A NOVEL DATASET FOR TIME-DEPENDENT HARMONIC SIMILARITY BETWEEN CHORD SEQUENCES

Franca Bittner **Jakob Abeßer** **Christon-Ragavan Nadar**

Hanna Lukashevich **Patrick Kramer**

Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany

jakob.abesser@idmt.fraunhofer.de

ABSTRACT

State-of-the-art algorithms in many music information retrieval (MIR) tasks such as chord recognition, multipitch estimation, or instrument recognition rely on deep learning algorithms, which require large amounts of data to be trained and evaluated. In this paper, we present the `IDMT-SMT-CHORD-SEQUENCES` dataset, which is a novel synthetic dataset of 45,000 chord progressions played on 45 different musical instruments. The dataset is organized in a triplet fashion and each triplet includes one "anchor" chord sequence as well as one corresponding similar and dissimilar chord progression. The audio files are synthesized from MIDI data using different sound fonts from FluidSynth. Furthermore, we conducted a benchmark experiment on time-dependent harmonic similarity based on learnt embedding representations. The results show that a convolutional neural network (CNN), which considers the temporal context of a chord progression outperforms a simpler approach based on temporal averaging of input features.

1. INTRODUCTION

In music production, it is common to integrate audio samples for different purposes in the final mix of a track. Samples can act as individual sound effects, audio loops, or ambient sound. Nowadays, large audio sample libraries are available, which complicates an efficient browsing and fast retrieval of matching audio samples in a given situation. As a solution, a recommendation system, which is integrated into a DAW could prevent enduring searches of the database and hence support the creative workflow.

If we consider audio samples with chord progressions, musical knowledge can facilitate the choice by pre-selecting harmonically matching samples. However, chords are inherently ambiguous as they heavily depend on the context they are played in. From a music production perspective, it is also important to consider the temporal order and rhythm of chords in chord sequences. Chord sequences in popular music genres like electronic dance music (EDM), rap, or pop are commonly looped after a few bars. This motivates to find similar samples not only based on harmonic, but also time-depending features, such as rhythm or tempo. If a matching sample was found, it often needs to be pitch shifted to precisely match the harmonic requirements. However, since algorithms for shifting the pitch of polyphonic audio recordings often introduce audible artifacts, only minor pitch differences between are preferred.

In this work, we present a licence-free dataset which contains 45,000 chord sequences organized in 15,000 triplets. Each triplet includes a reference chord sequence (anchor) as well as a similar chord sequence (positive) and a dissimilar counter part (negative). This triplet-based structure allows to use the dataset for instance for research on triplet loss based metric learning. For each chord sequence, the dataset includes the symbolic note annotation as MIDI file, a synthesized audio file, as well as additional metadata concerning the synthesis process. The dataset is mainly intended for research purpose. As a showcase, we conduct a baseline experiment to demonstrate the dataset's use for research on time-dependent harmonic similarity. Furthermore, the dataset is useful for other MIR tasks such as automatic chord recognition (ACR), multipitch estimation (MPE), as well as automatic instrument recognition (AIR).

2. RELATED WORK

Several datasets containing chord progressions already exist. The Lakh MIDI Dataset [1] consists not only of chords, but also includes the lead melodies of a song. The Freesound Loop Dataset [2] addresses the topic of audio loops for electronic music. However, this dataset includes no chord annotations but only metadata about instruments, tempo, and key. Another dataset of MIDI-based chord and melodic progression including different playing styles was proposed in [3]. Furthermore, several dataset including aligned audio and MIDI files were published for guitar [4], classical piano [5], and jazz performances [6, 7].

3. DATASET

The IDMT-SMT-CHORD-SEQUENCES [1] was created in three steps: the generation of MIDI data, synthesis of WAV files, and dataset splitting. To generate the MIDI data, we first created an anchor chord progression by randomizing parameters like tempo, number of chords per bar, key, instrument, and the chord progression itself. The cadence contains simple harmonic functions which are commonly used in popular music. The chords are selected from a pool with major and minor chords in a specific key. Then, we generated the similar chord progression (positive) by transposing the anchor by a maximum of two semitones in any direction. Finally, the dissimilar chord progression (negative) was created by replacing around 75% of all chords by other (unrelated) major and minor chords. The velocity (volume) and tempo is allowed to vary between anchor, positive and negative. The synthesis parameter settings are stored in a JSON format and the same random seed was used for anchor, positive and negative samples for reproducible results. This seed also acts as ID that indicates in the filename if a sample belongs to a triplet. The instrument type was not changed within a triplet to permit a dataset split on instrument level.

We used the software FluidSynth [2] with the SoundFont "GeneralUser GS" [3] to convert the MIDI files to WAV files. Finally, the dataset is split into a training, validation, and test subset on an instrument level with the approximate ratio of 80/10/10. The distribution is stored in a CSV file, along with other general metadata about all chord sequences in the dataset. The final dataset consists of 45,000 individual chord progressions, or 15,000 triplets, with a duration between 4 to 32 seconds. It comprises 45 different instruments and has a total size of about 3 GB.

4. EXPERIMENTS

4.1 Methods

Three approaches for time-dependent harmonic similarity estimation are evaluated in our baseline experiment. The first method is based on a convolutional neural network (CNN), which is trained as Siamese network using the triplet loss [8] and uses stacked chromagrams as input (as described below). The network has a VGG-style architecture with three pairs of convolutional layers with increasing filter numbers (32, 64, and 128) and intermediate max-pooling. Batch normalization and a non-linearity (ReLU) is applied after each convolutional layer. Finally, the output is flattened for the final embedding vector. As a baseline, we compared the triplet-based learning approach with a representation, which is obtained by simply averaging frame-wise chromagrams over the full duration of a chord sequence. Here, we compute the similarity between two embedding vectors as the highest cross-correlation for 4 (two in each direction) as well as all 12 possible pitch class transpositions. For both methods, we compare the two chromagram feature types Chroma Energy Normalized Statistics (CENS) [9] and Chroma DCT-Reduced log Pitch (CRP) [10], which has a higher timbre invariance. For comparability, we interpolated chromagram values for 64 equidistant frames to obtain a tempo-invariant feature representation. Furthermore, we vertically stacked two copies of the chromagram to enable the convolutional layers to learn transposition-invariant patterns.

[1] https://www.idmt.fraunhofer.de/en/publications/datasets.html

[2] https://www.fluidsynth.org. Default settings were used.

[3] https://www.schristiancollins.com/generaluser.php

Approach	Triplet Score
Siamese NN (CENS)	**99.63%**
Siamese NN (CRP)	**99.63%**
Averaged chroma (CENS, 4 shifts)	96.85%
Averaged chroma (CENS, 12 shifts)	96.67%
Averaged chroma (CRP, 4 shifts)	95.93%
Averaged chroma (CRP, 12 shifts)	95.19%

Table 1. Triplet score results for all methods and chromagram feature types.

4.2 Evaluation Metric

The *triplet score* specifies the percentage of correctly classified triplets based on the distance between anchor, positive, and negative. Each classification is considered as correct if the distance between anchor and positive is smaller than the one between anchor and negative. This metric is simple to implement and straight-forward to interpret.

4.3 Results

Table 1 summarizes the triplet scores obtained when applying both methods combined with both chromagram feature types on the 540 test set triplets. The Siamese neural networks (NNs) achieves the highest triplet score values above 99 %. Even though they omit the temporal feature progression, time-averaged chromagrams still yield a high triplet score between 95 % and 96 %. Eight more triplet pairs are detected correctly with the CENS feature in comparison to the CRP feature. The triplet score can slightly be improved by computing the cross-correlation only for 4 semitone shifts instead of all 12.

5. CONCLUSION

The proposed IDMT-SMT-CHORD-SEQUENCES has several application scenarios in MIR research. As one example, we presented in this paper the results of an initial experiment on time-dependent harmonic similarity estimation. The results indicate a potential of machine learning based approaches to learn temporal patterns which are useful to characterize the similarity between chord progressions. Since the audio samples are artificially generated with a synthesizer, a domain shift to real-world audio data is inevitable. Nevertheless, we believe that the dataset is useful for initial research on the given task due to the large amount of data triplets included.

6. REFERENCES

[1] C. Raffel, "Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching," Ph.D. dissertation, Columbia University, 2016.

[2] A. Ramires, F. Font, D. Bogdanov, J. B. L. Smith, Y.-H. Yang, J. Ching, B.-Y. Chen, Y.-K. Wu, H. Wei-Han, and X. Serra, "The Freesound Loop Dataset and Annotation Tool," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020, pp. 287–294.

[3] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, G. Bin, and G. Xia, "POP909: A pop-song dataset for music arrangement generation," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020, pp. 38–45.

[4] Q. Xi, R. M. Bittner, J. Pauwels, X. Ye, and J. P. Bello, "GuitarSet: A dataset for guitar transcription." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 453–460.

[5] Q. Kong, B. Li, J. Chen, and Y. Wang, "GiantMIDI-Piano: A large-scale MIDI dataset for classical piano music," *arXiv preprint arXiv:2010.07061*, 2020.

[6] V. Eremenko, E. Demirel, B. Bozkurt, and X. Serra, "Audio-Aligned Jazz Harmony Dataset for Automatic Chord Transcription and Corpus-based Research," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 483–490.

[7] T.-P. Chen, S. Fukayama, M. Goto, and L. Su, "Chord Jazzification: Learning Jazz Interpretations of Chord Symbols," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020, pp. 360–367.

[8] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France, 2015.

[9] M. Müller, F. Kurth, and M. Clausen, "Audio matching via chroma-based statistical features," in *Proc. of the 6th Int. Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 288–295.

[10] M. Müller, S. Ewert, and S. Kreuzer, "Making chroma features more robust to timbre changes," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.