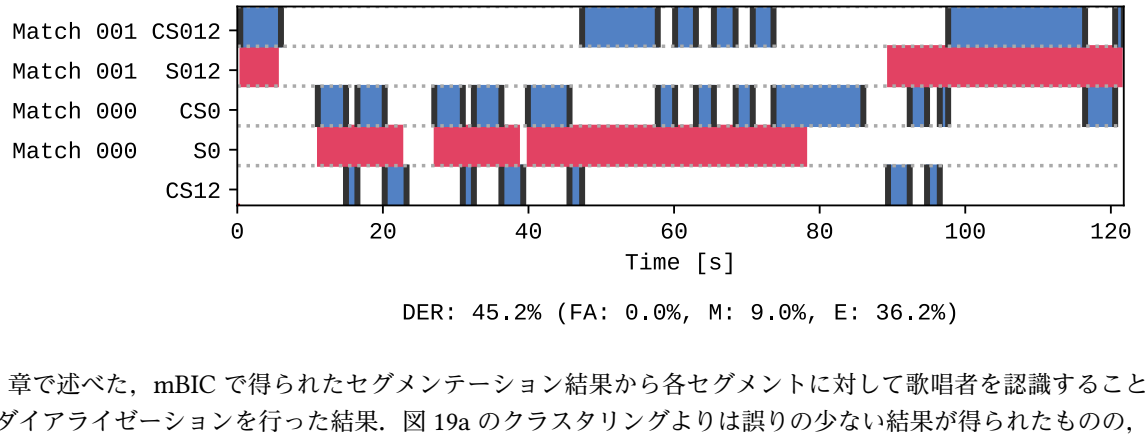
楽曲	パート割り	手法 1		手法 2		手法 3	
		平均	標準偏差	平均	標準偏差	平均	標準偏差
A	デュオなし	35.7	2.0	29.9	8.5	41.3	8.4
D		71.0	6.0	34.6	8.4	44.5	11.1
E		49.1	1.8	50.2	8.7	52.9	7.4
F		67.2	5.5	42.2	8.5	45.6	9.8
G		63.6	4.9	52.0	10.5	48.4	9.5
Lf		69.5	5.2	32.1	11.6	45.9	10.6
H	デュオあり	60.7	3.8	48.8	10.4	51.0	6.9
I		65.2	3.2	60.2	6.5	62.1	2.3
J		57.0	4.0	34.0	9.7	49.1	4.5
K		45.2	11.7	46.0	13.6	27.2	11.3
Lt		62.3	5.0	37.1	11.2	44.2	7.2
デュオなし		64.1	9.3	42.2	13.4	47.5	9.9
デュオあり		58.1	9.3	45.2	13.8	46.7	13.4
全体		61.1	9.7	43.7	13.6	47.1	11.7



(a) 第 6 章で述べた，標準的な歌唱者ダイアライゼーション手法を用いた結果．楽曲 J ではクラスタリングでの凝集が不十分な問題は見られないが，セグメンテーションが不十分で短時間の歌唱者の切り替わりを反映することができていない．



(b) 第 7 章で述べた，短時間ごとの歌唱者認識によって歌唱者ダイアライゼーションを行った結果．存在しない歌唱者 K のソロパートを認識しているなどの誤りはあるが，他の 2 手法と比較して細かな歌唱者の切り替わりが反映されており，DER も最も低い．



(c) 第 8 章で述べた，mBIC で得られたセグメンテーション結果から各セグメントに対して歌唱者を認識することにより歌唱者ダイアライゼーションを行った結果．図 19a のクラスタリングよりは誤りの少ない結果が得られたものの，セグメンテーションが不十分であることによる影響を大きく受けている．

図 19　同楽曲に対する手法によるダイアライゼーション結果の比較．楽曲は，歌唱者 B，H，K の歌唱した楽曲 J である．

第 9 章

同時歌唱者数推定

本章では，歌声から歌唱者の認識を行うことなく，同時に何人が歌っているかのみを推定する手法について議論する．同時に歌唱している人数を推定できれば，候補となる歌唱者の組み合わせを絞り込むことが可能になり，歌唱者ダイアライゼーション全体の性能に貢献できると考えられる．本章では，歌唱者ダイアライゼーションと組み合わせず，同時に発話している人数を推定することのみに着目して議論する．

9.1　同時発話者数推定の現状

同時に発話している人数やその中で話している話者の聴取は分散的聴取と呼ばれ，話し声について広く研究がなされてきた [75,76]．これらの研究によれば，10 名程度までの同時発話においては，このうち 2 名程度しか分離できない傾向がある．一方，4 人程度以上の同時発話においては，実際の発話者数よりも少なく聴取するものの，聴取した話者数と実際の話者数に相関が見られ，話者を分離することなく話者数を聴取することが可能であるとされている．

同時に発話している人数を，単一チャンネルの音声から推定する手法も検討されている [77]．しかしフレーム毎で人数を推定する問題は，1 人と 2 人の識別であっても困難であることが指摘されている．ダイアライゼーションの文脈でも，種々の特徴量を用いて話者のオーバーラップを推定する手法が試みられている [78,79]．

フレーム毎での人数推定は話者が既知である場合にのみ有効であることが示唆されている [77]．これは，同時発声の音響モデルが単一話者の音響モデルと同様に構築されているだけに過ぎず，本質的には単一話者か複数話者かを区別できていない．同時に発話している人数を推定するためには，既存の音響特徴量に固執せず，音声に対して多角的な分析を行う必要がある．

9.2　聴取実験による話者数の推定

複数の話者による話し声については，聴取実験によって聴覚による分離の性能が議論されてきた．一方，歌声に対して議論されたことはなく，特に本研究で扱うような発話内容も音高も同じ音声に対しては議論されたことがない．そこで予備実験として，聴覚によってどの程度歌声の人数を推定できるかを調査する実験を行った．

9.2.1　実験条件

各被験者は，0.5 秒間の歌声を聴取して，その音声が 1 人による歌声か 2 人による歌声かを分類した．本研究で整備したデータセットの歌声を用い，パワーにもとづいて無音でない区間を生成した．予備実験として行ったため，各被験者は任意の問題数について回答することを許容した．また，各被験者についてデータセッ

トとして利用したゲームである『アイドルマスター』についての知識があるかを質問した．被験者は 12 名とし，このうち 6 名以上が回答した 42 音声について分析の対象とした．

9.2.2　実験結果と考察

『アイドルマスター』の知識の有無にかかわらず，85% 以上の割合で正解することができた．提示した音声のうち，1 人の歌声なのにも関わらずリバーブの影響で 2 人の歌声に聞こえた音声，リバーブやブレスのみしか聞こえず 2 人の歌声を知覚できなかった音声，2 人の歌声の基本周波数がほとんど一致しており 1 人に知覚した音声，の 3 種類の音声について特に誤りの割合が高かった．実験に用いた音声はあくまでパワーにもとづき抽出しているため，0.5 秒間に有声区間が存在しない音声も含まれていた．これを考慮すると，被験者は 0.5 秒の音声から基本周波数の異なりを知覚することで複数の歌唱者を認知していたと考察できる．

9.3　同時歌唱音声に関する考察

聴取実験の結果から，基本周波数の違いによって同時に歌唱している人数を知覚していることが推測される．そこで，合成音声を用いて，まったく同じ基本周波数の歌声を 2 つ重ねた場合と，一方の音声を 20 セント（1/60 オクターブ）高くした歌声を用いて重ねた場合の 2 つの音声に対して同一の分析を行い，分析結果の違いを考察する．どちらも 2 つの音声を重ねているが，基本周波数がまったく同じ場合には 1 人による歌声に聞こえ，異なる場合には 2 人による歌声に聞こえることを聴取により確認した．ここでは，歌唱者 C，K が歌唱した楽曲「フタリの記憶」の最後の歌詞「いつまでも忘れないでいるよ ずっとずっと空で見守っているよ」の部分を切り出して実験を行った．この楽曲は歌唱者ダイアライゼーションのデータセットとしては利用していないが，テンポが遅く，またリバーブなどのエフェクトがほとんど適用されていないため，詳細な検討に適している．

分析合成には WORLD [80]（D4C edition [81]）を用いた．基準の基本周波数は歌唱者 C のものとし，歌唱者 K の歌声は歌唱者 C の基本周波数と歌唱者 K のスペクトル包絡から合成した．分析を容易にするため非周期成分はゼロとして再合成を行ったが，ボコーダの特性上無声子音などがすべて取り除かれた音声が合成されるわけではない．

9.3.1　スペクトル

ある時刻におけるスペクトルを図 20 に示す．基本周波数が異なるため調波構造に違いが生じているが，スペクトル包絡のような大域的な形状はほぼ同一である．MFCC などのスペクトル包絡に着目した音響特徴量では，同時に歌唱している人数の推定が難しいことを示している．

9.3.2　音声波形

短時間での音声波形の比較を行う．母音を発している定常的な区間と音素が移り変わる非定常的な区間の 2 つについて波形を切り出した．これらの波形を図 21 に示す．図 21a に示す定常的な音声では，基本周波数が異なる音声に比べて同一の音声はより周期的な波形であることが観察される．一方，図 21b に示すような非定常的な区間ではどちらの場合にも波形が周期的でない．信号波形から識別を行うには，その信号波形による識別の信頼性を評価する必要がある．

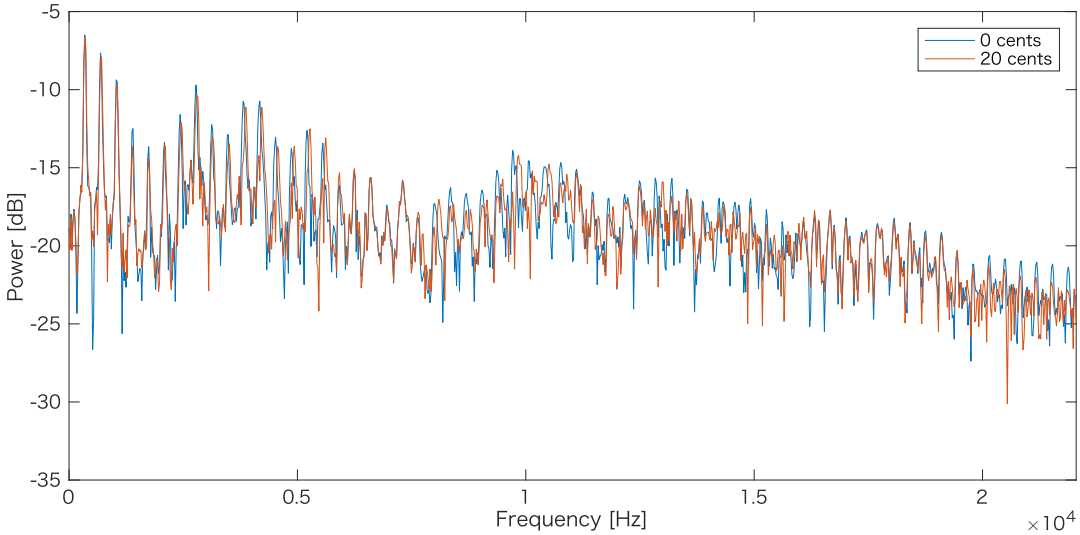


図 20　同時歌唱音声のスペクトルの比較．歌詞のうち「見守っているよ」の「よ」の母音部分から抽出した．青のスペクトルは 2 人の歌声が同一の基本周波数とした場合，オレンジ色のスペクトルは異なる場合のスペクトルである．

9.3.3　自己相関関数

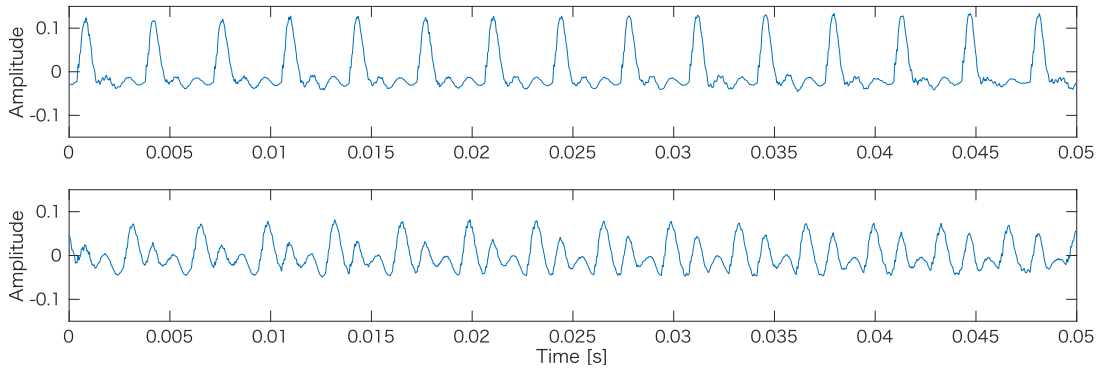
音声の周期性を特徴量として抽出する手法として，自己相関関数が挙げられる．周期的な信号に対する自己相関関数は，基本周波数の逆数を周期とした周期的な関数となる．一方，周期的でない音声に対する自己相関関数では，周期性が得られにくく減衰しやすい．

図 21 に示した同時歌唱音声の波形から自己相関関数を計算した結果を図 22 に示す．音声が定常的な場合，基本周波数が異なる音声では減衰が早く後の周期ほど形状が大きく変化していることが観察される．一方，音素が遷移している音声から得た自己相関関数は，どちらも減衰が早く区別できる違いを見出すことは難しい．これらの性質は波形の場合と同様であり，自己相関関数の場合も音声の定常性を判断して用いる必要がある．

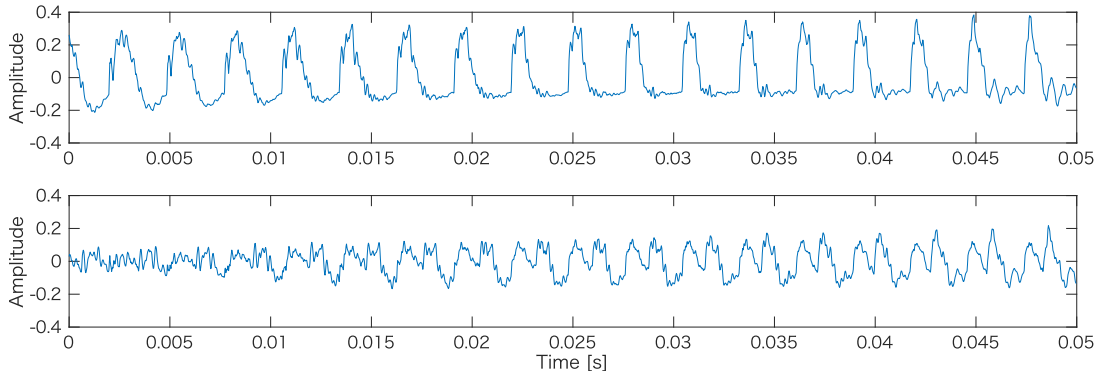
9.3.4　ピッチマーク

長い時間幅から波形の周期性を評価する手法として，ピッチマークを取ることが考えられる．ピッチマークとは，波形の各周期に対して割り当てられる点で，ピッチマーク同士の時間差によって基本周波数を推定することができる．ここでは SPTK[*5] に実装されている REAPER（Robust Epoch And Pitch EstimatoR）[*6] を用いたピッチマーク検出を利用した．ピッチマークの推定結果から基本周波数を計算した結果を図 23 に示す．本来の歌声の基本周波数に対して，基本周波数が異なる音声では推定誤りが多く，基本周波数が同一の音声では同様の誤りが少ない．ピッチマークの結果の誤りを評価することができれば，同時歌唱音声の識別を行うことができると考えられる．

[*5] http://sp-tk.sourceforge.net/
[*6] https://github.com/google/REAPER



(a) 母音が定常的に発されている時間の波形．歌詞のうち「見守っているよ」の「よ」の母音部分から抽出した．



(b) 音素が遷移している時間の波形．歌詞のうち「見守っているよ」の「るよ」が遷移する部分から抽出した．

図 21　同時歌唱音声の波形の比較．ともに，上段が基本周波数が同一の場合，下段が基本周波数が異なる場合の波形である．

9.4　実音声に対する歌唱者数の推定

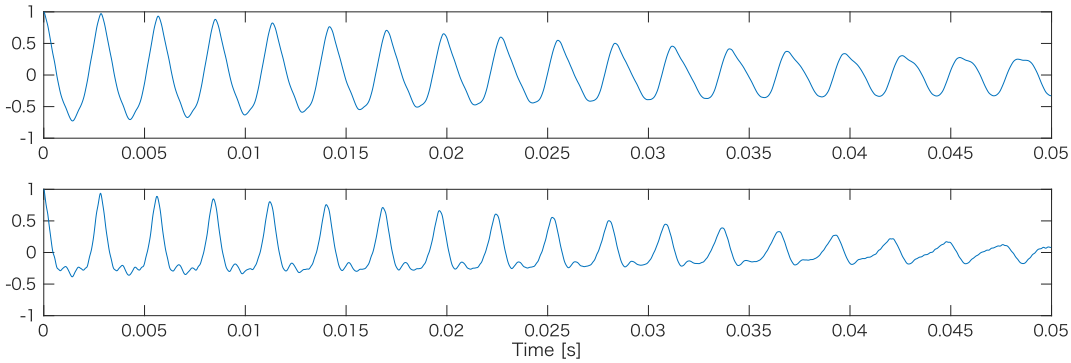
本節では，前節で述べた基本周波数の異なりによる音響的な変化をもとに，歌声に対して実際にソロ音声か 2 人の歌唱者によるユニゾン音声かを識別する実験を行った．

9.4.1　実験条件

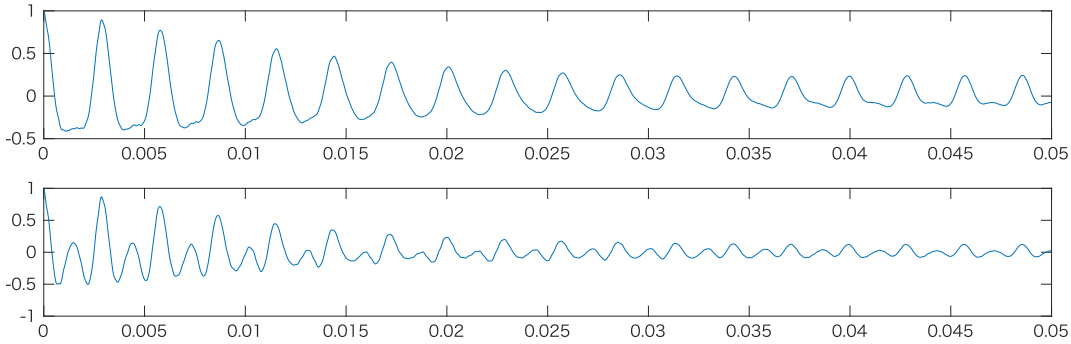
音声のサンプリング周波数は，ダウンサンプリングを行わず 44 100 Hz とした．学習には楽曲 A，D，E，G–J，L–N の 10 曲の音声を用い，評価には楽曲 O，P の 2 曲を用いた．楽曲 O では 12 通りずつ，楽曲 P では 4 通りずつ用いた．評価に用いた楽曲のうち，楽曲 O はどの歌唱者も学習データに含まれており，楽曲 P はどの歌唱者も学習データに含まれていない．本節では，この 2 条件の歌唱者をそれぞれ in，out と示す．フレーム長は 1024 サンプル，フレームシフトは 10 ms とした．

9.4.2　i-vector を用いた同時歌唱者数推定

i-vector による話者認識を用いて，人数のみを推定した．すなわち，話者で識別を行うのではなく，i-vector を用いて 1 人か 2 人かのみを推定した．ここでは 16 次 MFCC と，その Δ 特徴量および ΔΔ 特徴量を用いた．i-vector は 1 秒の音声から抽出した．識別にはコサイン類似度および SVM を用いた．



(a) 母音が定常的に発されている時間の波形から得た自己相関関数．波形は図 21a に対応する．



(b) 音素が遷移している時間の波形から得た自己相関関数．波形は図 21b に対応する．

図 22　同時歌唱音声から計算した自己相関関数の比較．ともに，上段が基本周波数が同一の場合，下段が基本周波数が異なる場合の信号から得た自己相関関数である．

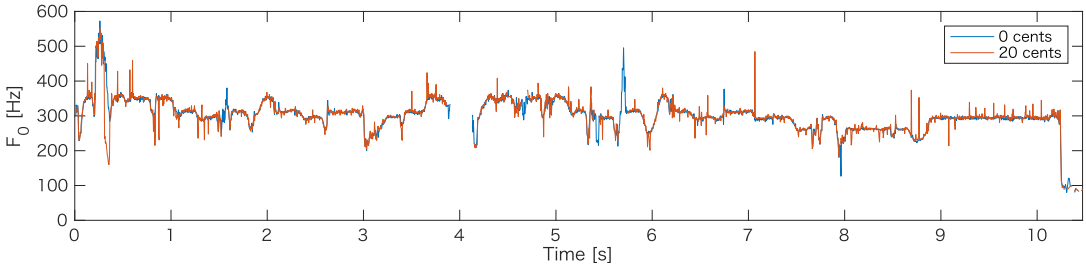


図 23　同時歌唱音声に対するピッチマークの比較．ここではピッチマークから推定した基本周波数を示している．青色の系列は 2 人の歌声の基本周波数を同一とした場合，オレンジ色の系列は異なる場合に推定された基本周波数である．ここでは 80 Hz 未満の基本周波数を無声部分として扱い空白として示している．

この結果を表 8 に示す．これによれば，学習データにある歌唱者の歌声のほうが正解率が高いものの，ともに 75% 程度の割合で音声を分類することが可能であった．MFCC のようなスペクトル包絡に着目した音響特徴量では同時歌唱者数の推定を行うことが難しい．一方，i-vector のように多数のフレームから MFCC を抽出してその統計値に対して評価する場合，各フレームの MFCC からは得られない変調スペクトルなどの系列特徴が現れ，同時歌唱の識別を可能にしていると考えられる．しかし，どのような系列特徴によって識別が可能になったかは明らかでなく，これを解明すればさらに性能の高い同時歌唱者数の推定を行うことができると考えられる．また，判別法によって正解率が大きく異なるため，複数の判別法を組み合わせることで

表 8　i-vector を用いて同時に歌唱している人数のみを推定した結果．Solo および Unison は，正解の歌唱者数がそれぞれ 1 人と 2 人の場合の正解率を表し，平均は全体の正解率を示す．

楽曲	判別法	Solo	Unison	平均
O（in）	コサイン類似度	78.1%	75.8%	77.0%
P（out）	コサイン類似度	72.1%	77.8%	74.9%
O（in）	SVM	89.1%	67.6%	78.4%
P（out）	SVM	87.5%	60.6%	74.1%

表 9　自己相関関数の周期性を評価する手法で同時に歌唱している人数を推定した結果．Solo および Unison は，正解の歌唱者数がそれぞれ 1 人と 2 人の場合の正解率を表し，平均は全体の正解率を示す．

楽曲	Solo	Unison	平均
O（in）	78.6%	78.0%	78.3%
P（out）	80.3%	96.7%	88.5%

より高い性能の推定が可能になると推測される．

9.4.3　自己相関関数の周期性評価による同時歌唱者数推定

9.3.3 節で述べた自己相関関数の周期性を，1 周期ごとに切り出すことで評価した．各フレームのコストは，自己相関関数の 1 周期目に対する，2 周期目以降のコサイン距離の和とした．コサイン距離を用いているため，減衰の大きさについては評価していない．計算対象とする周期は 8 周期とし，フレーム長は 1 周期の長さに応じて 8 周期が得られるよう可変とした．1 周期のサンプル数は周期によって異なる場合があるため，自己相関関数を切り出したのち同じサンプル数になるよう補間を行った．また，1 周期の長さから基本周波数を計算し，80 Hz 以上 800 Hz 以下のフレームのみ選択して学習および評価を行った．判別法には SVM を用いた．

結果を表 9 に示す．歌唱者が訓練データに含まれているかどうかに関わらず，i-vector を用いた手法よりも高い推定性能が得られた．一方，歌唱者が訓練データに含まれている場合と含まれていない場合で推定性能に大きな差があり，楽曲によって推定性能が大きく異なることを示唆している．

第 10 章

結論

10.1　本研究の成果

本論文では，パート割り構造を持つような複数の歌唱者による楽曲に対し，誰がいつ歌唱しているかを推定する歌唱者ダイアライゼーションについて議論した．本論文では，実音声を直接用いず歌声のみを抽出した音声を用いることで，問題を明確にし基礎的な検討を行った．また，このような音声を用いることで，評価において適切な正解ラベルを与えることを可能にした．

本論文では，次の 3 つの手法を用いて歌唱者ダイアライゼーションを試みた．

1. よく知られた歌唱者の音響モデルを用いない手法
2. 短時間ごとに歌唱者認識を行う手法
3. 上の 2 手法を組み合わせた手法

手法 1 では，歌唱者の組み合わせに対する変動が少ないが，楽曲ごとに適切なハイパーパラメータが異なり，パラメータの選択によっては性能が大きく低下することが確認された．手法 2 では，歌唱者の組み合わせやエフェクトなどの環境に影響を受けやすい一方，多くの楽曲で手法 1 よりも高い性能が得られた．手法 3 では，両手法の欠点を補う手法として提案したが，単純な組み合わせでは各々の手法の問題点によって期待した性能が得られないことを確認した．

さらに本論文では，歌唱者ダイアライゼーションにおいて声の重なりを検出するため，音声の歌唱者が 1 人か 2 人かを推定する問題に対しても基礎的な検討を行った．従来はフレーム毎のスペクトルやケプストラム領域の音響特徴量を用いて推定されていたが，本論文では波形や自己相関関数などさらに信号に近い特徴量を観察し，それにもとづき推定を行った．実験から，同時歌唱者の推定においては歌唱者ごとの基本周波数の異なりを検出することが必要であり，そのためには従来の音響特徴量よりも長い時間の信号から特徴量を抽出する必要があることが示された．また，自己相関関数の形状を評価することで，従来用いられていた MFCC に比べ高い精度で同時歌唱の識別を行えることが示された．

10.2　今後の展望

本論文の歌唱者ダイアライゼーションでは，重なり合った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，さらに歌唱者が増え歌唱者の組み合わせが増えれば，考慮する歌唱者の数に対して指数関数的に増加してしまう．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，同数の単一話者から話者認識を行うよりも困難になり，推定精度が大きく下がることが考えられる．そのため，本論文で検討した同時歌唱者数の推定をさらに多人数に対しても適用できるよう拡張し，歌唱者ダイアライゼーションと組み合わせる必要がある．

また，本論文では複数人が歌唱している音響モデルを，実際に波形の加算によって得た音声から構築していた．一方，この構築手法には同一の楽曲を歌唱した歌声が必要であり，適用できる歌唱者が限られる．MFCCはケプストラム領域の特徴量であるため加法性が適用できないが，ベクトルテイラー系列を用いて同時発声をモデル化する手法が提案されている [82]．さらに幅広い楽曲に対して歌唱者ダイアライゼーションを行うためには，このような各歌唱者のみの音響モデルから同時歌唱モデルを構成する手法を検討する必要がある．

本論文で扱った音声は伴奏音のない歌声のみの音声である．歌声のみに対しての歌唱者ダイアライゼーションを十分に行える手法を構築するとともに，伴奏音やエフェクトに頑強な歌唱者ダイアライゼーション法を構築することが求められる．また，多くの楽曲では同時歌唱の音声にパンニングの処理を用いており，マルチチャンネル処理によりパンニングによる情報を利用する手法も検討する必要がある．

歌唱者ダイアライゼーションは，それ自体が最終目的ではない．歌唱者ダイアライゼーションを多様な音楽情報処理技術と組み合わせ，本研究が多様な人々の音楽鑑賞をさらに豊かにすることを目指さなければならない．

謝辞

指導教員である本学大学院工学系研究科の齋藤大輔講師には，熱心なご指導をいただき，また多くの助言を賜りました．また，直接的な研究の指導だけでなく，研究室のサーバ管理を始めとして私たちの研究の環境を整え支えてくださいました．心より感謝申し上げます．

本学大学院工学系研究科の峯松信明教授には，第二の指導教員として研究に関わってくださり，普段から多くの助言をいただきました．深く感謝いたします．

本研究の出発点は産業技術総合研究所でのインターンシップにありました．産業技術総合研究所の後藤真孝博士，深山覚博士，中野倫靖博士には，インターンシップ中に本研究のテーマ提案から研究に関する多くの助言をいただき，またインターンシップ終了後も継続して本研究に関わってくださいました．この先生方との出会いがなければ，本研究が生まれ形を成すこともありませんでした．深く感謝を申し上げます．また，インターンシップは，濱崎雅弘博士をはじめとした産業技術総合研究所の方々，また同時期にインターンシップを行っていた 5 人の学生のみなさんに支えられておりました．インターンシップを有意義なものにしてくださった皆さまに感謝いたします．インターンシップに関連し，本研究の一部は JST ACCEL（JPMJAC1602）の支援を受けました．

日頃の研究室生活は，齋藤研究室の同期である是松優作氏をはじめとした峯松・齋藤研究室の多くの同期や後輩によって有意義で楽しいものになりました．また，電気系工学専攻の同期である小谷岳氏には，音響学会研究発表会や APSIPA での研究発表をはじめとして声質変換研究の多くで関わっていただき，研究を共に築き上げていただきました．さらに，高橋登技術専門員，事務補佐員である押田美智子氏および池上恵氏には，峯松・齋藤研究室のメンバーとして研究室を支えていただきました．ここに挙げた方々はじめとする，研究室で共に生活してくださった皆さまに心から感謝いたします．ありがとうございました．

最後に，本研究は『アイドルマスター』という規模の大きく歴史あるゲームがなければ成し得ないものでした．『アイドルマスター』の存在によって，本研究の課題を提案し，そして複雑な実験系の構築と評価を行うことができました．キャラクターに息を吹き込んでくださる声優のみなさま，楽曲の作り手である作詞家・作曲家・編曲家のみなさま，そのほか制作に携わっているみなさま，ユーザでありアイドルたちを支え続けている世界中のプロデューサーのみなさま，そして『アイドルマスター』のキャラクターであるアイドルたちに心から感謝を申し上げます．

平成 31 年 1 月 31 日

須田 仁志

参考文献

[1] Masataka Goto, Kazuyoshi Yoshii, Hiromasa Fujihara, Matthias Mauch, and Tomoyasu Nakano. Songle: A web service for active music listening improved by user contributions. In *International Society for Music Information Retrieval Conference*, pp. 311–316, 2011.

[2] Jun Kato, Masa Ogata, Takahiro Inoue, and Masataka Goto. Songle Sync: A large-scale web-based platform for controlling various devices in synchronization with music. In *ACM International Conference on Multimedia*, pp. 1697–1705, 2018.

[3] Wei-Ho Tsai and Hsin-Min Wang. Automatic singer recognition for popular music recordings via estimation and modeling of solo vocal signals. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 1, pp. 330–341, 2006.

[4] Wei-Ho Tsai, Shih-Jie Liao, and Catherine Lai. Automatic identification of simultaneous singers in duet recordings. In *International Society for Music Information Retrieval Conference*, pp. 115–120, 2008.

[5] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 2, pp. 356–370, 2012.

[6] Marwa Thlithi, Claude Barras, Julien Pinquier, and Thomas Pellegrini. Singer diarization: Application to ethnomusicological recordings. In *International Workshop on Folk Music Analysis*, pp. 124–125, 2015.

[7] Toshihiko Tokizane. The formant construction of japanese vowels. *The Japanese Journal of Physiology*, Vol. 1, pp. 297–308, 1950.

[8] Bishnu S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, Vol. 55, No. 6, pp. 1304–1312, 1974.

[9] Douglas A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, Vol. 17, No. 1-2, pp. 91–108, 1995.

[10] Douglas A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *EUROSPEECH*, pp. 963–966, 1997.

[11] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, Vol. 10, pp. 19–41, 2000.

[12] Olivier Thyes, Roland Kuhn, Patrick Nguyen, and Jean-Claude Junqua. Speaker identification and verification using eigenvoices. In *International Conference on Spoken Language Processing*, pp. 242–245, 2000.

[13] William M. Campbell, Douglas E. Sturim, and Douglas A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, Vol. 13, No. 5, pp. 308–311, 2006.

[14] Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *INTERSPEECH*, pp. 1559–1562, 2009.

[15] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 3, pp. 345–354, 2005.

[16] Man-Wai Mak. Lecture notes on factor analysis and i-vectors. Technical report, Department of Electronic and Information Engineering, the Hong Kong Polytechnic University, 2016.

[17] Patrick Kenny. Bayesian speaker verification with heavy tailed priors. In *Odyssey Speaker and Language Recogntion Workshop*, 2010.

[18] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *INTERSPEECH*, pp. 999–1003, 2017.

[19] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5329–5333, 2018.

[20] Ondřej Novotný, Oldřich Plchot, Pavel Matějka, Ladislav Mošner, and Ondřej Glembek. On the use of x-vectors for robust speaker recognition. In *Odyssey Speaker and Language Recogntion Workshop*, pp. 168–175, 2018.

[21] Chunlei Zhang and Kazuhito Koishida. End-to-end text-independent speaker verification with triplet loss on short utterances. In *INTERSPEECH*, pp. 1487–1491, 2017.

[22] Hervé Bredin. TristouNet: Triplet loss for speaker turn embedding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5430–5434, 2017.

[23] Athanasios K. Noulas and Ben J. A. Kröse. On-line multi-modal speaker diarization. In *International Conference on Multimodal Interfaces*, pp. 350–357, 2007.

[24] Jacob Benesty, M. Mohan Sondhi, and Yiteng Haung, editors. *Springer Handbook of Speech Processing*. Springer, 2007.

[25] Lei Sun, Jun Du, Chao Jiang, Xueyang Zhang, Shan He, Bing Yin, and Chin-Hui Lee. Speaker diarization with enhancing speech for the first DIHARD challenge. In *INTERSPEECH*, 2018.

[26] Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and Sanjeev Khudanpur. Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In *INTERSPEECH*, 2018.

[27] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florêncio, and Mark Hasegawa-Johnson. Speech enhancement using Bayesian wavenet. In *INTERSPEECH*, pp. 2013–2017, 2017.

[28] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5069–5073, 2018.

[29] Scott Shaobing Chen and Ponani S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, 1998.

[30] Herbert Gish, Man-Hung Siu, and Robin Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 873–876, 1991.

[31] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard. 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.

[32] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, Vol. 6, No. 2, pp. 461–464, 1978.

[33] Margarita Kotti, Emmanouil Benetos, and Constantine Kotropoulos. Computationally efficient and robust BIC-based speaker segmentation. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 5, pp. 920–933, 2008.

[34] Jan Prazak and Jan Silovsky. Speaker diarization using PLDA-based speaker clustering. In *The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems*, pp. 347–350, 2011.

[35] Jan Silovsky, Jan Prazak, Petr Cerva, Jindrich Zdansky, and Jan Nouza. PLDA-based clustering for speaker diarization of broadcast streams. In *INTERSPEECH*, pp. 2909–2912, 2011.

[36] Nicholas Evans, Simon Bozonnet, Dong Wang, Corinne Fredouille, and Raphaël Troncy. A comparative study of bottom-up and top-down approaches to speaker diarization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 20, No. 2, pp. 382–392, 2012.

[37] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. Partitioning and transcription of broadcast news data. In *International Conference on Spoken Language Processing*, pp. 1335–1338, 1998.

[38] Jitendra Ajmera and Charles Wooters. A robust speaker clustering algorithm. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 411–416, 2003.

[39] Sylvain Meignier, Jean-François Bonastre, and Stéphane Igounet. E-HMM approach for learning and adapting sound models for speaker indexing. In *Odyssey Speaker and Language Recogntion Workshop*, pp. 175–180, 2001.

[40] Corinne Fredouille, Daniel Moraru, Sylvain Meignier, Laurent Besacier, and Jean-françois Bonastre. The NIST 2004 spring rich transcription evaluation: Two-axis merging strategy in the context of multiple distance microphone based meeting speaker segmentation. In *RT2004 Spring Meeting Recognition Workshop*, 2004.

[41] James L. Flanagan, James D. Johnston, Rolf Zahn, and Gary W. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America*, Vol. 78, No. 5, pp. 1508–1518, 1985.

[42] Xavier Anguera, Chuck Wooters, Barbara Peskin, and Mateu Aguiló. Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. In *Machine Learning for Multimodal Interaction*, pp. 402–414, 2005.

[43] Charles H. Knapp and G. Clifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 24, No. 4, 1976.

[44] Earl Vickers. Frequency-domain two- to three-channel upmix for center channel derivation and speech enhancement. In *Audio Engineering Society Convention*, 2009.

[45] 後藤真孝, 齋藤毅, 中野倫靖, 藤原弘将. 歌声情報処理の最近の研究. 日本音響学会誌, Vol. 64, No. 10, pp. 616–623, 2008.

[46] Takeshi Saitou, Masashi Unoki, and Masato Akagi. Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech Communication*, Vol. 46, No. 3-4, pp. 405–417, 2005.

[47] Johan Sundberg. The acoustics of the singing voice. *Scientific American*, Vol. 236, No. 3, pp. 82–84, 86, 88–91, 1977.

[48] Tong Zhang. Automatic singer identification. In *International Conference on Multimedia and Expo*, pp. I–33–36, 2003.

[49] Wei Cai, Qiang Li, and Xin Guan. Automatic singer identification based on auditory features. In *International Conference on Natural Computation*, pp. 1624–1628, 2011.

[50] Annamaria Mesaros, Tuomas Virtanen, and Anssi Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *International Society for Music Information Retrieval Conference*, 2007.

[51] Hiromasa Fujihara, Masataka Goto, Tetsuro Kitahara, and Hiroshi G. Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 3, pp. 638–648, 2010.

[52] Wei-Ho Tsai and Hao-Ping Lin. Background music removal based on cepstrum transformation for popular singer identification. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 5, pp. 1196–1205, 2011.

[53] Hiromasa Fujihara and Masataka Goto. A music information retrieval system based on singing voice timbre. In *International Conference on Music Information Retrieval*, 2007.

[54] Zhiyao Duan, Yungang Zhang, Changshui Zhang, and Zhenwei Shi. Unsupervised single-channel music score separation by average harmonic structure modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 4, pp. 766–778, 2008.

[55] Jean-Louis Durrieu an Gaël Richard, Betrand David, and Cédric Fevotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 3, pp. 564–575, 2010.

[56] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 57–60, 2012.

[57] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, Vol. 58, No. 3, 2011.

[58] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pp. 556–562. 2001.

[59] Paris Smaragdis and Judith C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[60] Beiming Wang and Mark D. Plumbley. Musical audio stream separation by non-negative matrix factorization. In *Digital Music Research Network Summer Conference*, 2005.

[61] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, pp. 1066–1074, 2007.

[62] Angkana Chanrungutai and Chotirat Ann Ratanamahatana. Singing voice separation for mono-channel music using non-negative matrix factorization. In *International Conference on Advanced Tech-*

nologies for Communications, 2008.

[63] Eri Ochiai, Takanori Fujisawa, and Masaaki Ikehara. Vocal separation by constrained non-negative matrix factorization. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 480–483, 2015.

[64] Tomohiko Nakamura and Hirokazu Kameoka. Shifted and convolutive source-filter non-negative matrix factorization for monaural audio source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 489–493, 2016.

[65] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 12, pp. 2136–2147, 2015.

[66] Andrew J. R. Simpson, Gerard Roma, and Mark D. Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 429–436, 2015.

[67] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 258–266, 2017.

[68] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-Net convolutional networks. In *International Conference on Music Information Retrieval*, pp. 745–751, 2017.

[69] Sue E. Tranter and Douglas A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp. 1557–1565, 2006.

[70] Pierre-Alexandre Broux, Florent Desnous, Anthony Larcher, Simon Petitrenaud, Jean Carrive, and Sylvain Meignier. S4D: Speaker diarization toolkit in python. In *INTERSPEECH*, 2018.

[71] Anthony Larcher, Kong Aik Lee, and Sylvain Meignier. An extensible speaker identification SIDEKIT in Python. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5095–5099, 2016.

[72] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp. 1505–1512, 2006.

[73] Seyed Omid Sadjadi and Malcolm Slaney. MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research. *IEEE Speech and Language Processing Technical Committee Newsletter*, 2013.

[74] Ahilan Kanagasundaram, Robbie Vogt, David Dean, Sridha Sridharan, and Michael Mason. I-vector based speaker recognition on short utterances. In *INTERSPEECH*, pp. 2341–2344, 2011.

[75] Makio Kashino and Tatsuya Hirahara. One, two, many—judging the number of concurrent talkers. *The Journal of the Acoustical Society of America*, Vol. 99, No. 4, pp. 2596–2603, 1996.

[76] Takayuki Kawashima and Takao Sato. Perceptual limits in a simulated "cocktail party". *Attention, Perception, & Psychophysics*, Vol. 77, No. 6, pp. 2108–2120, 2015.

[77] Michael A. Lewis and Ravi P. Ramachandran. Cochannel speaker count labelling based on the use of cepstral and pitch prediction derived features. *Pattern Recognition*, Vol. 34, No. 2, 2001.

[78] Martin Zelenák and Javier Hernando. Speaker overlap detection with prosodic features for speaker

diarization. *IET Signal Processing*, Vol. 6, No. 8, pp. 798–804, 2012.

[79] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4353–4356, 2008.

[80] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, Vol. E99.D, No. 7, pp. 1877–1884, 2016.

[81] Masanori Morise. D4C, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, Vol. 84, pp. 57–65, 2016.

[82] Pranay Dighe, Marc Ferras, and Herve Bourlard. Modeling overlapping speech using vector taylor series. In *Odyssey Speaker and Language Recogntion Workshop*, pp. 194–199, 2014.

発表文献

[1] 須田仁志, 齋藤大輔, 峯松信明. ソースフィルタ非負値行列因子分解によるボコーダを用いない声質変換の実験的検討. 日本音響学会 2017 年秋季研究発表会, pp. 301–304, 2017.

[2] 須田仁志, 小谷岳, 高道慎之介, 齋藤大輔. 高品質声質変換のための特徴量分析再訪. 日本音響学会 2018 年春季研究発表会, pp. 337–340, 2018.

[3] Hitoshi Suda, Gaku Kotani, Shinnosuke Takamichi, and Daisuke Saito. A revisit to feature handling for high-quality voice conversion based on Gaussian mixture model. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2018*, pp. 816–822, 2018.

[4] 須田仁志, 深山覚, 中野倫靖, 齋藤大輔, 後藤真孝. 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検討. 情報処理学会 音楽情報科学研究会第 121 回研究発表会, pp. 1–6, 2018.

付録 A

本研究で用いたグループアイドルソングデータセットの詳細

本章では，実験に用いるため構築したグループアイドルソングのデータセットについて述べる．

A.1　用いた楽曲群および歌唱者群

本研究では，グループアイドルソングのデータとして，ゲーム『アイドルマスター』[*7]の楽曲を用いた．これは，複数の歌唱者が同一の楽曲をそれぞれソロで歌った音声が多数用意できたためである．

データセットとして利用した歌唱者を表 10 に，利用した楽曲と各楽曲の歌唱者を表 11 に示す．コンテンツの性質上，すべての歌唱者がすべての楽曲を歌っているのではなく，各楽曲についてその楽曲を歌っていない歌唱者がある．なお，楽曲 M–P は第 9 章の実験のみに用いており，ダイアライゼーションの対象として用いていない．

A.2　歌声音声の処理

本論文では歌唱者ダイアライゼーションの基礎的な検討を行うため，伴奏音のない歌声をダイアライゼーションの対象として用いた．一方，伴奏音のない加工されていない歌声（ドライボーカル）は入手することが困難である．そこで，歌声のない伴奏音のみの音声（カラオケ音源）を用いて伴奏音を除去することで，ドライボーカルに近い音声を合成した．伴奏音の除去には，フリーソフトウェアである歌声りっぷ[*8]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おり，また歌声抽出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を用いることはできない．

ダイアライゼーションの対象にはパート割りのある歌声が必要なため，抽出したソロの歌声音声からパート割りのある音声にミックスしている．どのような歌声の組み合わせでも適切にミックスができるよう，データセット中の歌声はすべてカラオケ音源との時刻同期を取っている．ミックスは波形の加算によって行っており，n 人が同時に歌っている部分については加算後 $1/\sqrt{n}$ 倍することでパワーを正規化している．

A.3　パート割りの構成

より実際に用いられている構成に近いパート割りを再現するため，本論文ではゲーム内で用いられているパート割りを参考に構成した．そのため，歌唱者の切り替わりの頻度や各歌唱者の歌唱している時間が楽曲により異なる．また，どの楽曲もソロの歌声からミックスしているため，複数人で歌っている箇所は全歌唱者が同じ音高で歌うユニゾンとなる．本データセットでパート割りを行った楽曲 A，D–L について，各楽曲

[*7] https://idolmaster.jp/
[*8] http://www.vector.co.jp/soft/win95/art/se127635.html

表 10　データセットに含まれる歌唱者の一覧．歌唱者 P1–P4 は楽曲 P のみを歌唱しているため，他の歌唱者とは異なる番号付けを行っている．

	歌唱者 キャラクター名	声優
A	天海春香	中村繪里子
B	如月千早	今井麻美
C	萩原雪歩[†1]	長谷優里奈
D	萩原雪歩[†1]	浅倉杏美
E	高槻やよい	仁後真耶子
F	秋月律子	若林直美
G	三浦あずさ	たかはし智秋
H	水瀬伊織	釘宮理恵
I	菊地真	平田宏美
J	双海亜美・真美[†2]	下田麻美
K	星井美希	長谷川明子
L	四条貴音	原由実
M	我那覇響	沼倉愛美
P1	佐々木千枝	今井麻夏
P2	櫻井桃華	照井春佳
P3	市原仁奈	久野美咲
P4	赤城みりあ	黒沢ともよ

[†1] 本論文で用いたデータセットには 2 人の異なる声優が担当した萩原雪歩が存在するため，これらを区別する．
[†2] 双海亜美・真美は双子のキャラクターだが，同一の声優が担当し，また共通の歌声が用いられているため，本論文では 1 人として扱う．

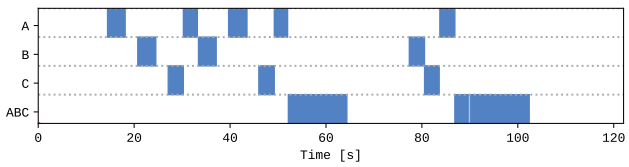
のパート割りの概要を表 12 に，パート割りの構成を図 24 に示す．なお，楽曲 L のパート割り Lf は，Lt における 50–57 秒付近のデュオ部分を 3 人歌唱にしたものである．

表 11　データセットとして利用した楽曲の一覧と，各楽曲の歌唱者の一覧．歌唱区間長は，楽曲のうちすべての歌唱している区間をあわせた時間を表す．

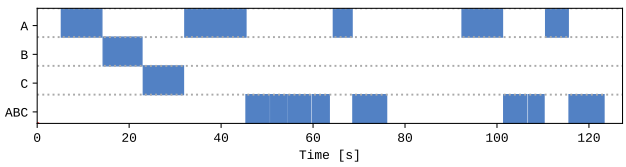
	楽曲名	楽曲長	歌唱区間長	歌唱者
A	おはよう!! 朝ご飯	2:02	1:04	A B C E F G H I J K
B	スタ→トスタ→	2:01	1:28	A B C E F G H I J K L M
C	MEGARE!	2:03	1:45	A B D E F G H I J K L M
D	蒼い鳥	2:07	1:38	A B C E F G H I J K
E	きゅんっ! ヴァンパイアガール	2:02	1:19	A B D E F G H I J K L M
F	Honey Heartbeat	2:02	1:51	A B D E F G H I J K L M
G	隣に…	2:05	1:28	A B C E F G H I J K L M
H	9:02 pm	2:03	1:34	A B C E F G H I J K
I	エージェント夜を往く	2:05	1:19	A B C E F G H J K
J	キラメキラリ	2:02	1:48	A B C E F G H I J K L M
K	THE IDOLM@STER	2:04	1:40	A B C E F G H I J K
L	relations	2:05	1:21	A B C E F G H I J K
M	Little Match Girl	2:09	1:34	A B D E F G H I J K L M
N	迷走 Mind	2:05	1:18	A B C E F G H I J K L M
O	Kosmos, Cosmos	2:02	1:35	A B C E F G H I J K L M
P	ハイファイ☆デイズ	3:49	3:10	P1–P4

表 12　本論文で用いた各楽曲に対するパート割りの概要．ここでのチャンスレートとは，全音声を 1 つのクラスタと認識したと仮定した error の割合を指す．すなわち，歌唱区間の検出が完全にできたと仮定した上で，全音声を 1 つのクラスタと認識した場合の DER となる．このチャンスレートは，歌唱区間検出さえ行うことができれば必ず達成できる DER を意味し，この値が楽曲によって大きく異なることに留意する．

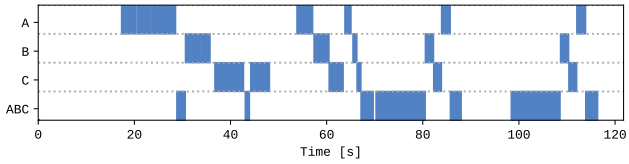
楽曲	楽曲長	歌唱区間長	デュオの有無	セグメント数	クラスタ数	チャンスレート
A	2:02	1:04	なし	14	4	57.2%
D	2:07	1:38	なし	15	4	58.5%
E	2:02	1:19	なし	26	4	61.3%
F	2:02	1:51	なし	23	4	38.2%
G	2:05	1:28	なし	9	4	55.0%
H	2:03	1:34	あり	20	5	53.8%
I	2:05	1:19	あり	20	5	71.0%
J	2:02	1:48	あり	26	3	55.5%
K	2:04	1:40	あり	9	3	11.6%
Lf	2:05	1:21	なし	15	4	57.5%
Lt	2:05	1:21	あり	15	5	60.3%



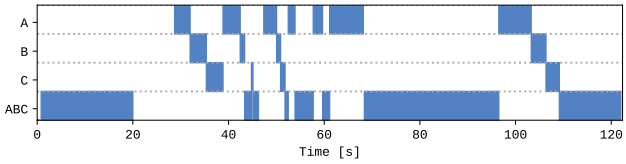
(a) 楽曲 A



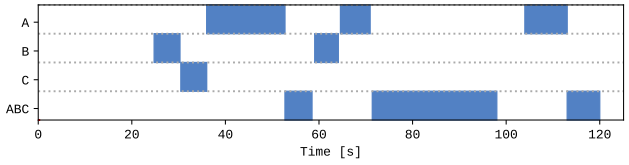
(b) 楽曲 D



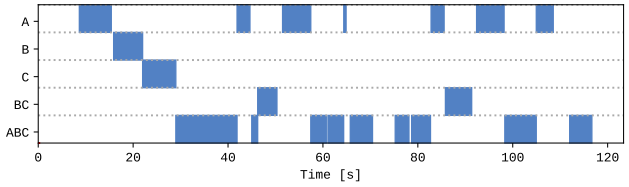
(c) 楽曲 E



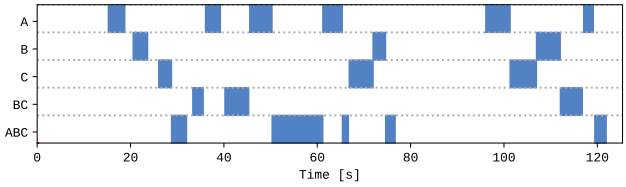
(d) 楽曲 F



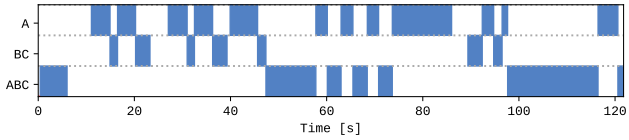
(e) 楽曲 G



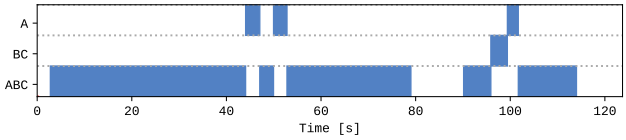
(f) 楽曲 H



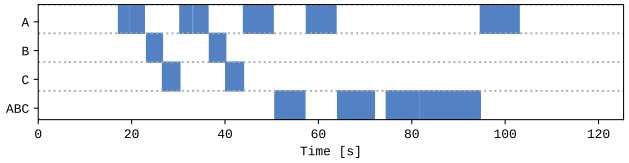
(g) 楽曲 I



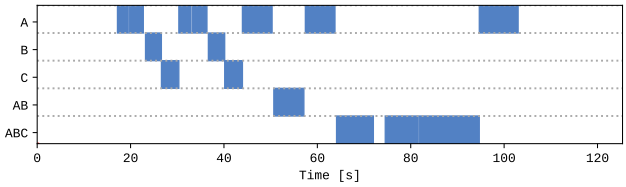
(h) 楽曲 J



(i) 楽曲 K



(j) 楽曲 L（パート割り Lf）



(k) 楽曲 L（パート割り Lt）

図 24　本論文で用いた各楽曲に対するパート割りの構成．ここでは歌唱者を A，B，C とした場合のラベルを用いている．楽曲 J，K においては歌唱者 B，C がソロで歌うパートが存在しない．