歌声の分離と音響モデルの分離歌声への適応に基づく音楽音響信号と歌詞の時間的対応付け手法*

◎藤原弘将 (京大), 後藤真孝, 緒方淳 (産総研), 駒谷和範, 尾形哲也, 奥乃博 (京大)

1 はじめに

本稿では，市販 CD 等の歌声と伴奏を含む音楽音響信号と歌詞の時間的対応付け手法について述べる．つまり，音楽音響信号と対応する歌詞のアラインメントをとることで，歌詞の各フレーズの開始時刻と終了時刻を推定する．本手法は，音楽ビデオのテロップ自動作成や，歌詞を用いた頭出しなどに応用できる．

関連する先行研究として，Wang らが開発した LyricAlly [1] がある．彼らは，歌声の音韻的特徴を考慮せず，歌詞中の各音素の持続長のみを用いて時間的対応関係を推定していた．しかし，音素の持続長は楽曲中の登場位置によって大きく異なるので，正確な対応付けはできなかった．

本研究では，音声認識で用いられる強制アラインメントに基づき，歌声の音韻的特徴を用いて時間的対応関係を推定する．しかし，現在の音声認識で用いられるアラインメント手法は，背景音等を含まないクリーンな話し声しか対象としていないので，歌声と共に伴奏音が演奏されている場合や歌が歌われない間奏部が存在する場合の正確な対応付けの実現が課題となる．この問題を解決するため，まず我々が以前開発した**伴奏音抑制** [2] を適用する．この手法では，メロディの調波構造を抽出・再合成することで，歌声を含むメロディのみを分離する．次に，図 1 に示した歌声・非歌声状態を行き来する隠れマルコフモデル (HMM) に基づく**歌声区間検出**を用いて，実際に歌声が存在する区間を検出する．最後に，**強制アラインメント**を用いて，分離歌声と歌詞を対応づける．その際，音響モデルを特定歌手の分離歌声に適応させることも可能である．

2 歌声と歌詞の時間的対応付け手法

本研究では，音楽音響信号と歌詞の時間的対応関係を推定し，歌詞の各フレーズの開始時間と終了時間を求めるために，音声認識で使われる強制アラインメントを適用する．伴奏を含む音楽に対して強制アラインメントを適用するために本研究では，伴奏音抑制，歌声区間検出，音響モデルの適応と強制アラインメント，という 3 つの手法を考案し，この問題を解決した．以下，それぞれについて述べる．

2.1 伴奏音抑制 [2]

混合音からメロディのみの音響信号を得るため，メロディの調波構造を抽出し，再合成する．具体的な処理は以下の通りである．

まず，PreFEst[3] を用いて，混合音中のメロディの基本周波数 (F0) を推定する．PreFEst は，制限された周波数帯域における最も優勢な調波構造の F0 を EM 法によって推定する．次に，推定された F0 に基づき，調波構造の各倍音成分 (調波構造) のパワーを抽出する．最後に，抽出された調波構造を正弦波重畳モデルに基づき再合成する．このとき元の信号中の位相情報は無視し，各フレーム間の周波数が線形に変化するように，位相の変化を 2 次関数で近似する．また，各フレームの振幅の変化は 1 次関数で近似する．

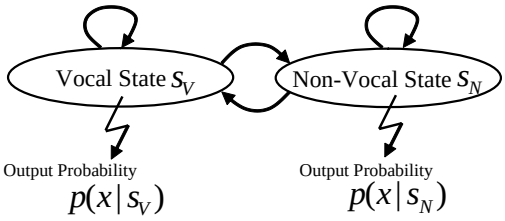

Fig. 1 Vocal activities detection based on HMM

2.2 歌声区間検出

伴奏音抑制によって得られたメロディの音響信号は，間奏部などでは歌声以外の楽器音を含んでいる．それらの非歌声区間を歌声状態と非歌声状態を行き来する HMM(図 1) に基づく歌声区間検出手法を用いて除去する．

本手法は，入力音響信号から抽出された特徴ベクトル列に対して，歌声・非歌声状態の最尤経路 $\hat{S} = \{\hat{s_1}, \cdots, \hat{s_t}, \cdots\}$ の探索問題として次式でモデル化できる．

$$\hat{S} = \underset{S}{\operatorname{argmax}} \sum_t \{\log p(\boldsymbol{x}|s_t) + \log p(s_{t+1}|s_t)\}, \quad (1)$$

ただし，$p(\boldsymbol{x}|s)$ は状態 s の出力確率を，$p(s_i|s_j)$ は状態 s_j から状態 s_i への遷移確率を表す．各状態の出力確率は混合ガウス分布 (GMM) の確率密度関数を用いて近似する．歌声，非歌声状態の GMM のパラメータは，それぞれ，予め学習データの歌声区間と非歌声区間を用いて学習しておく．

GMM の特徴量は，歌声の音韻的特徴を表現する LPM メルケプストラムと，歌声の動的な性質を表現する ΔF0 (F0 の微分係数) を用いる．

2.3 音響モデルの適応と強制アラインメント

アラインメントに用いる音響モデルを，入力音響信号中の特定歌手に適応させた後，分離歌声から抽出された特徴ベクトル入力された歌詞を用いて，強制アラインメントを行う．

まず，入力歌詞からアラインメントに用いる音素レベルのネットワークを作成する．本研究では，アラインメントの際に，調波構造が安定して抽出できる母音のみを用いる．具体的には，まず歌詞を音素列に変換し，その後，以下の三つの規則を用いて音素ネットワークに変換する．

1. 撥音 (ん) を表す音素以外の子音を削除
2. 文やフレーズの境界を複数回のショートポーズに変換

*Automatic synchronization between music and lyrics based on segregation of vocal and phoneme model adaptation for segregated vocal. by Hiromasa Fujihara (Kyoto University), Masataka Goto, Jun Ogata (AIST), Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno (Kyoto University)

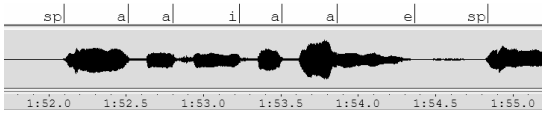

Fig. 2　An example of phoneme label for adaptation

3. 単語境界を一回のショートポーズに変換

アラインメントに用いる音響モデルは，話し声用の音素 HMM を元に，分離歌声に適応させたものを使用する．音響モデルとしては，大量の歌声のデータから学習されたモデルを使用することが理想的であるが，現段階ではそのようなデータベースは構築されていない．そのため，本研究では初期モデルとして話し声用のモデルを使用した．

適応手法は，以下の 3 段階からなる．

(1) 話し声用の音響モデルを単独歌唱の歌声に適応
(2) 単独歌唱用の音響モデルを伴奏音抑制手法によって抽出された分離歌声に適応
(3) 分離歌声用の音響モデルを入力楽曲中の特定楽曲に適応

(1) と (2) は教師あり適応で，事前に行われる．一方，(3) は教師無し適応で，実行時に行われる．ここで，教師情報とは，各音素の時間情報 (開始時刻，終了時刻) を意味する．つまり，教師あり適応では，手動で付与した時間情報 (図 2) により正確にセグメンテーションされたデータを用いる．適応時のパラメータ推定には，MLLR と MAP を組み合わせた手法を用いた．

最後に，歌詞を元に作成された音素ネットワーク，分離歌声の信号から抽出された特徴量，特定歌手に適応された音響モデルを用いて，強制アラインメントを行う．特徴量は，MFCC，ΔMFCC，Δ パワーを用いた．

3　評価実験

3.1　実験条件

「RWC 研究用音楽データベース: ポピュラー音楽」(RWC-MDB-P-2001) [4] の日本語の楽曲の中から，評価には楽曲 10 曲を用い，歌声区間検出の学習にはそれらとは異なる 19 曲を用いた．楽曲中に一部登場する英語の音素は類似した日本語の音素の音響モデルを用いて近似した．これらの楽曲に対して，性別毎の 5 fold cross-validation 法で評価をした．

強制アラインメント初期音響モデルとしては，CSRC ソフトウェア [5] 中の性別非依存モノフォンモデルを用いた．また，歌詞から音素列への変換には，日本語形態素解析システム茶筌 (ChaSen)[6] を実行し，その際に出力される読みの情報を用いた．

評価は，フレーズ単位のアラインメントを元に行った．本実験では，フレーズとは歌詞中のスペースや改行で区切られた一節を意味するものとする．楽曲の全体長の中で，フレーズ単位のラベルが正解していた区間の割合を精度として評価した．

3.2　結果と考察

図 3 に本実験の結果を示す (横軸は RWC-MDP-P-2001 の楽曲番号)．本手法により 10 曲中 8 曲について 90%以上の精度で時間的対応を推定することができた．007 番と 013 番の楽曲は，歌声区間検出の精度

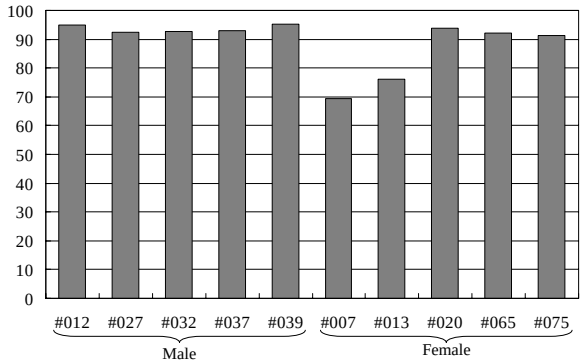

Fig. 3　Experimental result

が低く，曲の先頭部分や間奏部分の非歌声区間を正しく棄却出来なかったことが原因で，他の楽曲より精度が低かった．また，男声の精度が女性の精度に比べて高いことが見て取れる．これは，高い F0 を持つ声は，MFCC などのスペクトル特徴量を抽出するのが困難であるからである．各楽曲の内部では，歌詞が英語で歌われている部分付近で誤りが多く発生していた．今後は，日本語の音響モデルと英語の音響モデルを組み合わせることで，この問題に対処する．その他の代表的な誤りは，歌詞に書かれていない発声 (シャウト等) の部分で発生していた．各手法の単体での評価や，より詳細な考察は，文献 [7] に記されている．

4　おわりに

本稿では，音楽と歌詞の時間的対応付けを実現するための，伴奏音抑制手法，歌声区間検出手法，音響モデルの分離歌声への適応手法について述べた．評価実験により，10 曲中 8 曲に対して，十分な精度で時間的対応が推定できることを確認した．今後は，楽曲構造などの高次の情報を統合することで，より高度化を目指す．本研究の一部は，科研費，21 世紀 COE プログラム，CREST の支援を受けた．また，本研究の実験において，RWC 研究用音楽データベース (RWC-MDB-P-2001) [4] を使用した．

参考文献

[1] Wang et al., LyricAlly: Automatic synchronization of acoustic musical signals and textual lyrics, Proc. ACM Multimedia, pp.212–219, 2004.
[2] 藤原 他, 伴奏音抑制と高信頼度フレーム選択に基づく楽曲の歌手名同定手法, 情処論, Vol.47, No.6, pp.1831–1843, 2006.
[3] Goto, A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass line in real-world audio signals, Spe. Comm., Vol.43, No.4, pp.311-329, 2004.
[4] 後藤 他, RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース, 情処論, Vol.45, No.3, pp.728–738, 2004.
[5] Kawahara et al., Recent progress of open-source LVCSR engine Julius and japanese model repository – Software of continuous speech recognition consortium –, Proc. ICSLP2004, 2004.
[6] Matsumoto et al., Japanese morphological analysis system ChaSen, http://chasen.naist.jp/, 2000.
[7] Fujihara et al., Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals, Proc. ISM2006, pp.257–264, 2006.