Analyzing Byte-Pair Encoding on Monophonic and Polyphonic Symbolic Music: A Focus on Musical Phrase Segmentation

Dinh-Viet-Toan Le[1], Louis Bigo[2] and Mikaela Keller[1]

[1]Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille
[2]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence
dinhviettoan.le@univ-lille.fr

Abstract

Byte-Pair Encoding (BPE) is an algorithm commonly used in Natural Language Processing to build a vocabulary of subwords, which has been recently applied to symbolic music. Given that symbolic music can differ significantly from text, particularly with polyphony, we investigate how BPE behaves with different types of musical content. This study provides a qualitative analysis of BPE's behavior across various instrumentations and evaluates its impact on a musical phrase segmentation task for both monophonic and polyphonic music. Our findings show that the BPE training process is highly dependent on the instrumentation and that BPE "supertokens" succeed in capturing abstract musical content. In a musical phrase segmentation task, BPE notably improves performance in a polyphonic setting, but enhances performance in monophonic tunes only within a specific range of BPE merges.

1 Introduction

A major similarity between text and music lies in their nature as semiotic systems, as they can be represented as sequences of elements (Lerdahl, 2013). This common characteristic has led to numerous adaptations of Natural Language Processing (NLP) methods in the domain of symbolic music analysis and generation (Le et al., 2024). Mirroring the view of a text as a sequence of tokens representing words, subwords or characters, *tokenization* practices have also been adopted to process symbolic music. Several choices of types of musical "characters" and various tokenization algorithms to segment the sequence of musical "characters" have been proposed (Kumar and Sarmento, 2023).

However, music profoundly differs from text, notably because of some structural characteristics such as rhythm or polyphony (Jackendoff, 2009). We can, thus expect tokenization algorithms such as Byte-Pair Encoding (BPE) (Sennrich et al., 2016) to behave differently when applied to text or music. The aim of this study is to highlight some commonalities and differences in BPE behaviors with multiple types of music as compared to text. This work is twofold: we first propose a statistical description of the vocabulary of tokens obtained when BPE is applied to text compared to the vocabularies obtained with various types of music. This comparison highlights some musical properties captured by this tokenization algorithm (Section 3). Informed by these observations, we then focus on a downstream task, musical phrase segmentation, to quantitatively compare the impact of BPE on monophonic and polyphonic music (Section 4).

2 Subword tokenization in symbolic music

Subword tokenization, where tokens are subwords instead of characters or words, is a common practice in NLP. It is used to deal with out-of-vocabulary words that are obtained by combining multiple subwords. Multiple algorithms have been proposed to build from a corpus the most representative vocabulary of subwords, including Byte-Pair Encoding (BPE) (Sennrich et al., 2016), WordPiece (Schuster and Nakajima, 2012) or Unigram (Kudo, 2018). BPE was initially developed as a compression algorithm (Gage, 1994) before being applied to text as a tokenization method. The algorithm relies on creating new subword tokens by iteratively merging the most recurring pairs of successive tokens in a corpus until a chosen vocabulary size is reached. In the following, we call *atomic elements* the tokens from the initial vocabulary and *supertokens* the tokens added through BPE.

Some recent MIR studies have applied these algorithms to symbolic music (Kumar and Sarmento, 2023). BPE was first implemented to shorten token sequences (Liu et al., 2022). Fradet et al. (2023) specifically analyzed BPE for MIDI gen-

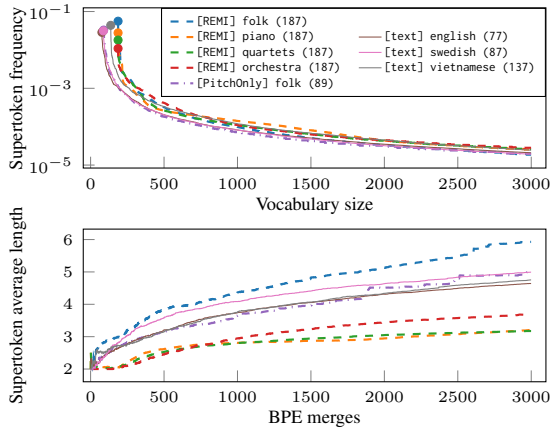

Figure 1: *(Top)* Frequency of the created supertokens through the vocab size increasing with the BPE steps, for different styles of music and multilingual text data. *(Bottom)* Average length of already created supertokens through BPE iterations for musical and text data. The initial vocabulary size of each tokenization is indicated.

eration purposes and showed that the learned embedding spaces are more structured. However, when applied to piano analysis tasks, a BPE with 4 times the initial vocabulary size does not seem to show any downstream improvement in model performance (Zhang et al., 2023). In contrast, Park et al. (2024) focus on specifically applying BPE on monophonic tunes using a pitch/duration-only representation and show that BPE enables the retrieval of style-specific motifs.

To date, research on BPE for symbolic music has focused on its evaluation on generation or global sequence classification tasks. Its behaviour has not been analyzed in depth, in particular when applied to various instrumentations. This work specifically focuses on these issues, with a descriptive analysis of BPE vocabularies followed by a quantitative evaluation of BPE on monophonic and polyphonic music on a musical phrase segmentation task.

Our experiments rely on the MidiTok package (Fradet et al., 2021) to handle the tokenization process and the HuggingFace library (Wolf et al., 2020) implementing Transformer models. We publicly release the datasets and source code, which are available at `http://algomus.fr/code/`.

3 Analyzing music BPE

In this section, we present analyses of the vocabulary produced by Byte-Pair Encoding when applied to text and music. We first analyse supertokens induced by various instrumentations as well as their relation to high-level or abstract musical features.

3.1 Comparing text and music BPEs

Musical notes are often compared to text at the level of characters (Hirata et al., 2022). Deep learning models have been shown to be more efficient when dealing with characters grouped into (sub)words (Shapiro and Duh, 2018; Tay et al., 2022). Therefore, we study the BPE results when processed, on text and music, in order to observe common or distinctive operating regime on such data with various languages and instrumentations. Text data includes alphabetic[1] languages from various regions, extracted from the XLNI dataset (Conneau et al., 2018) on which we run BPE on 100k premises. For music, we compare monophonic folk tunes, classical piano, string quartet, and orchestral corpora with similar sizes and tokenize these datasets using REMI (Huang and Yang, 2020) from which `Velocity` tokens are removed.

We first study the occurrence frequency of the newly created supertoken within the corpus, at each step of the training (Figure 1, top). To make the corpora and vocabularies comparable, supertoken frequencies are normalized by the initial corpus length, and the BPE iterations are aligned with the resulting vocabulary size. Interestingly, the vocabularies obtained on music or text through BPE do not show major differences with respect to the decay rate or the order of magnitude of the frequencies.

We also compute the mean length of the supertokens through the BPE steps (Figure 1, bottom). The evolution of supertoken length differs between text and music, depending on the instrumentation. While monophonic supertokens are generally longer than polyphonic ones, orchestra supertokens surprisingly appear to be longer than piano or string quartet ones. An in-depth study of the constructed vocabulary shows that the orchestral vocabulary predominantly consists of "harmonic" supertokens formed of simultaneous notes. In contrast, piano and string quartet vocabularies include both simultaneous and consecutive notes. This difference causes BPE to struggle to build long piano or string quartet supertokens. On a separated experiment, we observed that it takes over 10 times more steps on a piano corpus to get an average length comparable to that of the vocabulary obtained on the monophonic corpus. Moreover, when considering an alphabet which only keeps pitch tokens, we

[1]Experiments have also been conducted on syllabic (Japanese) and logographic (Chinese, Korean) languages, that show major differences due to the different nature of the atomic elements of their initial vocabulary.

show that monophonic supertoken lengths have a regime closer to that of text for this range of BPE merges (Figure 1, "PitchOnly" curve), while polyphonic curves still stand out. We can thus posit that the differences between the music and text curves might be due to simultaneity and timing information, which are inherent to music.

3.2 Musical content carried by supertokens

So far, we have drawn a broad characterisation of the BPE vocabularies, let us now zoom in and try to delineate which supertokens are present in a specific context. Borrowed from text, the terms "musical phrase" or "musical sentence" (Nattiez, 1990) denote a part of the music which can give the impression of a complete statement by its own. The TAVERN dataset (Devaney et al., 2015) include such phrase annotations.

Using a Structured (Hadjeres and Crestel, 2021) tokenization with pitches encoded as intervals (Kermarec et al., 2022) we analyzed the segmentation induced on the sequences by a 1024-merge BPE. This tokenization allows taking advantage of both Structured's relative encoding of rhythm with time-shifts and the relative encoding of pitches through intervals. A first observation is that only 4.2% of the supertokens among the tokens of the sequences do overlap phrases. In contrast, randomly splitting the piece into the same number of chunks as BPE segmentation results in 71% overlap ratio, indicating that supertokens are unlikely to span across phrase boundaries.

We then analysed the supertokens occurring at the beginning and end of musical phrases. In particular, our chosen tokenization allows this analysis to be key signature-independent and bar position-independent. The most recurrent start-of-phrase supertoken appears to be a melodic rising perfect fourth (Figure 2, top), which follows musicology studies (Meyer, 1973, p.145): "*an upbeat interval of a perfect fourth, moving to the tonic [...] may be understood as a rhythmic-harmonic event emphasizing the tonic on which the melody proper begins.*" Most represented end-of-phrase supertokens include descending arpeggio patterns on the tonic chord (Figure 2). This also verifies some musicological observations (Huron et al., 1996): "*Melodic passages tend to exhibit an arch shape where the overall pitch contour rises and then falls over the course of a phrase or an entire melody*". Therefore, similar to how BPE can capture syntactic rules in text, we observe that musical supertokens also convey high-level musical information.

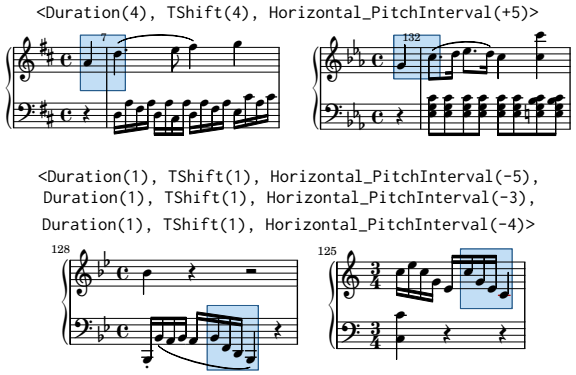

Figure 2: *(Top)* First most common start-of-phrase supertoken from Mozart's K.25 and Beethoven's WoO.68. *(Bottom)* 9-long common ending supertoken (10th most common) from Beethoven's WoO.73 and Mozart's K.179. The tokenization is Structured + intervals.

4 Evaluating BPE on musical phrase segmentation

BPE applied to MIDI-derived tokenization has been mainly evaluated through classification tasks with composer classification, on a general multi-track dataset (Fradet et al., 2023) or specifically piano music (Zhang et al., 2023). Inspired by sentence segmentation tasks in NLP (Read et al., 2012) and given our preliminary results showing that supertokens can play a role in musical phrase boundaries, we aim to quantitatively evaluate BPE on a task of musical phrase segmentation for monophonic and polyphonic datasets.

4.1 Musical phrase segmentation

We consider a *musical phrase segmentation task*, where a model is trained to tag each token of a sequence as being a start-of-phrase or not (Guan et al., 2018). For BPE sequences, if a start-of-phrase occurs within a supertoken, the whole supertoken is annotated as being a start-of-phrase.

We first performed this task on the MTC dataset (Van Kranenburg et al., 2014) composed of monophonic Dutch folk tunes and including phrase annotations. The MTC dataset contains 100 times more phrase annotations than TAVERN. Moreover, the nature of classical-style musical phrases, generally based on cadences (Spencer and Temko, 1994), may differ from folk music phrases, based on melodic contours (Huron et al., 1996). Therefore, for a fairer comparison, we discard TAVERN as our polyphonic dataset and we build and release a synthetic dataset of folk music piano arrange-

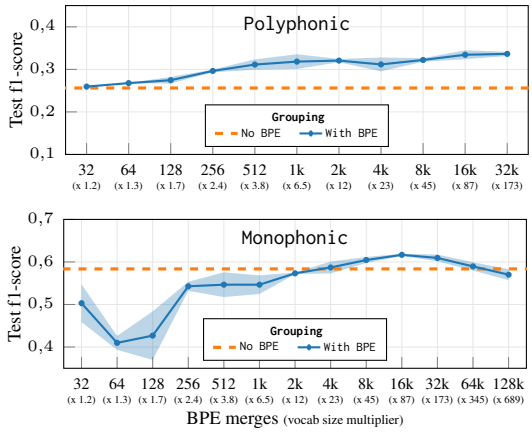

Figure 3: f1-score for start-of-phrase classification on the polyphonic *(top)* and monophonic dataset *(bottom)*.

ments from the MTC dataset generated by the AccoMontage model (Zhao and Xia, 2021) aligned with the original phrase annotations. We tokenize both datasets using REMI (Huang and Yang, 2020) and remove the `Velocity` tokens, for simplicity.

Note that the non-BPE dataset is by design more unbalanced than the BPE one. In the polyphonic setting, the proportion of start-of-phrases increases from 1.2% in the whole dataset to 3.3% after 128k merges, respectively from 2% to 27% in the monophonic dataset.

4.2 Experiments

We trained a 2-layer Transformer encoder-only model with 8 heads per layer and a common embedding size between BPE and non-BPE vocabularies on each dataset. We evaluate each model on 3 different splits of the datasets, using the F1-score of the start of phrase label prediction. As our experiments focus on representation impact, we chose to have light models rather than ones achieving optimal performance.

The polyphonic setting of our experiment seems to indicate that BPE can have an impact on performance. Indeed, unlike Zhang et al. (2023) also focusing on piano music, who demonstrated on a sequence global classification task that a BPE (with the initial vocabulary size $\times 4$) does not result in significant improvements, we see on this local classification task that the performance increases with the number of merges (Figure 3, top).

Our results on the monophonic dataset show even that BPE with too few number of merges can degrade the performance (Figure 3, bottom). This surprising behavior also occurs in NLP tasks, where character-based models can outperform

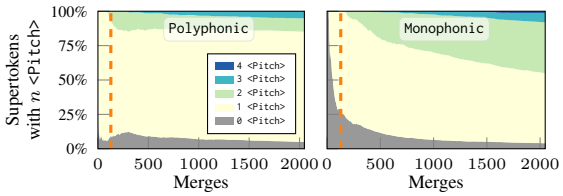

Figure 4: Ratio of supertokens containing n `<Pitch>` atomic elements in the vocabulary for each number of BPE merges.

subword-based models (Chung et al., 2016).

Figure 4 describes the "melodic" content of the supertokens created along BPE steps. An analysis of supertokens reveals that early merges tend to produce *structural* supertokens, such as combinations of `Bar` and `Beat` (Figure 4 gray area: proportion of created supertokens with 0 `<Pitch>` atomic element), while melodic patterns emerge later, and at different rates for monophonic and polyphonic datasets. At 128 merges (Figure 4, dashed line), 26% of monophonic supertokens do not include any `<Pitch>` atomic element (gray area) while this ratio is only 9% for polyphonic and 7% contain 2 `<Pitch>` atomic element (green area). Fewer melodic patterns, which are more likely to indicate phrase boundaries in monophonic tunes (Huron et al., 1996), may explain why the BPE model performs better only after a certain number of merges.

In the monophonic dataset we also see that, after too many merges, the model performance drops. An analysis of the supertoken length shows that, after 128k merges, monophonic supertokens are on average 38.6-long (compared to 8.4 for polyphonic ones). Indeed, the smaller size of the monophonic dataset ($3\times$ smaller than the polyphonic one) leads late steps supertokens to capture long but rare patterns that might be less relevant for this task of phrase segmentation.

5 Conclusion

In this work, we show that Byte-Pair Encoding behaves differently depending on the type of music it is trained on. With a descriptive approach, we highlight that the resulting vocabulary highly depends on the type of instrumentation, and supertokens can carry high-level musical content. On a downstream task, we confirm the impact of instrumentation on the model performance and show that the number of BPE merges should be chosen carefully. For future work, we think the initial tokenization impact over BPE performance should be investigated.

Acknowledgments

This work was supported by grant ANR-20-THIA-0014 program "AI_PhD@Lille". The authors would like to thank Zih-Syuan Lin for providing feedbacks on earlier versions of the paper.

References

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703, Berlin, Germany. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Johanna Devaney, Claire Arthur, Nathaniel Condit-Schultz, and Kirsten Nisula. 2015. Theme and variation encodings with roman numerals (TAVERN): A new data set for symbolic music analysis. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Nathan Fradet, Jean-Pierre Briot, Fabien Chhel, Amal El Fallah-Seghrouchni, and Nicolas Gutowski. 2021. MidiTok: A Python package for MIDI file tokenization. In *International Society for Music Information Retrieval Conference (ISMIR), Late-Breaking Demo Session*.

Nathan Fradet, Nicolas Gutowski, Fabien Chhel, and Jean-Pierre Briot. 2023. Byte pair encoding for symbolic music. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2001–2020, Singapore. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Yixing Guan, Jinyu Zhao, Yiqin Qiu, Zheng Zhang, and Gus Xia. 2018. Melodic phrase segmentation by deep neural networks. *Preprint*, arXiv:1811.05688.

Gaëtan Hadjeres and Léopold Crestel. 2021. The piano inpainting application. *Preprint*, arXiv:2107.05944.

Keiji Hirata, Satoshi Tojo, and Masatoshi Hamanaka. 2022. Music as formal language. In *Music, Mathematics and Language: The New Horizon of Computational Musicology Opened by Information Science*, pages 51–78, Singapore. Springer Nature Singapore.

Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1180–1188, New York, NY, USA. Association for Computing Machinery.

David Huron et al. 1996. The melodic arch in western folksongs. *Computing in Musicology*, 10:3–23.

Ray Jackendoff. 2009. Parallels and nonparallels between language and music. *Music Perception: An Interdisciplinary Journal*, 26(3):195–204.

Mathieu Kermarec, Louis Bigo, and Mikaela Keller. 2022. Improving tokenization expressiveness with pitch intervals. In *International Society for Music Information Retrieval Conference (ISMIR), Late-Breaking Demo Session*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Adarsh Kumar and Pedro Sarmento. 2023. From words to music: A study of subword tokenization techniques in symbolic music generation. *Preprint*, arXiv:2304.08953.

Dinh-Viet-Toan Le, Louis Bigo, Mikaela Keller, and Dorien Herremans. 2024. Natural language processing methods for symbolic music generation and information retrieval: A survey. *Preprint*, arXiv:2402.17467.

Fred Lerdahl. 2013. Musical Syntax and Its Relation to Linguistic Syntax. In *Language, Music, and the Brain: A Mysterious Relationship*. The MIT Press.

Jiafeng Liu, Yuanliang Dong, Zehua Cheng, Xinran Zhang, Xiaobing Li, Feng Yu, and Maosong Sun. 2022. Symphony Generation with Permutation Invariant Language Model. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Leonard B. Meyer. 1973. *Explaining Music: Essays and Explorations*, DGO - Digital original edition. University of California Press.

Jean-Jacques Nattiez. 1990. *Music and discourse: Toward a semiology of music*. Princeton University Press.

Saebyul Park, Eunjin Choi, Jeounghoon Kim, and Juhan Nam. 2024. Mel2word: A text-based melody representation for symbolic music analysis. *Music & Science*, 7.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India. The COLING 2012 Organizing Committee.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Pamela Shapiro and Kevin Duh. 2018. BPE and CharCNNs for translation of morphology: A cross-lingual comparison and analysis. *Preprint*, arXiv:1809.01301.

Peter Spencer and Peter M Temko. 1994. *A practical approach to the study of form in music*. Waveland Press.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations (ICLR)*.

Peter Van Kranenburg, MJ de Bruin, Louis P Grijp, and Frans Wiering. 2014. The Meertens tune collections. *Meertens Online Reports*, 2014(1).

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Huan Zhang, Emmanouil Karystinaios, Simon Dixon, Gerhard Widmer, and Carlos Eduardo Cancino-Chacón. 2023. Symbolic music representations for classification tasks: A systematic evaluation. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Jingwei Zhao and Gus Xia. 2021. Accomontage: Accompaniment arrangement via phrase selection and style transfer. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, pages 833–840.