Vol.2018-MUS-119 No.39
Vol.2018-SLP-122 No.39
2018/6/17

音符区切り位置の推定誤りに頑健な高精度歌唱音声認識

鈴木 基之[1,a)]　富田 翔[1,†1]

概要：歌唱音声から歌詞を認識する際，一部の母音の長音化に対応させるため，音符の区切り時刻の情報を用いた歌唱音声認識法を以前提案した。しかしこの方法では，区切り時刻の推定誤りに弱く，精度が低下してしまう。そこで本研究では，音符の区切り時刻らしさを連続値として扱うことで，推定誤りに頑健な歌唱音声認識法を提案する。27 名が歌唱した童謡歌唱データによる実験では，単語正解精度が 89.6%となり，誤りを含む区切り時刻による従来方法での認識精度（72.8%）を大きく上回ることが示された。

キーワード：歌唱音声認識，音符の区切り時刻，楽曲検索システム．

1. はじめに

歌声を用いて楽曲を検索するシステム（Query-by-Singing）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して検索に用いることで，検索精度を大幅に高めることができる [1]。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り，通常の（話し声用に開発された）音声認識システ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 [2]。

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たり，認識可能な文章を検索対象曲の歌詞に限るように言語モデルに制限を加えることで精度向上を実現している [3]。また，長音化した母音に対応させるため，発音辞書を修正する方法 [4] も提案されている。

しかし，こうした方法では根本的な解決にはなっておらず，また精度も十分とは言えない。この問題に対し，歌詞とメロディの対応関係に注目し，歌唱音声中の音符の区切り時刻情報を用いて認識精度を向上させる方法が提案された [5]。この方法では，特徴量ベクトル系列において，別途推定された音符の区切り時刻に特殊な「マーカーベクトル」を挿入し，これに対応する「マーカー HMM」を発音辞書上でモーラ区切りに挿入することで，1 音符と 1 モーラを強制的に対応させた認識結果を導くものである。このようにすることで，不自然に長く発音された母音に対する挿入誤りを削減でき，結果として認識精度を向上させることができた。

[1] 大阪工業大学 情報科学部
[†1] 現在，システムズデザイン株式会社
[a)] moto@m.ieice.org

この方法では，事前に音符の区切り時刻を推定する必要があるが，その精度は 100%ではないため，誤った位置へのマーカーベクトルの挿入を避けることはできない。これに対し，パラメータを調整することで，ある程度はマーカーベクトルと通常の音素 HMM（あるいは，マーカー HMM と通常の特徴量ベクトル）を対応させることはできるが，推定誤りの影響を無くすことはできない。実際に行った歌唱音声認識においても，正しい音符の区切り時刻を用いた時は単語認識率が 92.0%であったのに対し，誤りを含む推定時刻を用いた場合は 72.8%まで精度が低下した。

音符の区切り時刻を推定する多くのアルゴリズム [6] では，なんらかの指標となる値をフレームごとに計算し，その値が閾値を越えた時刻を「区切り時刻」と推定している。そのため，推定された区切り時刻の情報だけを用いるのではなく，その元となった「区切り時刻らしさ」の指標の値を直接用いた方が，より柔軟で適切な対応づけを行うことができると思われる。

そこで本論文では，音符の区切り時刻を推定するための指標の値を用い，区切り時刻らしさをスコア計算に含めることで，音符の区切り時刻の推定誤りにも頑健な歌唱音声認識法を提案する。

2. 音符の区切り時刻らしさを利用した音声認識アルゴリズム

マーカーベクトルを用いる方法 [5] では，推定された音符の区切り時刻にマーカーベクトルを挿入する。この時，マーカーベクトルは「挿入する」か「挿入しない」かの 2 択であり，その推定位置の「確からしさ」は考慮されていない。そのため，たまたま音符の区切り時刻ではない位置

を区切り時刻として推定してしまうと，そこにマーカーベクトルが挿入されることからモーラとの対応関係が崩れ，認識結果に大きな影響を与えてしまう。

一般に音符の区切り時刻はなんらかの指標（ODF : Onset Detection Function）となる値（ODF 値）を元に推定している。そのため，音符の区切り時刻であると誤推定した時も，ODF 値は正しく推定した時と比べて，比較的低い値である（閾値と比較して，ぎりぎり越えていた）ことが予想される。また，区切り時刻であるにもかかわらず，「区切り時刻ではない」と誤推定した場合も，同様にその指標は比較的高い値であることが多いと予想される。

そこで，推定された区切り時刻の情報を用いるのではなく，その前段階の ODF 値を直接利用することで，閾値の設定にとらわれず，推定の「確からしさ」も考慮した認識を実現する。

具体的なアルゴリズムは以下のとおりである。

(1) 特徴量ベクトルの次元拡張
音声認識に用いる特徴量ベクトルの次元を拡張し，音符の区切り時刻らしさの指標を含んだ特徴量ベクトルを生成する。
まず，入力された歌唱音声に対して音符の区切り時刻推定を行い，ODF 値をすべてのフレームに対して計算する。その後，ODF 値を 0 ～ 1 の範囲（1 に近ければ区切り時刻らしい）となるように正規化し，各フレームごとに別途計算された特徴量ベクトルに 1 次元追加することでひとつのベクトルとして表す。

(2) 音素 HMM の出力確率分布を拡張された次元に対応
すべての音素 HMM の出力確率分布の次元数を 1 つ増やし，平均値を 0 に設定する。また分散の値は，他の次元の分散の値を参考に，平均的な分散値を設定する。

(3) マーカー HMM の作成
音符の区切り時刻に対応させる特殊な HMM（マーカー HMM）をひとつ定義する。この HMM は 1 状態で自己ループがなく，出力確率分布は全音声データを用いて（平均的な音声を表すように）推定する。その後，音素 HMM と同じように 1 次元拡張し，平均値には 1 を，分散は音素 HMM と同様に平均的な分散値を設定する。
こうすることで，マーカー HMM は音符の区切り時刻らしい特徴量ベクトルを高い尤度で出力する（対応づけられる可能性が高くなる）一方，その他の（通常の音声から得られる）特徴量ベクトルに対しては低い尤度を出力することとなる。

(4) 発音辞書へのマーカー HMM の記述
マーカー HMM がモーラ間の遷移時刻に対応するよう，発音辞書中に記述されているすべての単語について，モーラの切れ目となる位置にマーカー HMM を挿入する。例えば「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（ “#” はマーカー HMM を表す音素記号）。

このようにすることで，推定された ODF 値を認識仮説の尤度に反映させることができるようになる。発音辞書でマーカー 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，必ずマーカー HMM が使用されることになる。この時，マーカー HMM の出力確率分布において拡張された次元の値は 1 となっているため，特徴量ベクトルでの同次元の値が 1 に近い（音符の区切り時刻らしいと推定されている）ほど，高い尤度となる。このようにして音符の区切り時刻らしさが認識仮説の尤度に反映されるため，音響的な類似度（音素らしさ）とあわせ，総合的に最も妥当である認識仮説を探索することができる。一方マーカーベクトルを挿入する方法では，強制的にマーカーベクトルとマーカー HMM を対応づけることから，推定誤りの影響を強く受けてしまう。そのため，次元を拡張する方法はマーカーベクトルを挿入する方法と比較して，推定誤りに頑健に認識を行うことが可能となる。

表 1 実験条件

歌唱音声データベース	徳島大学歌唱データベース
使用した曲	童謡 48 曲
歌唱者	男性 19 名，女性 8 名
データ数	198 データ
音声認識システム	julius[7]
音響モデル	歌唱から学習した monophone
言語モデル	歌詞のみから学習した trigram
語彙数	314 単語（未知語なし）
音符の区切り時刻推定	OnsetsDS[8], [9]
ODF	Rectified complex deviation[10]

表 2 音符とモーラの対応別のデータ数内訳

		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

3. 歌唱音声の認識実験

3.1 実験条件

実験条件を表 1 に示す。実験には，徳島大学で収録された歌唱音声データベース [11] 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。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。

音響モデルは，評価用歌唱者 1 名のデータを除いた 26

Vol.2018-MUS-119 No.39
Vol.2018-SLP-122 No.39
2018/6/17

表 3　単語正解精度

区切り時刻情報	推定値	正確な値	未使用
次元拡張	89.6%	—	85.7%
マーカーベクトル	72.8%	92.0%	

名分の全歌唱データから monophone を学習し，評価用歌唱者を変えながら 27 回の認識実験を行い，結果を平均した。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。

3.2　歌唱音声の認識実験結果

歌唱音声の認識結果を表 3 に示す。この表において，「次元拡張」は本論文で提案している特徴量ベクトルの次元を拡張し，そこに ODF 値を入れたもの，「マーカーベクトル」は推定された音符の区切り時刻にマーカーベクトルを挿入する方法 [5] である。また，「区切り時刻情報」欄において「推定値」は自動で推定した ODF 値を用いたもの，「正確な値」は人手で付与した正確な音符の区切り時刻を利用したもの，また「未使用」は区切り時刻の情報を使用せず，通常の音声認識アルゴリズムで認識を行なったものである。

この表を見ると，音符の区切り時刻の情報を利用しない（通常の音声認識と同じ）方法に比べ，マーカーベクトルを用いた方法は，正確な値を用いると認識率が 6.3 ポイント向上していることがわかる。このことから，音符の区切り時刻の情報を利用することの有効性が見られる。しかし，誤りを含むような情報を用いると，認識率は 72.8%と，大幅に低下してしまい，音符の区切り情報を用いない方が認識率が高い，という結果になってしまう。

それに対し，次元拡張を行う方法では 89.6%と，通常の音声認識を行った場合と比較して 3.9 ポイント認識率が向上しており，正確な値を用いた場合と比較しても，認識率の低下を 2.4 ポイントに抑えていることがわかる。この事から，次元拡張を行う方法では，音符の区切り時刻の推定誤りに頑健な認識を行うことができていることがわかった。

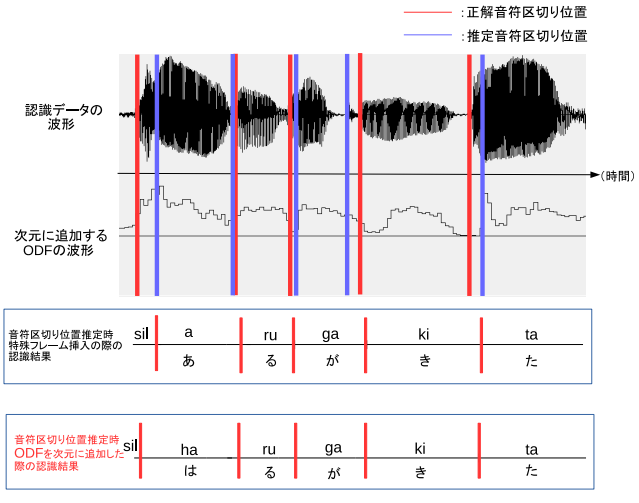

図 1　認識結果の例

表 4　複数次元拡張を行った時の認識率

次元数	$N+1$	$N+2$	$N+3$	$N+5$	$N+10$
認識率	89.6%	89.5%	89.1%	89.0%	88.0%

提案する方法で得られた認識結果の例を図 1 に示す。この図において，一番上は歌唱音声の時間波形，その下は推定された ODF 値を示す。また，青い線は ODF 値を元に自動で推定された音符の区切り時刻，赤い線は人手で与えられた正確な区切り時刻を示す。

この図を見ると，いくつかの場所で推定された音符の区切り時刻がずれていることがわかる。特に歌唱音声の最初の部分では，/ha/ の /h/ の部分が落ちてしまい，/a/の先頭部分の時刻を音符の区切り時刻として推定している。その結果，マーカーベクトルを挿入する方法では，先頭の /h/ を認識することができず，「あ」と誤認識してしまっている。一方次元拡張を行う方法では，/h/ の区間も ODF 値はそこそこの値を示しているため，/h/ の音響的な類似性とあわせ，「は」と正しく認識することができた。

このように，「最も音符の区切り時刻らしい時刻のみ」にマーカーベクトルが挿入されるのに対し，「ある程度区切り時刻らしい区間」全体で ODF 値が高くなっていることから，推定された区切り時刻のずれに頑健に認識することが可能である。

3.3　尤度に対する重みの変更

本方法では，次元を 1 次元拡張することで音符の区切り時刻情報を尤度に反映させている。その尤度に対する重みは，元々の特徴量ベクトルが N 次元であるとすれば，おおよそ $\frac{1}{N+1}$ となる。この重みを変更することで，より区切り時刻の情報を重視した認識も可能となると思われる。そこで，ODF 値に対する重みを変更し，認識率にどのような影響が出るかを検証した。

ODF 値に対する重みを変更する方法はいくつか考えられるが，ここでは簡易的に実装できる方法として，複数次元への拡張を行った。次元を 1 次元ではなく，2 次元，3 次元と複数次元拡張し，そのすべての次元に ODF 値をコピーして設定した。こうすることで，n 次元拡張すると，その重みは $\frac{n}{N+n}$ となり，次元を増やせば増やすほど重みをかけた認識を行うことができる。

複数次元拡張した時の認識率を表 4 に示す。これを見ると，10 次元まで拡張していくと認識率は微減していくことがわかる。ODF 値の推定精度によって最適な拡張次元数は変化すると思われるが，今回用いたデータベースと ODF 値では，拡張する次元は 1 次元が最適であることがわかった。

4. まとめ

歌唱音声中の歌詞情報を高精度に認識するため，音符の区切り時刻情報を用いて，歌唱中で母音が長音化している区間での挿入誤りを減らす方法が提案されている。この方法では，推定された音符の区切り時刻に特殊なマーカーベクトルを挿入することで，歌詞中の 1 モーラを 1 音符と対応させている。そのためこの方法は音符の区切り時刻の推定誤りに弱く，精度が低下してしまう。

そこで本研究では，音符の区切り位置らしさを連続値として扱うことで，推定誤りに頑健な歌唱音声認識法を提案した。区切り時刻を推定する際に計算される「音符の区切り時刻らしさ」を表す ODF 値を用い，それを特徴量ベクトルの次元を拡張して設定する。こうして得られた特徴量ベクトルを用いて認識することで，音響的な類似性に加えて音符の区切り時刻らしさも考慮した認識仮説を導出することが可能となる。

27 名が歌唱した童謡歌唱データによる実験では，単語正解精度が 89.6%となり，誤りを含む区切り時刻による従来方法での認識精度（72.8%）を大きく上回ることが示された。

この方法では，特徴量ベクトルの次元を拡張することで，音符の区切り時刻の情報を特徴量ベクトルに含有させている。今回は 1 種類の ODF 値をそのまま用いたが，それ以外にも複数の種類の ODF 値を用いたり，ピッチ系列といった音楽的な指標を用いることも可能である。こうした複数の値を組みあわせ，更なる性能向上を検討していく予定である。

参考文献

[1] 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).
[2] 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).
[3] 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, pp. Article ID 38727, 8 pages (2007). doi:10.1155/2007/38727.
[4] 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).
[5] 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).
[6] 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).
[7] Lee, A., Kawahara, T. and Shikano, K.: Julius — an open source real-time large vocabulary recognition engine, *Proc. EUROSPEECH*, pp. 1691–1694 (2001).
[8] Stowell, D. and Plumbley, M.: Adaptive whitening for improved real-time audio onset detection, *Proc. International Computer Music Conference*, pp. 312–319 (2007).
[9] Cannam, C. and Stowell, D.: OnsetsDS (2007). `https://code.soundsoftware.ac.uk/projects/vamp-onsetsds-plugin`.
[10] Dixon, S.: Onset detection revisited, *Proc. 9th International Conference on Digital Audio Effects*, pp. 133–137 (2006).
[11] 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).