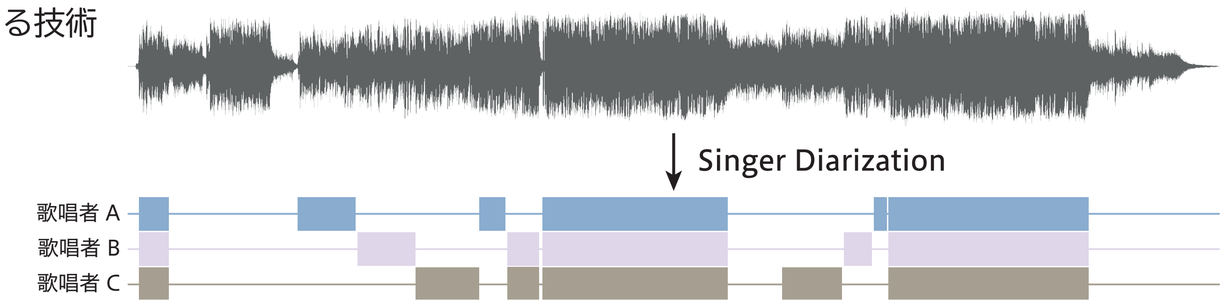
グループアイドルソングを対象とした歌唱者ダイアライゼーション手法の検討

須田仁志[*]、深山覚[†]、中野倫靖[†]、齋藤大輔[*]、後藤真孝[†]

[*]東京大学　[†]産業技術総合研究所

はじめに

- 話者ダイアライゼーション＝会話音声から「誰がいつ話しているか」を推定する技術
  - ▶ 歌唱者ダイアライゼーション＝歌声に対するダイアライゼーション
- 「誰がいつ歌っているか」がわかることで
  - ▶ 音楽データへのメタデータの付与
  - ▶ 同時に歌っている人数からの盛り上がり推定
  - ▶ 歌唱者の声質に応じた演出
  - ▶ 歌唱者ごとに音響モデルを適応することによる歌詞認識や伴奏音抽出の高性能化

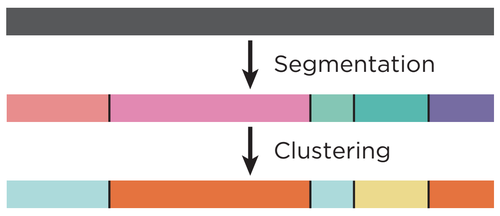


…などが期待できる

- 話者ダイアライゼーションの手法をそのまま適用できるだろうか？
  - ▶ 伴奏音をなくしたある程度理想に近い条件で検討したい

歌唱者の音響モデルが未知の手法 [X. Anguera+, 2012]

- 従来の話者ダイアライゼーションに用いられている手法と同様
- 主な 4 手順から構成される
  - ▶ 発話区間検出などの前処理
  - ▶ 単一話者と思われる区間でセグメンテーション
  - ▶ 各セグメントに対して話者でラベリング（クラスタリング）
  - ▶ HMM などを用いて推定結果の修正



- セグメンテーション・クラスタリングの規準には修正ベイズ情報量など
  - ▶ 情報量が大きい仮説を推定結果として選ぶ
  - ▶ 修正ベイズ情報量における定数は実験的に選択する
- 音響モデルが未知の手法は難しい
  - ▶ 短時間での音響的特徴に影響されやすい
    → 継続長の長い音素が現れやすい歌声では顕著に影響が出る
  - ▶ 後段の修正過程では前段での推定結果を初期値とするため初期値による影響を避けられない

歌唱者の音響モデルが既知の手法

- 短時間の歌唱者識別を繰り返す
  - ▶ i-vector を用いた話者認識法 [D. Reynolds+, 2000] を用いる
  - ▶ 本稿では 1 秒間の音声から 100 次元の i-vector を抽出し認識する
- 短時間で歌唱者が入れ替わるなど不自然な推定結果が得られやすい
  - ▶ 一定時間の窓内で最も多く推定された歌唱者を選ぶ平滑化により改善できる（majority vote）
  - ▶ 本稿では 3.1 秒の窓幅で平滑化している

同時歌唱に対する処理

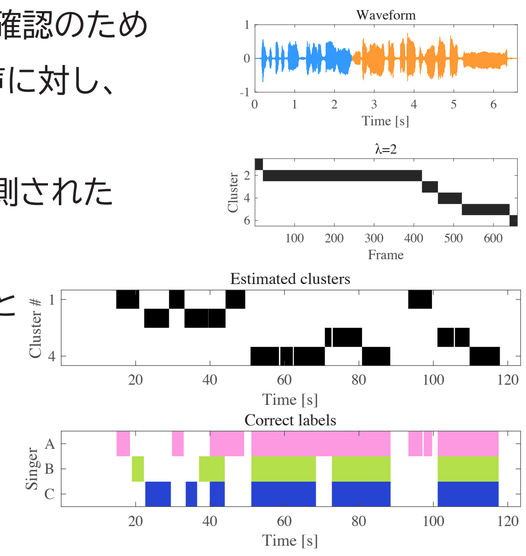
- 上のどちらの手法も、複数人が同時に発話することを考慮していない
  - ▶ 既存のダイアライゼーションでは同時発話を考慮することもあるが、手法を今回の音声に適用することは難しい
- 本稿では、同時に歌っている歌声がソロそれぞれの歌声と同様の音響モデルで表されることを仮定する
  - ▶ A さんモデル、B さんモデル、C さんモデル、A+B+C モデル

実験条件など

- 対象楽曲群として、ゲーム『アイドルマスター』内の楽曲を利用
  - ▶ 歌唱者の入れ替わり（パート割り）が存在する楽曲が多く、同じ楽曲を各歌唱者がソロで歌った音源や歌声なしの音源が市販されているため
  - ▶ 伴奏音の除去には歌声りっぷ [*1] を用いており、リバーブが残っているなど理想的なドライボーカルの音源ではない
- サンプリング周波数：16 kHz　音響特徴量：12 次 MFCC ＋Δ
- ダイアライゼーションの対象楽曲は 3 名の歌唱者による歌唱とし、3 人のうち同時に 2 人のみが歌っている状況は仮定しない
  - ▶ 実験のため伴奏なしの音源から作成したもの
- 誰も歌っていない状態はパワーにもとづき検出

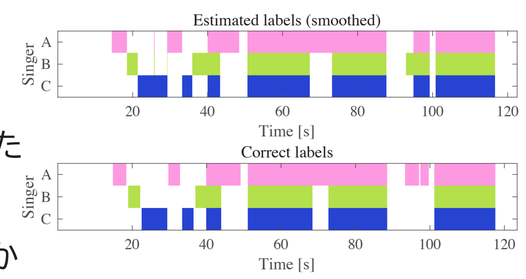
実験 1 : 歌唱者の音響モデルが未知の条件

- （予備実験）セグメンテーション性能の確認のため前半と後半で異なる歌唱者が歌った歌声に対し、前述のセグメンテーションを行った
  - ▶ 歌唱者ではなく音素による影響が観測された
- ツールキットに LIUM [*2] を用いた
  - ▶ 認識は行っていないため、クラスタと歌唱者の対応が取られていない
- 3 つめと 4 つめのクラスタがともにユニゾン部に割り当てられている
- 誤り率を示す DER [*3] は 35%



実験 2: 歌唱者の音響モデルが既知の条件

- UBM には APPBLA から学習した 2048 混合の GMM を利用
- 判別器は LDA とし、認識対象を含まない 15 曲を利用して学習
  - ▶（予備実験）認識対象の楽曲について、認識の正解率は 85%
- 平滑化前の DER は 28%、平滑化後の DER は 12%
- 93 秒から 100 秒付近に誤りが見られた
  - ▶ セリフであることにより特徴量の統計的性質が影響を受けたか



おわりに・今後の課題

- 音響モデルの有無による合理的な差であるとはいえ、音響モデルが未知の手法に大きな改善の余地があることが認められた
- 既知の場合でも平滑化に大きく依存しており、認識手法そのものと平滑化の両手法に対してさらなる検討が必要
- 3 人中 2 人のみが歌っているなどの複雑な場合の考慮
  - ▶ 同時に歌唱している人数をあらかじめ推定するなどの手法で対応したい
- 楽曲や歌唱者の組み合わせを様々に選び、さらに多くの検証を行いたい
- 伴奏音に頑健な手法の検討

*1 https://www.vector.co.jp/soft/win95/art/se127635.html　*2 http://www-lium.univ-lemans.fr/diarization [M. Rouvier+, 2013]　*3 Diarization Error Rate [S. E. Tranter+, 2006]