

楽曲パターンマイニングによる歌唱合成パラメータの自動推定

清水 豪[†] 浅田 学史[‡] 竹内 和広[†]

大阪電気通信大学 情報通信工学部[†] 大阪電気通信大学大学院 工学研究科[‡]

1. はじめに

歌唱合成ソフト Vocaloid[1]では、複数のパラメータを人手で編集することで、人間らしい歌声を再現する。本研究では、パラメータ編集の時間短縮を目的に、系列パターンマイニングにより獲得したパターンに基づいたパラメータ推定手法を提案する。

2. 楽曲の表現方法

楽曲の多くはコンピュータで操作するために、MIDIという規格で保存されている。MIDI規格はどの音をどの時刻にどれだけの長さを生じさせるかを記述するための規格であり、リアルタイム性・同期性を重視している。また、MIDI規格では区切りなどは一切なく、音符の系列である。

音符の特性を決めるもっとも重要な要素として音程と長さがある。そこで、本研究では図1のように、「音程」「長さ」「音程・長さ」を文字列として表現する。

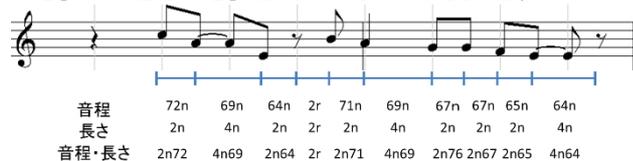


図1 音符を文字列として表現した例

3. 楽曲パターンマイニング

MIDI規格のボーカルパート 6,930 件の楽曲に対して、「小節線区切り」「n-gram」「系列パターンマイニング(頻度)」「系列パターンマイニング(系列信頼度)」を用いて楽曲パターンマイニングを行う。よって、文字列表現(3通り) × マイニング手法(4通り)の12通りを検討する。

小節線区切りとは、図2のように、小節線により等間隔に区切られた区間をパターンとして取り出したものである。

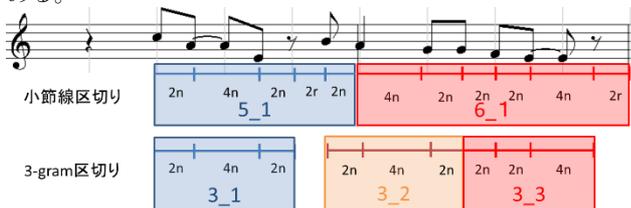


図2 小節と3-gramによるパターン化の例

n-gramとは音符列をnという定数間隔で区切り、パターン化したものである。n=3の例を図2に示す。事前実験において、8以上のパターンの頻度が著しく低かったため、nは2-7に設定する。

Mining Music Patterns for Automated Parameter Determination of Singing Synthesizer
Go Shimizu[†]
Satoshi Asada[‡]
Kazuhiro takeuchi[†]
[†]Department of Information and Communication Engineering, Osaka Electro-Communication University
[‡]Graduate School of Engineering, Osaka Electro-Communication University

系列パターンマイニング[2]とは、時系列に配置された系列から頻出パターンを抽出するためのパターンマイニング手法である。本研究では、系列パターンマイニングを行う手法の一つである PrefixSpan [3]を(図3)を用いる。

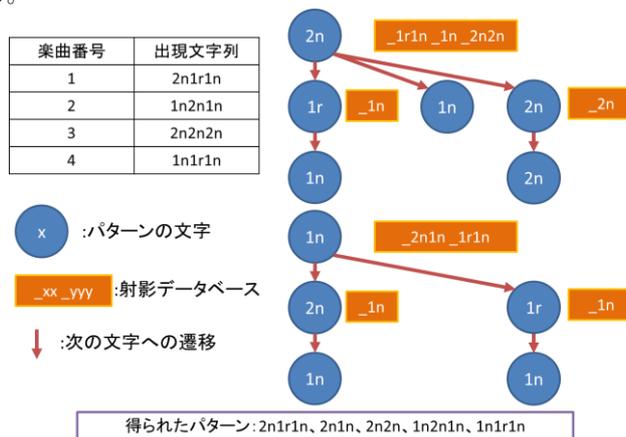


図3 PrefixSpanの例

系列パターンマイニングによって得られたパターンの頻度と、得られたパターンをp、その部分集合をp_{sub}としたときの式(1)で表される系列信頼度によるランキングを用いる。系列信頼度は文字が連続するかの信頼度を表している。文字が連続しやすい場合、大きな値となる。

$$\text{conf}(p) = \max_{p_{\text{sub}} \in \mathcal{P}} (p_{\text{sub}} \text{を含む系列の数}) \quad (1)$$

4. 複数尺度によるパターン化の評価

得られたパターンが楽曲内でどの程度使用されているかを調査するために、3章で得られた12通りのパターンを頻度または信頼度でスコア付した、上位100パターンを長さ1の特別な文字列に置き換え、式(2)の圧縮率と比較する(表1)。これは、符号化の考えを用いており、文字列に対して、効率的に符号長を短くするための方法である。ここで、符号化はパターンを多くすると符号長は短くなるが、パターンが多くなるため、効率が悪くなる。また、パターンが少なすぎると、曲長がとてま長くなってしまふ。そこで、パターン数と曲長のバランスを取らなくてはならない。

$$\text{圧縮率} = \frac{\text{適応されたパターン数} + \text{パターンではなかった音符数}}{\text{音符総数}} \quad (2)$$

表1 100パターンによる圧縮率

		圧縮率		
		音程	長さ	音程・長さ
マイ ニ ン グ 手 法	小節線	0.973	0.913	0.989
	2-gram	0.793	0.707	0.916
	3-gram	0.865	0.601	0.963
	4-gram	0.889	0.709	0.982
	5-gram	0.940	0.740	0.994
	6-gram	0.966	0.821	0.996
	7-gram	0.980	0.839	0.999
	PrefixSpan(頻度)	0.710	0.567	0.947
	PrefixSpan(信頼度)	0.731	0.584	0.960

表1より音程・長さでは、ほとんど圧縮できていない。

また、固定長の手法に対して、可変長である PrefixSpan の圧縮率が高かった。

次に、左記のパターン集合 $P = \{p_{i,1}, \dots, p_{i,k}, \dots, p_{i,100}\}$ がボーカルパートにおいて特徴的であるかの検討を行う。具体的には、パターン集合 P に基づいて、ボーカルパート以外の単音楽器パート 3,317 件を含む 10,247 件の楽曲 $M = \{m_1, \dots, m_i, \dots, m_{10247}\}$ から、ある楽曲 m_i がボーカルパートか否かを推定する SVM モデルを構築する。モデル構築における特徴ベクトルは、各楽曲 m_i に対して各パターン $p_{i,k}$ の有無を表現した特徴ベクトル $V_i = \{v_{i,1}, \dots, v_{i,k}, \dots, v_{i,100}\}$ を式(3)に基づいて作成した。ここで、SVM の実装には R に提供されている「e1071」パッケージを用いた。

$$v_{i,k} = \begin{cases} 1 & p_k \in m_i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

また、得られたパターンが出現せずに評価できない楽曲の量を調べるため、式(4)で求めた特徴出現率を表 2 に示す。表 2 は「ボーカル」と「楽器」の特徴出現率である。表 2 より、PrefixSpan を用いると、ボーカルパートの特徴ベクトル化ができていていることがわかる。

$$\text{特徴出現率} = \frac{|V_i| = 0 \text{ となる楽曲数}}{\text{総楽曲数}} \quad (4)$$

表 2 100 パターンによる特徴出現率 (ボーカル/楽器)

		特徴出現率(ボーカル/楽器)		
		音程	長さ	音程・長さ
マイニング手法	小節線	0.511/0.063	0.616/0.607	0.045/0.005
	2-gram	0.896/0.174	0.958/0.833	0.773/0.143
	3-gram	0.732/0.107	0.956/0.926	0.496/0.084
	4-gram	0.656/0.115	0.885/0.737	0.264/0.034
	5-gram	0.473/0.102	0.841/0.797	0.115/0.011
	6-gram	0.299/0.075	0.720/0.584	0.070/0.007
	7-gram	0.169/0.050	0.661/0.670	0.018/0.002
	PrefixSpan(頻度)	0.924/0.245	0.961/0.922	0.692/0.100
	PrefixSpan(系列信頼度)	0.916/0.277	0.960/0.923	0.623/0.127

ボーカルパート推定における十交差検定の平均値を表 3 に示す。表 3 より、小節線や 2-gram は PrefixSpan より判別率が高い、またはほぼ同じ値であった。しかし、表 1 より、小節線や 2-gram よりも PrefixSpan のパターンは出現率が高いことから、PrefixSpan が総合的にボーカルのパターン化に適しているといえる。

表 3 100 パターンによるボーカル判別率

		判別率		
		音程	長さ	音程・長さ
マイニング手法	小節線	0.676	0.705	0.676
	2-gram	0.835	0.708	0.769
	3-gram	0.759	0.684	0.676
	4-gram	0.713	0.676	0.676
	5-gram	0.677	0.682	0.676
	6-gram	0.678	0.681	0.676
	7-gram	0.676	0.684	0.676
	PrefixSpan(頻度)	0.842	0.718	0.761
	PrefixSpan(系列信頼度)	0.834	0.707	0.678

5. パラメータの自動推定

4 章の結果より、音程の PrefixSpan(頻度)と長さの PrefixSpan(信頼度)のパターンを組み合わせた、音程と長さをもつ 10 パターンを使用する。10 パターンに対して、ダイナミクス、ポルタメントの 2 つのパラメータをアノテーションする。ダイナミクスは音量を、ポルタメントは音程の変化のタイミングを調整するパラメータである。そして、パターンが一致する部分にアノテーション済みのパラメータを適用することでパラメータ推定を行う。1 パターンにつき 2 フレーズ、計 20 フレーズのパ

ラメータ推定を行った。なお、歌詞による発音の変化は考慮しないため、歌詞はすべて「ラ」で統一している。

6. 評価実験

パラメータ推定の有効性を比較するために、アンケートにより、心理実験を行った。被験者の 10 人に対して、パラメータ推定を行ったフレーズと、パラメータを全く変化させていないフレーズの組を 2 つずつ用意し、「どちらがよく聞こえたのか」という尺度で評価させた。被験者 10 人の内訳は Vocaloid の楽曲をよく聴く被験者 2 人と Vocaloid の楽曲をあまり聴かない被験者 8 人となっている。アンケート結果の平均値を表 4 に示す。表 4 より、Vocaloid の楽曲をよく聴く被験者 2 人はパラメータ推定したフレーズがよく聞こえ、Vocaloid の楽曲をあまり聴かない被験者 8 人に対してはパラメータ推定前の方がよく聞こえるという結果になった。Vocaloid の楽曲をよく聴く被験者は違いを聞き取ることができたが、Vocaloid の楽曲をあまり聴かない被験者は Vocaloid のパラメータを多少調整しても違いが分かりづらいためこのような結果になったと考えられる。

表 4 アンケート結果の平均値

	推定前の支持率	推定後の支持率
Vocaloidをよく聴く被験者	0.250	0.750
Vocaloidをあまり聴かない被験者	0.607	0.393

さらに、パラメータにより、編集時間の削減に貢献できているかを評価するため、パラメータ調整時間の比較を行う。調整時間の比較ではパラメータを推定したフレーズと、パラメータを変化させていないフレーズの組を 2 つ用意し、ダイナミクス、ポルタメントの 2 つのパラメータを調整し、調整に要した時間を測定し比較する。パラメータ推定ありとなしの平均作業時間と差を表 5 に示す。表 5 より、パラメータ推定ありのフレーズの方が 1.7 倍程時間短縮できるという結果になった。

表 5 パラメータ推定ありとなしの平均作業時間と差

	パラメータ推定あり	パラメータ推定なし	差	検定結果
平均	18.47(s)	32.42(s)	13.95(s)	p<0.01
分散	69.663	80.006	10.345	

7. おわりに

MIDI データに対してのパターンマイニングにより、メロディ分割に適したパターンを検討し、圧縮率やパート推定によって評価した。また、パターンマイニングによって得られたパターンに対してのパラメータ推定ルールを策定した。編集時間の短縮という観点では有効性を示せた。それらが機械的な評価でも、小節線や n-gram よりも有効であることを確認した。しかし、機械学習の導入や、歌詞の影響の考慮、他の感情コントロールの推定は今後の課題としたい。

参考文献

- [1] 剣持秀紀, 大下隼人: 歌声合成システム VOCALOID-現状と課題, 情報処理学会研究報, pp. 51-56 (2008)
- [2] Agrawal, R. and Srikant, R.: *Mining Sequential Patterns, Data Engineering, 1995. Proceedings of the Eleventh International Conference on, IEEE*, pp. 3-14 (1995)
- [3] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. and Hsu, M.: *PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth*, ICDE'2001, April, pp. 215-224 (2001)