Vol.2019-MUS-123 No.52
Vol.2019-SLP-127 No.52
2019/6/23

主旋律に注目したクラシック音楽の自動擬音語変換

鈴木 基之[1,a)]　竹中 智美[1,†1]

概要：人は歌詞のない楽曲を歌唱する際，擬音語表現をよく用いる。そこで用いられる擬音語には様々なバリエーションが考えられるが，ある程度の共通性があることが知られている。そのため，楽曲を自動で擬音語表現に変換できれば，擬音語歌唱音声からの楽曲検索や，聴覚障害者のための BGM の可視化，といったシステムに応用可能である。
そこで本論文では，クラシック音楽を入力とし，擬音語による表現を出力するシステムを開発した。楽曲の主旋律部分を入力とし，音声認識と同じ枠組みで擬音語表現へと変換する。更に入力楽曲の音符の区切り時刻情報を用いることで，より主旋律に同期した擬音語表現を得ることができた。変換結果が元楽曲にふさわしいか人間による評価を行ったところ，4 点満点で平均 2.96 点を得ることができた。

Automatic conversion method from melody of instrumental music into onomatopoeia

1. はじめに

近年音楽データがインターネットを介して配信・販売されるようになり，手軽に大量の音楽を聞くことができるようになってきた。こうしたデータはスマートフォンや MP3 プレーヤーといった携帯プレーヤーに何百曲と保存可能であるが，プレイリスト等を用いて流し聞きするのではなく，目的の曲を聞こうとした場合，検索のために曲名やアルバム名といった情報を入力する必要がある。この時，曲名等の情報を忘れてしまった場合は検索できず，また一般にキーボードといった入力手段があるわけではないため，情報の入力も手軽とはいい難い。

こうした問題を解決するため，目的の曲を歌唱することで検索を行うシステムが開発されてきた。初期のシステム [1], [2], [3] では，入力をハミング歌唱に限定しており，そこからメロディ情報（音符の高さや長さの情報）を抽出することで検索キーとしていた。しかしこの方法ではメロディ情報の抽出誤りに弱く，また歌唱法もハミングに限定されている事から自然な入力とはいえない。

そこで，より自然で高精度な検索を行うため，歌詞による歌唱音声を入力としたシステムが提案された [4], [5]。こうしたシステムでは，入力された歌唱音声からメロディ情報だけでなく，歌詞の情報を音声認識によって抽出し，その両方から検索を行うことで，最終的な検索精度を向上させている。しかし，こうしたシステムは歌詞が付与されている曲に対してしか用いることはできない。いわゆるクラシック音楽といった歌詞のない曲を検索しようとした場合，当然ではあるが，メロディ情報だけから検索する必要がある。

こうした歌詞のない曲をユーザが歌唱する場合，一般には「ちゃらら〜」といった擬音語が用いられる。どのような音に対してどのような擬音語を用いるかは一意には決められないため，歌詞と同じように擬音語からある曲を検索できる，というものではない。しかし，擬音語の使い方にはある程度の法則性があることが知られている [6], [7]。例えばピアノによる単音を表現した擬音語で「ピン」と「ポン」があれば，後者の方が低い音であることがわかる。また打楽器において「ドン」と「トン」と表現されれば，前者の方が大きな音であることが容易に想像できる。

そこで，入力された擬音語による歌唱音声から擬音語系列を音声認識し，歌詞のかわりに検索に用いるシステムを考える。このシステムを構築するためには，データベースに登録されている楽曲を事前に擬音語系列へと変換し，検索に用いられるようにしておく必要がある。そこで本論文では，楽曲を入力とし，それを表現するような擬音語系列

[1] 大阪工業大学 情報科学部
[†1] 現在，株式会社 メクゼス
[a)] moto@m.ieice.org

に自動変換する方法を提案する。

2. 擬音語自動変換システムの構築

2.1 システムの概要

擬音語自動変換システムは，ある楽曲データを入力とし，その曲を表現するような擬音語系列を出力するものである。これは，音を入力すると，それに関連するテキストが出力される，と見れば，一般の音声認識システムと同じ構造である。そこで，音声認識システムの枠組みをそのまま流用し，音響モデルと言語モデルをそれぞれ擬音語変換用に学習したものを用いることで，自動変換システムを構築する。

音響モデル，言語モデルを構築するための学習データベースには，擬音語歌唱データベース [8] を用いた。このデータベースには，クラシック音楽の一部（1 曲 20 秒程度）を人間に聞かせ，それを擬音語で歌唱してもらった音声データが収録されている。1 曲あたり 5 名程度の歌唱が収録されており，同じ曲に対しても歌唱者によって異なる擬音語が用いられている場合もある。

また，音響モデルの構築単位（ひとつの HMM で何を表現するか）を決める必要があるが，今回は「どのような音がどのような擬音語表現になるか」を学習する必要があるため，ひとつの擬音語をひとつの HMM で表すこととする。また，言語モデルは擬音語の系列を n-gram で学習する。つまり，通常の音声認識システムとの対応でいえば，ひとつの擬音語がひとつの音素であり，またひとつの単語はひとつの音素だけからなる，という形となる。

ここで，「ひとつの擬音語」の定義を決める必要がある。例えば「ちゃらりらり」という擬音語歌唱があった時，「ちゃら」「り」「ら」「り」と分割されるのか，それとも「ちゃらり」「らり」と分割されるのか，様々な可能性が考えられる。そこで本論文では，文献 [8] で採用されている階層 Pitman-Yor 言語モデルを用いて文字系列を単語に分割する方法 [9], [10] を用い，「ひとつの擬音語」を定義する。

2.2 音響モデルの学習

2.2.1 HMM の構造

音響モデルには HMM を用い，ひとつの HMM をひとつの擬音語に対応させて学習する。この時，ひとつの擬音語の長さには非常に大きなバリエーションがある（文献 [8] によれば，1 音節だけの擬音語から，最大で 35 音節からなる擬音語が抽出されている）。そのため，非常に長い擬音語に対応する HMM は，非常に長い音楽と対応することが予想される。一般に HMM において長い時間長の音をモデル化するには，その分状態数を増やす必要がある。そこで HMM の状態数をすべて一定にするのではなく，擬音語の長さ（音節数）にあわせて設定することとする。

擬音語に含まれる音節数を n とすると，HMM の状態数 S を式 (2.2.1) で定義する。

$$S = \begin{cases} 3n+2 & (n<4) \\ 2n+4 & (4 \leq n < 10) \\ n+13 & (10 \leq n) \end{cases} \tag{1}$$

このようにすることで，長い時間長の音をモデル化すると思われる HMM は多くの状態を持つことになる。また状態数と音節数を単純に比例関係にするのではなく，その対応を徐々にゆるやかにしていくことで，無駄に状態数が多くなることを避けた。

2.2.2 学習データの準備

音響モデルの学習データは楽曲データである。しかし，通常人間が擬音語で歌唱する時は，楽曲の主旋律に注目して歌唱していると思われる。そのため，学習データも（伴奏等が含まれた）楽曲ではなく，主旋律のみを抽出したデータを用いた方が，擬音語との対応関係がよくなるものと思われる。そこで，学習用データとして主旋律のみからなる楽曲データを作成した。

この時，楽曲中のどの音を「主旋律」として歌唱するか，ということも一意に決められるものではなく，歌唱者によって変わる可能性がある。そこで，データベース中の歌唱音声を実際に聞き，そこで歌唱されているメロディを「主旋律」として抽出した。個々の歌唱データごとに歌唱されているメロディを人手で MIDI 形式で抽出し，DTM ソフトウェアを使って音データへと変換した。この時，演奏している楽器の音色によって擬音語が変化することが考えられることから，抽出したメロディを演奏している楽器を指定し，MIDI データを音データへと変換した。

このようにすると，同じ曲に対しても歌唱者が異なれば別データとなる可能性がある。また，仮に同じメロディを歌唱していたとしても，その擬音語表現は異なる可能性がある。今回利用したデータベース [8] は，同じ曲に対して複数の歌唱者が歌唱しているため，1 曲の楽曲に対し，歌唱者数分だけの学習データ（音データと，それに対応する擬音語系列の組）を作成した。

2.2.3 学習の方法

一般に HMM の学習は，音声データとそれの書き起こしテキストが与えられ，書き起こしテキストに従って HMM を連結した上で音声データ全体で学習する，いわゆる連結学習で行われる。しかし今回の場合，同じ音データに異なる擬音語系列が書き起こしテキストとして与えられていたり，逆に音響的特徴の大きく異なる音データ同士に同じ擬音語系列が与えられていたりと，ひとつの HMM が表現する音響的特徴がまとまっているとは言い難い。そのため，連結学習を行うと各モデルの境界時刻の推定を誤ることが多くなり，よいモデルが学習できないと思われる。

そこで，個々の HMM と音データの時間的な対応を明示

的に与えた学習を行う。まずデータベース中の擬音語歌唱を聞き，どの擬音語が抽出した主旋律のどの部分を歌唱しているのかを人手で確認する。その後，音データにおいてある擬音語の開始時刻と終了時刻をすべて記録し，その情報をもとに個々の HMM を対応した音データだけから学習する，ということを行なった。

全 237 データのうち，30 データに対してこういったラベル情報を付与し，それのみを用いて HMM の学習を行った。こうして得られた HMM を初期モデルとして，残りのデータもあわせて連結学習を行った。

3. 擬音語と音符との対応関係を制約とした変換法

2.1 節で説明した方法では，ひとつの擬音語がひとつの HMM に対応づけられ，その HMM はある程度の長さの音楽の特徴量をモデル化している。その擬音語が複数音節からなるものであれば，それに対応する HMM も複数の音符を含む音楽をモデル化していると予想される。

一般に楽曲を擬音語で歌唱する場合，ひとつの音符をひとつの音節で表現することが多いと思われる。しかし 2.1 節の方法では，（複数音節からなる）ひとつの擬音語をモデル化しているため，音符と音節の対応づけは明確ではない。その結果，入力楽曲の各音符と出力擬音語の各音節が対応していないような変換（例えばひとつの音符に対して「らら」と 3 音節が対応する等）が行われる可能性がある。

こうした変換は，ユーザに対して強い違和感を与えるため，抑制される必要がある。そこで本節では，擬音語内の各音節が入力楽曲の各音符に対応するよう，擬音語への変換時に制限をかけた認識方法を提案する。

3.1 音符と音節を対応させた変換アルゴリズム

歌詞による歌唱において，歌詞中の各モーラが楽曲中の各音符と対応していることが多いことから，この関係を明示的に導入することで，歌唱音声から歌詞を高精度に音声認識する方法が提案された [11]。この方法では，まず入力された歌唱音声を分析し，そのパワーやピッチの変動等から音符の区切り時刻を推定する。その後，音響分析された歌唱音声の特徴量ベクトル系列に対し，推定された区切り時刻の位置に「マーカーベクトル」と呼ばれる（一般の歌唱音声では出てこないような特徴量を持つ）特殊な特徴量ベクトルを挿入する。一方，音響モデルにも「マーカーベクトル」が持つ特徴量を平均ベクトルとして持つ特殊な HMM（マーカー HMM）を定義し，発音辞書において，すべての登録単語のモーラ区切り位置にマーカー HMM を挿入する。こうする事で，音声認識時に特徴量ベクトル系列中のマーカーベクトルとマーカー HMM が対応づき，ひとつの音符がひとつのモーラと対応づく，という制約を与えた認識を行うことができる。

本論文でもこのアルゴリズムを流用し，入力楽曲のひとつの音符が変換された擬音語中のひとつの音節に対応させるよう制限を加える。入力される特徴量ベクトル系列に対しては，もともと MIDI データから生成した主旋律のみの楽曲が入力であるため，その音符の区切り時刻は MIDI データから自動で抽出が可能である。そこで，その位置にマーカーベクトルを挿入する。また，マーカー HMM も同じ方法で定義する。

一方，発音辞書については特別な対応が必要となる。擬音語自動変換においては，もともと（複数の音節からなる）ひとつの擬音語がひとつの HMM で表現され，それがそのままひとつの単語として（言語モデルで）扱われている。そのため，マーカー HMM を挿入する場合，ひとつの HMM の中に挿入する必要があり，どこの位置に挿入するかを決める事は簡単ではない。

そこで，複数の音節からなる擬音語については，ひとつの HMM でモデル化するのではなく，複数の（ひとつの音節からなる擬音語をモデル化している）HMM を連結して表すこととした。例えば「ちゃららん」という擬音語があった場合，それをひとつの HMM でモデル化するのではなく，「ちゃ」「ら」「らん」という 3 つの音節にわけ，それぞれを別の HMM でモデル化する。具体的には，すでに（2.1 節の方法で）学習されている擬音語モデルの中から単音節の擬音語に対応している HMM のみを抽出し，それらを連結するように発音辞書を記述することで，複数音節からなる擬音語に対応している HMM を使わないように変更した。その上で，各 HMM の後ろにマーカー HMM を挿入した発音辞書を作成した。

4. 擬音語変換実験

4.1 実験の概要

人手で主旋律のみを抽出した楽曲データを入力とし，自動で擬音語に変換を行った。擬音語への自動変換は，Julius[12] を用いた。特徴量は通常の音声認識と同様，MFCC とその差分（ΔMFCC），対数パワーの差分からなる 25 次元を用いた。認識結果に対し Viterbi アルゴリズムを用いて各 HMM と特徴量ベクトルとの対応をとり，それぞれの擬音語がどの区間に対応しているかを決定した。

その後，変換された擬音語系列を画面に表示（図 1）し，カラオケの歌詞表示のように，音楽にあわせて色を変えることで音楽との対応がわかるようにした。この時，擬音語変換システムへの入力は主旋律だけからなる楽曲データであるが，変換された擬音語にあわせて再生する音楽は元と

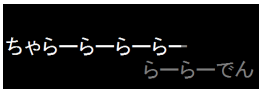

図 1 評価用擬音語表示システムの例

表 1　擬音語変換の妥当性評価結果

評価		評価の割合
ほぼ適切である	4	36.3 %
	3	31.1 %
	2	25.1 %
あまり妥当でない	1	7.5 %
平均点		2.96 点

なった（伴奏等も含む）楽曲データとした。

こうして得られた擬音語表示動画を 60 データ作成し，表示されている擬音語の妥当性について評価してもらった。10 秒から 20 秒程度の動画を 1 名あたりおよそ 20 クリップ提示し，それぞれ「ほぼ適切である」から「あまり妥当でない」までの 4 段階で評価してもらった。評価者は日本人大学生 25 名（男性 17 名，女性 8 名）であり，1 クリップあたりおよそ 10 名が評価している。なお，評価の際は動画を何回再生してもよいものとした。

4.2　評価結果の分析

擬音語変換の妥当性の評価結果を表 1 に示す。4 点の「ほぼ適切である」が 36.3%，3 点とあわせると 67.3%となり，3 分の 2 程度の曲は，良い変換ができていることがわかる。マーカーベクトルを挿入することで，音符と擬音語の対応や時間的タイミングが揃い，結果として高い評価に継がったものと思われる。なお，全体の平均点は 2.96 点であり，本変換方法の有効性が示された。

曲ごとに平均点を見てみると，「惑星」が 3.70 点，「白鳥の湖」が 3.26 点とよい性能を示していた。この 2 曲は，他の楽曲と比べると比較的メロディーがはっきりとしており，メロディーラインの動くスピードもゆっくりであった。したがって出力された擬音語がどの音符に対応しているのかが特にわかりやすいため評価が上がったのではないかと考える。

一方，「フィガロの結婚」は 2.38 点，「カルメン」は 2.40 点と，かなり低いものとなった。その原因を調べるため，評価が 3 点以下であるデータについて，その理由を調査した。理由として考えられそうなものをリストアップし，被験者には，それらのうちから選択してもらうことでなぜ妥当ではないと感じたかを回答してもらった。得られた理由の内訳を表 2 に示す。なおこの結果は，複数回答を許している。

これを見ると，擬音語のスピードや数が合わないといった理由はそれぞれ 20.5%と 12.1%となっており，音符と音節の対応制約を入れたにもかかわらず，一部の曲では問題が残っていることがわかる。どのような曲でこうした指摘が多いか調べたところ，曲のテンポが早く，短い長さの音符が多数使われている曲が多いことがわかった。特に「フィガロの結婚」では「タイミング・スピード感」についても「擬音語の数」についても，どちらも 3 割程度選択されていた。また「ハンガリー舞曲」や「くりみ割り人形」は「タイミング・スピード感」が合わないといった回答が 4 割程度あった。この 2 曲は常に早く演奏されているわけではないが，途中で早く演奏される部分があり，そこでのずれが気になったものと思われる。

また，「擬音語が曲にあっていない」と指摘された曲が 22.1%あった。こちらは擬音語の選択に違和感があった曲と思われる。こうした曲には「鱒」や「ジュ・トゥ・ヴ」「ラデッキー行進曲」等があり，いずれも主旋律をバイオリンが担当している。バイオリンの音色を表す擬音語は多様である [8] ことから，変換結果も安定せず，様々な擬音語を出力してしまったものと思われる。一方，「カノン」も同じように擬音語が曲にあっていないと評価されることが多かったが，こちらは主旋律と呼べるメロディがふたつあるため，評価者がどちらを主旋律と感じたかで違和感が生まれたのではないかと思われる。

4.3　伴奏付きの楽曲からの変換結果

前節の結果は，入力として主旋律を人手で抽出した楽曲データを使用していた。しかし実際に擬音語変換システムを使用する場面では，そうしたデータを準備することは難しく，伴奏等がはいった元楽曲をそのまま入力することになる。そこで本節では，音響モデルの学習は主旋律のみの楽曲データを用いるが，自動変換を行う際には元楽曲を入力し，そのまま擬音語に変換した結果を評価した。また，音符と音節との対応制約の効果を見るため，制約を導入せずに変換を行ったデータに対する評価も行った。

被験者による評価結果を表 3 に示す。ただし，制約を導入して変換を行った結果については，時間の都合上被験者

表 2　擬音語変換が妥当でないと感じる理由

間違っていると感じた理由	選択率
擬音語表示のタイミング・スピード感が合っていない	28.8 %
曲に対する擬音語の数が合っていない	17.0 %
変換しないでよいと感じる音（伴奏など）まで擬音語変換されている	7.8 %
濁音・半濁音（ば、ぱなど）が適切に表現されていない	1.2 %
はねる音 (小さいつなど) が適切に表現されていない	7.2 %
変換してほしいと感じる音 (歌いたい部分) が擬音語変換されていない	15.8 %
擬音語が曲に合っていない	22.1 %

表 3 元楽曲から変換を行なった結果の評価

評価		音符と音節の対応制約	
		有	無
ほぼ適切である	4	23.7 %	9.0 %
	3	39.0 %	20.5 %
	2	23.7 %	39.0 %
あまり妥当でない	1	13.4 %	31.6 %
平均点		2.73 点	2.07 点

1 名のみで行った結果である。そのため，両者の直接の比較はできず，また結果信頼性は低いことに注意していただきたい。

この表を見ると，「ほぼ適切である」の項目が 9.0%から 23.7%へ 15 ポイント程度向上していることがわかる。また平均点も 2.07 点から 2.73 点へと向上している。これらのこととから，音符と音節の対応関係の制約は非常に有効であることがわかる。

一方，表 1 と比較すると，「ほぼ適切である」と評価されたデータ数が 36.3%から 23.7%と 13 ポイント程度下回っている。被験者からの内省報告によると，「音に対しての擬音語の種類が妥当でない」と感じる結果が多かった，とのことである。入力楽曲を主旋律のみではなく伴奏等もついた楽曲にしたことで，複数の音色が混ざった音を擬音語変換することになり，音色が異なる擬音語表現へと変換されてしまった事が考えられる。一般に「主旋律」はその音響パワーも大きく，目立って聞こえてくるため，主旋律のみから HMM を学習してもある程度変換できると思われるが，やはりより高精度に変換を行うには，入力楽曲から主旋律のみを自動で抽出する，といった方法を組み合わせる必要があると思われる。

5. 結論

本論文では，擬音語歌唱による楽曲検索を実現するため，楽曲を擬音語系列へと自動変換する方法を提案した。音声認識と同じ枠組みを用い，主旋律のみを抽出した楽曲データを入力として音響モデルの学習を行った。更に入力された楽曲の音符と擬音語内の音節が対応するようにマーカーベクトルを挿入する方法を導入し，変換精度の向上をはかった。

20 秒程度のクラシック音楽を入力とし，変換された擬音語の妥当性を人間によって評価したところ，4 点満点で平均で 2.96 点と，その妥当性が示された。特に比較的ゆっくりしており，メロディがはっきりしている曲は高精度に変換されることがわかった。一方で早く音符の数が多い曲は擬音語のタイミングや個数があわず，違和感のある擬音語が出力されることがわかった。また，入力を主旋律を抽出した楽曲データではなく，元の楽曲データとすると，変換精度が落ちてしまうため，別途主旋律の自動抽出法等が必要であることもわかった。

謝辞 本研究の一部は，JSPS 科研費 JP18K11321 の助成を受けて行われた。

参考文献

[1] Kosugi, N., Nishihara, Y., Sakata, T., Yamamoto, M. and Kushima, K.: A Practical Query-By-Humming System for a Large Music Database, *ACM Multimedia 2000*, pp. 333–342 (2000).
[2] Ghias, A., Logan, J., Chamberlin, D. and Smith, B. C.: Query By Humming: Musical Information Retrieval in An Audio Database, *Proc. ACM Multimedia*, pp. 231–236 (1995).
[3] Liu, B., Wu, Y. and Li, Y.: A Linear Hidden Markov Model for Music Information Retrieval Based on Humming, *Proc. ICASSP 2003*, Vol. V, pp. 533–536 (2003).
[4] 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 (2007). Article ID 38727, 8 pages, doi:10.1155/2007/38727.
[5] Suzuki, M., Hosoya, T., Ito, A. and Makino, S.: Music Information Retrieval from a Singing Voice Based on Verification of Recognized Hypotheses, *Proc. ISMIR*, pp. 168–171 (2006).
[6] 田中基八郎，松原謙一郎，佐藤太一：異音の表現における擬音語の検討（衝突音等の単発音やうなり音の場合），日本機械学会論文集 C，Vol. 61, No. 592, pp. 4730–4735 (1995).
[7] 比屋根一雄，澤部直太，飯尾　淳：単発音のスペクトル構造とその擬音語表現に関する検討，技術研究報告（音声）SP97-125，電子情報通信学会 (1998).
[8] Suzuki, M. and Hisaoka, A.: Development of Singing-by-Onomatopoeia corpus for Query-by-Singing Music Information Retrieval system, *International Journal of Advanced Intelligence*, Vol. 9, No. 1, pp. 63–75 (2017).
[9] Mochihashi, D., Yamada, T. and Ueda, N.: Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling, *Proc. ACL*, pp. 100–108 (2009).
[10] Neubig, G., Mimura, M., Mori, S. and Kawahara, T.: Learning a Language Model from Continuous Speech, *Proc. INTERSPEECH*, pp. 1053–1056 (2010).
[11] 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).
[12] Lee, A., Kawahara, T. and Shikano, K.: Julius — an open source real-time large vocabulary recognition engine, *Proc. EUROSPEECH*, pp. 1691–1694 (2001).