2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Neural network for musical data mining for phrase boundary detection

1st Daniel Henel
Dept. of Applied Computer Science
AGH University of Krakow
Krakow, Poland
danielhenel@student.agh.edu.pl

2nd Aleksander Mazur
Dept. of Applied Computer Science
AGH University of Krakow
Krakow, Poland
olekmazur@student.agh.edu.pl

3rd Marcin Retajczyk
Dept. of Applied Computer Science
AGH University of Krakow
Krakow, Poland
retajczyk@student.agh.edu.pl

4th Weronika T. Adrian
Dept. of Applied Computer Science
AGH University of Krakow
Krakow, Poland
wta@agh.edu.pl

5th Krzysztof Kluza
Dept. of Applied Computer Science
AGH University of Krakow
Krakow, Poland
kluza@agh.edu.pl

6th Adrian Horzyk
Dept. of Biocybernetics and
Biomedical Engineering
AGH University of Krakow
Krakow, Poland
horzyk@agh.edu.pl

***Abstract*—The surge of interest in artificial intelligence systems has sparked new directions of research in the realm of music data analysis. At the same time, the exploration and exploitation of intrinsic musical structures and sequences, a timeless endeavor, continue to captivate scholars and practitioners alike.**

In this context, the fusion of computational techniques with music analysis emerges as a natural progression. One of the pivotal crossroads in this convergence is the identification of musical phrase boundaries, pivotal demarcations that underpin the organizational fabric of a musical composition.

This article pioneers an inventive approach to address the challenge of detecting these musical phrase boundaries, harnessing the power of artificial neural networks. However, the innovation does not stop there; the pinpointed phrase boundaries undergo a comprehensive dissection utilizing pattern mining techniques. The focus of this analysis is on unveiling recurrent motifs and classifying phrases into coherent clusters, predicated on the repetitions and similarities exposed through these neural network-driven techniques.

This exploration was conducted using an extensive repository of folk songs, a treasure trove of foundational musical expressions that indelibly shape the stylistic contours of musical compositions. We claim that the presented approach not only opens avenues for penetrating and nuanced analysis, but also enriches our comprehension of the intricate interplay of musical components and their manifestations inspired by neural networks.

***Index Terms*—data mining, neural networks, computational intelligence, music analysis, musical phrases detection.**

I. Introduction

Music, as an expressive art form, has captivated humanity for centuries. It serves as a universal language that communicates emotions, stories, and cultural identities. Understanding the structure and organization of music is vital for musicians, musicologists, and music enthusiasts alike, as it provides insight into the creative choices made by composers and helps unravel the intricate layers of meaning embedded within a composition.

In recent years, computational music analysis has gained significant attention in the field of Music Information Retrieval (MIR), propelled by remarkable advances in deep learning techniques [1]–[8]. However, while these techniques have shown promise, the potential of integrating neural network approaches for musical phrase boundary detection remains relatively underexplored. To the best of our knowledge, there is only a single paper [9] that presents the usage of process mining methods for music analysis; however, it relies on bar-level segmentation. However, this perspective poses a limitation since musical compositions often transcend strict adherence to repeated bars, encompassing broader and more intricate musical ideas.

This paper seeks to bridge this gap by proposing an innovative approach that synergizes neural network techniques with the principles of music theory. Specifically, we introduce a novel methodology for segmenting musical compositions into phrases, leveraging the capabilities of artificial neural networks. These phrases can then be subjected to various analyses, such as discovering similarities between phrases with pattern mining. In this context, a phrase is defined as a substantial musical thought created through an interaction of melody, harmony, and rhythm and ended with a cadence [10]. In Fig. 1, number 1 is used to signify the beginning of a musical phrase, while number 2 denotes its end. Number 3 represents a musical measure or bar, and number 4 indicates its conclusion. Number 5 refers to the entire musical phrase, while number 6 represents a single musical note.

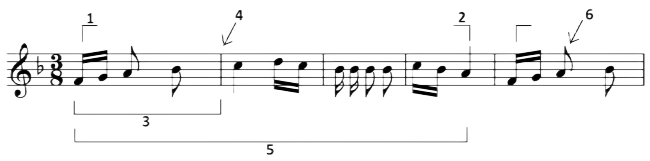

Fig. 1. Example of the phrased song "Mości gospodarzu" (polish), i.e. "Lord, my host" (in translation)".

Accurate identification of musical phrase boundaries bears

significance for various music-related applications, including transcription, analysis, and automated composition. However, the process of partitioning a musical piece into phrases using neural network-driven techniques remains relatively nascent [11]. Moreover, the subjective nature of phrase boundaries, where two humans rarely annotate the boundaries at exactly the same positions, adds complexity to the problem [12].

In this paper, we propose an innovative approach that not only advances the detection of musical phrase boundaries, but also enriches the landscape of computational music analysis. By harnessing the power of artificial neural networks, we endeavor to provide a more nuanced and insightful understanding of the intricate dynamics that shape musical compositions.

II. Related Works

Various methods have been developed to tackle the task of music segmentation; however, they introduced different definitions of a phrase than in this paper. An effective approach involves using convolutional neural networks (CNNs) to identify distinct sections in music compositions. An example of this approach is the use of CNNs in the SALAMI dataset, which contains a large collection of annotated songs in audio format (SALAMI dataset[1]). While SALAMI is renowned for its comprehensive annotations and focus on popular music, its segmentation process is primarily influenced by shifts in tonality, timbre, coloration, and instrumentation. Consequently, this segmentation method does not identify musical phrases, but instead delineates segments of a musical composition characterized by similar sonic attributes.

In the contribution [13], a sequence-to-sequence autoencoder is introduced, harnessing acoustic features such as Mel-Frequency Cepstral Coefficients to represent variable-length audio segments with fixed-length vectors. This approach effectively captures the phonetic nuances of sound and finds applications in speech-processing domains, facilitating tasks such as speaker identification, emotion recognition, and phrase retrieval. The autoencoder excels in distinguishing segments marked by minimal phonetic differentiations. Similarly, this solution is not designed to identify musical phrases.

McFee and Lanckriet [14] propose a novel approach called Tree Measures (T-measures) inspired by Schenkerian analysis [15]. Unlike previous music segmentation methods, which often focus on shallow divisions, T-measures embrace the intricate hierarchical structure present in music. While highly effective for hierarchical musical contexts, this method faces difficulties with under- or over-segmentation in different scenarios. Likewise, this methodology is not specifically devised for the identification of musical phrases.

Nieto [16] proposed a unique approach to identify recurring themes in polyphonic music using the JKU Pattern Development Dataset. They employed computational methods to analyze music by creating multi-dimensional representations and transposition-invariant self-similarity matrices. Although these efforts showcased computational potential, they often dealt with less structured audio data and varied interpretations of musical phrases.

In contrast, our approach emphasizes symbolic music notation, which is more suitable for processing and analysis. We use carefully curated professional data, aligned with music theory principles. This ensures well-defined phrase boundaries, enriching compositions with clarity and coherence.

III. Exploring the Folk Songs Dataset

Music's essence resonates uniquely across diverse cultures, with an array of styles and genres evolving over centuries, each painting a distinct auditory tapestry. In our pursuit, we direct our attention to the realm of Polish traditional folk music, a vibrant heritage nurtured over generations. In particular, our focus is on the invaluable collection curated by Oskar Kolberg (1814-1890), a Polish ethnographer, folklorist, and composer [17].

This treasure trove of folk melodies, meticulously amassed by Kolberg, has been transcribed into the Essener Assoziativ Code (EsAC) format [18], [19] by the Institute of Art of the Polish Academy of Sciences. The ingeniously conceived EsAC format stands as a concise and machine-readable representation of musical scores, encompassing crucial information about the boundaries of musical phrases. Given that the EsAC format is not readily compatible with standard libraries, our efforts have been directed toward a transformative process aimed at facilitating the integration of this valuable musical heritage into our analysis. This endeavor involves allowing the use of the music21 library for this purpose. First, the EsAC files gracefully transitioned into Humdrum format [20], a crucial intermediary in our quest. From there, a final metamorphosis led us to the Music Encoding Initiative (MEI) format [21], ultimately facilitating the engagement with the music21 library [22].

The employment of the music21 library enabled the transformation of musical notation into numeric vector representations, a pivotal step elaborated upon in the following section, with a focus on machine learning applications.

IV. Data Preparation for Training

A. Musical Sequences Vectorization

Within the scope of this paper, our attention is moved to the core of our method, located at the intersection of music and neural networks. After parsing each musical piece into objects that encapsulate abstract classes from the music21 library, we arrive at the pivotal juncture of data transformation. This transformation endsows musical narratives with vectorized form, a prerequisite for deep learning. In this transformation, every musical piece undergoes a transmutation into a quartet of vectors, each housing distinct strands of musical information:

- **Notes:** Encoding the pitches of consecutive notes according to the MIDI scale (ranging from 0 to 127), augmented by special markers – 128 denoting rests and 129 signifying the song's inception.
- **Durations:** Capturing the durations of successive notes, adhering to the conventions of the music21 library.

[1]https://ddmal.music.mcgill.ca/research/salami/

- **Bars:** A rhythmic heartbeat, resonating 1 when the note at the corresponding position in the vector concludes a musical measure and 0 otherwise.
- **Phrases:** This vector serves as a representation of the musical structure, with a value of 1 at the positions where a note signifies the culmination of a musical phrase, and 0 elsewhere.

The transformation process for this data set was exhaustively documented and resides in our GitHub repository[2].

B. Data Cleaning

In the context of machine learning, the step of data processing must be performed meticulously, with data cleaning, standing as a pivotal prelude. In the domain of music recognition, this process takes on a cadence, where the removal of outliers and the meticulous selection of a representative subset converge to improve the quality and reliability of our machine learning model.

In our melodic odyssey, drawn from a rich set of more than 19 thousand songs, we embark on a quest to distil the essence. This entails selecting compositions that resonate with the common patterns enshrined within our dataset, thus cultivating a representative ensemble. The criteria guiding this selection encompass:

- the length of a piece should be such that more than k% of all pieces from the dataset have the same length,
- the piece has only phrases that length is such that more than p% of all phrases have the same length (from the pieces with the right length),
- the piece has only phrases that end regularly with the end of a measure.

This selection is attuned to parameters k and p, enabling a balance that reverberates with optimal harmony. Our explorations culminated with k set at 2% and p attuned to 3.5%.

As outliers are removed and our selected compositions align with our framework, our dataset becomes more refined and prepared for the effective operation of our machine learning ensemble, akin to a well-tuned instrument ready for skilled performance.

C. Network Input and Output

In the orchestration of our model, a sequential ensemble takes center stage, imbuing the essence of sequential processing within the realm of machine learning. Our architectural framework processes notes, durations, bars, and phrases sequentially, integrating these elements into a cohesive neural network for analysis and prediction.

For training choreography, our ensemble seamlessly concatenates the vectors representing notes, durations, bars, and phrases from all compositions, converging them into four contiguous sequences. This quartet of sequences is then partitioned into smaller parts of 36 elements each, a number that emerged as harmonious through empirical experimentation. To maintain the distinctiveness of each piece, we insert neutral elements as separators.

In this orchestral symphony of neural harmony, the network input unfurls as an intricate tapestry. Each element within the vector unravels into dual sub-elements: a sequence of notes and bars. Each of these sub-elements blossoms into a duet of vectors, a musical conversation between the elements.

The climax of our neural ensemble is reached in the output phase, where a single numerical value signifies the anticipation of a phrase's conclusion. As the neural conductor guides the process, the output can surge to 1, indicating the end of a phrase, or gracefully remain at 0, signaling the continuation of our melodic journey.

$$[[\text{sequence\_of\_notes}], [\text{sequence\_of\_bars}]] \rightarrow 0 \text{ or } 1$$

V. The Architecture of the Model

Within our computational framework, the architecture of our model emerges as a masterpiece. It intricately weaves neural harmonies with a central attention mechanism. Drawing inspiration from Foster's work on music generation [23], our journey is a testament to innovation and refinement.

Our initial steps involved the fusion of two vectors, uniting notes and durations. Over time, simplicity emerged as a compelling aspect. Through experimentation, we realized that freeing durations from the training process led to superior results. Merging notes with durations made things more complicated, but when we introduced the essential concept of "bars," it brought clarity to our music's rhythm and structure. This led us to create a musical canvas with two separate input layers for notes and bars.

Building upon these inputs, a sequence of transformations unfolded. Discrete notes and bars embarked on a journey, passing through dedicated embedding layers. Here, computational alchemy transformed discrete values into continuous, dense vectors.

At the core of our model, the attention mechanism took center stage. It was composed of two LSTM layers working in tandem, enriched by a dense layer with tanh activation, followed by a softmax layer. The flow of computation continued through permutations, multiplications, and lambda layers as our model surged forth.

As the music piece ended, we added a finishing touch with a dense layer that uses a technique called softmax. Our GitHub repository[3] contains the score of our work.

The visual representation present in Figure 2, a testament to the intricate architecture that underpins our melodic odyssey.

Within the intricate neural network architecture, layers collaboratively unveil patterns and transform data:

Insightful Transformation: With embedding layers, the data is transmuted into 100-dimensional vectors. With 4400 and 1900 parameters, they capture data nuances.

Embedding Fusion: The concatenate layer merges embeddings seamlessly, crafting a 200-dimensional unity without new parameters.

[2]https://github.com/Music-Miners

[3]https://github.com/Music-Miners/Music-Analysis-with-Process-Mining

Sequential data handling: LSTM layers manage time-related patterns. Through 256 units, a (None, None, 256) journey materializes, propelled by 467968 and 525312 parameters.

Expressive Culmination: The dense layer, supported by 257 parameters, its output unit distils profound insights in (None, None, 1) dimensions.

The model is guided by the RMSprop optimizer with a learning rate of 0.001. The binary cross-entropy loss function conducts this journey of mastery in binary classification.

For further exploration, the Music Analysis with Process Mining GitHub repository[4] awaits.

VI. Results

In the process of annotating boundaries in musical phrases, there is often a lack of consensus among individuals, primarily due to the inherent subjectivity of musical perception. This intricate phenomenon has previously been observed and documented, as indicated in [12]. It can be attributed to the diverse musical sensibilities that each individual possesses. Each person has their own set of expectations and preferences that influence how they perceive the boundaries of musical phrases.

Furthermore, the absence of clear regulations governing the placement of phrase endings contributes to this variability. This issue heavily depends on the specific musical counterpoint cultivated in various regions and historical periods. As a result, the boundaries of musical phrases can be perceived and interpreted in numerous ways.

What is particularly noteworthy is that our model has demonstrated exceptional performance in this task. It achieved an impressively high F1 score of 68% for both the training and the test datasets. This underscores the resilience and reliability of our approach, which allows for the accurate annotation of musical phrase boundaries in an extremely precise and efficient manner. Thus, our model makes a significant contribution to the field of musical analysis.

VII. Further Analysis Using Pattern Mining Techniques

Having successfully divided the musical piece into distinct phrases, the next step is to delve into the analysis of connections and similarities between these phrases. Typically, it may not be immediately apparent that some of these phrases share similarities, as they often exhibit subtle differences, such as a single-note variation. However, experienced musicians can discern these similarities and may consider grouping such phrases together. This grouping not only preserves the musical integrity, but also allows for the exploration of more intricate patterns and the derivation of deeper insights from the data.

To automate the process of identifying and grouping similar phrases, we employ a pattern mining technique known as the Apriori algorithm, implemented in the library[5]. This algorithm is designed to identify frequent sets of items, in our case, musical notes, across the database, which, in this context, refers to the musical piece.

It is imperative to emphasize that this approach aligns with established principles in music theory. Furthermore, it is noteworthy that the algorithm is capable of identifying similarities between phrases that adhere to different imitation rules:

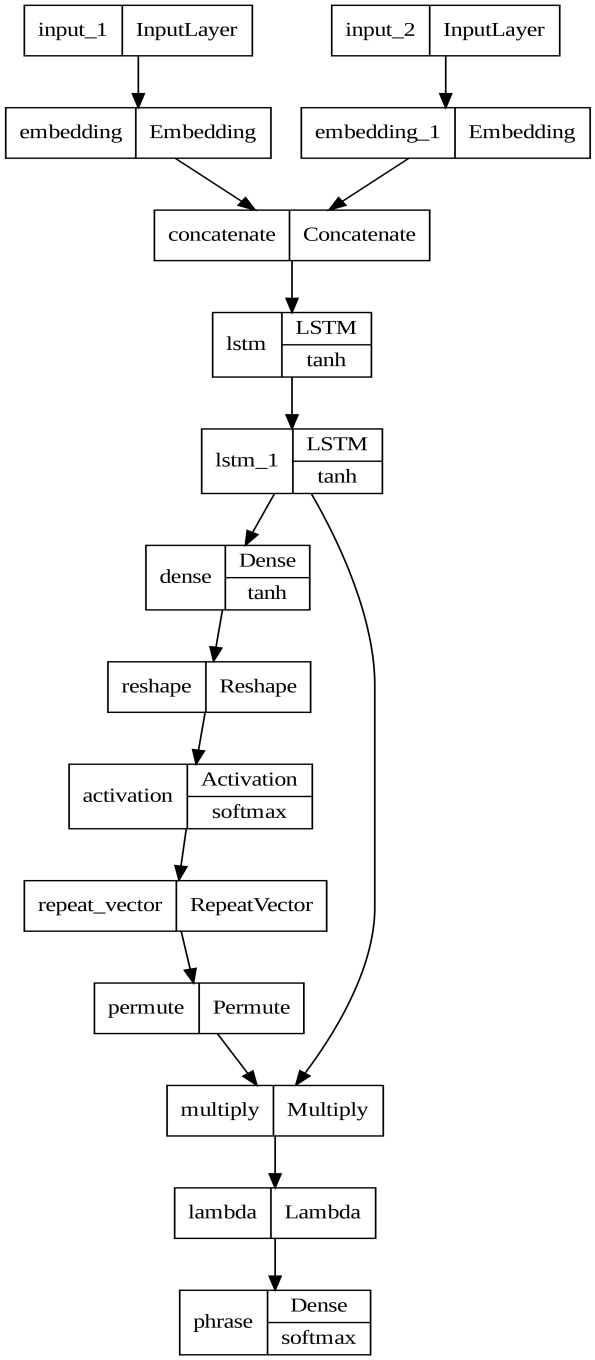

Fig. 2. The architecture of the model.

[4]https://github.com/Music-Miners/Music-Analysis-with-Process-Mining

[5]https://github.com/udayRage/PAMI

- simple imitation (every note in a phrase is moved by the same number of semitones) – Fig. 3 and 4,

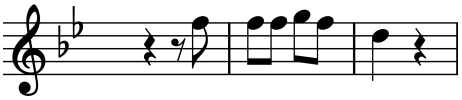

Fig. 3. Base phrase

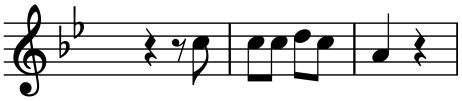

Fig. 4. Simple imitation

- inversion (reversing the direction of the interval; ascending to the corresponding falling interval and viceversa)[6],
- augmentation (a phrase is presented in longer note-values than were previously used)[7] – Fig. 5 and 6,

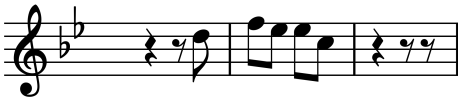

Fig. 5. Base phrase

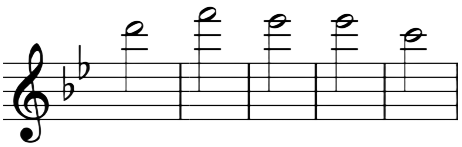

Fig. 6. Augmentation

- diminution (a phrase is presented in shorter note-values than were previously used)[8] – Fig. 7 and 8,

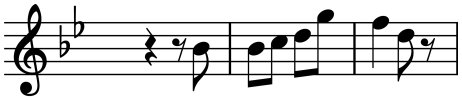

Fig. 7. Base phrase

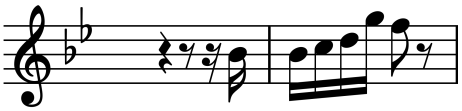

Fig. 8. Diminution

- crab imitation / retrograde (walking backward, the pitches and rhythms are in reverse)[9],
- imitation with a variety of ornaments – Fig. 9 and 10.

[6]https://en.wikipedia.org/wiki/Inversion_(music)
[7]https://en.wikipedia.org/wiki/Augmentation_(music)
[8]https://en.wikipedia.org/wiki/Diminution
[9]https://en.wikipedia.org/wiki/Crab_canon

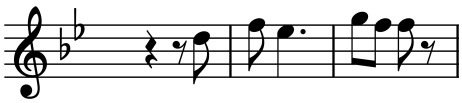

Fig. 9. Base phrase

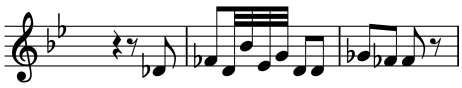

Fig. 10. Imitation with a variety of ornaments

These imitation techniques are particularly common in fugues, so the algorithm could be very helpful in analyzing, first of all, this musical form.

VIII. Conclusion

Music analysis has fascinated humanity for centuries. We have explored various methods, each providing a partial view of the rich musical landscape. However, these approaches often failed to capture the deep essence of music itself — the blend of melody, harmony, and rhythm that defies simple computational frameworks.

Amidst this exploration, we have introduced an innovative technique that marries technology and artistry to revolutionize music analysis. Our method employs neural networks, acting as conductors, to identify the core of musical compositions, the boundaries between phrases. Our goal is to achieve a balance between accuracy and authenticity, crafting a comprehensive understanding of the structure and narrative of a piece.

As we conclude this work, we stand on the threshold of a new era in music analysis. Neural networks emerge as instruments that bridge human creativity with musical comprehension. Our approach, the fusion of technology and imagination, paves the way for more nuanced and insightful analyses.

Our efforts mark a significant stride in unraveling music's complexities. This convergence of data and art pays homage to timeless melodies while embracing the boundless possibilities of the future.

IX. Future work

In our ongoing journey, we see a future full of new ideas and possibilities. Our model's neural capabilities allow us to explore a wide range of musical styles and genres. It can be adapted to different types of music, making it a versatile tool for analysis.

We are now delving into embeddings, where we are finding patterns in melodies and identifying complex phrases. This exploration helps us to understand music on a deeper level.

Looking ahead, we are focusing on transitions between sounds and mapping melodies. By studying these aspects, we aim to better understand the core of musical pieces, and even turn dividing phrases into an art form itself. This will improve the way we analyze and create music, fostering more creativity and insight.

As we continue, we are venturing into unexplored territory. Our goal is to better grasp musical understanding and creativity. With innovation as our guide, we are ready to shape the future of music analysis and composition, one note at a time.

REFERENCES

[1] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. Universal onset detection with bidirectional long-short term memory neural networks. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 589–594, 2010.

[2] Philippe Hamel and Douglas Eck. Learning features from music audio with deep belief networks. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, volume 10, pages 339–344. Utrecht, The Netherlands, 2010.

[3] Xinquan Zhou and Alexander Lerch. Chord detection using deep learning. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, volume 53, page 152, 2015.

[4] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 121–126, 2015.

[5] Siddharth Sigtia, Nicolas Boulanger-Lewandowski, and Simon Dixon. Audio chord recognition with a hybrid recurrent neural network. In *International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 127–133, 2015.

[6] Dawen Liang, Minshu Zhan, and Daniel PW Ellis. Content-aware collaborative music recommendation using pre-trained neural networks. In *International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 295–301, 2015.

[7] Thomas Grill and Jan Schluter. Music boundary detection using neural networks on spectrograms and self-similarity lag matrices. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1296–1300. IEEE, 2015.

[8] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Accurate tempo estimation based on recurrent neural networks and resonating comb filters. In *International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 625–631, 2015.

[9] István Koren. Music score analysis with process mining. In *ICPM 2022 Doctoral Consortium and Demo Track 2022*, pages 123–127, 2022.

[10] Bruce Benward. *Music in Theory and Practice Volume 1*. McGraw-Hill Higher Education, 2014.

[11] Reed James Perkins. *Musical Phrase Segmentation via Grammatical Induction*. PhD thesis, Brigham Young University, 2022.

[12] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary detection in music structure analysis using convolutional neural networks. In *International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 417–422, 2014.

[13] Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder, 2016.

[14] Brian McFee, Oriol Nieto, and Juan Pablo Bello. Hierarchical evaluation of segment boundary detection. In *International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 406–412. Citeseer, 2015.

[15] Thomas Pankhurst. *SchenkerGUIDE: a brief handbook and website for Schenkerian analysis*. Routledge, 2008.

[16] Oriol Nieto and Morwaread M Farbood. Identifying polyphonic patterns from audio recordings using music segmentation techniques. In *Proc. of the 15th International Society for Music Information Retrieval Conference*, pages 411–416, 2014.

[17] Filip Wróblewski et al. Agata Skrukwa, „Oskar Kolberg. 1814-1890". *Lud*, 98(1):388–391, 2014.

[18] Marcin Werla, Jacek Jackowski, Madgalena Chudy, Ewa Łukasik, Ewa Kuśmierek, and Ewa Dahlig-Turek. Developing music digital library based on polish traditional music archives and dlibra. In *5th International Conference on Digital Libraries for Musicology*, 2018.

[19] Ewa Dahlig. *EsAC. An Integrated System of Encoding, Analysing and Processing of One-part Melodies*. Universitatsverlag Rasch, Osnabruck, 2000.

[20] David Huron. Music information processing using the humdrum toolkit: Concepts, examples, and lessons. *Computer Music Journal*, 26(2):11–26, 2002.

[21] Perry Roland. The music encoding initiative (mei). In *Proceedings of the First International Conference on Musical Applications Using XML*, volume 1060, pages 55–59. Citeseer, 2002.

[22] Michael Scott Cuthbert and Christopher Ariza. music21: A toolkit for computer-aided musicology and symbolic music data. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010), August 9-13, 2010, Utrecht, Netherlands*, pages 637–642, 2010.

[23] David Foster. *Generative deep learning*. " O'Reilly Media, Inc.", 2022.