INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2025, VOL. 71, NO. 3, PP. 1-6
Manuscript received February 27, 2025; revised July 2025. doi: 10.24425/ijet.2025.153608

Classification of Music Structural Functions using Deep Learning

Dina Al-Daloo, and Piotr Bilski

***Abstract*—Music Structure Analysis (MSA) is crucial for understanding and leveraging the arrangement of musical compositions in various applications, such as music information retrieval, multimedia description, and recommendation systems. The following paper presents a novel approach to MSA that aims to predict labels for structural music segments (such as verse or chorus), thereby it would enhance any MSA-based applications. This is the supervised approach in contrast to clustering-based methods. For the task, selected pre-trained Convolutional Neural Networks (CNNs), such as VGG, ResNet or MobileNet were applied to classify the segments of musical structures (verse, chorus, etc.). Results demonstrated that ResNet50 and DenseNet121 achieved the highest performance in terms of classification accuracy, with ResNet50 reaching 87% and DenseNet121 reaching 85.16%. This highlights the potential of deep learning models for accurate and efficient music structure segment labeling, opening possibilities for advanced applications in both offline and real-time music analysis scenarios.**

***Keywords*—deep learning; MSA; music segments**

I. Introduction

MUSIC Structure Analysis (MSA) is a fundamental task in Music Information Retrieval (MIR), essential for understanding the organization and narrative flow of musical compositions. It plays a pivotal role in various applications, such as summarization, and recommendation systems. The proper classification is specifically important in the latter, as it allows for proposing the list of new songs attractive for the particular listener. To do that, either collaborative filtering [1],[2], or the acoustic analysis [3] may be employed. The first approach requires access to the listening history of the users. Its outcome depends on the ability to extract lists of songs used by different users and find the common ground. This method is versatile, as it can be used for any type of commodities (books, movies, games, etc.) [4] and requires finding the user preferences from their history. Its disadvantage is that the reasons of playing (even repeatedly) particular song are not known and are not automatically linked with preferences. The already applied solutions include Spotify or Tidal recommendation modules [5].

On the other hand, detailed analysis of the listener's taste requires deep insight into his/her psychological profile. The acoustic analysis is an attractive alternative, providing information about the songs' structures and their melodical aspects. This case covers multiple applications, such as genre identification, instruments detection or sentiment analysis. The difficulty in applying this approach is the need to isolate the significant features of the songs, which describe the particular (usually well defined) fragments. However, this way it is possible to bind the user's preferences with the musical structure of the song. This makes the acoustics-based approaches potentially attractive for many applications.

The music structure analysis is based on the concept that a song is decomposed into non-overlapping (possibly redundant) segments that can be labeled. This is true for most of the contemporary musical genres, including pop, rock, disco, or metal. The fragmentation process is comprised of two parts: the first one involves finding boundaries of segments (for instance, based on tempo). The second one involves structural grouping (labeling) the segments into their corresponding categories [6]. This operation may incorporate symbols (like 'A', 'B', 'C'), or functional labels such as 'verse', 'chorus', 'intro', etc. [7]. Focusing on the structural functional labeling, i.e. identifying song structures like verses, choruses, and bridges, allows for the more precise song analysis, for instance, clustering the similar songs as a whole, but based on the similarity between the corresponding parts. The manual decomposition is a time-consuming process requiring the expertise of musicians to annotate songs. Development of automated systems speeds up the whole process, making it applicable for music creation and production, music recommendation, automatic music generation, audio visualization, and enhance the user experience in real-time scenarios (live concerts, video games, broadcasts) [8].

The following paper investigates the utilization of deep learning networks, specifically Convolutional Neural Networks (CNN), for labeling music structural segment functions. By evaluating a range of pre-trained CNN models, including VGG16, VGG19, ResNet50, ResNet50V2, ResNet101, DenseNet121, MobileNet, and MobileNetV2, it was possible to compare their efficiency during the musical segments' identification. The performance measure includes classification accuracy calculated based on the ground truth. While many prior studies [9],[10] focus on segmentation or boundary detection, this paper considers explicit functional labeling, assigning roles such as "verse" or "chorus" to each segment in a supervised manner. It leverages transfer learning with pre-trained CNNs, which enables the model to generalize and reduces the need for large, already annotated datasets. Multiple CNN architectures are explored to find optimal models regarding accuracy and efficiency measures (including precision, recall and F1-score). Obtained results show CNNs are applicable for understanding the organization and narrative flow of musical compositions.

The content of the paper is as follows. In Section II the state-of-the-art in the structural approaches to the musical content is

Dina Al-Daloo and Piotr Bilski are with Warsaw University of Technology, (e-mail: dina.al-daloo.dokt@pw.edu.pl, Piotr.bilski@pw.edu.pl).

presented. Section III presents the architecture of the MSA module, exploiting the deep neural networks. In Section IV experimental results on the SALAMI (Structural Annotations for Large Amounts of Music Information) dataset are presented. The last section contains conclusions and future prospects.

II. Background and related work

Music Structure Analysis has grown significantly due to the developments in signal processing, machine learning, specifically deep learning. Early approaches relied on audio signal processing techniques, self-similarity matrices, and clustering techniques to detect boundaries and segment musical structures [8],[11],[12].

Conventional methodologies for music segmentation mostly relied on audio signal processing techniques such as self-similarity analysis, spectral graph theory, and time series similarity measures to analyze and segment audio.

In [13] one of the earliest techniques for audio segmentation using audio novelty detection based on self-similarity analysis is introduced. It examines changes in audio features over time to identify boundaries in the music and speech. In [14] an approach to find recurrent patterns and analyze musical structure in acoustic music data was proposed. It uses the constant Q transform (CQT) to extract musical information and propose a new similarity measure. Significant repeating patterns were identified using an adaptive threshold on a self-similarity matrix, and then a heuristic-based approach further defines the segmented boundaries. A three-level hierarchy to analyze the structure of music was presented in [15] Pitch class profile characteristics are extracted at the note level. At the measure level, a similarity matrix is created. Dynamic time warping improves audio segment similarity calculations. In [16] a spectral graph theory was used to analyze repeated patterns in musical recordings, through Laplacian's eigenvectors and identify structural segments. Spectral clustering partitions the repetition graph, while k-means clustering further refines the segment boundaries. In [17] an unsupervised method was utilized, combining the structure features that capture local and global properties with time series similarity measures to detect boundaries and annotate segments in music. The proposed method in [18] uses path-enhanced self-similarity matrices (SSMs), applies non-negative matrix factor 2-D deconvolution (NMF2D) to convert them into block-enhanced SSMs, and fuses multiple SSMs to improve boundary detection and segmentation accuracy.

Other approaches leverage statistical learning and AI-driven algorithms for structural audio segmentation and boundary detection, such as Hidden Markov Models (HMMs), clustering, and probabilistic systems.

In [19] proposed a method to segment musical signals by leveraging HMM where each state corresponds to a distinct audio texture representing steady statistical properties of the music (e.g., instrumentation, polyphonic timbre). The use of dynamic features from audio signal was proposed in [20]. This approach details a multi-pass system involving segmentation of the music signal, grouping potential states using K-means clustering for unsupervised segments grouping, and applying a hidden Markov model. In [21] a musical structure analysis system that employs a cost function to detect repeated parts of music was introduced. It generates candidate descriptions from acoustic input signals, and a method determines the best descriptions based on cost. A method segmenting musical audio into structural sections by labeling audio frames with HMM is shown in [22], the frames are then clustered into segment types based on state distributions. Temporal continuity is ensured through constraints modeled by a Hidden Markov Random Field. In [23] two methods to structure segmentation were combined: timbral novelty measure segmentation and harmonic analysis. The combination of these methods improves the simultaneous estimation of keys, chords, and structure boundaries.

A new era has begun with the introduction of deep learning, specifically CNNs. They are known for their ability to automatically extract features from raw data and have been effectively used in a variety of music retrieval tasks [24]-[26].

In [24] CNNs trained directly on mel-scaled magnitude spectrograms were applied to automatically identify boundaries in audio signals. These networks were employed in [25] for the detection of musical boundaries. They were trained using Mel-scaled log-magnitude spectrograms and similarity lag matrices (SSLMs). Automatic songs segmentation based on their musical structure using CNNs was presented in [26]. It employs a small-scale architecture inspired by VGGNet, trained on Mel-scale spectrograms to predict segment boundary scores via regression. Post-processing (e.g., peak-picking) is then applied to identify discrete boundary times.

While these studies aim to segment music and detect boundaries, they do not emphasize the functional labeling of segments (e.g., intro, verse, chorus). Further works specifically aim to detect the "chorus," which is important in applications like thumbnailing [27]-[32]. There has been limited work on functional labeling of all structural segments [33]-[36], which impacts the effectiveness of MSA for music recommendation, retrieval, and annotation systems by reducing their ability to identify and suggest relevant song parts based on function.

In [27] a method for identifying repeated sections in music, particularly chorus, was proposed. It employs chroma-based representations to capture harmonic relationships, which are especially effective for the structure of popular music. In [28] a method called RefraiD introduced, which utilized chroma to identify chorus sections in popular music recordings by analyzing repeated patterns and handling modulations. In [32] DeepChorus, a comprehensive chorus detection model was introduced that minimizes engineering effort and prerequisite knowledge. It uses a multi-scale network to capture both global and local structural information and a self-attention convolution network to model correlations between segments, enabling end-to-end learning from Mel-spectrograms without complex handcrafted features or post-processing.

A model for segmentation and labeling of music structures combining Long Short-Term Memory (LSTM) networks and Hidden Semi-Markov Models (HSMM) was presented in [33]. This hybrid approach leverages the sequence modeling capability of LSTMs with the segmentation advantages of HSMMs, enabling both the segmentation and functional labeling of music structures. The approach explicitly addresses homogeneity, repetitiveness, and regularity in music sections, aligning with the described objectives. In [34] a multi-task deep learning framework introduced using a Transformer-based model (SpecTNT) to directly estimate semantic structural labels (e.g., "verseness," "chorusness") from audio via activation

curves. It employs a 7-class taxonomy, consolidates annotations across datasets, and integrates Connectionist Temporal Localization (CTL) loss, achieving precise structural analysis. In [35] the effectiveness of the Convolutive Block-Matching (CBM) algorithm in achieving unsupervised music segmentation was verified. In [36] an all-in-one model for hierarchical music structure analysis, integrating beat tracking, downbeat tracking, segmentation, and functional structure labeling into a unified framework. Their model leverages demixed audio (source-separated spectrograms) and employs a neighborhood attention mechanism, including 1D Dilated Neighborhood Attention (DiNA) and 2D Neighborhood Attention (NA) to handle long-term and local dependencies in the music data.

Though MSA has made significant progress recently, it often lacks the ability to assign functional roles to structural segments like "verse" and "chorus." Traditional approaches primarily focus on boundary detection or segment clustering but often lack the ability to assign meaningful labels. Deep learning techniques, while promising in segmentation tasks, often focus on identifying transitions or boundaries rather than the human-readable musical passages. This gap hinders the full potential of MSA in advanced applications, such as music recommendation as it enables deeper understanding of songs.

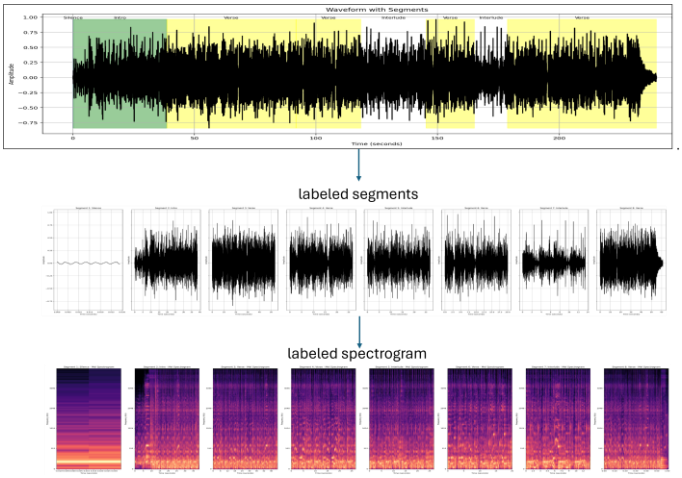
The proposed method addresses these challenges by using CNNs trained in a supervised manner to identify already assigned labels for musical segments. Transfer learning was used for the task to reduce the dependency on extensive annotated datasets. The pre-trained network additionally trained with the application-specific data usually provides reasonable performance with much shorter training convergence time.

III. METHODOLOGY

A. Dataset Preparation

The proposed method employs CNNs to classify musical segments. The network requires a labelled image dataset for training, where each label represents the structural function of the corresponding music segment. In the presented research, the SALAMI (Structural Annotations for Large Amounts of Music Information) dataset was utilized. It is a well-known dataset in MIR. It includes structural annotations of full-length tracks from a variety of genres, such as: popular music, jazz, classical music and world music [37]. Overall, 468 full-length audio recordings were obtained from Internet and downloaded following [38]. However, 29 of them were excluded due to incomplete annotations. Each file was then split into labelled segments using annotations included in the set metadata. The annotation specifies segments such as "silence," "intro," "verse," "chorus," and so on, with precise time intervals (e.g., 0.0 - silence, 28.746 - intro, etc.). Afterwards, audio data are converted into image representations known as mel-spectrograms, that represent the frequency distribution of the signal as it changes over time [39]. Mel-spectrograms are utilized in music genre classification, voice recognition, and sound event detection. Here, they represent the input data for the deep learning models.

To simplify the analysis and improve clarity, the several labelled classes that were initially present in the dataset were combined into a smaller collection of seven categories, as described in [34]. They are “verse”, “silence”, “outro”, “intro”, “inst” (i.e., “instrumental”), “chorus”, and “bridge”. Figure 1 illustrates the procedure for data processing and preparation.



The obtained images may then be used to train the CNN.

Fig. 1. Dataset preparation

B. Applying Convolutional Neural Networks for Supervised Classification

The CNNs in the transfer learning mode were employed for supervised classification. This addressed the limitation of the requirement for large amount of training data [40]. The exploited set consisted of 6,408 files categorized into seven distinct types. The distribution of images in each class is as follows: 'chorus' has 1234 images, 'bridge' has 280 images, 'inst' has 1685 images, 'verse' has 1342 images, 'outro' has 490 images, 'silence' has 941 images, and 'intro' has 436 images. Following preparation, the set was partitioned into training and testing subsets using an 80%-20% split, resulting in 161 training batches and 40 test batches each with 32 samples each. This distribution indicates a class imbalance, which is common in many real-world datasets. To address this class weighting and data augmentation techniques were experimented. However, these approaches did not significantly impact the results, so they were neglected.

In the experiment, eight transfer learning architectures were utilized, including VGG16, VGG19, ResNet50, ResNet50V2, ResNet101, DenseNet121, MobileNet, and MobileNetV2. These architectures differ in their design. For instance, VGG models use simple sequential architectures with small filter sizes, ResNet models incorporate residual connections, DenseNet employs densely connected layers and MobileNet are lightweight models. By incorporating these commonly used pre-trained CNN architectures, the goal was to determine which model performed best for the task of classifying music structural functions. While this was not intended as a formal comparative study, provided insights into the effectiveness of various CNN models on the given dataset.

Figure 2 illustrates the proposed model, employing the specific variant of CNN. It was trained using image inputs that were resized to 224 x 224 pixels. A preprocessing layer preceding the applied CNN model was added to ensure that images were scaled appropriately according to the requirements of the chosen transfer learning architecture. Each base model was initialized with pre-trained weights from ImageNet (which represents a large-scale image dataset) [41] excluding the top

classification layers. To leverage the pre-trained features, the initial layers of the base model were frozen. Specifically, all layers before layer 100 were set to be non-trainable (frozen). However, for models with fewer than 100 layers (e.g., VGG16), only the available layers were frozen. Additional layers were inserted to enhance performance and prevent overfitting. They consist of GlobalAveragePooling2D, Dropout with rates of 0.4 and 0.2, Dense layer with 128 units and ReLU activation function, and Output Dense layer with seven outputs (equal to the number of identified fragments) and Softmax activation function. Each model was compiled with an Adam optimizer of 0.0001 learning rate and trained for 100 epochs (though training of some networks, i.e. VGG16 and VGG19, terminated earlier based on early stopping criteria) with the hold-out validation (i.e. one-time division of the available data into the training and testing set). Finally, the proposed models were evaluated using the test dataset to determine their effectiveness.

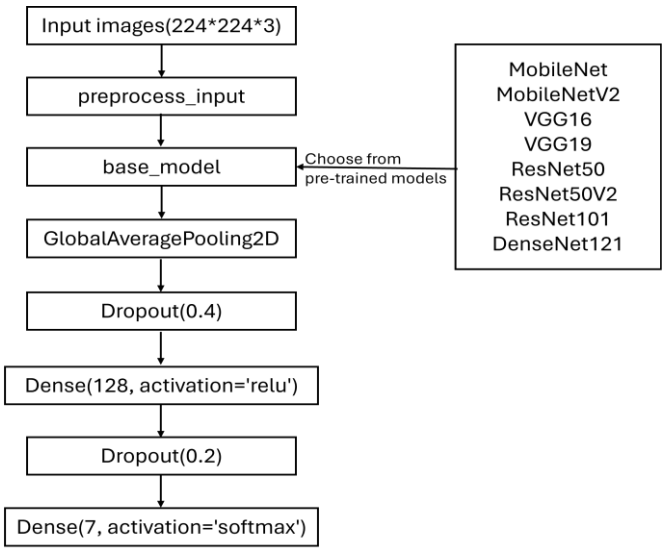


Fig. 2. The proposed model

IV. Experimental results

Experiments were conducted on a computer equipped with an Intel Core i7 processor, 32 GB of RAM, and a 1 TB hard disk drive (HDD). In terms of software, the Spyder integrated development environment (IDE) for Python language was used. To check the ability of the traditional computer system for the transfer learning no GPU was used for the task, all off-line and on-line computations have been performed on the CPU.

Table I shows the averaged computational time for each model per epoch. Notably, the MobileNet, MobileNetV2, and ResNet101 models exhibited the lowest average computational time, not exceeding 105 seconds. This is because their networks are relatively simpler, therefore modifying their weights is faster. In contrast, the VGG19 and VGG16 models had the longest computation times, with values of 821 seconds and 677 seconds, respectively. All other networks have training durations significantly longer (between three to five times) than MobileNet, which may limit their applications when the learning phase must be performed from scratch, without any prior knowledge transfer.

Regarding the accuracy, ResNet50 is the best (despite its long training time), but should be preferred when the efficiency is the top priority. with DenseNet121 scoring the second place, highlighting the effectiveness of deeper architectures with advanced connectivity patterns. Conversely, VGG19 and VGG16 produced the lowest testing accuracies, likely due to their simpler sequential architecture which lacks advanced connectivity mechanisms, which can limit their performance on complex tasks. Other models slightly lag behind the best performing one. These results suggest that while deeper models generally perform better, lightweight architectures like MobileNetV2 can still achieve competitive results.

TABLE I
MODELS' PERFORMANCE METRICS AND TRAINING TIME

Model	Testing accuracy	Training time (per epoch) [s]
MobileNet	0.7875	**102**
MobileNetV2	0.7922	103
DenseNet121	0.8516	450
ResNet101	0.7945	104
VGG19	0.4930	821
ResNet50	**0.87**	503
VGG16	0.4914	677
ResNet50V2	0.8375	389

In addition to testing accuracy, the precision (Table II), recall (Table III), F1-score (Table IV), and a confusion matrix [42] of the models were used to evaluate their performance. Identification of the particular parts of songs differs, as usually "silence" and "intro" are easiest to detect. All other elements are identified with varying performance, depending on the applied architecture.

Figure 3 containing confusion matrices confirms the superior performance of ResNet50 and DenseNet121 for each song fragment.

TABLE II
PRECISION OF THE MODELS ACROSS THE SEVEN CATEGORIES

Model	Bridge	Chorus	Inst	Intro	Outro	Silence	Verse
MobileNet	0.51	0.78	0.79	0.71	0.72	0.82	0.90
MobileNetV2	0.88	0.68	0.88	0.57	0.76	0.81	0.90
DenseNet121	0.88	0.85	0.84	0.85	0.79	0.89	0.87
ResNet101	0.67	0.75	0.78	0.80	0.78	0.87	0.82
VGG19	0	0.39	0.47	0.62	0.50	0.83	0.46
ResNet50	0.94	0.84	0.86	0.81	0.82	0.88	0.92
VGG16	0	0.40	0.41	0.25	0.49	0.84	0.49
ResNet50V2	0.77	0.76	0.90	0.79	0.77	0.87	0.89

The obtained results prove the efficiency of the more complex architectures, though their training is related with the longer training times. When computational resources are not limited, they are the most preferable architectures. When time matters, simple networks (MobileNet and ResNet). The additional problem to solve is the detection of the change in the song fragment, that should precede the identification operation, presented in this research. This can be done through the tempo analysis of the original song.

TABLE III
RECALL OF THE MODELS ACROSS THE SEVEN CATEGORIES

Model	Bridge	Chorus	Inst	Intro	Outro	Silence	Verse
MobileNet	0.65	0.76	0.80	0.70	0.78	0.97	0.72
MobileNetV2	0.55	0.86	0.75	0.71	0.75	0.95	0.79
DenseNet121	0.65	0.81	0.87	0.82	0.77	0.95	0.87
ResNet101	0.62	0.81	0.75	0.55	0.82	0.90	0.85
VGG19	0	0.35	0.66	0.05	0.37	0.81	0.53
ResNet50	0.75	0.87	0.88	0.72	0.90	0.94	0.87
VGG16	0	0.17	0.74	0.02	0.46	0.83	0.58
ResNet50V2	0.61	0.88	0.86	0.73	0.86	0.93	0.80

Fig.3. Confusion matrix (a) DenseNet121 (b) MobileNet (c) MobileNetV2 (d) ResNet50 (e) ResNet50V2 (f) ResNet101 (g) VGG16 (h) VGG19

CONCLUSION

The presented research demonstrated the effectiveness of deep learning models (i.e. CNN) in classifying music structural segments. The effectiveness of models such as ResNet50 and DenseNet121 highlights their potential to improve music analysis tasks. Exploration of transfer learning techniques (even using CPU only) proved the adaptability of pre-trained models, offering promising avenues for future research and applications in the music recommendation systems.

Future research will cover the combination of the proposed models with the procedure of detecting the song fragments prior to their identification. The second step is the usage of the fragment detection to calculate similarities between songs seen not as a whole, but as sets of segments which could be compared. This will lead to create the complete music recommendation system.

TABLE IV
F1-SCORE OF THE MODELS ACROSS THE SEVEN CATEGORIES

Model	Bridge	Chorus	Inst	Intro	Outro	Silence	Verse
MobileNet	0.57	0.77	0.80	0.71	0.75	0.89	0.80
MobileNetV2	0.67	0.76	0.81	0.63	0.75	0.87	0.84
DenseNet121	0.75	0.83	0.85	0.83	0.78	0.92	0.87
ResNet101	0.65	0.78	0.77	0.65	0.80	0.89	0.83
VGG19	0	0.37	0.55	0.10	0.43	0.82	0.49
ResNet50	0.84	0.85	0.87	0.76	0.86	0.91	0.89
VGG16	0	0.24	0.53	0.04	0.47	0.84	0.53
ResNet50V2	0.68	0.81	0.88	0.76	0.81	0.90	0.84

REFERENCES

[1] Y. Chen, "A music recommendation system based on collaborative filtering and SVD," in 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dalian, China, 2022, pp. 1510-1513. https://doi.org/10.1109/TOCS56154.2022.10016210

[2] D. Sánchez-Moreno, A. B. Gil González, M. D. Muñoz Vicente, V. F. López Batista, and M. N. Moreno García, "A collaborative filtering method for music recommendation using playing coefficients for artists and users," Expert Systems with Applications, vol. 66, pp. 234-244, 2016. https://doi.org/10.1016/j.eswa.2016.09.019

[3] R.B. Dannenberg and M. Goto, "Music structure analysis from acoustic signals," in Handbook of Signal Processing in Acoustics, Springer, 2008, pp. 305-331. https://doi.org/10.1007/978-0-387-30441-0_21

[4] R. Zhang, Q. Liu, C. Chun-Gui, J. Wei, and Huiyi-Ma, "Collaborative filtering for recommender systems," in 2014 Second International Conference on Advanced Cloud and Big Data, Huangshan, China, 2014, pp. 301-308. https://doi.org/10.1109/CBD.2014.47

[5] L. Colley et al., "Elucidation of the relationship between a song's Spotify descriptive metrics and its popularity on various platforms," in 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), Los Alamitos, CA, USA, 2022, pp. 241-249. https://doi.org/10.1109/COMPSAC54236.2022.00042

[6] O. Nieto and J. P. Bello, "Systematic Exploration Of Computational Music Structure Research," in Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR). New York City, NY, USA, 2016. ISMIR2016-NietoBello.pdf

[7] Salami Annotator Guide, [Online]. https://github.com/DDMAL/salami-data-public/blob/master/SALAMI%20Annotator%20Guide.pdf

[8] O. Nieto et al., "Audio-based music structure analysis: Current trends, open challenges, and applications," Transactions of the International Society for Music Information Retrieval, vol. 3, no. 1, pp. 246–263, 2020. https://doi.org/10.5334/tismir.54

[9] K. Jensen, "Multiple scale music segmentation using rhythm, timbre, and harmony," EURASIP Journal on Advances in Signal Processing, vol. 2007. https://doi.org/10.1155/2007/73205

[10] M. C. McCallum, "Unsupervised learning of deep features for music segmentation," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 346-350. https://doi.org/10.1109/ICASSP.2019.8683407

[11] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in Proc of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010. pp. 625-636. https://ismir2010.ismir.net/proceedings/ismir2010-107.pdf

[12] X. Li, R. Liu, and M. Li, "A review on objective music structure analysis," in 2009 International Conference on Information and Multimedia Technology, Jeju, Korea (South), 2009, pp. 226-229. https://doi.org/10.1109/ICIMT.2009.20

[13] J. Foote, "Automatic audio segmentation using a measure of audio novelty," 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of

Multimedia (Cat. No.00TH8532), New York, NY, USA, 2000, pp. 452-455 vol.1. https://doi.org/10.1109/ICME.2000.869637

[14] L. Lu, M. Wang, and H. Zhang, "Repeating pattern discovery and structure analysis from acoustic music data," In Proc. of the 6th ACM SIGMM international workshop on Multimedia information retrieval (MIR '04). Association for Computing Machinery, New York, NY, USA, pp. 275–282. https://doi.org/10.1145/1026711.1026756

[15] Y. Shiu, H. Jeong, and C. J. Kuo, "Musical structure analysis using similarity matrix and dynamic programming," in Proc. SPIE 6015, Multimedia Systems and Applications VIII, 601516, 24 October 2005. https://doi.org/10.1117/12.633792

[16] B. McFee and D. P.W. Ellis, "Analyzing song structure with spectral clustering," in Proc. of 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014, pp. 405-410.

[17] J. Serra, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," in IEEE Transactions on Multimedia, vol. 16, no. 5, pp. 1229-1240, Aug. 2014. https://doi.org/10.1109/TMM.2014.2310701

[18] T. Cheng, J. B. Smith, and M. Goto, "Music structure boundary detection and labelling by a deconvolution of path-enhanced self-similarity matrix," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 106-110. https://doi.org/10.1109/ICASSP.2018.8461319

[19] M. Sandler and J. J. Aucouturier, "Segmentation of musical signals using hidden Markov models," in Audio Engineering Society. 110, 2001.

[20] G. Peeters, A. La Burthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," in Proceedings International Conference on Music Information Retrieval, 2002.

[21] J. Paulus and A. Klapuri, "Music structure analysis by finding repeated parts," in Proc. of the 1st ACM workshop on Audio and music computing multimedia (AMCMM '06). Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/1178723.1178733

[22] M. Levy and M. B. Sandler "Structural segmentation of musical audio by constrained clustering," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 2, pp. 318-326, Feb. 2008. https://doi.org/10.1109/TASL.2007.910781

[23] J. Pauwels, F. Kaiser, and G. Peeters, "Combining harmony-based and novelty-based approaches for structural segmentation," in International Society for Music Information Retrieval, 2013.

[24] K. Ullrich, J. Schlüter, and T. Grill, "Boundary detection in music structure analysis using convolutional neural networks," in Proc.15th International Society for Music Information Retrieval Conference, 2014.

[25] T. Grill and J. Schlüter, "Music boundary detection using neural networks on spectrograms and self-similarity lag matrices," in 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 2015, pp. 1296-1300. https://doi.org/10.1109/EUSIPCO.2015.7362593

[26] T. O'Brien, "Musical structure segmentation with convolutional neural networks," in Proc. of the 17th International Society for Music Information Retrieval Conference, 2016.

[27] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in Proc. 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575), New Platz, NY, USA, 2001, pp. 15-18. https://doi.org/10.1109/ASPAA.2001.969531

[28] M. Goto, "A chorus-section detecting method for musical audio signals," in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proc. (ICASSP '03)., Hong Kong, China, 2003, pp. V-437. https://doi.org/10.1109/ICASSP.2003.1200000

[29] A. Eronen, "Chorus detection with combined use of MFCC and chroma features and image processing filters," in Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), Bordeaux, France, September 10-15, 2007.

[30] S. Gao and H. Li, "Popular song summarization using chorus section detection from audio signal," in 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP), Xiamen, China, 2015, pp. 1-6. https://ieeexplore.ieee.org/document/7340798

[31] J. C. Wang, J. B. Smith, J. Chen, X. Song, and Y. Wang, "Supervised chorus detection for popular music using convolutional neural network and multi-task learning,". https://doi.org/10.48550/arXiv.2103.14253

[32] Q. He, X. Sun, Y. Yu, and W. Li, "Deepchorus: A hybrid model of multi-scale convolution and self-attention for chorus detection," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 411-415. https://doi.org/10.1109/ICASSP43922.2022.9746919

[33] [G. Shibata, R. Nishikimi, and K. Yoshii, "Music structure analysis based on an LSTM-HSMM hybrid model," in Proc. of the 21st Int. Society for Music Information Retrieval Conf., Montréal, Canada, 2020.

[34] J. Wang, Y. Hung, and J. B. L. Smith, "To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 416-420. https://doi.org/10.1109/ICASSP43922.2022.9747252

[35] A. Marmoret, J. E. Cohen, and F. Bimbot, "Convolutive block-matching segmentation algorithm with application to music structure analysis," 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2023, pp. 1-5. https://doi.org/10.1109/WASPAA58266.2023.10248174

[36] T. Kim and J. Nam, "All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio,". 10.48550/arXiv.2307.16425

[37] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie, "Design and creation of a large-scale database of structural annotations," in Proc. of the International Society for Music Information Retrieval Conference. Miami, FL. 555-60, 2011.

[38] SALAMI Dataset, [Online]. Available: https://github.com/DDMAL/salami-data-public

[39] B. Zhang, J. Leitner, and S. Thornton, "Audio recognition using mel spectrograms and convolution neural networks,", 2019.

[40] E. Waisberg et al., "Transfer learning as an AI-based solution to address limited datasets in space medicine," Life Sciences in Space Research, Vol. 36, 2023, pp. 36-38. https://doi.org/10.1016/j.lssr.2022.12.002

[41] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[42] A. A. Dina, T. S. Asmaa, T. N. Marwa, and H. J. Ali, "Classification of COVID-19 from CT chest images using convolutional wavelet neural network," International Journal of Electrical and Computer Engineering Vol. 13, No. 1, 2023. https://doi.org/10.11591/ijece.v13i1.pp1078-1085