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

2008－MUS－74 (1)
2008－SLP－70 (1)
2008／2／8

音楽と歌詞の時間的対応付けシステムLyricSynchronizerを改良する3つの手法

藤原弘将　後藤真孝

産業技術総合研究所

本稿では，以前開発した音楽と歌詞の時間的対応付けシステムLyricSynchronizerを改良するための3つの手法について述べる．本システムは，伴奏を含む音響信号から歌声の母音とその区間を抽出し，歌詞から作成した音素ネットワークを用いて強制アラインメントを実行することで，最先端の性能を実現していたが，改善の余地があった．そこで，1つめの改良手法では，摩擦音が存在しない区間を検出し，その区間に歌詞中の摩擦音が割り当てられないようにする．2つめの手法では，音素ネットワークの各フレーズの間にフィラーモデルを挿入する．このモデルにより，歌詞に書かれていない発声を無視することができ，システムの性能が向上する．3つ目の手法では，歌声区間検出のための新しい特徴量を導入し，高調波構造同士の距離をスペクトル包絡を推定することなく計算できる．評価実験により，3つの手法それぞれが性能の向上に寄与することを確認した．

Three Techniques for Improving a System "LyricSynchronizer" for Automatically Synchronizing Music and Lyrics

HIROMASA FUJIHARA and MASATAKA GOTO

National Institute of Advanced Industrial Science and Technology (AIST)

In this paper, we describe three techniques that improves LyricSynchronizer, which is our previous system for automatically synchronizing lyrics with musical audio signals. Although this system achieved state-of-the-art accuracy by extracting a vocal melody from polyphonic sound mixtures, and using forced alignment between the vocal melody and a phoneme network of the lyrics, there was still room for improvement. The first technique detects regions in which fricative consonant sounds do not exist, and prohibits the alignment of fricative phonemes in the lyrics to those regions. The second technique inserts a filler model between phrases in the phoneme network. This model can ignore inter-phrase vowel utterances that are not included in the lyrics and improve a system performance. The third technique introduces novel feature vectors for vocal activity detection, which enable a distance calculation between two sets of the harmonic structure without estimating their spectral envelopes. Experimental results showed that all three techniques contribute to improved synchronization.

1. はじめに

市販CD等の音楽音響信号と歌詞の時間的対応付けは，音楽ビデオのテロップ自動生成や，ユーザが聴きたい箇所を歌詞中から選んで再生できる音楽再生インタフェースなど，様々な用途に有用である．しかし，歌声は他の楽器の伴奏音と混ざった状態で提供される場合がほとんどなため，音楽と歌詞の時間的対応付けは困難な課題であった．Wang[1]らは，LyricAllyと呼ばれるシステムを開発し，歌声を伴奏音から抽出することなく音楽と歌詞の時間的対応付けを実現しようとした．LyricAllyでは，対応付けのための手がかりとして，主に各音韻の持続時間長の情報を利用した．しかし，発声された音素の持続時間長は，歌詞中の登場位置によって大きく異なる場合があるため，必ずしも有効ではなかった．Wongら[2]は，広東語のポピュラー音楽を対象に，音楽と歌詞の時間的対応付けシステムを開発した．彼らは，音の高低で意味を区別する声調言語である広東語の性質を利用し，歌詞の各単語の高さを歌声の基本周波数（F0）と比較するというアプローチをとった．

彼らの手法は，声調言語でない他の多くの言語に直接適用するのは困難であった．Loscos ら[3] と Wang ら[4] は，音声認識器を使って歌声のアラインメントと認識を試みた．しかし，彼らは伴奏を含まない単独歌唱を対象にしていた．Gruhne ら[5] は，伴奏を含む歌唱からの音素の認識に取り組んだ．彼らは，音素境界は既知であるとの仮定の下，様々な識別手法を適用し比較した．しかし，彼らの実験は予備的なものであり，そのまま実際の楽曲に適用するのは困難であった．

我々は以前，伴奏を含む音楽音響信号から抽出された歌声と対応する歌詞を，時間的に対応付けるシステム LyricSynchronizer を開発した[6]．歌詞の各フレーズの開始時間と終了時間を推定するため，まず伴奏を含む音響信号から各時刻で最も優勢な音を，調波構造に基づいて分離した（伴奏音抑制）．最も優勢な音は，歌唱が存在する区間（歌声区間）では，多くの場合歌声の母音を含んでいる．そして，それらの分離された音響信号から歌声区間を検出した（歌声区間検出）．さらに，分離歌声に適応された音響モデルを使用し，音声認識で用いられる強制（Viterbi）アラインメント手法により，歌詞と分離された歌声の対応関係を推定した．その際，母音のみを使用し，子音は無視した．評価実験によりこの手法は有効であることを確認した一方で，1) 子音（特に無声子音）が正しくアラインメントできない問題，2) 歌詞に書かれていない発声（例えば歌手のシャウトなど）に歌詞を割り当ててしまう問題，3) 歌声の F0 が高い場合は歌声区間検出が必ずしも正確でないという問題，の 3 つの課題があり，これらを解決することでさらに性能を改善できることを確認した．

本稿では，これらの 3 つの課題を解決する 3 つの手法を提案し，本システムの性能の向上を図る．1 つ目は摩擦音検出である．これは摩擦音が存在しない区間を検出し，歌詞中の摩擦音の音素がその区間に割り当てられないようにするものである．この手法の特徴は，存在しない区間を検出するという点である．一般に，存在する区間を全て正確に検出することは困難であるが，確実に存在しない区間のみを検出することは比較的容易である．2 つ目は，フィラーモデルの導入である．フィラーモデルにより，歌詞に書かれていない単語等を歌手が発声した場合に，その区間を無視させることができる．我々の以前のシステムは，誤ってそのような発声区間に歌詞を当てはめてしまうことがあった[6]．本研究では，あらゆる母音の登場を吸収するフィラーモデルを各フレーズの間に挿入することで，そのような誤りを低減させる．3 つ目は，歌声区間検出のための新しい特徴量である．これは，歌声の高調波構造の F0 の値と，各倍音のパワーの対数値を特徴量として使用したものである．ケプストラムや LPC などの特徴量は，スペクトル包絡を推定するため，高い F0 を持つ音に対しては適切に機能しない場合があった．本稿で提案する特徴量を使用することで，スペクトル包絡を推定することなく高調波構造同士を比較することができる．

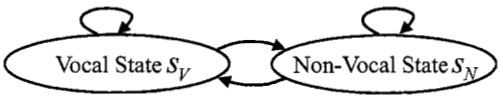


図 1　歌声区間検出のための隠れマルコフモデル

2. LyricSynchronizer の概要

文献 6) で述べた音楽と歌詞の時間的対応付けシステムの基本アプローチは，音声認識で用いられる強制（Viterbi）アラインメントを適用することである．音声認識の場合は雑音が少ない音声を対象とするが，我々は伴奏を含む信号中の歌声に適用することを可能にするため，(1) 歌声を含む最も優勢な音を分離する（伴奏音抑制），(2) 歌声区間を検出する（歌声区間検出），(3) 音声用の音響モデルを分離された歌声に適応させる，という 3 つのステップを用いた．

2.1 伴奏音抑制

まず，入力音響信号中の，メロディ（各フレームの最も優勢な F0[7]）の高調波構造を抽出・再合成することで，伴奏音の影響を低減させる．この処理は以下のように行われる．

(1) PreFEst[7] を使用して，入力音響信号のメロディの F0 を推定する．
(2) 推定された F0 に対応する高調波構造を抽出する．
(3) 正弦波重畳モデルを用いて再合成する．

2.2 歌声区間検出

次に，歌声を含まない領域（非歌声区間）を除去する．これは，分離されたメロディの音響信号は，間奏などの区間では歌声以外の楽器音を含んでいるからである．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．

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 1）を用いて歌声区間を検出する．ここでの問題は，分離されたメロディから抽出された特徴量 x 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．

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

ここで，$p(x|s)$ は状態 s の出力確率を表し，$p(s_i|s_j)$ は

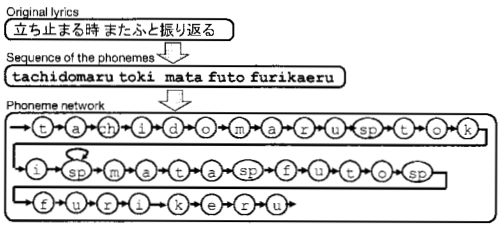


図 2　音素ネットワークの作成の例．この例では子音の音素も用いている．

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，歌声・非歌声混合ガウス分布（GMM）を用いて近似する．歌声・非歌声 GMM のパラメータは，それぞれ学習データの歌声区間，非歌声区間から抽出された特徴量を用いて推定される．特徴量として，文献 6) では LPMCC と ΔF0 を用いた．LPMCC は LPC 法とメルケプストラム法に基づく，スペクトル包絡を表現する特徴量である．なお，本稿では，LPMCC の代わりとなる新しい特徴量を 3.3 節で提案する．

2.3　音響モデル適応と強制アラインメント

次に，歌詞と分離歌声を時間的に対応づける．まず，与えられた入力音響信号に対応する歌詞から，下記の 2 つのルールを用いて，アラインメントに用いる音素ネットワークを作成する．

- 歌詞中の文（改行で区切られた区間）やフレーズ（空白で区切られた区間）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．

この処理の例を図 2 に示す．文献 6) では，母音のみを用いていたが，本稿では，3.1 節で述べるように，子音も使用する．SP は一般には単語等の間の短い無音区間を表すモデルであるが，ここでは非歌声区間一般を表現するモデルとして使用する．前奏や間奏などの長い非歌声区間は歌声区間検出によって除去されるが，フレーズ間の休符のような短い非歌声区間は強制アラインメントの際に SP を用いることで対応する．

強制アラインメントを実行する前に，音響モデルを歌声に対して適応させる．適応は，MLLR と MAP を組み合わせた手法により，下記の手順で行われる．

(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歌手に適応させる．

(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．(1) と (2) の適応で用いるデータに対しては，音素ラベルを手動で付与した．

最後に，音素ネットワークと特徴ベクトル列（MFCC，ΔMFCC，Δパワー），適応された音響モデルを用いて，強制アラインメントを実行する．

3. LyricSynchronizer を改良する新しい手法

本章では，我々の以前のシステムにおける以下の 3 つの弱点を克服するための 3 つの新しい手法について述べる．

(1) 子音を正しくアラインメントできない問題
伴奏音抑制手法は，調波構造に基づくため，無声子音を分離できない．そのため，母音の HMM のみを使用して歌詞と音楽の対応付けを行っていた．その場合，様々な子音の音を母音の HMM によって表現することになるため，子音の開始時間を正しく推定することが困難であった．この問題は，摩擦音検出を用いて，子音の情報も統合することで解決する．

(2) 歌詞に書かれていない発声の問題
ポピュラー音楽などでは，「イェー」や「ララ」のような歌詞には書かれていない言葉を，曲の間奏や，フレーズ間の休符の部分で発声することがよくある．そのようなフレーズの間で発声される母音が歌詞の他の部分に間違って対応付けられるため，システムの精度が低下する場合がある．この問題点は，フィラーモデルを導入することで解決する．

(3) 歌声区間検出誤りの問題
歌声区間検出結果に誤りがあった場合，歌声があるはずの区間に正しく歌詞が対応付けられないという問題がある．この問題は高い F0 を持つ女性歌手の楽曲でよく見られた．F0 が高い音は，スペクトル包絡を推定するのが難しいためである．この問題は，各調波構造の倍音比に基づく新しい特徴量を使用することで解決する．

3.1　摩擦音検出

子音の情報も使用するための最も単純なアプローチは，強制アラインメントで使われる音素ネットワークを作成する際に，子音も用いることである（図 2 参照）．しかし，我々が使用する伴奏音抑制手法は調波構造に基づいているため，無声子音を分離することができない．そのため，それだけでは無声子音を正しくアラインメントする精度には限界がある．我々は，無声子音の中でも摩擦音の候補を分離前の入力音響信号から直

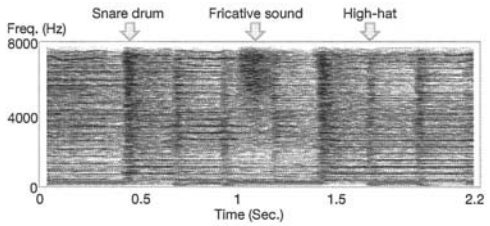


図 3 スネアドラムの音，摩擦音，ハイハットシンバルの音を含むスペクトログラムの例．

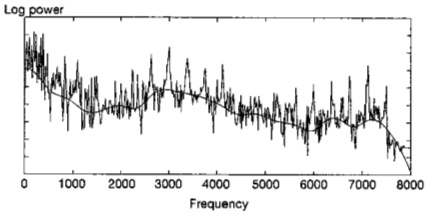


図 4 スペクトルのボトムエンベロープの例

接推定する手法を開発し，その情報もアラインメントの手がかりとして用いる．ここで摩擦音のみを用いたのは，摩擦音の持続時間長は一般に他の無声子音より長く，検出がし易いためである．

3.1.1 非存在区間の検出

摩擦音の音響的特性はシンバルやスネアドラムなどの音と似ることがあるため，その存在を正確に検出することは困難である．そこで我々は，反対のアプローチ，つまり摩擦音が存在しない区間（非存在区間）を検出するというアプローチをとる．そして，強制アラインメントの際に，歌詞中の摩擦音音素が非摩擦音区間に配置されるのを禁止する．摩擦音の非存在区間を検出することは容易であるので，検出誤りが強制アラインメントの精度を低下させるのを防ぐことができる．反対に，存在する区間を検出するアプローチでは，検出誤りが発生するのを避けられず，精度が低下してしまう場合がある．

3.1.2 摩擦音検出

図 3 はスネアドラム，摩擦音，ハイハットシンバルの音などの非定常音を含むスペクトログラムの例である．非定常音はスペクトログラム中の縦方向に広がる周波数成分として表れ，定常音は横方向に広がる周波数成分として表れる．各時刻の周波数スペクトル上では，縦方向の成分は平坦な周波数成分として，横方向の成分はピークを持つ成分として表れる．

非定常音に起因する平坦な周波数成分を検出するためには，スペクトル中のピークを持つ成分を除去するのがよい．そこで本研究では，亀岡ら[8]によって提案されたスペクトルのボトムエンベロープを推定する手法を用いる．ボトムエンベロープとは，図 4 のように，スペクトルの谷周辺を通るエンベロープ曲線のことである．ボトムエンベロープの関数クラスを，

$$g(f,a)=\sum_{i=1}^{I} a_i \mathcal{N}(f;400\times i,200^2) \tag{2}$$

のように定義する．ただし，f は Hz が単位の周波数を表し，$\mathcal{N}(x;m,\sigma^2)$ はガウス分布を表す．また，$a=(a_1,\cdots,a_I)$ は各ガウス分布の重みを表す．そして，次式の目的関数を最小化する a を推定することで，ボトムエンベロープが推定できる．

$$J=\int\left(\frac{g(f;a)}{S(f)}-\log\frac{g(f;a)}{S(f)}\right)df \tag{3}$$

ここで，$S(f)$ は各フレームのスペクトルを表す．この目的関数は，正の誤差と比べて負の誤差により重いペナルティを課す非対称な距離尺度である．この目的関数に基づいて $\hat{a}$ を推定するためには，以下の 2 つの式を反復計算する．

$$\hat{a}_i=\frac{\int m_i(f)df}{\int\frac{\mathcal{N}(f;400\times i,200)}{S(f)}df} \tag{4}$$

$$m_i(f)=\frac{a_i'\mathcal{N}(f;400\times i,200)}{\sum_{\forall i}a_i'\mathcal{N}(f;400\times i,200)} \tag{5}$$

ここで，a_i' は，前回の繰り返し時の推定値を表す．このようにして，スペクトル $S(f)$ のボトムエンベロープは $g(f,\hat{a})$ として推定される．

摩擦音の周波数成分は，スペクトルの特定の周波数帯域に集中している．そのため，ボトムエンベロープのその周波数帯域のパワーと，その他の帯域のパワーの比を用いて，摩擦音を検出する．現在の我々の実装ではサンプリング周波数は 16kHz であり，摩擦音の中でも，ナイキスト周波数である 8kHz 以下の帯域に成分が集中する/SH/の音素のみを扱う．6kHz から 8kHz の帯域に強い成分を持つ/SH/の存在度合いを，

$$E_{\mathrm{SH}}=\frac{\int_{6000}^{8000}g(f,\hat{a})df}{\int_{1000}^{8000}g(f,\hat{a})df} \tag{6}$$

として定義する．E_{SH} が閾値 0.4 を下回る区間を音素/SH/の非存在区間として検出する．0.4 という閾値の値は実験的に定められた．なお，バスドラムに起因する周波数成分の影響を避けるため，1kHz 以下の周波数帯域を計算に用いなかった．

3.2 フィラーモデル

歌詞中に書かれていない発声が原因のエラーを低減させるため，本研究では，フィラーモデルを使用する．図 5 のように，フィラーモデルは連続する 2 つのフレーズ間にあらゆる母音が複数回登場することを許容する．以前のシステムでは，ショートポーズを表す音素/SP/を用いて，そのような短時間の非歌声区間を表現して

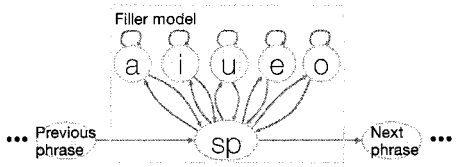


図 5　歌詞中の各フレーズの間に挿入されるフィラーモデル

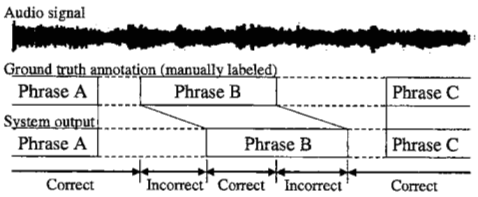
いたが，しかし，歌手が歌詞に書かれていない単語を非歌声区間で歌った場合，非歌声区間を用いて学習された/SP/では表現しきれなかった．そのため，以前のシステムではそのような非歌声区間に，他の箇所の歌詞を誤って当てはめてしまっていた．フィラーモデルの導入により，そのようなフレーズ間の発声は，フィラーモデル中の母音音素によって表現される．

3.3　歌声区間検出のための新しい特徴量

高調波構造の抽出に基づく伴奏音抑制手法の後に歌声区間検出を行うことは，抽出された高調波構造が歌声かどうかを判定する問題と捉えることができる．以前のシステムでは，抽出された高調波構造のスペクトル包絡を推定し，学習データ中のスペクトル包絡との距離を計算していた．しかし，ケプストラムや LPC を用いて，高い F0 を持つ音のスペクトル包絡を推定した場合，推定結果は各倍音成分の間の谷の部分に大きく影響を受ける．そのため，いくつかの楽曲（とくに女性歌手の楽曲）では，歌声区間検出が適切に機能しない場合があった．

この問題はつまり，抽出された高調波構造から推定されたスペクトル包絡は，各倍音成分付近の周波数帯域以外は必ずしも信頼できないということである．なぜなら，ある高調波構造に対応するスペクトル包絡は，いくつもの可能性が考えられるからである．そのため，高調波構造からスペクトル包絡を完全に復元することは，困難である．また，MFCC や LPC などのスペクトル包絡推定手法は，ある 1 つのスペクトル包絡の可能性のみを推定するため，元は同じスペクトル包絡であっても F0 が異なる二つの高調波構造同士の距離が，適切に小さくならない場合がある．この問題を解決するためには，距離を計算する際に，高調波構造の各倍音成分上の信頼できる点のみを使用するとよい．

本研究では，2 つの高調波構造の F0 がほとんど等しい場合は，各倍音のパワーを直接比較できることに着目する．我々のアプローチは，各倍音成分のパワーの値を特徴量として直接使用し，学習データベース中で近い F0 を持つ高調波構造のみと比較することである．このアプローチは，スペクトル包絡を推定する必要がないため，学習データが十分に存在する場合は，高い



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

図 6　評価基準

周波数の音に対しても頑健である．

さらに，近い F0 を持つ高調波構造のみと比較するため，F0 の値自体も特徴量として追加する．そして，その特徴ベクトルを GMM を使用してモデリングすることで，GMM の各ガウス分布それぞれが，F0 が近い特徴量をカバーする．GMM の尤度を計算する際は，F0 が大きく異なるガウス分布の影響は極めて小さくなる．それにより，近い F0 の値を持つ高調波構造のみとの比較が実現できる．

しかし，高調波構造の各倍音パワーの絶対値は，各楽曲ごとの音量の違いにより，バイアスがかかっている．そのため，各楽曲ごとに倍音パワーを正規化する必要がある．本研究では，時刻 t における h 次倍音のパワーを p_h^t とした場合に，正規化された倍音パワー p_h'' を，

$$\log p_h'' = \log p_h^t - \frac{\sum_t \sum_h \log p_h^t}{T \times H}, \tag{7}$$

のように表現する．なお，T はフレームの総数を表し，H は抽出された倍音数を表す．

4. 評価実験

4.1　実験条件

評価用のデータとして，「RWC 研究用音楽データベース：ポピュラー音楽」（RWC-MDB-P-2001）[9] から選択した 10 曲を使用し，5 fold cross-validation 法により評価した．歌声区間検出のための歌声・非歌声 GMM の学習データとして，同じく RWC 研究用音楽データベースから選択した別の 19 曲を使用する．

評価はフレーズ単位で行った．ここでフレーズとは，元の歌詞で空白または改行で区切られた一節のことを指す．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 6）．

実験は下記の 5 つの条件で行われた．

(i) 比較法： 以前のシステム[6] をそのまま使用．

(ii) 摩擦音検出： 以前のシステムに加えて摩擦音検出を使用（3.1 節）．

表 1 実験結果（単位：%）

曲	性別	(i) 比較法	(ii) 摩擦音	(iii) フィラー	(iv) 新特徴量	(v) 提案法
12	男	95.7	95.1	96.3	97.8	95.7
27	男	87.4	87.6	86.3	90.2	91.2
32	男	66.4	69.6	70.2	81.3	71.7
37	男	83.7	85.9	89.5	89.5	89.5
39	男	93.6	93.2	92.4	93.9	93.3
7	女	62.8	62.5	67.4	79.9	70.0
13	女	63.6	70.4	67.2	46.0	68.0
20	女	93.3	93.3	93.1	92.7	94.0
65	女	73.7	85.4	91.6	91.2	92.0
75	女	90.6	88.2	90.3	85.9	87.8
平均		81.1	83.1	84.4	84.8	85.3

番号は RWC-MDB-P-2001[9] の曲番号

(iii) フィラーモデル: 以前のシステムに加えてフィラーモデルを使用（3.2 節）.

(iv) 新しい特徴量: 以前のシステムに加えて，歌声区間検出用の新しい特徴量を使用（3.3 節）.

(v) 提案法: 以前のシステムに加えて，3 つの手法を全て使用.

4.2 結果と考察

結果を表 1 に示す☆．本稿で提案した新しい手法（表 1 中の (ii) と (iii)，(iv)）を用いることで，平均の認識精度がそれぞれ 2.0，3.3，3.7 ポイント向上した．さらに，全ての手法を使用した場合（表 1 中の (v)）が，最も認識精度が高かった．3 つの手法の中でも，歌声区間検出のための新しい特徴量が，最も効果的であった．また，フィラーモデルを使用した際の出力結果を見ると，フィラーモデルは歌詞に出てこない発声を吸収しているだけでなく，歌声区間検出で除去しきれなかった非歌声区間も吸収していることがわかった．評価基準がフレーズ単位であるため，摩擦音検出の効果は十分には確認できなかったが，音素単位のアラインメントを見ると，フレーズ途中での音素のずれが削減できている例が見られた．今後は，音素単位での評価を検討している．

5. 終わりに

本稿では，歌詞を音楽と同期させるシステム LyricSynchronizer を改善するため，摩擦音検出，フィラーモデル，歌声区間検出のための新しい特徴量の 3 つの手法を提案した．それぞれの手法は，特定の言語や楽曲構造に依存しないため，有用性が高い．

本研究では，摩擦音が存在する区間を残さず正確に検出することは困難だが，非存在区間ならば検出が比較的容易であることを利用し，その情報を統合することで性能向上を実現した．これにより，ここで提案した「非存在区間を検出する」というアイディアの有効性も確認できた．次に，フィラーモデルの導入は，単純なアイディアではあるが効果が高かった．このモデルは，元の歌詞をスキップすることは許容しない一方で，歌詞に書かれていない様々な発声を取り除き，性能を向上させる効果があった．また，F0 と倍音パワーに基づく新しい特徴量は，スペクトル包絡を推定する必要がないため，高い F0 の音に対して頑健に機能する汎用性の高い提案と言える．今回は歌声区間検出のみに用いたが，十分な量の学習データを準備することで，強制アラインメントの特徴量としても使用できる可能性がある．本特徴量は，今後，歌声だけでなく一般の音声を対象とした研究（音声認識など）にも使用することを検討している．また，今後の課題としては，/SH/ 以外の摩擦音も使用することや，英語の音響モデルを用いて英語楽曲でも評価することが挙げられる．

謝辞 本研究の一部は，科学技術振興機構 CrestMuse プロジェクトによる支援を受けた．また，本研究の実験において，「RWC 研究用音楽データベース：ポピュラー音楽」（RWC-MDB-P-2001）[9] を使用した．

☆ 文献 6) での実験は，男声楽曲に使用する音響モデルの適応には他の男声楽曲のみを用い，女声楽曲には他の女声楽曲のみを用いていた．また，文献 6) では子音を用いていなかったが，今回は子音を用いている．そのためベースラインの結果が前回のものと異なっている．

参考文献

1) Ye Wang *et al.*, "Lyrically: Automatic synchronization of acoustic musical signals and textual lyrics," in *Proc. ACM Multimedia 2004*, pp. 212–219, 2004.
2) Chi Hang Wong *et al.*, "Automatic lyrics alignment for Cantonese popular music," *Multimedia Systems*, vols. 4-5, no. 12, pp. 307–323, 2007.
3) Alex Loscos *et al.*, "Low-delay singing voice alignment to text," in *Proc. ICMC99*, 1999.
4) Chong-kai Wang *et al.*, "An automatic singing transcription system with multilingual singing lyrics recognizer and robust melody tracker," in *Proc. Eurospeech 2003*, pp. 1197–1200, 2003.
5) Matthias Gruhne *et al.*, "Phoneme recognition in popular music," in *Proc. ISMIR 2007*, pp. 369–370, 2007.
6) Hiromasa Fujihara *et al.*, "Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals," in *Proc. ISM 2006*, pp. 257–264, 2006.
7) Masataka Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Spe. Comm.*, vol. 43, no. 4, pp. 311–329, 2004.
8) 亀岡弘和 他，スペクトル制御エンベロープによる混合音中の周期および非周期成分の選択的イコライザ，情報処理学会研究報告，2006-MUS-66-13, pp. 77–84, 2006.
9) 後藤真孝 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，情報処理学会論文誌，vol. 45, no. 3, pp. 728–738, 2004.