Morphing-Based Reharmonization using LSTM-VAE

Aiko Uemura and Tetsuro Kitahara*

Nihon University
{uemura, kitahara}@chs.nihon-u.ac.jp

Abstract. We present a method of morphing between reharmonized chord progressions using long short-term memory (LSTM) and a variational autoencoder (VAE). If the reharmonization is automated, beginner musicians can easily change impressions of a harmony even without having expert musical knowledge. Morphing between many instances of reharmonization can be a practical approach to generating various reharmonizations. This study specifically examines interpolation in the latent space of an LSTM-VAE and achieves morphing between two chord progressions. Experimental results demonstrate that our method generates chord progressions that fit the melody based on interpolation between prepared chord progressions in the latent space obtained through the model.

Keywords: chord progression, reharmonization, morphing, VAE

1 Introduction

It is common knowledge that more than one musically suitable chord progression exists for a certain melody. Musicians carefully choose different chord progressions for the same melody to gain different impressions of music. For example, when a musician covers a song in a jazz or blues style, they often change the chord progression. In jazz music, it is particularly common to change the chords of a song by adding extended notes or by replacing them with alternative chords, which is called "reharmonization."

Reharmonization is not easy for beginner musicians. Therefore, its automation can enable anyone to enjoy playing songs in various styles. The biggest difficulty in designing an automatic reharmonization system is controllability of the musical style (i.e., controlling the style of generated chord progressions). The controller should be intuitive, musically understandable, and user-friendly for people with no musical knowledge.

Therefore, for this study, we specifically examine morphing in a latent space as a controller for reharmonization. Morphing is a technique of generating the medial data between two given instances. Although morphing is mostly common in image processing, the morphing of melodies (Hirata, Tojo, & Hamanaka,

* This work was supported by JSPS KAKENHI Grant Numbers 16H01744, 17H00749, 19K12288, 20K19947 and Casio Sci. Promotion Foundation.

2011), chords (Murata, Bando, Itoyama, & Yoshii, 2018), basslines, and drum patterns (Simon et al., 2018; Masuda & Iba, 2018) has also been conducted. A promising approach to morphing two musical instances is mapping them into a certain latent space and interpolating them in that space. In fact, this approach is commonly used by MUSIC-VAE (Roberts, Engel, & Eck, 2017; Simon et al., 2018) and MIDI-VAE (Brunner, Konrad, Wang, & Wattenhofer, 2018). The generated chord must fit the same melody in the reharmonization. However, interpolation of reharmonized chord progressions has never been reported.

In this paper, we propose a method of reharmonization in which we morph multiple reharmonized chord progressions based on a variational autoencoder (VAE). In an earlier study (Uemura & Kitahara, 2018), we developed a method of morphing reharmonized chord progressions using a VAE. That method generates intermediate chord progressions by interpolating and extrapolating in latent space, but the chord progressions tend to switch instead of changing gradually because we did not consider the time series in the model. Moreover, the distributions among the songs in the latent space were separated, making continuous change by interpolation unobtainable. To solve this problem, our new method improves the existing VAE model by using long short-term memory (LSTM) and melody conditions. A long short-term memory variational autoencoder (LSTM-VAE) maintaining the temporal connection and melody conditioning allows us to change chord progressions while keeping the musical texture.

2 Morphing Using LSTM-VAE

A VAE is a neural network that acquires characteristics that express data; it assumes a multivariate standard normal distribution for latent variables(Kingma & Welling, 2014a). It models the specific data generated based on the abstract representation of the latent variable. For this study, we assume latent variable $\boldsymbol{z}$, which represents the chord progression character. As usual, let $\boldsymbol{z}$ obey the multivariate standard Gaussian distribution of the mean vector $\mathbf{0}$, and covariance matrix I. We use $\boldsymbol{x}$ as song data and $\boldsymbol{z}$ as the latent variable corresponding to $\boldsymbol{x}$ to maximize the marginal likelihood $p_\theta(\boldsymbol{x})$ for training the VAE. We make a distribution of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, which generates a latent variable distribution in the VAE. θ and ϕ are sets of model parameters of the encoder and the decoder respectively. A conditional VAE (CVAE)(Kingma & Welling, 2014b) improves the VAE by conditioning the encoder and decoder to y. We treat the melodic pattern as a condition because we incorporate the constraint that the melody is identical in reharmonization. The variational lower bound of the log likelihood is in the following equation.

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}) + \log p(\boldsymbol{y})] - D_{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{x})||p(\boldsymbol{z})] \quad (1)$$

We regard song $\boldsymbol{x}$ as a chord progression sequence. We also combine 12-dimensional chroma vectors so that the same chord names are close to each other in the latent space. The onset series of note pitches is represented in units of quarter

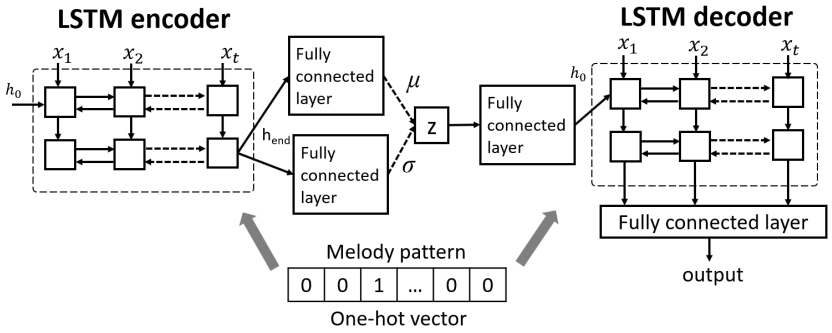

Fig. 1. Network structure.

notes. Here, we treat four octaves of C2 to B5. The number of notes contained in a chord is limited to five, and one note is represented as a one hot vector. It includes the silent state. The number of input dimensions x_t is assumed to be $49 \times 5 + 12$, using data of four bars, excluding the anacrusis of each song.

Figure 1 shows the network structure. The encoder has two LSTM layers, and the decoder layer has two LSTM layers. All of them have a bidirectional structure, and the hidden size is 128. Adaptive moment estimation (Adam) was the optimizer. The parameters follow the explanation in the document (Kingma & Ba, 2015). The latent variable $\boldsymbol{z}$ has 64 dimensions.

3 Experimental Results

We compared this model to a simple VAE we used in our previous studies(Uemura & Kitahara, 2018), in which the input vector was 1-hot vectors and chroma vectors, ignoring the time series. Its dimensions are $(49 \times 5 + 12) \times 16$. The hidden layer was 2,048 dimensions to accommodate the 128-dimensional model of 16 sequences of LSTM cells. We set the dimensions of the latent variable to 64 as well. The models were trained 1,331 chord progressions from common pattern data[1] and manually reharmonized chords by a professional musician. We did not consider changes in dynamics in the songs. We split notes longer than quarter notes and omitted notes shorter than quarter notes. The model was trained with a size of 128 and 50K epochs. In the loss function, we used SmoothL1loss for the reconstruction loss. We incorporated KL divergence annealing, so we used linear weighting to 2,500 epochs. Since the decoded results were $[0, 1]$, we found the maximum value of each component note and output the corresponding note number or rest, respectively. Each velocity was set to 100 to convert to a MIDI signal.

[1] `https://alfanote.jp/item/anb011/`, `https://www.rittor-music.co.jp/product/detail/3117317134/`, `https://www.rittor-music.co.jp/product/detail/3117317120/`

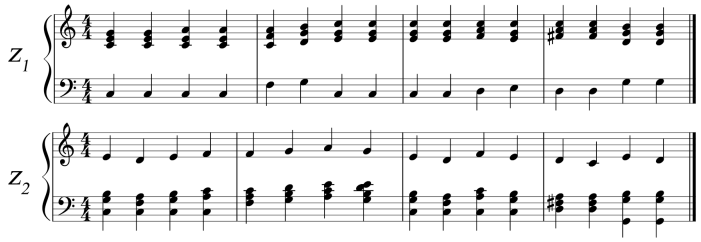

Fig. 2. First four bars of trained songs. Top: Original chord progression[4] z_1. Bottom: Reharmonized chord progression z_2.

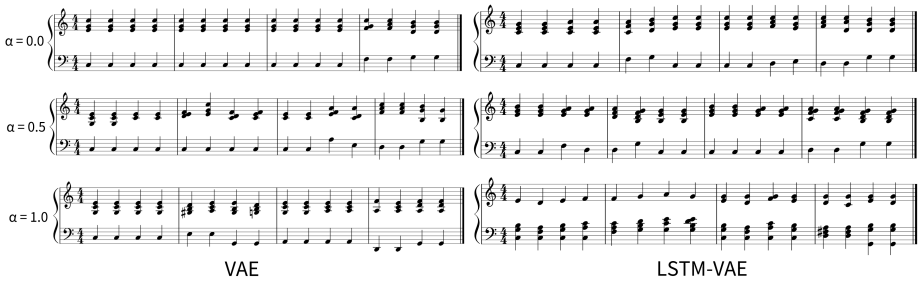

Fig. 3. Reconstruction and morphing results using VAE and LSTM-VAE.

3.1 Chord Morphing

Assuming that the latent variables of two songs are z_1 and z_2, the latent variable $\boldsymbol{z}$ of the morphed song is expressed as shown below:

$$\boldsymbol{z} = (1 - \alpha)\boldsymbol{z}_1 + \alpha \boldsymbol{z}_2, \tag{2}$$

where $\alpha \in [0, 1]$. Here, we conducted morphing between z_1 and z_2 in Fig. 2. We used the same condition vector.

Figure 3 shows the results of chord morphing for $\alpha = 0.5$[2]. The results of the reconstruction are mostly reproduced by the LSTM-VAE. The VAE has some matching chords, but they seem to be interpolated with different chord progressions influenced by the training data in the latent space. The VAE morphing results showed dissonant or unnatural chords in the second and third bars. In comparison, the LSTM allowed more consideration of chord connections than the VAE morphing. This is because the LSTM could deal with time-series data. Chord progressions might be sparse in latent space, so they are suitably complemented by nearby data.

[2] Other morphing results are available at `https://drive.google.com/drive/folders/1HsnzVwQdFyPzgO2AxKHIhUbfyupgNaEm?usp=sharing`

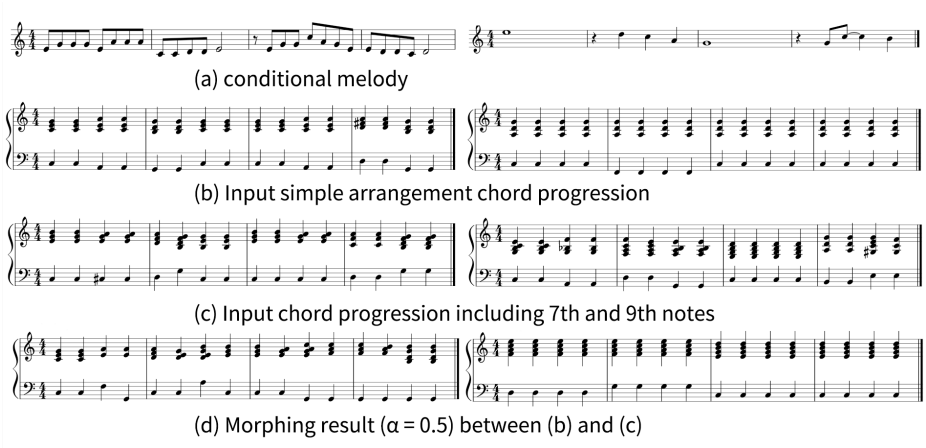

Fig. 4. Decoded results from changing melody condition.

3.2 Changing Melody Conditions

We kept the chord progressions and chroma features intact, but changed the melody conditions to decode them. Figure 4 shows the decoded results from changing the melody condition[5]. The chord progression changed from the conditions and fit the melodies. Focusing on the notes that compose chords, the simple chord progression became a progression in which the bass note changed slightly after decoding. Also, we found that the progression that included seventh and ninth notes was converted into one that consisted of seventh, ninth, and tension notes. The decoded results fit the melody and reflect the component notes of the input chord progressions.

4 Conclusion

Our study addressed supplementation of the latent space of VAE. We analyzed chord progressions generated by training a number of reharmonized chord progressions. Experimental results confirmed that an LSTM-VAE has a higher reconstructed quality than a VAE. By changing the condition of the melody, a chord progression was generated that fit the melody and reflected the style of the input.

We should consider musical constraints on the model. The chord progression is separated every four bars, and the structure needs to be considered since imperfect resolutions are mixed with perfect resolutions. Therefore, we must examine them in future studies.

[5] Other results are available at `https://drive.google.com/drive/folders/1HsnzVwQdFyPzgO2AxKHIhUbfyupgNaEm?usp=sharing`

References

Brunner, G., Konrad, A., Wang, Y., & Wattenhofer, R. (2018). Midi-vae: Modeling dynamics and instrumentation of music with applications to style transfer. In *Proceedings of the 19th international society for music information retrieval conference (ismir).*

Hirata, K., Tojo, S., & Hamanaka, M. (2011). Melodic morphing algorithm in formalism. In *Proceedings of third international conference mathematics and computation in music (mcm).*

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of international conference on learning representations* (pp. 1–13).

Kingma, D. P., & Welling, M. (2014a). Auto-encoding variational bayes. In *Proceedings of international conference on learning representations.*

Kingma, D. P., & Welling, M. (2014b). Semi-supervised learning with deep generative models. In *Proceedings of advances in neural information processing systems (nips).*

Masuda, N., & Iba, H. (2018). Musical composition by interactive evolutionary computation and latent space modeling. In *Proceedings of ieee international conference on systems, man and cybernetics.*

Murata, S., Bando, Y., Itoyama, K., & Yoshii, K. (2018). Melody and chord morphing using variational auto encoder. In *The 80th national convention of ipsj.*

Roberts, A., Engel, J., & Eck, D. (2017). Hierarchical variation autoencoders for music. In *Proceedings of 31st conference on neural information processing system.*

Simon, I., Roberts, A., Raffel, C., Engel, J., Hawthorne, C., & Eck, D. (2018). *Learning a latent space of multitrack measures.* arXiv preprint arXiv:1806.00195.

Uemura, A., & Kitahara, T. (2018). Preliminary study on morphing of chord progression. In *Proceedings of third international conference on computer simulation of musical creativity (csmc).*