社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－MUS－66（7）
2006／8／7

音楽音響信号と歌詞の時間的対応付け手法: 歌声の分離と母音の Viterbi アラインメント

藤原　弘将[†]　後藤　真孝[‡]　緒方　淳[‡]
駒谷　和範[†]　尾形　哲也[†]　奥乃　博[†]

[†] 京都大学大学院 情報学研究科 知能情報学専攻　　[‡] 産業技術総合研究所

本稿では，伴奏音を含む音楽音響信号と対応する歌詞の時間的な対応付け手法について述べる．クリーンな音声信号とその発話内容の時間的対応付けを推定をする Viterbi アラインメント手法はこれまでも存在したが，歌声と同時に演奏される伴奏音の悪影響で市販 CD 中の歌声には適用できなかった．本稿では，この問題を解決するため，歌声の調波構造を抽出・再合成することで混合音中の歌声を分離する手法，歌声・非歌声状態を行き来する隠れマルコフモデル (HMM) を用いた歌声区間検出手法，音響モデルを分離歌声に適応させることで Viterbi アラインメントを適用する手法を提案する．日本語のポピュラー音楽を用いた評価実験を行い，本手法により 10 曲中 8 曲について十分な精度で音楽と歌詞の対応付けが出来ることを確かめた．

Automatic synchronization between musical audio signals and their lyrics: vocal separation and Viterbi alignment of vowel phonemes

HIROMASA FUJIHARA[†], MASATAKA GOTO[‡], JUN OGATA[‡], KAZUNORI KOMATANI[†], TETSUYA OGATA[†] and HIROSHI G. OKUNO[†]

[†] Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
[‡] National Institute of Advanced Industrial Science and Technology (AIST)

This paper describes a method that can automatically synchronize between polyphonic musical audio signals and corresponding lyrics. Although there were methods that can synchronize between monophonic speech signals and corresponding text transcriptions by using Viterbi alignment techniques, they cannot be applied to vocals in CD recordings because accompaniment sounds often overlap with vocals. To align lyrics with such vocals, we therefore developed three methods: a method for segregating vocals from polyphonic sound mixtures by extracting and resynthesizing the vocal melody, a method for detecting vocal sections using a Hidden Markov Model (HMM) that transitions back and forth between vocal and non-vocal state, and a method for adapting a speech-recognizer phone model to segregated vocal signals. Experimental results for 10 Japanese popular-music songs showed that our system can synchronize between music and lyrics with satisfactory accuracy for 8 songs.

1. はじめに

歌声とその歌詞は，楽曲の主題やストーリーを表現し，楽曲を特徴づける要素の一つであるため，ポピュラー音楽を始めとする多くのジャンルの音楽で重要な役割を果たしている．実際，人が楽曲を聴く際には，歌声のメロディを聴き，その歌詞を耳で追うことが多い．そのため，音楽ビデオやテレビの音楽番組の中には，演奏の映像と同期して歌詞を表示することで，視聴者の楽曲鑑賞の手助けをするものもある．

本稿では，伴奏を含む音楽とその歌詞の時間的な対応付け手法を提案する．つまり，与えられた音楽音響信号と対応する歌詞をアラインメントすることで，歌詞の各フレーズの開始時刻と終了時刻を推定する．歌詞は多くの場合 Web などの情報源から手に入るので，我々は，歌詞の認識とは異なるアプローチを取った．本

手法は，音楽ビデオのテロップの自動作成や，楽曲中のユーザーの聴きたい歌詞の部分にジャンプ出来る機能を持つ音楽再生インタフェースなどに応用出来る．

Wang ら[1] は，我々と同様の問題に取り組んでいた．彼らは，ビートトラッキングやサビ区間検出などの高次の情報と，低次の歌詞アラインメント手法を統合するというアプローチを取った．しかし，彼らの低次の歌詞アラインメント手法は，歌詞中の各音素の発声長の情報のみを用いていた．各音素の発声長は，楽曲中での登場位置によって大きく異なるため，この手法は不十分であった．また，高次の情報を推定するための手法は，楽曲の構造や拍子に対して強い仮定を必要としていたため，適用できる楽曲に対して制限が大きかった．その他の関連研究としては，音声認識器を用いて歌詞の認識に取り組んだ研究があげられる．[2)~4)]．しかし，それらの研究は伴奏音を含まない単独歌唱の歌声を対象としているため，本研究が対象とする市販 CD などの音楽音響信号に適用するのは困難であった．

現行の音声認識で用いられるアラインメント手法はクリーンな話し声を対象とするため，伴奏を含む音楽とその歌詞の時間的な対応付けを取ることは出来なかった．そのため，本稿では下記の 3 つの手法を用いてこの問題に対処した．混合音中の歌声の音響信号を分離する手法，歌声が歌われている区間を検出する手法，音響モデルを分離歌声に適応させることで音声認識器を音楽と歌詞の時間的な対応付けに適用する手法である．

以下，第 2 章では，本研究の問題設定と提案手法の全体像について述べる．第 3 章から第 5 章では，提案手法の詳細について説明する．第 6 章では提案手法の有効性を確かめるために評価実験を行い，第 7 章ではまとめと今後の展望について述べる．

2. 音楽音響信号と歌詞の時間的対応付け手法

本稿では，与えられた音楽音響信号とその歌詞に対して時間的な対応付けを推定することで，歌詞の各フレーズの開始時間と終了時間を求めることを目指す．対象データは，市販 CD などの実世界の音楽音響信号であり，歌声だけでなく様々な楽器音を含んでいる．対象楽曲に対して，歌声の主要な部分は (コーラスを除いて) 一人の歌手によって歌われるという仮定を設けるが，伴奏音の音源の種類や数については仮定を設けない．

この目的のための本研究のアプローチは，音声認識で使われる Viterbi アラインメント (強制アラインメント) を適用することである．しかし，この手法は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には，適切に機能しない．この問題に対処するため，本研究ではまず我々が以前提案した，**伴奏音抑制**[5] 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が分離された音響信号を得る．次に，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，分離されたメロディから実際に歌声が存在する区間を検出する．最後に，**Viterbi アラインメント**を用いて，分離歌声と歌詞のアラインメントをする．また，ここでは，音響モデルを特定歌手の分離歌声に適応させる手法についても述べる．

3. 伴奏音抑制

混合音中の歌声の音韻的特徴を表す特徴量を抽出するためには，伴奏音の影響を低減させる必要がある．我々は，以前提案した伴奏音抑制手法[5]，つまり音響信号中のメロディの調波構造を抽出・再合成することで，伴奏音の影響を低減させる手法を用いる．伴奏音抑制手法は，以下の 3 つの処理からなる．

(1) 後藤の PreFEst[6] を用いて，メロディ(歌声) の基本周波数を推定する．
(2) 推定された基本周波数に基づき，メロディの調波構造を抽出する．
(3) 抽出された調波構造を，正弦波重畳モデルを用いて音響信号に再合成する．

以上の処理で，楽曲中のメロディのみの音響信号を得ることが出来る．図 1 に本手法の概要を示す．本手法により得られたメロディの音響信号は，間奏などの区間では (歌声でない) 楽器音も含んでいる．この問題は，4 節で述べる歌声区間検出手法によって対処する．

3.1 F0 推 定

基本周波数推定 (F0 推定) には，後藤の PreFEst[6] を用いる．PreFEst は，制限された周波数帯域において最も優勢な調波構造を持つ F0 を推定する手法である．メロディは中高域の周波数帯域において最も優勢な調波構造を持つ場合が多いため，周波数帯域を適切に制限することで，メロディの F0 を推定することが出来る．

以下，PreFEst の概要を記す．以後，x は cent の単位で表わされる対数周波数軸上の周波数で，(t) は時間を表わすとする．cent は，本来は音高差 (音程) を表す尺度であるが，本論文では文献 6) に従い，$440 \times 2^{\frac{3}{12}-5}$ Hz を基準として，次式のように絶対的な音高を表す

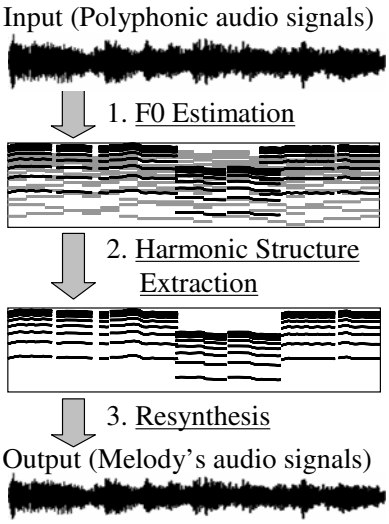


図 1 伴奏音抑制

単位として用いる．

$$f_{\text{cent}} = 1200 \log_2 \frac{f_{\text{Hz}}}{440 \times 2^{\frac{3}{12}-5}} \tag{1}$$

パワースペクトル $\Psi_p^{(t)}(x)$ に対して，メロディの周波数成分の多くが通過するように設計された帯域通過フィルタを適用する．本研究では，文献 6) に従い，4800 cent 以上の成分を通過させるフィルタを用いた．フィルタを通過後の周波数成分は $BPF(x)\Psi_p^{(t)}(x)$，と表わされる．ただし，$BPF(x)$ はフィルタの周波数応答である．以後の確率的処理を可能にするため，フィルタを通過後の周波数成分を確率密度関数 (PDF) として，以下のように表現する．

$$p_\Psi^{(t)}(x) = \frac{BPF(x)\Psi_p^{(t)}(x)}{\int_{-\infty}^{\infty} BPF(x)\Psi_p^{(t)}(x)dx} \tag{2}$$

その後，周波数成分の PDF が，全ての可能な F0 に対応する音モデルの重みつき和からなる確率モデル，

$$p(x|\theta^{(t)}) = \int_{\text{Fl}}^{\text{Fh}} w^{(t)}(F) p(x|F) dF, \tag{3}$$

$$\theta^{(t)} = \{w^{(t)}(F) | \text{Fl} \le F \le \text{Fh}\} \tag{4}$$

から生成されたと考える．ここで，$p(x|F)$ は，それぞれの F0 についての音モデルとし，Fh と Fl を取り得る F0 の上限と下限とする．また，$w^{(t)}(F)$ は音モデルの重みで，

$$\int_{\text{Fh}}^{\text{Fl}} w^{(t)}(F) dF = 1 \tag{5}$$

を満たす．音モデルとは典型的な調波構造を表現した確率分布である．そして，EM アルゴリズムを用いて $w^{(t)}(F)$ を推定し，それを F0 の PDF と解釈する．最終的に，$w^{(t)}(F)$ の中の優勢なピークの軌跡を，マルチエージェントモデルを用いて追跡することで，メロディの F0 系列を得る．

3.2 調波構造抽出

推定された F0 に基づき，メロディの調波構造の各倍音成分のパワーを抽出する．各周波数成分の抽出には，前後 r cent ずつの誤差を許容し，この範囲で最もパワーの大きなピークを抽出する．l 次倍音 ($l = 1,\ldots,L$) のパワー A_l と周波数 F_l は，以下のように表される．

$$F_l = \underset{F}{\text{argmax}} |S(F)| \quad (l\overline{F} \cdot (1 - 2^{\frac{r}{1200}}) \le F \le l\overline{F} \cdot (1 + 2^{\frac{r}{1200}})), \tag{6}$$

$$A_l = |S(F_l)|, \tag{7}$$

ここで，$S(F)$ はスペクトルを，$\overline{F}$ は PreFEst によって推定された F0 を表す．本稿では，r の値として 20 を用いた．

3.3 再 合 成

抽出された調波構造を正弦波重畳モデル[7] に基づき再合成することで，メロディの音響信号を得る．時刻 t における l 次倍音の周波数を $F_l^{(t)}$ と，振幅を $A_l^{(t)}$ と表す．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，各フレーム間の振幅の変化は 1 次関数で近似する．再合成された音響信号 $s(k)$ は，以下のように表現される．

$$\theta_l(k) = \frac{\pi(F_l^{(t+1)} - F_l^{(t)})}{K} k^2 + 2\pi F_l^{(t)} k + \theta_{l,0}^{(t)}, \tag{8}$$

$$s_l(k) = \left\{ (A_l^{(t+1)} - A_l^{(t)}) \frac{k}{K} + A_l^{(t)} \right\} \sin(\theta_l(k)), \tag{9}$$

$$s(k) = \sum_{l=1}^{L} s_l(k), \tag{10}$$

ここで，k は時間 (単位: 秒) を表し，時刻 t において $k = 0$ とする．また，K は，(t) と $(t+1)$ の時間の差，つまりフレームシフトを秒の単位で表す．$\theta_{l,0}^{(t)}$ は，位相の初期値を表し，入力信号の先頭のフレームでは，$\theta_{l,0}^{(t)} = 0$ とする．以後のフレームでは，$\theta_{l,0}^{(t)}$ は，前フレームの l 次倍音の周波数 $F_l^{(t-1)}$ と，初期位相 $\theta_{l,0}^{(t-1)}$ を用いて，$\frac{\pi(F_l^{(t)} - F_l^{(t-1)})}{2K} + \theta_{l,0}^{(t-1)}$ で与えられる．

4. 歌声区間検出

伴奏音抑制手法によって得られたメロディの音響信号は，間奏部などの非歌声区間では楽器音を含んでいる．そのような非歌声区間の存在は，音響信号と歌詞をアラインメントする際に悪影響を与える．本稿では，この問題を解決するため，入力音響信号中の非歌声区間を歌声区間検出手法を用いて除去する．

一般に，歌声の検出は，正解率 (hit rate) と棄却率 (correct rejection rate) によって評価される．ただし，正解率とは実際に歌声を含む領域の内，正しく歌声区間

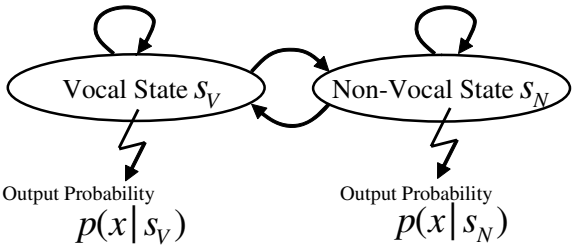


図 2 隠れマルコフモデル (HMM) に基づく歌声区間検出

として検出できた割合を指し，棄却率とは実際に歌声を含まない領域の内，正しく非歌声区間として棄却出来た割合を指す．我々が提案する歌声区間検出手法は，正解率と棄却率のバランスを調整することが出来る．なぜなら，それらの二つの基準はトレードオフの関係があり，適切な関係は用途によって異なるからである．例えば，本研究では歌声区間検出手法は Viterbi アラインメントの前処理であるため，正解率を高く保ち，歌声を含む可能性のある部分は余さず検出出来ることが望ましい．一方で，歌手名の同定などに用いる場合は，棄却率を高く保ち，確実に歌声を含む部分のみを検出するべきである．歌声の検出に関する先行研究[8)~10)]，では，正解率と棄却率のバランスを調整することは出来なかった．

4.1 定 式 化

歌声状態 (s_V) と非歌声状態 (s_N) を行き来する隠れマルコフモデル (図 2) を用いて歌声区間を検出する．歌声状態は歌声が存在する状態を表し，非歌声状態は歌声が存在しない状態を表す．ここでの目的は，次式のように，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 $\hat{S} = \{\hat{s_1}, \cdots, \hat{s_t}, \cdots\}$ を探索することである．

$$\hat{S} = \underset{S}{\operatorname{argmax}} \sum_t \{\log p(x|s_t) + \log p(s_{t+1}|s_t)\}, \quad (11)$$

ここで，$p(x|s)$ は状態 s の出力確率を表し，$p(s_i|s_j)$ は状態 s_j から状態 s_i への遷移確率を表す．

各状態の出力確率を，次式のように近似する．

$$\log p(x|s_V) = \log \mathrm{N}_{\mathrm{GMM}}(x;\theta_V) - \frac{1}{2}\eta, \quad (12)$$

$$\log p(x|s_N) = \log \mathrm{N}_{\mathrm{GMM}}(x;\theta_N) + \frac{1}{2}\eta, \quad (13)$$

ここで，$\mathrm{N}_{\mathrm{GMM}}(x;\theta)$ は混合ガウス分布 (GMM) の確率密度関数を表す．また，η は正解率と棄却率の関係を調整するパラメータである．歌声 GMM のパラメータ，θ_{V}，と非歌声 GMM のパラメータ，θ_{N} は，それぞれ，学習データの歌声区間と非歌声区間を用いて学習する．本稿では，混合数 64 の GMM を用いた．

4.2 閾値の設定

正解率と棄却率の関係は，式 (12) と (13) 中の η を変更することで調整する．しかし，GMM の尤度には楽曲によってバイアスがかかるため，全ての楽曲に適切な η を定めるのは困難である．そこで，本稿では η を バイアス調整値 $\eta_{\mathrm{dyn.}}$ とタスク依存値 η_{fixed} に分割する．

$$\eta = \eta_{\mathrm{dyn.}} + \eta_{\mathrm{fixed}} \quad (14)$$

タスク依存値，η_{fixed}，は用途に応じて手動で設定する．一方，バイアス調整値，$\eta_{\mathrm{dyn.}}$，は大津の閾値自動設定法[11)] を用いて楽曲毎に自動的に設定する．まず，入力音響信号から抽出された特徴ベクトル列に対して，次式のように歌声 GMM と非歌声 GMM の対数尤度差 $l(x)$ を計算する．

$$l(x) = \log \mathrm{N}_{\mathrm{GMM}}(x;\theta_V) - \log \mathrm{N}_{\mathrm{GMM}}(x;\theta_N). \quad (15)$$

そして，$l(x)$ のヒストグラムを作成する．最後に，そのヒストグラムをある閾値で 2 クラスに分割する場合に，クラス間分散が最小となるような閾値を決定し，$\eta_{\mathrm{dyn.}}$ の値として用いる．

4.3 特 徴 抽 出

本手法では，下記のような二種類の特徴量を用いる．

- LPC メルケプストラム (LPMCC)
  歌声・非歌声識別のためのスペクトル特徴量として，LPC メルケプストラム (LPMCC) を用いる．LPMCC は LPC スペクトル[12)] から計算されたメルケプストラム係数である．この特徴量は，我々が以前行った歌手名同定の実験で，メル周波数ケプストラム係数 (MFCC)[13),14)] と比べて，歌声の特徴をよく表現することを確認した．[15)]．本稿では，LPC スペクトルから MFCC を計算することで LPMCC を抽出した．
- ΔF0s:
  歌声の動的な性質を表現する特徴量として，F0 の微分係数 (ΔF0)[16)] を用いた．歌声は他の楽器音と比較して，ビブラートなどに起因する時間変動が多いので，F0 の軌跡の傾きを表す ΔF0 は，歌声と非歌声の識別に適していると考えられる．
  ΔF0 の計算には，次式のように 5 フレーム間の回帰係数を用いた．

$$\Delta f[t] = \frac{\sum_{k=-2}^{2} k \cdot f[t+k]}{\sum_{k=-2}^{2} k^2}, \quad (16)$$

  ここで，$f[t]$ は，時刻 t における周波数 (単位: cent) であるとする．

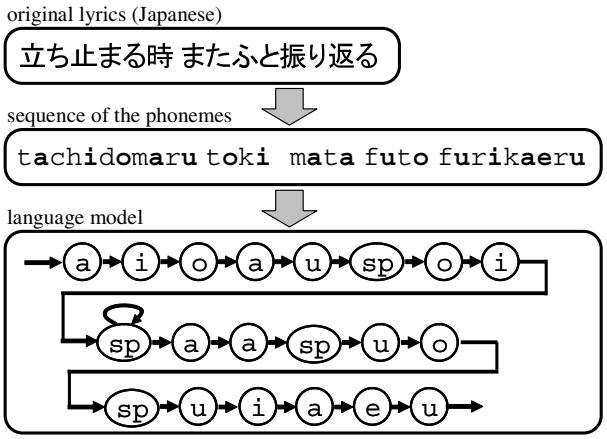


図 3 歌詞から文法への変換の一例

5. Viterbi アラインメント

本節では，歌詞と分離歌声を Viterbi アラインメント (強制アラインメント) する手法について述べる．まず，与えられた歌詞を元に，アラインメント用の文法を作成する．次に，分離歌声から特徴ベクトルを抽出する．最後に，それらの文法と特徴ベクトルを用いてアラインメントを行う．また，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る手法についても述べる．

5.1 歌詞のテキスト処理

与えられた入力音響信号に対応する歌詞を用いて，アラインメントに用いる文法を作成する．本研究では，アラインメントの際に，母音のみを用いる．これは，無声子音は調波構造を持たず，伴奏音抑制手法で抽出出来ないことと，有声子音も発声長が短いため安定して F0 を推定するのが難しいことが理由である．具体的な処理としては，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文法に変換する．

- 撥音，すなわち “ん” を表す音素以外の子音を削除する．
- 歌詞中の文やフレーズの境界を複数回のショートポーズ (sp) に変換する．
- 単語の境界を一回のショートポーズに変換する．

図 3 に，歌詞から文法への変換の例を示す．

5.2 音響モデルの適応

音響モデルを，入力楽曲中の特定歌手に適応させる．本研究のように，歌声に対してアラインメントを行う場合，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．そこで，本実験では初期音響モデルとしては，話し声用の音素 HMM を利用する．これは，主として音声認識を目的として構築されたモデルである．適応手法は，以下のように 3 段階からなる．

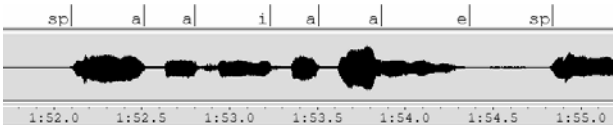


図 4 適応用音素ラベルの一例

(1) 話し声用の音響モデルを単独歌唱の歌声に適応させる．

(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させる．

(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させる．

(1) と (2) は教師あり適応で，事前に行われる．一方，(3) は教師なし適応で，認識時にオンラインで行われる．ここで，教師情報は，各音素ごとの時間情報 (音素の始端時間，終端時間) を指している．したがって，教師あり適応の場合は，時間情報により正確にセグメンテーションされた音素データを用いて適応が行われる．このときの時間情報は手動により付与した (図 4)．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．

5.3 アラインメント

歌詞を元に生成された文法，分離歌声の信号から抽出された特徴量と特定歌手に適応された音響モデルを用いて，Viterbi アラインメントを行う．特徴量は，MFCC[13]，Δ MFCC と Δ パワーを用いた．

6. 評 価 実 験

提案手法の性能を確認するために，評価実験を行った．

6.1 実 験 条 件

評価には，表 1 に示される 10 歌手 10 曲を用いた．これらの楽曲は，“RWC 音楽データベース:ポピュラー音楽 (RWC-MDB-P-2001)” から選んだ．[17] 楽曲の大半の部分は日本語で歌われているが，一部は英語で歌われている．本実験では，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．つまり，ある歌手によって歌われている楽曲を評価する際は，その歌手と同じ性別の歌手によって歌われている他の楽曲を用いて音響モデルを適応させた．

歌声区間検出手法の学習データには，表 2 に示される 11 歌手からなる 19 曲を用いた．これらの楽曲も “RWC 音楽データベース:ポピュラー音楽 (RWC-MDB-

表 1　評価用データ

Song #	Singer Name	Gender
012	Kazuo Nishi	Male
027	Shingo Katsuta	Male
032	Masaki Kuehara	Male
037	Hatae Yoshinori	Male
039	Kousuke Morimoto	Male
007	Tomomi Ogata	Female
013	Konbu	Female
020	Eri Ichikawa	Female
065	Makiko Hattori	Female
075	Hiromi Yoshii	Female

表 2　歌声区間検出の学習データ

Singer Name	Gender	Piece Number
Hiroshi Sekiya	M	048, 049, 051
Katsuyuki Ozawa	M	015, 041
Masashi Hashimoto	M	056, 057
Satoshi Kumasaka	M	047
Oriken	M	006
Tomoko Nitta	F	026
Kaburagi Akiko	F	055
Yuzu Iijima	F	060
Reiko Sato	F	063
Tamako Matsuzaka	F	070
Donna Burke	F	081, 089, 091, 093, 097

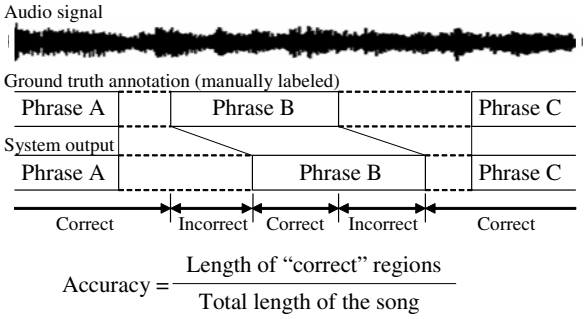
表 3　Viterbi アラインメントの分析条件

サンプリング	16 kHz, 16 bit
窓関数	Hamming 窓
フレーム幅	25 ms
フレームシフト	10 ms
特徴量	12th order MFCC 12th order ΔMFCC ΔPower

P-2001)" から選んだ．また，これらの 11 歌手は評価に用いられた 10 歌手には含まれていない．歌声区間検出手法の学習データにも，伴奏音抑制手法は適用した．また，η_{fixed} の値は 1.5 に設定した．

表 3 に，Viterbi アラインメントの分析条件を示す．初期音響モデルとしては，CSRC ソフトウェア[18] 中の性別非依存モノフォンモデルを用いた．また，歌詞から音素列の変換には，日本語形態素解析システム茶筅 (ChaSen)[19] を実行し，その際に出力される読みの情報を用いた．特徴抽出，Viterbi アラインメントと音響モデルの適応には，Hidden Markov Toolkit (HTK)[20] の HCopy，HVite，HEAdapt 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，元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．評価基準として，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計算した (図 5)．精



$$\text{Accuracy} = \frac{\text{Length of "correct" regions}}{\text{Total length of the song}}$$

図 5　評価基準

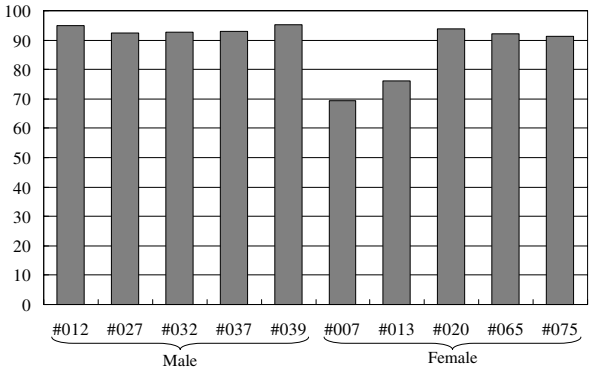


図 6　実験結果: システム全体の性能

度が 90%を超えていた場合に，その楽曲は正しくアラインメントされたと判断した．

6.2　システム全体の評価

提案手法全体での性能を評価するため，本稿で述べた提案手法を全て用いて実験を行った．図 6 に本実験の結果を示す．

6.3　音響モデル適応手法の評価

本実験の目的は，音響モデルの適応手法の効果を確認することである．具体的には，以下の 4 つの条件で実験を行った．

(i) 適応なし： 音響モデル適応を行わなかった．

(ii) 1 段階適応： 話し声用の音響モデルを直接分離歌声に適応させた．特定歌手への教師なし適応は行わなかった．

(iii) 2 段階適応： まず，話し声用の音響モデルを単独歌唱音声に適応させた後，分離歌声に適応させた．特定歌手への教師なし適応は行わなかった．

(iv) 3 段階適応 (提案手法)： まず，話し声用の音響モデルを単独歌唱音声に適応させた後，分離歌声に適応させた．最後に，入力音響信号の特定歌手への教師なし適応を行った．

また，本実験では全ての条件について歌声区間検出手法を使用した．図 7 に本実験の結果を示す．

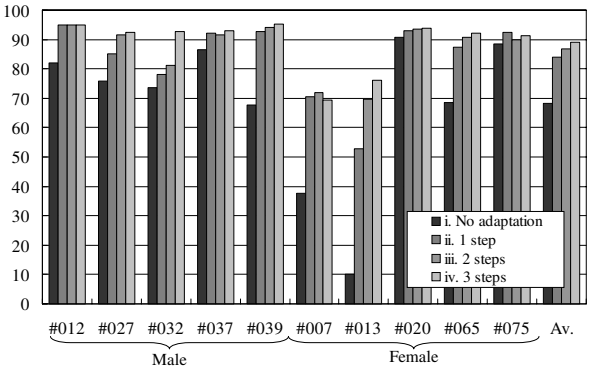


図 7 実験結果: 音響モデル適応の効果

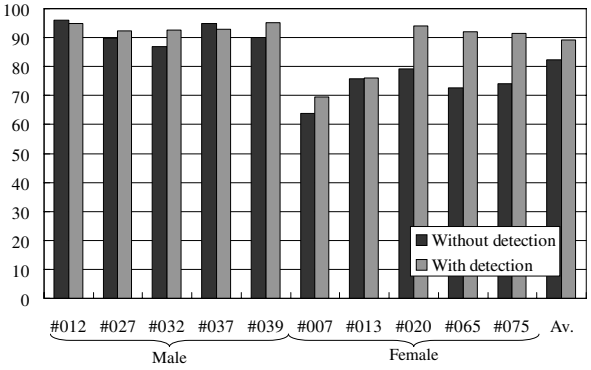


図 8 実験結果: 歌声区間検出の効果

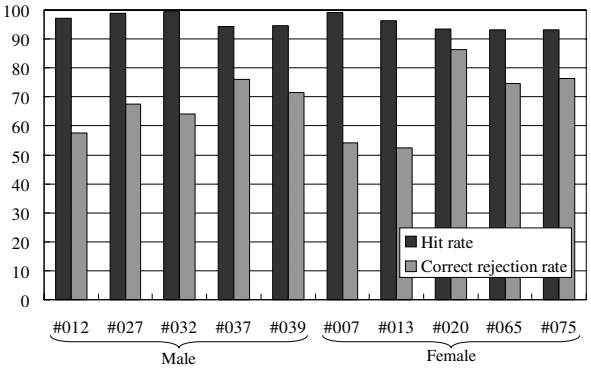


図 9 実験結果: 歌声区間検出の正解率と棄却率

6.4 歌声区間検出の評価

本実験の目的は，歌声区間検出の有効性を確認することである．また，歌声区間検出自体の性能の評価も行う．歌声区間検出を用いた場合と用いない場合の 2 通りの条件で実験した．本実験では，適応処理には全て 3 段階の適応手法を使用した．図 8 に本実験の結果を示す．また，図 9 に，歌声区間検出自体の正解率と棄却率を示す．

6.5 考　　察

図 6 を見ると，#007 と#013 を除き精度が 90%を超えていることがわかる．つまり，本手法により 10 曲中 8 曲について十分な精度で時間的対応を推定することが出来た．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[3]．各楽曲の内部での誤りを分析すると，歌詞が英語で歌われている部分付近では誤りが多く発生していた．これは，英語で発声された区間を，日本語の音素表記，日本語の音素モデルで近似することは困難な場合があるということを意味している．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，このような問題に対処する予定である．その他の代表的な誤りは，歌詞に書かれていないハミング等が歌われている部分で発生していた．

図 7 によると，音響モデル適応手法は，全ての楽曲で一定の効果があることがわかる．図 8 を見ることで，歌声区間検出手法は，比較的精度が低い楽曲に適用すると特に効果を発揮していることがわかる．しかし，#007 と#013 に関しては，元々の精度が低いにもかかわらず，歌声区間検出手法の効果が薄い．この理由は，これらの楽曲は，図 9 に見られるように，歌声区間検出の棄却率が高くないため非歌声区間を十分に除去できなかったからであると考えられる．また，歌声区間検出手法が，#012 や#037 など元々精度が高い楽曲に適用されると，精度が僅かながら低下している．これは，この手法で誤って除去されてしまった歌声区間は，必ず不正解と判定されてしまうからである．

7. ま と め

本稿では，音楽音響信号とその歌詞の時間的な対応付け手法について述べた．提案手法は，伴奏音抑制，歌声区間検出と Viterbi アラインメントの 3 つの処理からなる．また，音響モデルを特定歌手の分離歌声に適応させる手法についても述べた．評価実験により，様々な伴奏音を含む実世界の音楽音響信号に対して頑健にその歌詞を時間的に対応付けることが出来ることを確認した．

本研究には以下のような意義がある．

- 伴奏を含む楽曲と歌詞の時間的対応付けの問題に対して，混合音から歌声を分離し，母音を認識する手法を提案することで，初めて正面から取り組んだ．伴奏音による悪影響により，先行研究ではこの問題に音声認識の技術を適用することが出来

ていなかった．

- 正解率と棄却率のバランスを調整できる，新しい歌声区間検出を提案した．正解率と棄却率のバランスは用途によって異なるにもかかわらず，先行研究ではそのような観点に至っていなかった．本研究では，閾値をバイアス調整値とタスク依存値の 2 つの値に分割し，バイアス調整値を大津の閾値設定法[11] を用いて自動的に設定することで，可能にした．
- 話し声の音響モデルを特定歌手の分離歌声に適応させる手法を提案した．この手法は，音楽と歌詞のアラインメントの問題だけでなく，今まで扱われたことのなかった伴奏を含む音響信号に対する歌詞認識の問題に対しても重要な知見である．

今後は，日本語以外の言語で歌われる楽曲に対しても評価実験を行う予定である．また，楽曲構造などの高次の情報を統合することで，より高度な音楽と歌詞の時間的対応付け手法を目指す．

謝辞 本研究の一部は，科学研究費補助金（基盤研究 (A), 特定領域「情報学」），21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」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：ポピュラー」(RWC-MDB-P-2001)[17] を使用した．最後に，ご討論いただいた北原鉄朗氏，吉井和佳氏 (京都大学)，中野倫靖氏 (筑波大学) に感謝する．

参 考 文 献

1) Wang, Y., Kan, M.-Y., Nwe, T. L., Shenoy, A. and Yin, J.: LyricAlly: Automatic Synchronization of Acoustic Musical Signals and Textual Lyrics, *Proceedings of the 12th ACM International Conference on Multimedia*, pp.212–219 (2004).
2) Wang, C.-K., Lyu, R.-Y. and Chiang, Y.-C.: An Automatic Singing Transcription System with Multilingual Singing Lyric Recognizer and Robust Melody Tracker, *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech2003)*, pp. 1197–1200 (2003).
3) Sasou, A., Goto, M., Hayamizu, S. and Tanaka, K.: An Auto-Regressive, Non-Stationary Excited Signal Parameter Estimation Method and an Evaluation of a Singing-Voice Recognition, *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pp. I–237–240 (2005).
4) Hosoya, T., Suzuki, M., Ito, A. and Makino, S.: Lyrics Recognition from a Singing Voice Based on Finite State Automaton form Music Information Retrieval, *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pp. 532–535 (2005).
5) 藤原弘将, 北原鉄朗, 後藤真孝, 駒谷和範, 尾形哲也, 奥乃博: 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, 情報処理学会論文誌, Vol. 47, No. 6, pp. 1831–1843 (2006).
6) Goto, M.: A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals, *Speech Communication*, Vol. 43, No. 4, pp. 311–329 (2004).
7) Moorer, J. A.: Signal Procesing Aspects of Computer Music: A Survey, *Proceedings of the IEEE*, Vol. 65, No. 8, pp. 1108–1137 (1977).
8) Berenzweig, A. L. and Ellis, D. P. W.: Locating singing voice segments within music signals, *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2001).
9) Tsai, W.-H. and Wang, H.-M.: Automatic Detection and Tracking of Target Singer in Multi-Singer Music Recordings, *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, pp. 221–224 (2004).
10) Nwe, T. L. and Wang, Y.: Automatic Detection of Vocal Segments in Popular Songs, *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, pp. 138–145 (2004).
11) Otsu, N.: A Threshold Selection Method from Gray-Level Histograms, *IEEE Transaction on System, Man, and Cybernetics*, Vol. SMC-9, No. 1, pp. 62–66 (1979).
12) Atal, B. S.: 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).
13) Davis, S. B. and Mermelstein, P.: Comparison of parametric representation for monosyllabic word recognition, *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. 28, No. 4, pp. 357–366 (1980).
14) Logan, B.: Mel frequency cepstral coefficients for music modelling, *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2000)*, pp. 23–25 (2000).
15) Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G.: Singer Identification Based on Accompaniment Sound Reduction and Reliable Frame Selection, *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pp. 329–336 (2005).
16) Ohishi, Y., Goto, M., Itou, K. and Takeda, K.: Discrimination between Singing and Speaking Voices, *Proceedings of 9th European Conference on Speech Communication and Technology (Eurospeech 2005)*, pp. 1141–1144 (2005).
17) 後藤真孝, 橋口博樹, 西村拓一, 岡隆一: 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, 情報処理学会論文誌, Vol. 45, No. 3, pp. 728–738 (2004).
18) 河原達也, 武田一哉, 伊藤克亘, 李晃伸, 鹿野清宏, 山田篤: 連続音声認識コンソーシアムの活動報告及び最終版ソフトウェアの概要, 情報処理学会研究報告, pp. SLP–49–57 (2003).
19) Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K. and Asahara, M.: Japanese Morphological Analysis System ChaSen, http://chasen.naist.jp/ (2000).
20) The Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk/.