UDC 004.021:004.75

ADVANCING BPM DETECTION IN HIP-HOP AND R&B THROUGH AUDIO REPRESENTATIONS AND CONVOLUTIONAL NEURAL NETWORKS

K.V. Tushynskaya
Teacher assistant of the Faculty of Computer Systems and Networks BSUIR, master degree student
katetushkan@icloud.com

M.M. Zyranova
Last year student of the Faculty of Computer Systems and Networks BSUIR
z.y.r.y.a.n.o.v.a@mail.ru

A.E. Asadchy
Last year student of the Faculty of Computer Systems and Networks BSUIR
aleh.asadchy@gmail.com

K. Tushynskaya
Graduated from the Belarusian State University of Informatics and Radioelectronics. The areas of scientific interests are related to the investigation and analyzing of different music representations and their ability to give accurate information about the track.

M. Zyranova
Studying at the Belarusian State University of Informatics and Radioelectronics at the fourth course. The areas of scientific interests are related to the development and implementation of convolutional neural networks for music beats detection and analyzing.

A. Asadchy
Studying at the Belarusian State University of Informatics and Radioelectronics at the fourth course. The areas of scientific interests are related to the development and implementation of convolutional neural networks for music beats detection and analyzing.

Abstarct. This study introduces a method for improving BPM detection in hip-hop and R&B music by integrating audio representations with Convolutional Neural Networks (CNNs). Through analyzing tempograms, spectrograms, and onset features optimized for CNN processing, our approach demonstrates enhanced accuracy in BPM detection across these genres. Evaluated on the Million Song Dataset (MSD), our findings offer significant advancements for automated music analysis and applications in music recommendation and genre classification.

Keywords: Beats per Minute (BPM), Convolutional neural network (CNN), tempogram, beatgram, rhythmogram, tempo-invariant processing.

Introduction. The dynamic and complex nature of contemporary music, particularly within the hip-hop and R&B genres, presents unique challenges in the field of music information retrieval (MIR). These genres are renowned for their intricate rhythms, diverse tempos, and the prominent use of vocal elements, all of which play a pivotal role in defining their sonic identity. Accurately detecting Beats Per Minute (BPM) is crucial for analyzing, categorizing, and interacting with these musical forms, yet the variability and complexity inherent in hip-hop and R&B tracks often elude traditional BPM detection methods.

This paper introduces a focused exploration into different audio representations as a solution to improve BPM detection specifically within hip-hop and R&B music. By examining a variety of audio representations, including tempograms, rhythmograms, chromagrams, spectrograms, and onset and tempo features, we aim to identify those that are most effective in

capturing the unique rhythmic and melodic characteristics of these genres. The selected representations are further optimized for integration with Convolutional Neural Networks (CNNs), enhancing their ability to accurately detect BPM across a wide range of hip-hop and R&B compositions.

Our investigation not only addresses the technical challenges associated with BPM detection in these genres but also seeks to contribute to the broader understanding of their musical structure through advanced audio analysis techniques. By narrowing our focus to hip-hop and R&B, we aim to develop methodologies that are finely tuned to the nuances of these genres, thereby offering more precise and reliable tools for music analysis, genre classification, and the creation of engaging music.

Audio Representation Comparison Analysis. In our dedicated pursuit to refine and enhance the accuracy of Beats Per Minute (BPM) detection methodologies, particularly within the nuanced soundscapes of hip-hop and R&B music, we embarked on a comprehensive reevaluation of the spectrum of audio representations.

Our research endeavors have led us to place a significant emphasis on the examination and integration of rhythmograms and chromagrams, alongside the established utility of tempograms, spectrograms, and onset and tempo features. This strategic recalibration in our focus is driven by the recognition of the paramount importance of vocal and melodic elements in these genres, which often serve as the primary indicators of BPM.

Hip-hop and R&B, genres celebrated for their rich lyrical density and melodic complexity, demand a nuanced approach to BPM detection that goes beyond the conventional metrics. The rhythmic cadence of spoken word, the syncopation of beats, and the harmonic undercurrents present unique challenges and opportunities for audio analysis.

By including rhythmograms and chromagrams in our analytical arsenal, we aim to capture the essence of these elements more effectively. This initiative not only reflects our commitment to addressing the specific needs of hip-hop and R&B music analysis but also underscores our broader objective to advance the precision of BPM detection technologies.

Through this refined approach, we seek to contribute meaningful insights and methodologies that can adapt to the evolving landscapes of music genres, thereby enhancing the applicability and accuracy of BPM detection in music information retrieval systems.

Below is the comparison table 1 that summarizes our findings from the comparative analysis of the selected audio representations. This table highlights the pros and cons of each representation and indicates their selection for further investigation based on their potential to enhance BPM detection in hip-hop and R&B music.

The criteria for evaluation include the representation's ability to capture genre-specific rhythmic and melodic nuances, computational efficiency, and adaptability to convolutional neural network architectures for automated BPM detection.

Table 1. Comparative Analysis of Audio Representations

Representation	Function	Strengths	Weaknesses
Tempogram	Detecting and analyzing rhythmic patterns in music.	Emphasis on rhythmic periodicities	Sensitive to changes in timbre and instrument characteristics.
Rhythmogram	Detecting and analyzing beats and rhythm-related features in music.	Stable representations for detecting beats and rhythm-related features.	Can be computationally expensive.
Spectrogram	Analyzing the frequency content and temporal evolution of sounds	Unique information about its frequency content.	Temporal evolution tools like spectral filtering and morphing can be used for creative sound design
Chromagram	Identifying harmonic patterns and melodic characteristics	Similar to a spectrogram, but only shows the intensity of the twelve pitches that make up the equal-tempered scale.	Often used in music analysis
Onset and Tempo Features	Creating representations that are less sensitive to tempo variations.	Less sensitive to tempo variations when processed properly	May miss subtle rhythmic nuances.
Percussive and Harmonic Separation	Making the representation more robust to tempo changes.	Isolates rhythmic elements	Can be difficult to achieve perfect separation.
Pulse or Beat Tracking	Providing a stable foundation for analyzing rhythmic patterns.	Provides a stable foundation for analyzing rhythmic patterns	Can be inaccurate in complex or polyrhythmic music

This table presents the strengths and limitations of each audio representation, with all selected for further investigation to leverage their unique contributions towards enhancing BPM detection in hip-hop and R&B genres

Given the unique rhythmic and melodic characteristics of hip-hop and R&B, rhythmograms and chromagrams have been reintegrated into our analysis framework alongside tempograms, spectrograms. This selection aims to harness the strengths of each representation to accurately detect BPM, reflecting the complex interplay of beats, melodies, and vocals inherent in these genres.

Exploiting CNNs for Hierarchical Rhythm Analysis in Music. Convolutional Neural Networks (CNNs) have revolutionized the field of machine learning, offering robust frameworks for extracting and learning from hierarchical data structures. Originally popularized within the realm of image analysis, the principle behind CNNs – that of learning layered feature representations –translates effectively to the domain of audio analysis, particularly for the study of rhythm in music.

In rhythm analysis, CNNs exploit their hierarchical structure to dissect and understand the complex layers of musical composition. The initial layers of a CNN are adept at identifying fundamental rhythmic elements, such as individual beats or onset events, which can be thought of as the auditory equivalents of edges in image analysis. These basic components are the building blocks of rhythm, providing the tempo and the groove of the track.

As the information progresses through the network, intermediate layers amalgamate these basic components into more sophisticated rhythmic patterns. In the context of music, this might include the identification of rhythmic motifs, syncopation, or the interplay between different rhythmic instruments. These patterns are akin to the shapes and textures that CNNs learn to recognize in images, offering a deeper understanding of the music's rhythmic structure.

The efficacy of CNNs in music rhythm analysis is significantly enhanced by incorporating multi-dimensional audio representations. Spectrograms, for instance, offer a rich, time-frequency depiction of music, allowing CNNs to capture not only rhythm but also pitch and timbre information. Meanwhile, tempograms and rhythmograms provide targeted insights into temporal dynamics and rhythmic consistency, crucial for genres that thrive on rhythmic innovation.

This integration enables CNNs to not only recognize but also predict rhythmic patterns, facilitating advancements in automated music generation, remixing, and interactive music systems. It underscores the transformative potential of CNNs in music technology, bridging the gap between computational analysis and creative musical expression.

While the primary focus here is on rhythm analysis, the capability of CNNs extends to capturing harmonic and melodic features when trained on appropriate representations like chromagrams. This holistic approach to music analysis underscores the versatility of CNNs, enabling a comprehensive understanding of music beyond the rhythmic dimension.

CNN architecture for rhythm analysis. The architecture of our Convolutional Neural Network (CNN) for rhythm analysis is a multi-layered structure carefully designed to uncover intricate rhythmic patterns in music. The journey through this network begins with the input layer, represented as 2D matrices derived from tempograms, rhythmograms, or beatgrams. These representations encapsulate the temporal dynamics of the music and serve as the foundation for further analysis.

The initial convolutional layer embarks on the task of capturing local rhythmic features. It employs a sequence of convolutional layers, each equipped with filters of varying sizes to discern different rhythmic patterns. At this stage, batch normalization and activation functions (such as Rectified Linear Units – ReLU) are applied after each convolution, ensuring that the learned features are fine-tuned to highlight relevant rhythmic elements. Subsequently, max-pooling layers come into play, strategically downsampling the learned features while preserving essential rhythmic information.

As the data flows deeper into the network, additional convolutional layers are introduced to extract abstract and higher-level rhythmic patterns. The number of filters in these layers progressively increases to accommodate the growing complexity of the learned features. This hierarchical approach allows the network to discern intricate rhythmic nuances that may be deeply embedded within the music.

To summarize the hierarchical features learned across the entire representation, global average pooling is applied. This step effectively condenses the wealth of rhythmic information into a manageable form, facilitating subsequent analysis. The network then connects the global average pooling layer to a sequence of fully connected layers. To prevent overfitting and ensure robustness, the number of neurons in each layer is gradually reduced.

The final output layer is tailored to the specific task at hand, which may include BPM range prediction or genre classification. For classification tasks, a softmax layer is employed to predict a certain class label, offering insights into the rhythmic characteristics or genre affiliation of the analyzed music.

This CNN architecture is meticulously crafted to decode the intricate rhythmic fabric of music, gradually progressing from local features to more abstract patterns, and finally providing valuable insights through classification. It exemplifies the power of deep learning in unraveling the complexities of rhythm analysis.

Task results. For our research, we selected the Million Songs Dataset (MSD) as the primary source of information. The MSD is a comprehensive and freely-available collection of audio features and metadata for a vast repository of contemporary popular music tracks. At its core, the dataset encompasses feature analysis and metadata for an astounding one million songs, making it an invaluable resource for music-related research.

To extract the audio data from the MSD, we utilized the Python library librosa. Librosa is a versatile library tailored for audio and music processing, enabling us to access and manipulate the audio data effectively. Through librosa, we were able to represent each track as a 2D array of numerical values, capturing essential audio features for our analysis.

As part of our investigation into improving BPM detection in hip-hop and R&B music, we selected four distinct representation types for analysis. These representations serve as the foundational elements for our neural network's training dataset. The examples of representations are displayed at figure 1.

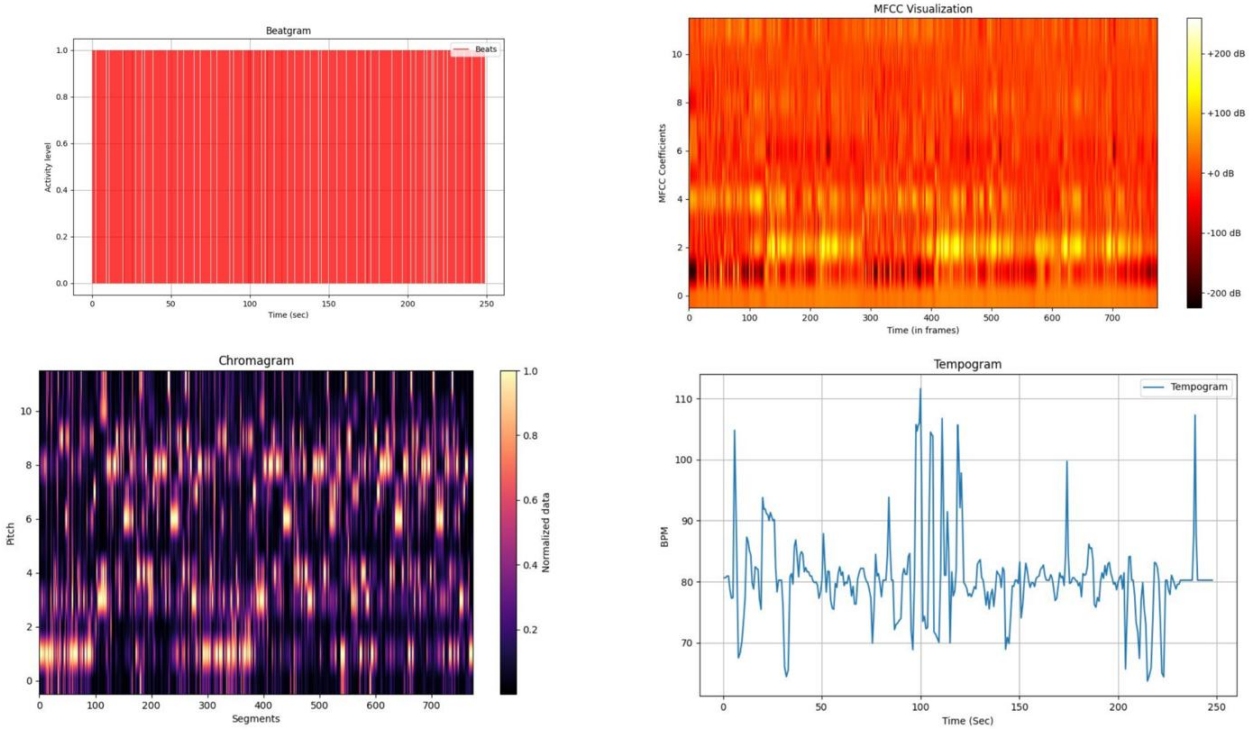

Figure 1. Examples of audio representation in 2d arrays: beatgram, chromagram, spectrogram, tempogram

In our endeavor, we assembled a dataset consisting of 10,000 plots for each of the four selected representation types: tempograms, beatgrams, chromagrams, and spectrograms. This dataset was thoughtfully divided into both training and testing subsets to facilitate the training and evaluation of our Convolutional Neural Network (CNN).

Following the dataset preparation, our CNN underwent extensive training over 1,000 epochs. During this training process, we closely monitored the learning curves to discern the representation type that yielded the most accurate results. It's important to note that the primary goal of this investigation was to establish correlations between different audio representations.

The outcome of this rigorous analysis is presented below at figure 2, shedding light on the performance of various representation types and their impact on the accuracy of our neural network. The plot unequivocally demonstrates that the tempogram representation outperforms others in terms of accuracy.

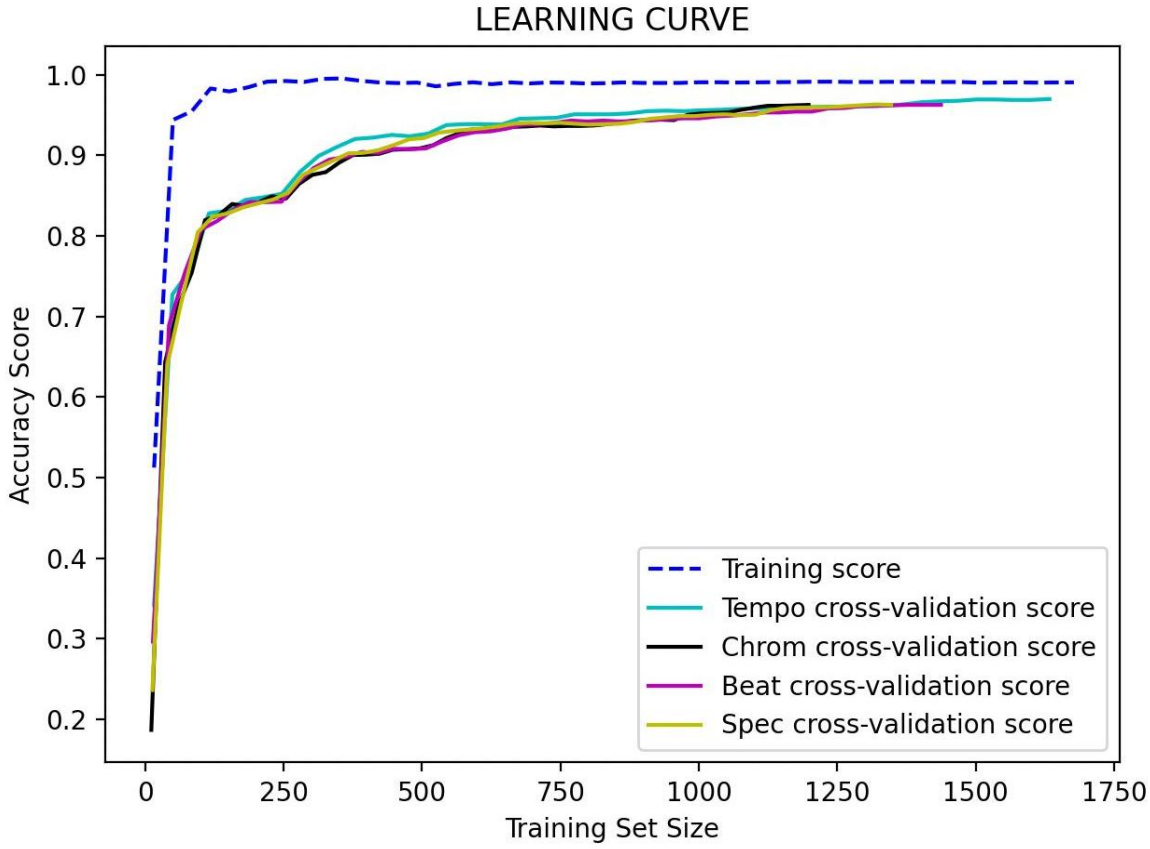

Figure 2. Validation results

Conclusion. Our research has ventured into the realm of hip-hop and R&B music, driven by the quest to refine Beats Per Minute (BPM) detection. Through meticulous analysis, we unveiled that the Tempogram representation emerges as the most accurate, particularly in handling tempo variations–a common characteristic of these genres.

The integration of Convolutional Neural Networks (CNNs) was crucial, resulting in superior BPM detection accuracy and robustness across diverse musical styles. Beyond BPM, our approach unearthed genre-specific insights, promising more nuanced music recommendations.

Our work demonstrates the transformative potential of deep learning in music analysis, inspiring a deeper exploration of rhythmic intricacies. In conclusion, this research not only advances BPM detection but also paves the way for a more profound understanding of the rhythmic diversity that defines hip-hop and R&B music.

Reference list

[1] E. Parada-Cabaleiro, M. Schmitt, A. Batliner, B. Schuller, and M. Schedl. Automatic recognition of texture in renaissance music. International Society for Music Information Retrieval Conference (ISMIR), 2021.

[2] D. Guigue and C. de Paiva Santana. The structural function of musical texture: Towards a computer-assisted analysis of orchestration. Journe es d'Informatique Musicale (JIM), 2018.

[3] Z.Wang, D.Wang, Y.Zhang, and G.Xia. Learning interpretable representation for controllable polyphonic music generation. International Society for Music Information Retrieval Conference (ISMIR), 2020.

Authors' contribution

Katsiaryna Tushynskaya – led the research on evaluation the quality of BPM detection and analyzing.

Maryia Zyranava – investigated and compared different audio representations, investigated methods of audio features' extracture, developed CNN and analyzed its output.

Aleh Asadchy – prepared a dataset of representations, developed and trained CNN as well as analyzed its output.

УЛУЧШЕНИЕ РАСПОЗНАВАНИЯ BPM В ХИП-ХОПЕ И R&B С ПОМОЩЬЮ АУДИОПРЕДСТАВЛЕНИЙ И КОНВОЛЮЦИОННЫХ НЕЙРОННЫХ СЕТЕЙ

Е.В. Тушинская
Ассистент кафедры информатики факультета компьютерных систем и сетей БГУИР, студент магистратуры

М.М. Зырянова
Студентка выпускного курса факультета компьютерных систем и сетей БГУИР

О.Э. Осадчий
Студент выпускного курса факультета компьютерных систем и сетей БГУИР

Аннотация. В данном исследовании представлен метод улучшения обнаружения темпа в хип-хоп и R&B музыке путем интеграции аудио представлений со сверточными нейронными сетями (CNN). Анализируя темпограммы, спектрограммы, хромограммы и битограммы оптимизированные для обработки CNN, разработанный нами подход демонстрирует повышенную точность обнаружения BPM в этих жанрах.

Ключевые слова: Количество ударов в минуту (BPM), сверточная нейронная сеть (CNN), темпограмма, битограмма, ритмограмма, обработка аудио информации.