6V-08

歌詞の自動生成における母音の出現頻度を考慮した検証※1

星　龍生※2　　横井　健※3

東京都立産業技術高等専門学校※4

1.　はじめに

作詞の際は文学的要素のみならず、メロディなどの音楽的要素も考慮する必要がある。

本研究ではポピュラー音楽における各メロディの母音の出現頻度を考慮した、メロディから歌詞を自動生成するシステムを提案し、生成される歌詞が人間から好まれる楽曲に近づくか否か検証する。システムは、メロディを入力し母音列を出力するルールベースのモデル(以降M2Vモデルと表記)と、母音列を入力し歌詞を出力する深層学習によるモデル(以降 V2L と表記)で構成する。ルールベースのモデルでは、母音生成のルールを変更することで母音列を複数生成する。生成された母音列を使って自動生成された歌詞をサイトで音声合成し、どの歌詞による曲が好ましいか被験者らの主観評価で本研究の検証を実施する。

2.　各提案モデル

M2V モデル、V2L モデルの構成について、それぞれ示す。

2.1.　M2V モデル構成

入力のメロディには UST ファイルを採用する。ここではメロディ全体のうち先頭 16 小節を扱う。本研究では既存の曲の UST ファイルを用いる。

メロディを複数のグループに分割し頂点音および非頂点音を推定する。橋田ら[1]による頂点推定ルールを用いて各メロディに対して評価ポイントを設け、最もポイントが大きい音を頂点音とし、それ以外を非頂点音とする。なお、UST ファイルですでに休符と定められている音はどちらでもない音とし、母音を割り当てないものとする。

岩橋ら[2]の分析結果による頂点音および非頂点音における母音出現率に従い、確率的に母音を割り当て続けて母音列を生成する。

次に母音出現率の数値を変更し、再度母音列を生成する。この工程を繰り返すことで一つのメロディに対して複数の異なる母音列を作成する。

本研究では母音の割り当て方のみを変えて検証を行うため、母音の割り当て方の違いが検証結果に因果するかを判定できる。

2.2.　V2L モデル構成

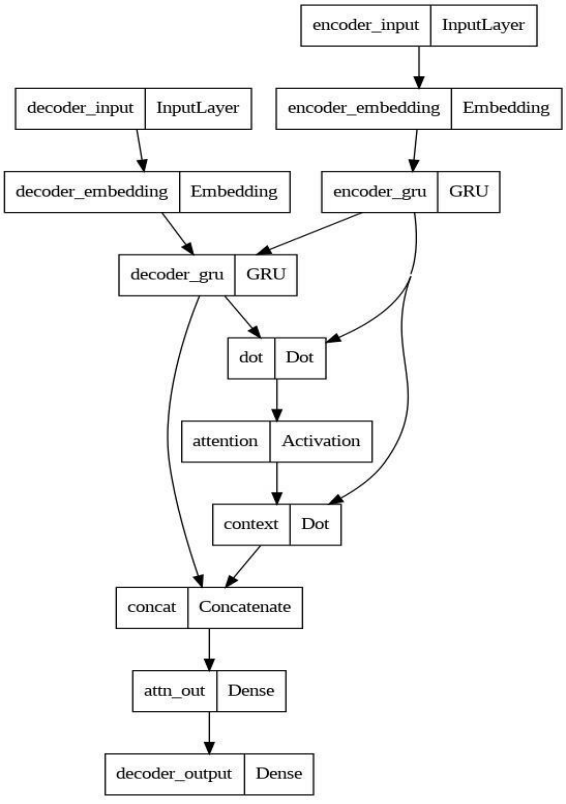

図 1: Seq2Seq モデル

※1Automatic lyrics generation considering vowel frequency

※2Ryusei Hoshi

※3Takeru Yokoi

※4Tokyo Metropolitan College of Industrial Technology

表1：M2V モデル実行結果の母音出現率の比較

	a	i	u	e	o
① 頂点音	0.476	0.286	0.143	0.048	0.048
② 頂点音	0.048	0.190	0.286	0.238	0.238
① 非頂点音	0.318	0.224	0.118	0.118	0.224
② 非頂点音	0.129	0.153	0.212	0.306	0.200

深層学習には Sequence-to-Sequence(以降 Seq2Seq と表記)という系列変換モデルを採用する。学習に用いる歌詞のデータは、Spotipy API を用いてダウンロードする。歌詞を形態素解析し、更に音素解析をすることで形態素・音素列のデータセットを作成する。

TensorFlow と Keras を用いて Seq2Seq の Encoder と Decoder を作成する。Keras の plot_model 関数を用いて出力したモデル図を図1に示す。Encoder と Decoder の RNN 層では GRU を用いている。また、入力である音素列は出力される歌詞の形態素の母音でありそれぞれの入力と出力は関連性が高いため、Attention も実装している。最後の Decoder の出力層では、活性化関数として Softmax 関数を用いている。

3. モデルごとの動作確認

M2V モデル、V2L モデルをそれぞれで実行し、その実行結果を述べる。

3.1. M2V モデルの動作結果

母音出現率のパターンに、岩橋ら[2]の分析結果によるもの(パターン①)、20%を基準点とした岩橋ら[2]の分析結果の反数(パターン②)の 2 つ設けて、それぞれ実行する。

実行結果について、出力全体における各母音の割合を表1に示す。M2V モデル内で設定した母音出現率の割合と相応しているため、このモデルは正しく構築できたと言える。

3.2. V2L モデルの動作確認

形態素解析を行う際の Ngram について、少量の N=2,3 と、N=8 の 2 パターンでそれぞれ実行し、20 組出力する。

N=2,3 で実行した場合は、20 組の入力の母音列と出力の形態素の音素が完全に一致していた。

対して N=8 で実行した場合は、20 組すべてにおいて入力の母音列とは音素がまったく異なる形態素を出力していた。原因として、学習データの不足が挙げられ、N=8 の音素列と一致するような形態素の組合せがほとんどないと考えられる。

4. まとめと展望

M2V モデルは正しく処理を行ったが、V2L モデルは少量の形態素数でしか処理できず、実用化には難しい。

今後は、学習に悪影響を及ぼさない程度の学習データの増大と、V2L モデルの再検討を行う必要がある。加えて、実用の際に Ngram が少量のままで通るかの検討も踏まえて取り組む。

参考文献

[1] 橋田光代, 片寄晴弘, 野池賢二, 保科洋, 河原英紀. (2004). LG-006 音楽聴取に関する一検討: グループと頂点の推定 (G. 音声・音楽). *情報科学技術レターズ, 3*, pp145-148.

[2] 岩橋亮人, 橋田光代, 片寄晴弘. (2016). ポピュラー音楽の頂点音における母音の出現頻度に関する分析. *研究報告音楽情報科学(MUS), 2016*(13), pp1-6.