情報処理学会論文誌 Vol.61 No.4 798–806 (Apr. 2020)

音符区切り情報を用いた高精度歌唱音声認識

鈴木 基之[1,a)] 杉田 裕亮[1,†1]

受付日 2019年7月6日, 採録日 2020年1月16日

概要：歌唱音声を入力として歌詞を認識しようとしたとき，通常の朗読音声に対する音声認識精度と比較して大きく劣化する．その原因の 1 つは，歌詞中の各モーラがそれぞれ音符と対応付けられているため，一部の母音が長音化し，挿入誤りが増加してしまうことである．そこで本論文では，別途推定した音符の区切り情報を用い，認識仮説のモーラ区切り時刻に制約を加えることで歌詞認識の高精度化を目指す．27 名が歌唱した童謡歌唱データを用いて認識実験を行ったところ，単語認識精度が 85.7%から 92.0%と大きく改善した．

キーワード：歌唱音声認識，音符区切り時刻，特殊フレームの挿入，楽曲検索システム

Singing Voice Recognition Using On-set Time Information of Notes

Motoyuki Suzuki[1,a)] Yusuke Sugita[1,†1]

Received: July 6, 2019, Accepted: January 16, 2020

Abstract: Lyrics recognition from singing voice is very difficult, even though speech recognition can be used for practical use. One of the main causes is long duration of vowel. Each mora in lyrics corresponds to each musical note. If a note has long duration (whole note, half note, and so on) then a mora assigned to the note is sung with long duration. It leads to increasing insertion errors. In order to solve this problem, we focus on note boundary information to use it as a constraint of making recognition hypotheses. It is estimated in advance, and speech recognizer considers hypotheses that every note corresponds to each mora. Since it is "hard" restriction, controlling parameters of correspondence are also introduced. The recognition experiment was performed using the Children's song sung by 27 singers. The word accuracy was greatly improved from 85.7% to 92.0%.

Keywords: singing voice recognition, note boundary information, insertion of marker vector, query-by-singing music information retrieval

1. はじめに

近年音楽圧縮技術が向上し，また楽曲配信の手段が CD 等の物理デバイスからネットワークを通した配信に変わったこと等から，大量の音楽データをスマートフォンや mp3 プレイヤー等の小型デバイスに保存，再生することが一般的となった．これらのデバイスは小型であるがゆえに，一般的にインタフェースが貧弱である．その一方で保存可能な曲数は何百曲，何千曲と膨大な数になっているため，保存してある曲の中から特定の 1 曲を指定して再生する，といった操作が非常にやり難い仕様となっている．

こうした問題を解決するため，歌声やハミングを用いて楽曲を検索するシステム（Query-by-Singing/Humming）が開発されてきた [1]．特に歌声を検索キーとして用いる検索システムでは，音の高さや長さといったメロディ情報に加えて，歌声から別途歌詞情報を抽出することで，両方の情報を用いて検索精度を大幅に高めることができる [2]．

しかし，一般に歌唱音声から歌詞情報を高精度に抽出することは困難である．音声波形から歌詞を認識する，という意味においては通常の音声認識とまったく同じ枠組みであるが，通常の（話し声用に開発された）音声認識システ

[1] 大阪工業大学情報科学部
Faculty of Information Science and Technology, Osaka Institute of Technology, Hirakata, Osaka 573–0196, Japan.
[†1] 現在，富士ソフト株式会社
Presently with FUJISOFT INCORPORATED
[a)] moto@m.ieice.org

ムをそのまま用いても，歌声からの歌詞認識精度は非常に低いことが知られている [3]．我々の実験においても，通常の音声認識システムで歌詞を朗読した音声を認識したときの単語正解精度は 88.5%であったが，同じ話者による同じ歌詞の歌唱音声データを認識したところ 12.1%となり，歌唱音声の認識は非常に困難であることが分かる．

その原因の 1 つは，歌唱することによる発声方法の違い [4] である．一般に歌唱音声は通常の話し言葉と比べて音高の変化が激しく，また歌唱フォルマントと呼ばれる独特のフォルマントが出現することがある等，発声方法自体にも違いが見られる．そのため，話し声から学習した音響モデルを用いた音声認識システムではミスマッチがおこり，認識率の低下を招いてしまう．

この問題に対しては，音響モデルを歌唱音声用に適応させることで改善することが可能である．MLLR 等の話者適応用アルゴリズムを用い，話し声用の音響モデルを歌唱音声に適応させることで，認識率が大幅に向上することが報告されている（文献 [3], [5], [6] 等）．

もう 1 つの問題点は，歌唱によって通常ではあり得ないほど長く発声する音素が出現する，ということである．歌唱においては，各モーラは基本的に各音符と対応付いているため，長い音符に対応付けられたモーラは長音として発話される．通常は HMM が音響モデルとして用いられているため，多少の時間変動は自己ループによる遷移で吸収することができるが，あまりにも長い発話の場合，その区間に別の音素が挿入され，結果的に誤認識してしまう場合が多い．尾関ら [3] は，4 分音符以上の長さに対応付けられたモーラを末尾に含む単語についての認識精度が 33.95%であり，全体の認識精度である 58.76%と比べて大幅に低下していると報告している．そのため，歌詞認識を高精度に行うためには，長く発音されたモーラを含む単語をいかに高精度に認識するか，という問題を解決する必要がある．

この問題に対し，主に言語モデルの観点から，いくつか対処法が提案されている．Hosoya ら [5], [7] は，言語モデルとして n-gram ではなく有限状態文法を用いた．楽曲検索システムは自身が持つデータベース中で最も類似した楽曲を検索することから，入力はデータベース中に存在する曲に限られ，歌詞もデータベース中の曲だけを考えればよい．そのため，事前にデータベースに登録されているすべての曲の歌詞を並列に並べた有限状態オートマトンを構築し，それを言語モデルとして用いている．こうすることで，不要な単語の挿入を抑えることができ，n-gram を用いた歌詞認識と比較して，おおよそ 15 ポイント程度の精度向上を実現している．しかしこの方法では，認識文法がデータベース中の歌詞から作成されているため，一部の歌詞を間違えた，といった入力に対しては対応することができず，n-gram 文法と比較して柔軟性に欠けるため，実用に用いるのには難点がある．

また川井ら [8] は，音響モデルの歌唱音声適応に加え，発音辞書に長音化した単語を追加で登録（たとえば「蝶」という単語に対し，/ch o u/だけでなく，/ch o u u/ や /ch o o u u/といった発音を登録）することで，20 ポイント程度の精度向上を果たしている．しかし，どの単語に対してどのような発音変形を加えればよいか，ということは実験的に決めていくしかなく，大規模なシステムを構築するうえでの問題点となる．

以上のように，言語モデルを修正する方法では実用化する上で問題点も多く，また間接的に挿入誤りを低減させようとしているに過ぎないことから，根本的な解決にはなっていない．そこで本論文では，音符の区切り時刻の情報を用い，明示的に各音素長を制限することで，挿入誤りを低減させた歌唱音声認識法 [9], [10] を提案する．

2. 認識仮説におけるモーラ長を制限した音声認識アルゴリズム

2.1 基本的な考え方

歌唱音声における各モーラの継続時間長は，メロディに制約される．各モーラはメロディを表す各音符に対応付いているため，長い音符に対応したモーラは長く，短い音符に対応したモーラは短く発話されることとなる．特に日本語の場合，歌詞を入力とした自動作曲システム（たとえば，文献 [11], [12], [13]）において，基本的にモーラ数と音符数を同数としてメロディを生成していることからも分かるように，1 つの音符に 1 つのモーラが対応付くことが多い．そのため，音符の長さが分かればモーラの長さが推定でき，音声認識時に余計なモーラが挿入されている仮説を棄却することが可能となる．

そこで，入力歌唱中の各音符の区切り時刻を推定し，その時刻でのみモーラ間の遷移を許すようにした音声認識法を提案する．音符の区切り時刻を推定する方法は “On-set detection” と呼ばれ，様々な方法が提案されている [14]．その結果を用い，音符の区切り時刻ではないフレームにおいては，次のモーラへの遷移を許さない（1 つのモーラ区間であるように制限する）ことで，長い音符に対応した区間も 1 つだけのモーラと対応付くことになり，挿入誤りを大幅に低減できると思われる．

2.2 実装の詳細

前節で提案した基本的な考え方は，既存の認識アルゴリズムを変更することなく，実装のレベルで実現可能である（**図 1**）．

まず，通常の音声データから得られる特徴量ベクトルとはかけ離れた値を持つ特殊な特徴量ベクトル（以下，「マーカベクトル」と呼ぶ）を定義する．このマーカベクトルを，入力された歌唱音声から計算された特徴量ベクトル系列中の，音符の区切り時刻に対応するフレーム位置に 1 つずつ

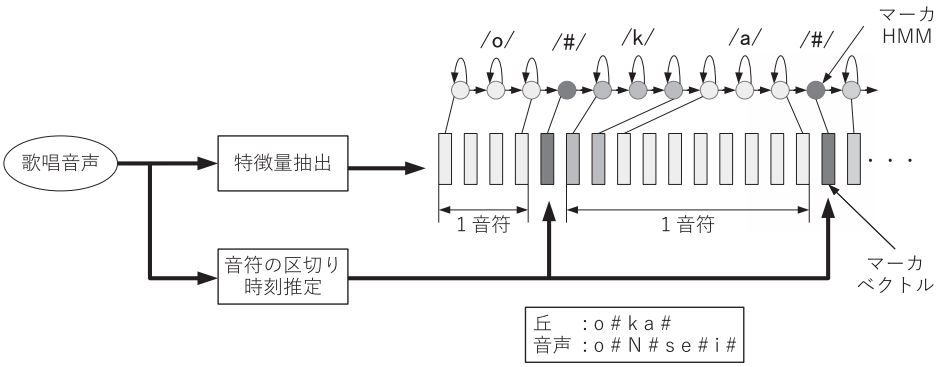

図 1 提案法の概要
Fig. 1 Overview of the proposed method.

挿入していく．こうすることで，特徴量ベクトル系列でどの位置に音符の区切りがあるか，判別可能としておく．

また，音声認識を行う音響モデルに，マーカベクトルに対応した特殊な音素 HMM（以下，「マーカ HMM」と呼ぶ）を 1 つ追加する．このマーカ HMM は自己ループを持たない状態 1 つだけからなる HMM であり，その状態が持つ出力確率分布の平均値はマーカベクトルが持つ特徴量と一致させ，分散は小さな値を定義する．こうすることで，マーカ HMM はマーカベクトルを非常に高い尤度で出力するようになるが，その他の（通常の音声から得られる）特徴量ベクトルは非常に低い尤度でしか出力しなくなる．

最後に発音辞書を変更する．各単語について，モーラの切れ目となる位置にマーカ HMM を挿入する．たとえば「音声」という単語であれば，/o N s e i/という定義を変更し，/o # N # s e # i #/ とする（“#” はマーカ HMM を表す音素記号）．

このようにすることで，認識仮説中のモーラの時間長を，音符長と一致させることが可能となる．発音辞書でマーカ HMM をモーラ間に挿入することで，モーラ間を遷移する際には，必ずマーカ HMM が使用されることになる．しかし，マーカ HMM は特殊な出力確率分布を持つため，マーカベクトルの位置と対応付けないと尤度が極端に低くなり，仮説が棄却されてしまう．結果として，モーラ間の時刻にはマーカベクトル（＝音符の区切り時刻）が対応している仮説のみが残るため，長音化による不要な単語挿入を抑制することが可能となる．

なお，促音はそれ単独で 1 つのモーラであるが，単独で発声することはできない．そのため，促音が単独で 1 つの音符と対応することはないと思われることから，促音は直前のモーラとあわせて「1 モーラ」と定義した．たとえば「あさって」という単語であれば，/a # s a Q # t e #/ と定義した（/Q/ は促音）．

3. 1 対多に対応している歌詞への対処

3.1 1 対多対応への対処法

前節で提案したアルゴリズムでは，歌唱音声中の 1 つの音符は，必ず 1 つのモーラと対応することとなる．しかし一般には，1 つの音符に複数のモーラが対応していたり，逆に複数の音符に 1 つのモーラが対応していたりする曲も存在する．こうした曲の歌唱が入力されると，前節のアルゴリズムでは必ず誤認識が起きてしまう．

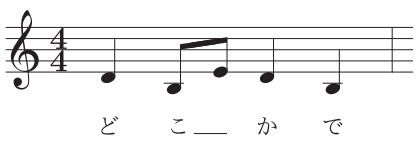

図 2 複数音符に 1 つのモーラが対応している楽曲の例
Fig. 2 Example of a mora corresponding to two notes.

そこで，マーカベクトルが持つ特徴量の値や，マーカ HMM が持つ出力確率分布の値を適切に設定することで，こうした 1 対多対応をした単語に対しても誤認識させない方法を提案する．

3.1.1 1 つのモーラに複数の音符が対応している場合

1 つのモーラに複数の音符が対応している曲（**図 2**）の場合，1 つのモーラ区間（マーカ HMM を含まず，すべて通常の音素 HMM が対応すべき区間）が複数の音符と対応付くため，音符の区切り時刻にマーカベクトルが出現することとなる．この場合，正しくは音素 HMM とマーカベクトルを対応付ける必要があるが，音素 HMM はマーカベクトルを非常に低い尤度でしか出力しないため，この仮説のトータル尤度は極端に低くなってしまう．その結果，この仮説は認識結果としては採用されず，誤認識をおこすこととなる．

こうした誤認識を避けるためには，通常の音素 HMM がマーカベクトルも通常の特徴量ベクトルと同じ程度の尤度で出力するようになればよい．このようにすると，音素 HMM とマーカベクトルを対応付けたとしても尤度の低下がないため，仮説が棄却されることを避けることができる．しかしこの場合，音素 HMM がマーカベクトルと容易に対応付くことが可能となるため，モーラ区切りと対応付けるべきマーカベクトルに対しても（マーカ HMM ではなく）音素 HMM が対応付くことが可能となり，結果として音符

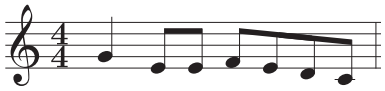

図 3 1 つの音符に複数モーラが対応している楽曲の例
Fig. 3 Example of two morae corresponding to a note.

の区切り時刻を考慮した認識を行うことができなくなる．

そこで，通常の音素 HMM がマーカベクトルを「通常の特徴量ベクトルを出力するときよりは低いが，極端に低くはない」程度の尤度で出力するよう設定する．具体的にはマーカベクトルが持つ特徴量の値を通常の音声とかけ離れた値とするのではなく，ある程度類似した値に設定する．まず，事前に大量の音声データから計算される全特徴量ベクトルから，各次元ごとに平均 μ_i と標準偏差 σ_i を計算しておき，マーカベクトルが持つ i 次元目の特徴量の値を $\mu_i + n\sigma_i$ と設定する．ここで，n は実験的に決められるパラメータである．このようにすることで，(n を適切に設定すれば) マーカベクトルが通常の音素 HMM から「適切な尤度」で出力されるようになる．

なお，ここでは特徴量の値を $\mu_i + n\sigma_i$ と平均値に対して正の方向へとずらしたが，$\mu_i - n\sigma_i$ と負の方向へずらす方法も考えられる．正規分布の対称性から，どちら方向へずらしても認識性能には同程度の影響を与えると考えられるため，今回は正の方向へずらした実験のみを行った．

3.1.2 1 つの音符に複数のモーラが対応している場合

一方，1 つの音符に複数のモーラが対応している曲 (**図 3**) の場合，音符区切りがない区間（マーカベクトルが存在しない区間）にマーカ HMM を対応付ける必要がある．そのためには，マーカ HMM が通常の音声から得られた特徴量ベクトルを極端に低い尤度で出力するのではなく，「通常の音素 HMM が出力するときの尤度よりは低いが，極端に低すぎない」程度の尤度で出力する必要がある．そこで，マーカ HMM が持つ出力確率分布を変更し，通常の特徴量ベクトルに対してもある程度の尤度を与えるようにする．

具体的には，マーカ HMM が持つ出力確率分布の平均値 (＝マーカベクトルが持つ特徴量の値) を 3.1.1 項で説明した方法で通常の特徴量ベクトルの値にある程度近づけたうえで，分散の値を大きくすることで，出力確率分布の平均値から離れたベクトルに対してもある程度の尤度で出力するように変更する．

なお，「マーカベクトルが存在しない区間にマーカ HMM を対応させる」という意味においては，「マーカ HMM 自体をスキップする遷移を付与する」方法も考えられる．しかし，この場合は任意の位置でマーカ HMM をスキップすることが可能となるため，「1 音符の区間は 1 モーラと対応させる」という制約を課すことができなくなる．こうしたことから，マーカ HMM 自体をスキップする遷移は付与せず，マーカ HMM の出力確率分布のパラメータを変更することで対処を行った．

表 1 特徴量と分散を変えることによる認識精度の変動
Table 1 Relationship between parameters of marker HMM and recognition accuracy.

平均ベクトル	通常音声に近い		かけ離れた値	
分散	狭い	広い	狭い	広い
1 音符対 1 モーラ	↑	↓	↑	↑
1 音符対複数モーラ	↓	↑	↓	↓
複数音符対 1 モーラ	↑	↑	↓	↓

3.2 マーカ HMM 等に設定する値と認識性能の関係

すべての曲において，音符とモーラが 1 対 1 の関係であれば，マーカベクトルの特徴量の値は通常の特徴量ベクトルとはかけ離れた値とし，またマーカ HMM が持つ出力確率分布の分散は狭くしておいた方がよい．しかし，3.1 節で指摘したように，実際には 1 対多の対応があるため，ある程度値を通常の特徴量ベクトルに近づけ，分散も広くする必要がある．

このようにした場合，逆に 1 対 1 の対応をしている部分においてはモーラ間の遷移時刻の制約が弱くなってしまうため，誤認識を発生させる可能性がでてくる．ここでは，マーカベクトルが持つ特徴量とマーカ HMM の出力確率分布の分散の設定を変更すると，認識精度がどのように変化するかを定性的に検討する．

マーカ HMM の出力確率分布の平均ベクトルの値 (＝マーカベクトルの値) と分散の設定によって認識精度がどのように変化するか，音符とモーラの対応ごとにまとめた結果を**表 1** に示す．ここで上矢印は認識率が向上する方向であることを，下矢印は低下する方向であることを示している．

まず，1 音符に対して 1 モーラの対応の場合，マーカベクトルが区切りの役割を果すときに精度が向上すると思われる．つまり，マーカベクトルが持つ特徴量の値が通常の特徴量ベクトルからかけ離れている場合と，マーカ HMM が持つ出力確率分布の分散が極端に狭く，通常の特徴量ベクトル（のほとんど）に対して低い尤度を与える場合に認識精度が向上する．

一方，1 音符に対して複数モーラが対応する曲の場合，マーカベクトルのない区間にマーカ HMM が対応付けられることが必要なため，マーカ HMM の出力確率分布が通常の特徴量ベクトルをある程度高い尤度で出力する必要がある．つまり，平均ベクトルが通常の音声に比較的近く，また分散も広くする必要がある．

最後に複数音符に対して 1 モーラが対応する曲の場合，マーカベクトルを含む区間と通常の音素 HMM を対応づける必要があるため，マーカベクトルを通常の音素 HMM がある程度高い尤度で出力する必要がある．そこで，マーカ

ベクトルが持つ特徴量の値を通常の特徴量ベクトルの値に近づける必要がある．

このようにまとめてみると，すべての曲に対して認識性能が向上すると思われる設定は存在しない．そこで，マーカベクトルが持つ特徴量をどの程度通常の特徴量ベクトルに近づけるのか，またマーカ HMM が持つ出力確率分布の分散を極端に狭くするのか，それとも広くするのかについて，実験的に検討する．

4. 歌唱音声の認識実験

4.1 実験条件

4.1.1 歌唱音声データベース

実験には，徳島大学で収録された歌唱音声データベース [15] を使用した．これは，童謡 48 曲について 27 名（男性 19 名，女性 8 名）がアカペラ歌唱を行ったもので，おおよそ 1 名あたり 7 曲，全 198 データを用いた．なお，1 データの長さは一定ではなく，4 小節程度歌唱されたものが多かった．短いもので 2 小節，長いもので 1 曲全体が歌唱されており，歌唱の長さは平均で 6.74 小節であった．これらのデータについて，音符と歌詞中の各モーラがどのように対応しているかを調べた結果を**表 2** に示す．童謡というジャンルの特性上，音符とモーラが 1 対 1 で対応している部分が非常に多くあった．一方で，1 対 2 や 2 対 1 対応している部分が所々見られた．1 対 3 以上の対応関係は存在しなかった．

また，音符の区切り時刻情報については，本来であれば入力歌唱から自動推定を行うべきであるが，今回は提案方法の有効性を確認するため，別途事前に正解となる区切り時刻情報を求め，それを用いた．正解となる区切り時刻情報は，まずデータベース中の全データを用いて monophone を学習し，それを用いて歌唱データの強制アライメントを行った．得られた結果（歌唱データ中の各音素区切り時刻）と楽譜の情報（歌詞と音符の対応関係）から，各音符の区切り時刻を人手によって求めた．

4.1.2 使用した音声認識システム

音声認識システムは julius [16] を用いた．音響モデルは，評価用歌唱者を含め，27 名全員分の全歌唱データから話者・発話ともにクローズな多数話者 monophone を学習した．言語モデルは 48 曲の歌詞データのみから trigram を学習して用いた．使用した単語数は 314 単語であり，未知語なしの条件である．この言語モデルはテストデータに対してクローズなものであることから，実用の際に用いられるであろう言語モデルと比較して認識性能の向上に強く寄与すると思われる．しかし，Hosoya ら [5], [7] も述べているように，音楽検索システムでは基本的に入力される曲はデータベース中の曲であるため，データベース中の曲のみから言語モデルを構築することは実用の場面でも有効であると思われること，また，一般の（歌詞に特化していない）言語モデルを用いると認識性能の大幅な劣化が予想されるため，今回の提案方法による性能変化が表れにくくなる可能性があることから，こうした言語モデルを用いた．なお，同じ言語モデルをすべての実験で用い，言語重み，挿入ペナルティといったパラメータも固定することで，認識性能の比較に影響を与えないようにした．今回は 48 曲の歌詞データから言語モデルの学習を行ったが，より大規模な歌詞データから学習した言語モデルによる性能評価，ならびに，学習曲に含まれない曲が歌唱されたときの性能評価は，今後の課題である．

なお，歌唱音声の特性上，休符や息継ぎ等で無音区間が挿入される可能性があるが，本実験では特別な対処は行わなかった．無音 HMM に対応する単語は発音辞書内で定義されているため，無音区間が単語境界にあれば，適宜無音 HMM が挿入される可能性があるが，単語の途中で息継ぎをする等の歌唱であれば無音 HMM は挿入されず，いずれかの音素の一部として認識される．こうしたことは認識率の低下を招くため，何らかの対処が必要であると思われる．この問題も今後の課題である．

また，マーカ HMM の出力確率分布は単一正規分布とし，その平均ベクトルは 3.1.1 項で説明したように，全歌唱データの特徴量ベクトルの平均とその標準偏差を元に，いくつかの場合を設定した．共分散行列は対角共分散とし，各次元の分散の値はすべての次元で 10^{-4} としたもの（以下，「狭い分散」と表記）と，すべての歌唱音声から学習したすべての音素 HMM のすべての共分散行列の値の平均値を各次元ごとに計算したもの（以下，「平均的分散」と表記）の 2 種類を設定した．本来は分散の値もより細かく設定し，その値と認識性能との関係について分析すべきであるが，今回は実験結果が煩雑になってしまうことから，その点については今後の課題とした．

その他の詳細設定は**表 3** に示す．なお今回の言語モデルはデータベース中すべての歌詞データから構築していることから，表中の「Perplexity」はいわゆる「テストセットパープレキシティ」ではなく，学習データ（すべての歌詞データ）に対するパープレキシティを示している．

表 2 音符とモーラの対応別のデータ数内訳
Table 2 Breakdown of the data by correspondence.

		2 音符対 1 モーラの対応	
		含まない	含む
1 音符対 2 モーラの対応	含まない	103	21
	含む	61	13

4.2 マーカベクトル挿入による効果

まず，音符の区切り時刻位置にマーカベクトルを挿入することによる効果を検証した．マーカベクトルの特徴量（＝マーカ HMM の出力確率分布の平均ベクトルの値）は，

通常の特徴量としては表れないような値（10^5）に設定し，またマーカ HMM の出力確率分布の共分散行列は「狭い分散」とした．

各条件における単語正解精度を**図 4** に，また認識結果の例を**表 4** に示す．これを見ると，1 音符に対して 1 モーラの対応しか含まない曲（“1 vs. 1”）に対しては目論見どおり認識精度が大幅に向上していることが分かる．認識結果の例を見ても，長音化している「ド」が，「ドーナツ」と認識されるのではなく，正しく「ド」と認識されていることが分かる．

一方で，1 音符に対して 2 モーラの対応を含む曲（“1 vs. 2”）や 2 音符に対して 1 モーラの対応を含む曲（“2 vs. 1”），また，それら両方の対応を含む曲（“Both”）については，大幅な認識精度の低下が見られる．これも 3.2 節で予想したとおりの結果であり，1 音符に対応している「どん」が「ド」と認識されたり，2 音符に対応している「こ」が「こお」と 2 モーラに認識されてしまったりしている．

最終的に 198 データ全体（“All data”）としては，対応付けたことによる効果よりも多対 1 の対応部分における弊害の方が大きな影響があり，認識精度はマーカベクトルを挿入することで低下してしまっている．

表 3 音声認識システムの詳細設定
Table 3 Experimental setup of speech recognizer.

認識ソフトウェア	julius ver.4.3.1
音響モデル	GMM-HMM monophone
モデル数	34 個
状態数	3 状態
混合数	16 混合
音響特徴量	25 次元
MFCC	12 次元
ΔMFCC	12 次元
Δ 対数パワー	1 次元
言語モデル	単語 triphone
平滑化	Back-off スムージング
ディスカウント	Witten-Bell
Perplexity	2.95
評価基準	単語正解精度

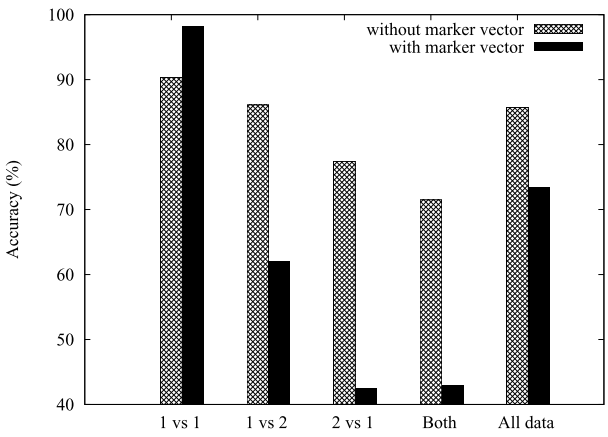

図 4 マーカベクトルの有無による認識精度比較
Fig. 4 Comparison of word accuracy with and without marker vector.

4.3 マーカ HMM の出力確率分布のパラメータ調整による精度変化

多対 1 対応の部分における誤認識を減らすため，マーカ HMM の出力確率分布の平均ベクトルと共分散行列の値を変化させ，そのときの認識精度を調べた．i 次元目の平均ベクトルの値は，通常の特徴量としては表れないような値（10^5）に加え，$\mu_i + n\sigma_i$ において n を 0〜3 まで変化させた値で実験を行った．一方共分散行列の値は，「狭い分散」と「平均的分散」の 2 通りを設定した．

4.3.1 1 音符と 2 モーラが対応している曲の認識精度

まず，1 音符に対して 2 モーラが対応している部分を含む曲（61 データ）についての単語正解精度を**図 5** に示す．この場合，マーカ HMM と通常の特徴量ベクトルが対応付く必要があるため，表 1 で示したとおり，分散を広くとり，平均ベクトルを通常の音素 HMM に近づけることで対処することができる．結果を見ると，「平均的分散」を設定したうえで，平均ベクトルの値が $\mu + 2\sigma$ 程度までであれば，高い認識性能を示すことが分かった．

なお，今回用いた楽曲において 1 音符に 2 モーラが対応していた箇所は全部で 236 カ所あったが，そのうち 109 カ所は 2 モーラ目が撥音であるもの（ドン，ラン等）であった．こうした音節は比較的 1 音符と対応付けられることが多いと予想されることから，モーラごとにマーカ HMM を挿入した発音辞書（/d o # N #/等）に加え，撥音との間のマーカ HMM を除いた発音辞書（/d o N #/等）も併記しておくことで，認識性能を向上させる可能性があると思

表 4 マーカベクトルの有無による認識結果の例
Table 4 Example of recognition results.

音符とモーラの対応	マーカベクトル	認識結果
1 音符対 1 モーラ	なし	ド は ドーナツ の ドーナツ で は レモン の レ
	あり	ド は ドーナツ の ド レ は レモン の レ
1 音符対 2 モーラ	なし	どんぐり コロコロ どんぶりこ
	あり	ド 栗 コロコロ どこ か と
2 音符対 1 モーラ	なし	どこ か で 春 が
	あり	どこ お か で 春 る が

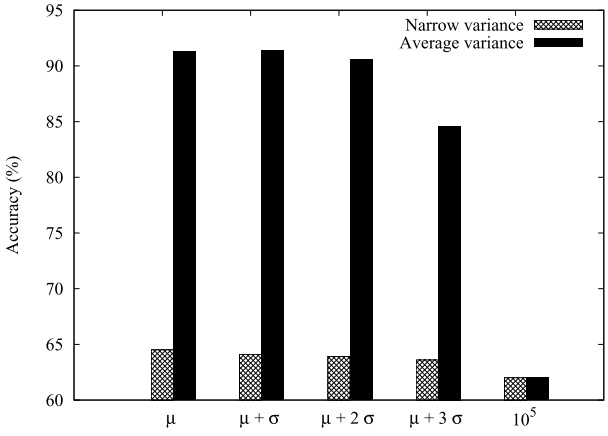

図 5 1 音符対 2 モーラの対応を含む曲の精度変化

Fig. 5 Word accuracy for songs containing a note corresponding to two morae.

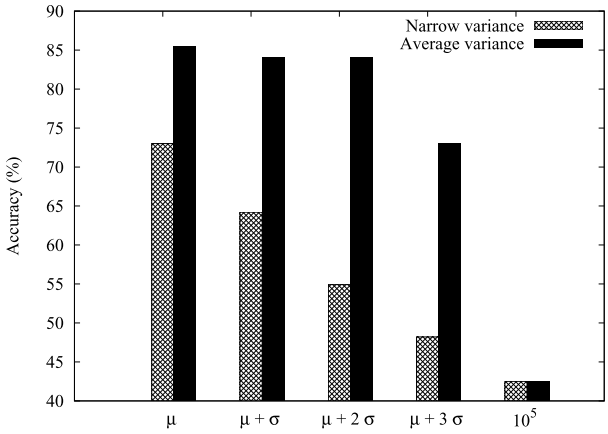

図 6 2 音符対 1 モーラの対応を含む曲の精度変化

Fig. 6 Word accuracy for songs containing two notes corresponding to a mora.

われる．

4.3.2 2 音符と 1 モーラが対応している曲の認識精度

次に，2 音符に対して 1 モーラが対応している部分を含む曲（21 データ）についての単語正解精度を**図 6** に示す．こちらは，マーカベクトルと通常の音素 HMM を対応付ける必要があるため，表 1 で示したとおり，マーカベクトルの値を通常の特徴量ベクトルに近づける必要がある．結果を見ると，平均ベクトルの値が $\mu + 2\sigma$ 程度までであれば，高い認識性能を示すことが分かった．また共分散行列については，平均ベクトルの値が平均値に近ければ，「狭い分散」でもそれなりの認識率を示すが，平均値から離れていくに従って認識率が低下していくことが分かった．これは，分散が狭いことでマーカベクトルに対してマーカ HMM が与える尤度が非常に高いものとなるため，平均ベクトルの値が平均値から離れていくに従って通常の音素 HMM がマーカベクトルに与える尤度との差が大きくなり，マーカベクトルとマーカ HMM が対応付く（1 音素に 1 モーラを対応付ける）仮説の方が尤度が高くなっていくためと思われる．そのため，分散の値は「平均的分散」とし，過度にマーカ HMM が与える尤度が高くならないようにしておく必要があることが分かった．

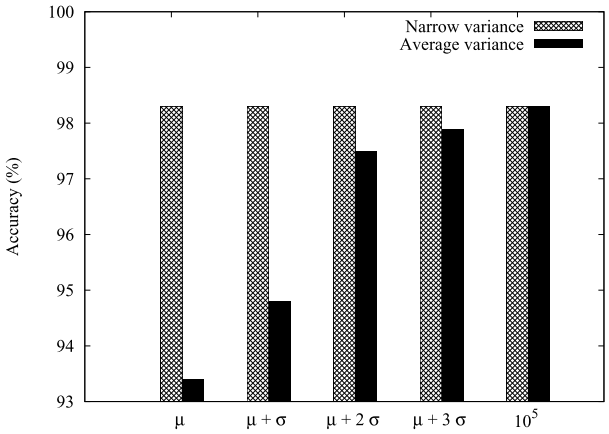

図 7 1 音符対 1 モーラの対応のみである曲の精度変化

Fig. 7 Word accuracy for songs comprising only notes each of which corresponds to a mora.

4.3.3 1 音符対 1 モーラの対応のみである曲への影響

平均ベクトルや共分散行列の値を変化させるのは，1 対多対応している部分への対処のためであり，1 対 1 対応している部分については，逆に悪影響が及ぶ可能性がある．そこで，1 対 1 対応しか含まない曲について，平均ベクトル等の値を変化させたときの影響を調べた．

1 対 1 対応しか含まない曲（103 データ）についての単語正解精度を**図 7** に示す．これを見ると，「狭い分散」を設定することで，平均ベクトルの値にかかわらず高い認識精度を示すことが分かる．「狭い分散」にすることでマーカベクトルに対して非常に高い尤度をマーカ HMM が与えるため，マーカベクトルが通常の音素 HMM と対応付くこともなく，安定して高い認識精度を出していると思われる．

一方で，「平均的分散」に設定すると，マーカ HMM が与える尤度が通常の音素 HMM が与える尤度と大きな差がなくなるため，平均ベクトルの値を通常の特徴量ベクトルと異なる値にしていかないと，マーカベクトルが区切りとしての役割を果たせなくなる．実験結果から，$\mu + 2\sigma$ 程度離せば区切りとしての役割を果たすことができ，高い認識精度を示すことが分かった．

4.3.4 全曲に対する結果

最終的に，全曲を用いたときの単語正解精度は，**図 8** のようになった．この結果は，評価データ中に多対 1 対応している部分がどの程度含まれるかによって変化すると思われるが，今回のデータベースの場合では，共分散行列を「平均的分散」としたうえで，平均ベクトルを $\mu + 2\sigma$ とすることで，92.0%と最も良い認識精度が得られた．音符の区切り情報を用いない場合の認識精度が 85.7%であったことから，適切なパラメータ設定を行うことで，6.3 ポイントの認識精度の改善を得ることができた．

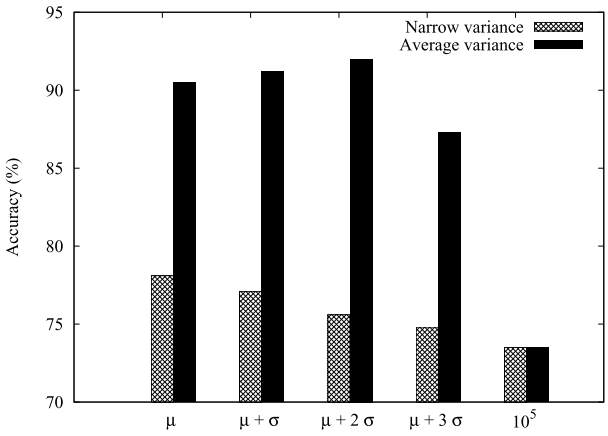

図 8 全曲まとめたときの精度変化
Fig. 8 Word accuracy for all songs.

表 5 音素を単位とした認識結果
Table 5 Phoneme accuracy for all songs.

	正解精度	置換誤り	脱落誤り	挿入誤り
従来法	93.6%	1.0%	0.9%	4.5%
提案法	96.7%	0.9%	0.6%	1.8%

また，本論文で提案した方法は，音符の区切り時刻情報を用いることで主にモーラの挿入誤りを抑制することを目的としていたため，音素に注目して認識結果の分析を行った．一番単語正解精度が高かった条件（平均ベクトルは $\mu+2\sigma$，分散の値は「平均的分散」）と，音符の区切り情報を用いない方法（従来法と表記）について，音素を単位とした正解精度等の結果を**表 5** に示す．

この結果を見ると，挿入誤りが大きく減少していることが分かる．このことから，音符の区切り時刻情報を用いることでモーラの挿入誤りを減らす，という目論見どおりの結果となっていることが分かった．また，脱落誤りも削減されていた．これは，今回の言語モデルに登録されている語彙数が少なかったことから，従来法においては音素脱落を含むような単語に置換された結果，音素の脱落誤りも増加してしまっていたと思われる．

4.4 自動推定した音符区切り時刻情報を用いた認識

前節までの実験では，音符の区切り時刻は人手によって求められた正解情報を与えていた．しかし実用の際は音符の区切り時刻を自動で推定する必要がある．その方法はいくつか提案されている [14] が，推定精度はいまだ 100%ではないため，マーカベクトルが誤った位置に挿入され，認識性能が低下してしまう可能性がある．そこで本節では，実際に自動推定された位置にマーカベクトルを挿入し，どの程度認識性能が低下するか実験によって検証した．

音符の区切り時刻の推定には，Rectified complex deviation [17] を指標として用いた OnsetsDS [18], [19] を用いた．また，共分散行列の値は「平均的分散」とし，平均ベクトルの値をいくつか変更しながら認識性能を比較した．

表 6 音符の区切り時刻を自動推定したときの単語正解精度
Table 6 Word accuracy using estimated note boundary.

平均ベクトルの値	音符の区切り時刻情報		
	使用せず	正解の値	自動推定
μ	85.7%	90.5%	89.4%
$\mu+\sigma$		91.2%	87.7%
$\mu+2\sigma$		92.0%	72.9%

各条件での単語正解精度を**表 6** に示す．これを見ると，前節で最も良い性能を示していた条件（平均ベクトルの値を $\mu+2\sigma$ としたもの）では，正解位置にマーカベクトルを挿入したときと比較して，20 ポイント近く正解精度が低下していることが分かる．このことから，音符の区切り時刻の推定誤りが，大きく影響を与えていることが分かった．

一方，平均ベクトルの値を平均値（μ）に近づけると，正解精度が改善している．これは，マーカベクトルの値を通常の音声から計算された特徴量ベクトルの値に近づけた結果，マーカベクトルと通常の HMM が対応しやすくなり，その結果マーカベクトルが誤った時刻に挿入されていたとしても，影響が小さくなったからだと思われる．一部の（区切り時刻が正しく推定された）マーカベクトルに関しては，マーカ HMM と対応させることで通常の HMM と対応させるよりは尤度が高く計算されるため，そうした認識仮説が採用されることで正解精度が向上したのではないかと推測される．

音符の区切り時刻情報を用いない方法と比較すると認識率は向上しているが，正解の区切り情報を用いた場合と比較して性能は低下しており，またこのような設定では，区切り時刻の情報を十分生かしているとはいえない．そのため，今後認識結果の詳細を分析し，より推定誤りに頑健な方法に改良していく必要があると思われる．

5. まとめ

本論文では，歌唱音声中の歌詞情報を高精度に認識するため，歌唱中で長音化している場所での挿入誤りを減らす方法を提案した．基本的に歌詞中の 1 モーラは 1 音符と対応していることに注目し，別途推定した音符の区切り時刻でのみモーラ間の遷移を許すように認識アルゴリズムを変更した．具体的には，通常の音声から得られる特徴量とはかけ離れた値を持つ「マーカベクトル」を，入力音声から得られた特徴量ベクトル系列中の音符の区切り時刻の位置に挿入する．また，マーカベクトルの値を出力確率分布の平均値に持つ 1 状態の「マーカ HMM」を定義し，発音辞書において，すべてのモーラ間にマーカ HMM を挿入する．こうすることで，強制的に 1 モーラが 1 音符と対応することとなり，長音化による挿入誤りを抑制することを可能とした．

また，曲によっては 1 音符が 2 モーラに対応したり，2 音符が 1 モーラに対応したりすることもあることから，マーカ HMM の出力確率分布の平均ベクトル（＝マーカベクトルが持つ特徴量の値）と共分散行列の値を適切に設定することで，こうした多対 1 の対応にも対応する方法を提案した．

27 名が歌唱した童謡歌唱データを用いて認識実験を行ったところ，マーカ HMM の出力確率分布に適切な値を設定することで，単語認識精度を 85.7%から 92.0%へと大きく改善させることができた．なお今回の実験では童謡を用いたが，J-POP 等，他ジャンルの楽曲に対する有効性も，今後検討していく予定である．

謝辞 本研究の一部は JSPS 科研費 JP25330140, JP18K11321 の助成を受け行われた．

参考文献

[1] Typke, R., Wiering, F. and Veltkamp, R.: A Survey of Music Information Retrieval Systems, *Proc. ISMIR*, pp.153–160 (2005).

[2] 伊藤彰則，鈴木基之，牧野正三：この曲、何だっけ？歌で音楽を探す「歌声検索」，*DTM MAGAZINE*, Vol.183, pp.102–103 (2009).

[3] 尾関弘尚，鎌田貴幸，後藤真孝，速水 悟：歌声の歌詞認識における音高の影響について，2003 年秋季音講論集 1–1–1，日本音響学会 (2003).

[4] Mesaros, A. and Virtanen, T.: Automatic Recognition of Lyrics in Singing, *EURASIP Journal on Audio, Speech, and Music Processing*, Vol.2010 (online), DOI: 10.1155/2010/546047 (2010).

[5] Hosoya, T., Suzuki, M., Ito, A. and Makino, S.: Lyrics Recognition from a Singing Voice based on Finite State Automaton for Music Information Retrieval, *Proc. ISMIR*, pp.532–535 (2005).

[6] 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, *Proc. ICASSP 2005*, Vol.I, pp.237–240 (2005).

[7] Suzuki, M., Hosoya, T., Ito, A. and Makino, S.: Music Information Retrieval from a Singing Voice Using Lyrics and Melody Information, *EURASIP Journal on Advances in Signal Processing*, Vol.2007 (online), DOI: 10.1155/2007/38727 (2007).

[8] 川井大陸，山本一公，中川聖一：DNN-HMM を用いた歌声の自動歌詞認識の検討，音楽情報科学研究会研究報告，Vol.2015-MUS-107, No.58, pp.1–6 (2015).

[9] 鈴木基之，杉田裕亮：音符区切り情報を用いた高精度歌唱音声認識，音楽情報科学研究会研究報告，Vol.2017-MUS-115, No.22, pp.1–6 (2017).

[10] Suzuki, M. and Sugita, Y.: Lyrics recognition from singing voice focused on note boundary, *Proc. 2017 International Conference for Top and Emerging Computer Scientists* (*IC-TECS 2017*), p.31 (2017).

[11] 深山 覚，中妻 啓，酒向慎司，西本卓也，小野順貴，嵯峨山茂樹：音楽要素の分解再構成に基づく日本語歌詞からの旋律自動作曲，情報処理学会論文誌，Vol.54, No.5, pp.1–12 (2013).

[12] 白石 隆，管 秀俊，内田 理，菊池浩明，中西祥八郎：言語情報と和声進行に基づく旋律の自動生成システム，第 70 回全国大会 3X-2，情報処理学会 (2008).

[13] 中村千紗，鬼沢武久：ユーザの好みを考慮した作詞作曲システム，第 24 回ファジィシステムシンポジウム WA1-3，日本知能情報ファジィ学会 (2008).

[14] Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. and Sandler, M.B.: A Tutorial on Onset Detection in Music Signals, *IEEE Trans. Speech and Audio Processing*, Vol.13, No.5, pp.1035–1047 (2005).

[15] 鈴木基之，岡松竜徳，任 福継：音程に注目した歌唱音声中の音符区間推定，音楽情報科学研究会研究報告，Vol.2010-MUS-85, No.9, pp.1–6 (2010).

[16] Lee, A., Kawahara, T. and Shikano, K.: Julius — An open source real-time large vocabulary recognition engine, *Proc. EUROSPEECH*, pp.1691–1694 (2001).

[17] Dixon, S.: Onset detection revisited, *Proc. 9th International Conference on Digital Audio Effects*, pp.133–137 (2006).

[18] Stowell, D. and Plumbley, M.: Adaptive whitening for improved real-time audio onset detection, *Proc. International Computer Music Conference*, pp.312–319 (2007).

[19] Cannam, C. and Stowell, D.: OnsetsDS (2007), available from ⟨https://code.soundsoftware.ac.uk/projects/vamp-onsetsds-plugin⟩.

鈴木 基之 （正会員）

1993 年東北大学工学部情報工学科卒業．1995 年同大学大学院博士前期課程修了．1996 年同大学院博士後期課程を退学し，同大学大型計算機センター助手．2006～2007 年英国エジンバラ大学客員研究員．2008 年徳島大学大学院ソシオテクノサイエンス研究部准教授，2012 年大阪工業大学情報科学部准教授，2017 年同教授．博士（工学）．音声認識・理解，音楽情報処理，自然言語処理，感性情報処理等の研究に従事．電子情報通信学会，日本音響学会，人工知能学会各会員．

杉田 裕亮

2017 年大阪工業大学情報科学部卒業．在学中は歌唱音声認識の研究に従事．