

# Pitchclass2vec: Symbolic Music Structure Segmentation with Chord Embeddings

Nicolas Lazzari<sup>1,\*</sup>, Andrea Poltronieri<sup>1,†</sup> and Valentina Presutti<sup>2,†</sup>

<sup>1</sup>Department of Computer Science and Engineering, University of Bologna, Mura Anteo Zamboni, 7, Bologna 40126, Italy

<sup>2</sup>LILEC, University of Bologna, Via Cartoleria, 5, Bologna 40124, Italy

## Abstract

Structure perception is a fundamental aspect of music cognition in humans. Historically, the hierarchical organization of music into structures served as a narrative device for conveying meaning, creating expectancy, and evoking emotions in the listener. Thereby, musical structures play an essential role in music composition, as they shape the musical discourse through which the composer organises his ideas. In this paper, we present a novel music segmentation method, pitchclass2vec, based on symbolic chord annotations, which are embedded into continuous vector representations using both natural language processing techniques and custom-made encodings. Our algorithm is based on long-short term memory (LSTM) neural network and outperforms the state-of-the-art techniques based on symbolic chord annotations in the field.

## Keywords

music structure analysis, structural segmentation, deep learning, chord embeddings

## 1. Introduction

One of the main factors that influence music perception is the hierarchical structure of music compositions. Regardless of their level of musical knowledge and harmonic sensitivity [1] or their cultural origins [2], listeners are able to use intuitive knowledge to organize their perception of musical structures [3]. Indeed, there is empirical evidence that neural activity correlates with musical structure in listeners' perception [4]. The structuring and predictability of musical compositions is also recognised as a viable therapy in the treatment and assessment of children and adolescents with autistic spectrum disorder [5].

Music structuring is one of the tools used by composers to tell a story. According to [6] "Music-making is, to a large degree, the manipulation of structural elements through the use of repetition and change." The repetition of harmonic progressions (sequences of chords), in particular in the context of western tonal music, gives to artists the ability to guide listeners through a journey that creates dramatic narratives, conveying a sense of conflict that demands a solution [7].

---

CREAI 2022 - Workshop on Artificial Intelligence and Creativity, November 28-December 2, 2022, Udine, Italy


\*Corresponding author.


†Alphabetical order.

✉ nicolas.lazzari2@studio.unibo.it (N. Lazzari); andrea.poltronieri2@unibo.it (A. Poltronieri);

valentina.presutti@unibo.it (V. Presutti)

ORCID 0000-0002-1601-7689 (N. Lazzari); 0000-0003-3848-7574 (A. Poltronieri); 0000-0002-9380-5160 (V. Presutti)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Structure of *Helter Skelter* by The Beatles. Chords are presented in Harte format [8].

Figure 1 shows the structure of *Helter Skelter* by The Beatles, highlighting the musical chords of the song. In the example, by means of the alternation between *verse* and *refrain* the artist establishes a common repetitive pattern. The addition of an instrumental section after the second *refrain* and the repetition of the *intro* reinforces the repetitive aspect of the composition. The upcoming *outro* section denies the expectation of a new *verse*, right before the song ends. Expectation and the way it is fulfilled or denied is an essential part in musical enjoyment [7]. In fact, it has been shown empirically that the emotional response to a musical composition varies as the degree of repetition changes [9]. Understanding musical structures is hence fundamental in music analysis and composition. Artists can benefit from the feedback provided by a system able to highlight possible hierarchical structures in their compositions.

Music structure segmentation is a broad term related to the study of musical form, which describes how musical pieces are structured. In particular it can be divided in two main categories: phrase-structure segmentation and global segmentation [10]. Phrase-structure consists in detecting sections from the melodic information of a piece. While the aim of phrase-structure is not to obtain a global segmentation, the detected sections provide valuable insights in the task of global segmentation. In the following, we will refer to music structure segmentation as the task of global segmentation. Music structure segmentation is a music information retrieval (MIR) task that consists in identifying and labelling key music segments (e.g. *chorus*, *verse*, *bridge*) of a music piece [11]. Given a musical composition, its musical segmentation consists in the identification of non-overlapping segments, which we will refer to as sections. Each section is characterized by a label that classifies its function such as *intro* or *verse* in figure 1. A correct segmentation does not necessarily assign the correct labels to each section of the composition, but rather focuses on the correct estimation of the boundaries of each section. Once boundaries has been accurately predicted, a labeling process is performed to obtain the final annotation [12].

Most of the recent methods and research approaches are based on audio analysis techniques [12], nonetheless harmonic information, isolated from tempo and rhythm, have been successfully used in several tasks in the field of music information retrieval (MIR) (e.g. [13, 14]).

In this paper, we focus on the music structure segmentation task by only taking into account harmonic information extracted from symbolic notations (music chord annotations). The assumption behind this approach is that the identification of harmonic sub-sequences (harmonic patterns) can be influential in defining the structure of a song and the sections of which it is composed. For instance, by taking a closer look at Figure 1 it is easy to notice how harmonic information can provide valuable information in the structure segmentation task: all *verses* are roughly based on the same harmonic progression (*E, G, A, E*) while *refrains* are based on a different harmonic progression (*A, E, A, E, E*). A segmentation strongly based on those recurrent patterns is likely to be coherent with the way the composer shaped the progression in the first place.

The objective of this paper is threefold: (i) we propose *pitchclass2vec*, a novel chord embedding method; (ii) we use this encoding with a recurrent neural network on a corpus of musical chords; and (iii) we compare the performance of the encoding with the state-of-the-art methods in the field.

The chord embedding method proposed, *pitchclass2vec*, encodes a chord using a one-hot encoding of the notes that compose it by making use of word embedding techniques. Each embedded chord is defined to be similar to the embedding of its neighbouring chords in an harmonic progression. This formalization is supposed to approximate the semantic meaning of a chord [15] and has been widely used in the natural language processing field [16, 17]. We use *pitchclass2vec* embeddings to train an LSTM neural network that predicts the section of each chord. Through its recurrent layers the neural network is able to learn relationships between the elements of a sequence. This allows the model to detect repetitive patterns of the harmonic progression and predict the a segmentation of the whole composition. The model provides a baseline to test the efficacy of the proposed chord embedding method. State-of-the-art results are achieved in the task of music structure segmentation on symbolic harmonic data, providing evidences that *pitchclass2vec* is able to provide accurate chord representations. However, the embedding method employed here for the music segmentation task, can be used in a variety of applications in the field of Music Information Retrieval, such as retrieving harmonically related pieces [13], automatic chord recognition [18] and music genre classification [19].

The paper is organised as follows: Section 2 introduces the related works, Section 3 describes the novel chord embedding method and the recurrent neural network used for the segmentation task. Section 4 presents the experiments performed and Section 5 gives an overview of the obtained results. Finally in Section 6 we discuss the results and new research directions to be explored.

## 2. Music segmentation: state of the art

Automatic segmentation on audio signal is a prolific research field in which many different solutions have been presented, ranging from self-similarity matrices [20, 21] to neural network based methods [22, 23, 24]. Harmonic content has been used to improve those methods both using probabilistic models [25] and transformer based models [26]. Significant research has been performed on phrase-level structural segmentation based on melodic [27, 28, 29] as well as polyphonic content [30]. However, to the best of our knowledge, the only approach proposed in literature for global music segmentation on symbolic harmonic content is FORM [14].

FORM performs structural segmentation by exploiting repeated patterns extracted from harmonic progressions encoded as sequences of strings. Each string represent a chord. In the original work chord labels are transformed into 24 class of chords, 12 major chords and 12 minor chords, while every other chord feature is removed. In this paper, FORM is re-implemented in order to compare the results of the proposed method with the current state of the art (see Section 5). FORM pattern detection algorithm is based on suffix trees. Each node on a suffix tree represents a (possibly recurrent) sub-sequence of a string. FORM extracts sub-sequences appearing in at least two position in the analyzed harmonic progression. A partial segmentation is obtained by labeling each sub-sequence as a new section. The final segmentation is obtained

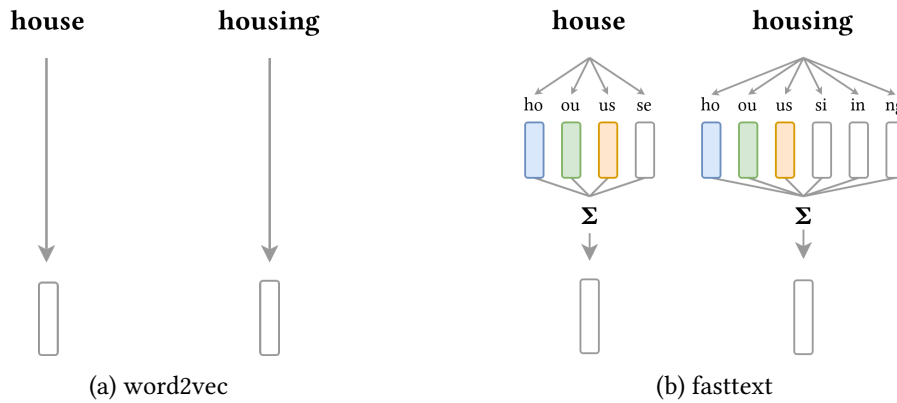
by labeling remaining sub-sequences as their preceding neighbouring section. The results are compared with a random baseline that generates arbitrarily long structures and to a heuristic that assigns to each composition the typical pop song structure *ABBBBCCBBBBCCDCCE* [14], in which each different label represent a structure in the chord progression and is stretched to fit the whole sequence. The main issue with FORM is in way chord labels are compared. The string representation does not take into account semantic similarity between chords nor the algorithm is able to detect near-similar patterns, i.e. patterns whose difference can be ignored in the context of music structure segmentation.

Our structure segmentation method is based on our novel chord representation method, *pitchclass2vec*, based on continuous word representation. The core idea of continuous word representation is based on the Distributional Hypothesis [15]: the semantic meaning of a word  $w$  can be approximated from the distribution of words that appear within the context of  $w$ . The objective of continuous word representation is the maximization of the following log-likelihood:

$$\sum_{t=1}^T \sum_{c \in \mathcal{C}_t} \log p(w_c | w_t),$$

where  $\mathcal{C}_t$  represents the indices of the words that appears as context of the word  $w_t$  and the function  $p(w_c | w_t)$  is parameterized using  $d$ -dimensional vectors in  $\mathbb{R}^d$ , respectively  $\mathbf{u}_{w_t}$  and  $\mathbf{v}_{w_c}$ . The problem can be framed as a binary classification task in which words are predicted to be present (or absent) from the context of  $w_t$ . A similarity function  $s(w_c, w_t)$  between two words  $w_c$  and  $w_t$ , can be computed as the scalar product  $\mathbf{u}_{w_t}^T \mathbf{v}_{w_c}$ . The representations obtained by training the described method on a large corpus correctly approximates the semantic meaning of words. In the last few years, continuous word representation has been applied in a growing number of application areas, achieving state-of-the-art results in the natural language processing field [31] in tasks such as part of speech tagging[32], named entity recognition[33] and document classification[34].

The described approach has been first proposed by the *word2vec* skipgram model [16]. *Word2vec*, however, is limited by the lack of morphological knowledge of a word. When computing the representation of a word, none of its morphological components are taken into account. Let's take for instance two morphologically similar words, *house* and *housing*. Their computation does not share any common element and the final representation of the words will not be influenced by their similarities. *Fasttext* [17] was presented as a solution to this issue and has proven to be more effective in the representation of a word using continuous representations. The novel aspect is in the way representations are computed. At first the n-grams that compose a word are extracted. For each n-gram a continuous vector representation is computed, using the same methodology as *word2vec*. The representation of the original words is finally obtained as the sum of its n-gram components. Using this technique, the final representation of a word is conditioned by its morphological structure. When two words share one or more n-grams their vectors will be the sum of at least one common element, which will bias both vectors in being more similar to each other. An additionally advantage of the *fasttext* approach is the way out-of-vocabulary words (words that never appear in the training corpus, and whose representation is hence unknown) are handled. When using fixed approach such as *word2vec*,



**Figure 2:** *Word2vec* (a) and *fasttext* (b) embedding methods. With *word2vec* each embedding is computed independently of its morphological structure. *Fasttext* instead compute the representation as the sum of the n-grams that compose a word. Words that share one or more n-grams have a similar representation as they are computed in a similar way. In the example 2-grams are represented but in general n-grams up to the length of the term are commonly used.

out of vocabulary words are represented as a static vector, usually randomly sampled from a normal distribution. *Fasttext* instead is able to compute the representation in a meaningful way, given that at least one of the n-grams in the out-of-vocabulary term has been computed previously in the training corpus. Figure 2 shows a visual comparison between *word2vec* (Figure 2a) and *fasttext* (Figure 2b).

Continuous word representations have already been applied to chord symbols with promising results. In *chord2vec* [35] the authors obtain state-of-the-art results on the log-likelihood estimation task. Log-likelihood estimation is the task of correctly estimating, given one element in a sequence, the probabilities of another element being the upcoming element in the sequence. *Chord2vec* is inspired by the *word2vec* method, in which chords are represented by the notes that they are composed of. The representation model proposed by *chord2vec* is similar to *pitchclass2vec*. Instead of computing chord representations with the notes that compose a chord, the representations of *pitchclass2vec* takes into account the relationship between the notes that compose each chord. An in depth discussion is presented in section 3.

More recently, *word2vec*-based approach on symbolic chord annotations has been analyzed by [36]. Chord representations are based on the chord label without taking into account the notes that compose it. The encoding is then used on two different tasks: chord clustering and log-likelihood estimation. The log-likelihood estimation task is used to investigate the historical harmonic style of different composers. The log-likelihood results strongly correlates with current musicological knowledge. For instance the model finds difficult to predict chords from artists that make a sporadic use of common harmonic progressions [36]. The chord clustering task highlights how it's possible to observe similarities between *functionally equivalent* chords (chords that shares notes between each other) and a well defined difference between *functionally different* chords. Continuous word representations are hence adequate to encode chords in the first place, and more importantly they are able to autonomously internalize relations between

chords that have been previously observed by domain experts.

### 3. Pitchclass2vec model

Embedding approaches used in natural language processing have obvious limitations when it comes to dealing with musical content, such as musical chords. While relying on purely syntactical representations has been shown to correctly encapsulate some forms of domain knowledge [36], more advanced representations are needed to obtain accurate results when dealing with harmonic progressions [35].

There are, however, some ambiguous cases in which both vector representations might introduce wrong similarities between chords. Let us take for instance the chords  $C:maj$  and  $C:maj13$ <sup>1</sup>, whose notes are respectively  $\mathcal{C}_{C:maj} = \{C, E, G\}$  and  $\mathcal{C}_{C:maj13} = \{C, E, G, B, D, A\}$ . Both chords' labels only differ by two characters, however the difference between the notes that they are composed of can't be neglected. A method exclusively based on syntactical information would wrongly represent the vectors as similar between each other. Conversely, only relying on the notes that compose a chord results in ambiguous representations of some particular classes of chords, called *enharmonic* chords. For instance, the *enharmonic* chords  $C:dim$  and  $Eb:dim$  share the exact same set of notes,  $\mathcal{C}_{C:dim} = \mathcal{C}_{Eb:dim} = \{C, Eb, Gb, A\}$  but need to be represented as different chords as they serve different harmonic purposes. *Chord2vec* would wrongly represent both chords as the same exact vector.

In order to overcome the aforementioned limitations, we propose an encoding which requirements can be summarised as follows: 1. it has to be based on the constituent notes of a chord, rather than its label; and 2. it must take into account the relation between those notes instead of the notes themselves.

The proposed encoding is grounded on tonal music theory: each chord  $c$  is composed of a set of notes  $\mathcal{C} \subset \mathcal{N}$ , where  $\mathcal{N}$  is the set of all notes and  $C$  is called the *pitch class* of a chord. An important distinction is represented by the *root* note, which names the chord and plays an important role in its harmonic function.

We encode each chord as the Cartesian product  $\mathcal{J}_c = root_c \times \mathcal{C}_c$  between the *root* note and the *pitch class* of the chord. The vector representation  $\mathbf{u}_c$  of a chord  $c$  is computed as

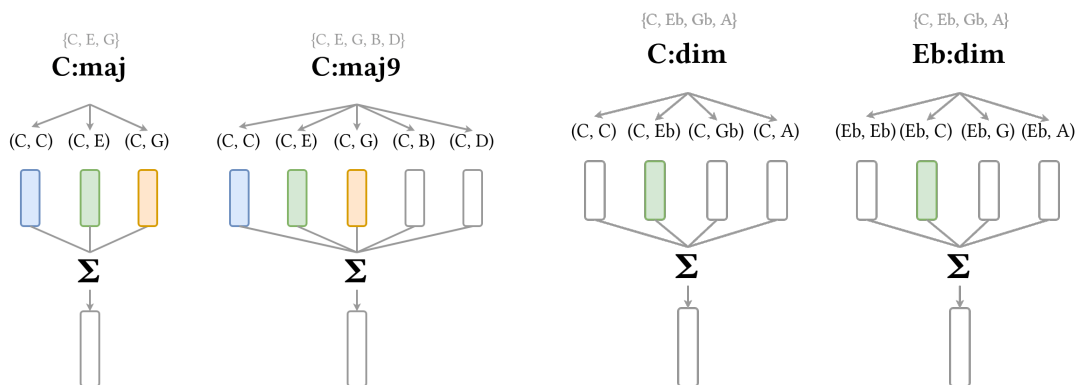
$$\mathbf{u}_c = \sum_{i \in \mathcal{J}_c} \mathbf{u}_i$$

where  $\mathbf{u}_i$  is the vector representation of the tuple  $x_i \in \mathcal{J}_c$ . See Figure 3b for a visual reference on how *pitchclass2vec* handles *enharmonic* chords and Figure 3a on how chords with common components are handled. This formalization can be seen as an extension of the *chord2vec* [35] method, in which the chord inner structure is taken into consideration as well.

Nevertheless, the label of a chord has a well-defined semantic. Chords composed of the same set of notes may have different harmonic functions. For example, the chords  $G:min7$  and  $Bb:6$ , despite different labels contain the exact same notes:  $\mathcal{C}_{G:min7} = \mathcal{C}_{Bb:6} = \{G, Bb, D, F\}$ . This problem is particularly evident in datasets containing annotations made by experts, where the

---

<sup>1</sup>In Harte[8] notation



(a) C:maj and C:maj9 chord embeddings. The final representation is computed from common elements and will hence share some aspects. (b) C:dim and Eb:dim chord embeddings. Both chords are composed of the same notes but using mostly different components.

**Figure 3:** Visual reference on pitchclass2vec embedding method.

choice of label is the result of a meticulous analysis. For this reason, we have implemented two different variants of *pitchclass2vec*: (i) a variant combining the approach proposed by *word2vec* with *pitchclass2vec*; and (ii) a variant combining *fasttext* with *pitchclass2vec*.

In order to obtain mixed embeddings we test different hybrid combinations before passing the new representation to the LSTM model:

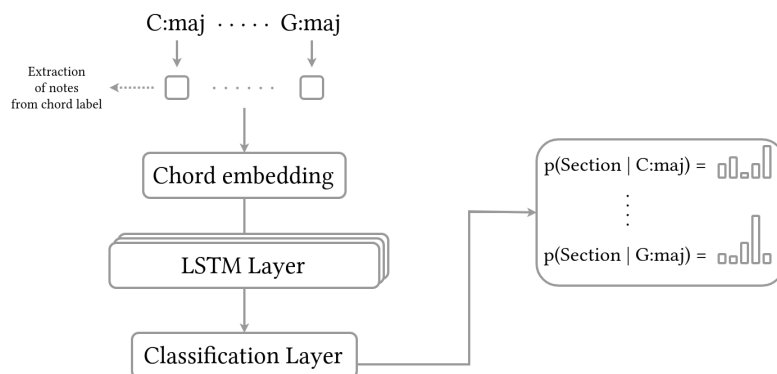
- (i) concatenating the embeddings;
- (ii) concatenating the embeddings and projecting the result in a  $N$ -dimensional vector, using a fully connected layer;
- (iii) projecting the embeddings in the same  $N$ -dimensional space by using two different fully connected layer and summing the  $N$ -dimensional vectors;
- (iv) computing a new representation of each embedding by using two separate LSTM layers and summing the resulting vectors;
- (v) computing a new representation of each embedding by using two separate LSTM layers and concatenating the resulting vectors.

None of the combination used proved to be able to outperform the others and we decided to stick to the first simpler and faster approach.

### 3.1. Implementation details

The model is implemented using pytorch. We train the model on a set of  $\approx 16000$  chord progressions (with a total number of over  $1M$  chord instances), taken from the Chord Corpus (ChoCo) dataset [37]. ChoCo is a chord dataset consisting of more than 20000 tracks taken from 18 different professionally curated datasets. All datasets have been parsed in JAMS [38]





**Figure 4:** Visual depiction of the implemented LSTM model.

format and converted in Harte Notation [8]. We train the model for at most 10 epochs on an *NVIDIA RTX 3090* with batch sizes of 512 harmonic progressions. We manually tune the batch size to efficiently train the model on our available resources. For each chord we take a window of 4 context chords as positive examples, 2 preceding and 2 succeeding, as it has been done in the original *fasttext* implementation [17]. Then, we sample 20 random chords as negative examples. Even though it has been shown that windows of different sizes yields different results depending on the task they are applied to [39] here we will rely on a fixed size window to better compare it to the related works. We subsample our corpus to obtain a more balanced one by removing some of the most frequent chords instances. We use a factor of  $t = 10^{-5}$  as suggested by [16] to allow a faster and more accurate training phase. The model is trained using a standard training procedure where a binary cross entropy loss between a chord and its positive and negative examples is minimized using Adam optimizer, with fixed learning rate of 0.025. We set the embedding dimension to 10 as the result of manual trials.

## 4. Experimental setup

This section shows how the proposed model compares to FORM [14], the state-of-the-art approach in the field of music segmentation. We develop a baseline model using a stacked LSTM-based neural network, depicted in figure 4. The model objective is to predict in which section each chord belongs to. We train our model on the Billboard dataset[40] provided by mirdata[41]. The dataset is composed of 889 expert annotated tracks. Each track is composed of a sequence of chords in Harte format[8], and a sequence of structure labels. Labels are provided in a similar format to the one presented by SALAMI[42]. 80 unique section labels are present in the whole dataset. We preprocess each label and reduce the number of unique labels to 11 by combining all those labels that fall under the same definition given by[42]. A complete reference of the label conversion step is given in table 1.

Although in literature there are neural network architectures that have proven to perform better in similar task [43, 44, 45], we deliberately decided to use a very straightforward architecture. This is due to the fact that the aim of this study is to compare different types of embedding,



**Table 1**

Label conversion reference. Each label is stripped out of numbers and symbols before the conversion.

Source labels	Converted label
[verse]	<i>verse</i>
[prechorus, pre chorus]	<i>prechorus</i>
[chorus]	<i>chorus</i>
[fadein, fade in, intro]	<i>intro</i>
[outro, coda, fadeout, fade-out, ending]	<i>outro</i>
[applause, bass, choir, clarinet, drums, flute, harmonica, harpsichord, instrumental, instrumental break, noise, oboe, organ, piano, rap, saxophone, solo, spoken, strings, synth, synthesizer, talking, trumpet, vocal, voice, guitar, saxophone, trumpet]	<i>instrumental</i>
[main theme, theme, secondary theme]	<i>theme</i>
[transition, tran]	<i>transition</i>
[modulation, key change]	<i>other</i>

rather than to achieve the best performance.

We split our dataset in the usual training, validation and test split (respectively 800, 178 and 89 elements) and fine-tune each model hyper-parameters (number of LSTM stacked layers, LSTM hidden size, dropout probability) to obtain the best results on the validation set. The final configuration of each model is summarised in Table 2. Training is performed using an *NVIDIA RTX 3090* with a batch size of 128. Each model takes at most few minutes to train and average less than 2 milion parameters.

We compare the proposed embedding model to *fasttext* and *word2vec* as well, in which both methods are trained on the string labels of chords in Harte format. Both the models are trained using the highly optimized gensim [46] implementation. The hyperparameters used are the same as the one described in section 3.1 except for the embedding dimension, which is set to 300.

## 5. Results

The results of the experiments are summarised in Table 3. We evaluate the segmentation results by computing pairwise precision, recall and F1-score ( $P$ ,  $R$  and  $F1$  in Table 3) [47] along with

**Table 2**

Best hyper-parameters obtained on the validation set for each model.

Model	Hidden size	Number of stacked layers	Dropout probability
word2vec	100	5	0.3
fasttext	100	5	0.5
pitchclass2vec	100	10	0
pitchclass2vec + word2vec	200	5	0.3
pitchclass2vec + fasttext	200	5	0

under-segmentation, over-segmentation and normalized cross entropy F1 ( $S_U$ ,  $S_O$  and  $S_{F1}$  in table 3) [48]. Every metric is computed using the standard MIR evaluation library `mir_eval` [49]. *Under-segmentation* and *over-segmentation* are two peculiar metrics for the evaluation of automatic music segmentation methods. When a method has an high over-segmentation measure, the final prediction accuracy is influenced mostly by false fragmentation. Conversely an high under-segmentation measure means that the prediction’s segments are the result of ground-truth segments being merged together [48].

Pairwise metrics are computed as the usual precision, recall and F1 scores on the set of identically labeled pairs in the sequence. Precision and recall can be interpreted as the amount of accuracy that is influenced respectively by under-segmentation and over-segmentation. On the other hand under and over-segmentation scores are computed by taking into account the normalized conditional entropy of the segmentation. In short,  $S_O$  gives a measure of how much information is missing in the predicted segmentation, given the ground truth segmentation while  $S_U$  gives a measure of how much noisy information are the result of the predicted segmentation [48]. A graphical explanation of these concepts is provided in Figure 5 (all the examples are taken from [48]).

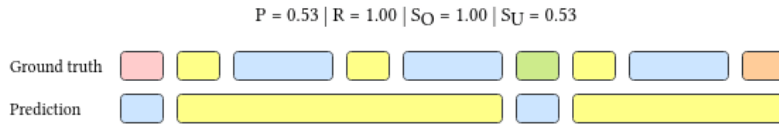
We evaluate our models based on the  $F1$  and  $S_{F1}$  scores of Table 3 since both metrics gives a balanced measure of over and under segmentation.

$FORM_{simple}$  detect repetitive patterns from simplified chord labels, as shown in [14]. The chord simplification process extracts the *root* note from the chord and classifies it either as *major* or *minor*.  $FORM_{raw}$  uses the same labels used by the neural approaches. The former performs better better than the latter. This is not surprising as more patterns between strings can be uncovered by only taking into account 24 labels (12 root notes, each of which can be either *minor* or *major*).  $S_O$  and  $S_U$  however suggests that the over-segmentation based approach of FORM ends up correctly segmenting only some particular portions of the whole composition, while the other ones are wrongly classified. This is an expected behaviour since FORM only relies on label-based repeated patterns. The presence of subtle differences in an harmonic progressions that belongs to the same section, such as the first and second *verse* in Figure 1, are not detected.

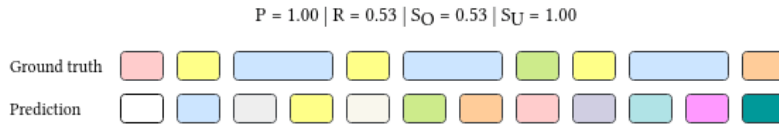
**Table 3**

Evaluation metrics on the test set.  $FORM_{raw}$  is computed on the same chord labels that are used for all the neural approaches.  $FORM_{simple}$  is computed on simplified chord representation as done in [14]: only the chord *root* and its quality (*major* or *minor*) are kept.

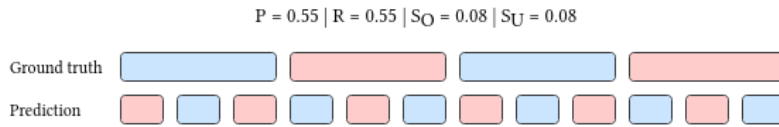
Method	$P$	$R$	$F1$	$S_U$	$S_O$	$S_{F1}$
$FORM_{raw}$	0.667640	0.338019	0.425610	0.667640	0.338019	0.425610
$FORM_{simple}$	<b>0.681281</b>	0.325479	0.416651	0.681281	0.325479	0.416651
word2vec	0.410190	0.823330	0.523692	0.605202	0.257582	0.360186
fasttext	0.373044	<b>0.993201</b>	0.526918	<b>0.947381</b>	0.154424	0.264553
pitchclass2vec	0.402290	0.953399	0.547694	0.719733	<b>0.431959</b>	<b>0.537879</b>
pitchclass2vec + word2vec	0.467940	0.664824	0.532202	0.544820	0.415806	0.471398
pitchclass2vec + fasttext	0.433019	0.835007	<b>0.553045</b>	0.539774	0.425738	0.473986



- (a) High over-segmentation example.  $R = 1$  since if we take each chord in the sequence pairwise then each chord that should be in the same section is indeed in the same section.  $S_O = 1$  since the accuracy of the prediction can be easily explained by the over-segmentation phenomena. Conversely,  $P = 0.53$  and  $S_U = 0.53$  clearly show how the prediction is not able to capture all the needed segments but rather merges ground truth segments together.



- (b) High under-segmentation example. The exact opposite of Figure (a) is displayed. The prediction is not able to capture segments and rather place each chord on its own segment.



- (c)  $P$ ,  $R$  and  $S_O$ ,  $S_U$  compared. In this edge case the main difference between the two measures is highlighted. While  $P$  and  $R$  suggests a decent segmentation  $S_O$  and  $S_U$  clearly states a completely wrong segmentation. Pairwise metrics can be misleading in absence of  $S_O$  and  $S_U$ .

**Figure 5:** Examples of metric computation on relevant instances.

All the neural models in table 3 outperforms FORM. Even though each neural model shows some differences in term of metrics, the clear trend is that syntactical-based models (*fasttext* and *word2vec*) yield overly segmented results, as the low  $S_O$  score suggests, while the approach taken by *pitchclass2vec* produces a more balanced segmentation, as suggested by the  $F1$  score. Surprisingly, *fasttext* and *word2vec* under-performs when compared to the FORM baseline on  $S_{F1}$  metric. The model is not able to generalize enough over the representation and cannot detect patterns that are detected by FORM. Finally, the combination of *pitchclass2vec* with either *fasttext* or *word2vec* doesn't bring any remarkable benefit to our novel representation. Even though an higher  $F1$  score is obtained by using an hybrid approach, the lower  $S_{F1}$  score suggests that it has an underlying less accurate segmentation, similar to example (c) in Figure 5.

## 6. Conclusion and Future Work

In this article, we presented a new embedding method for musical chords, *pitchclass2vec*, that considers the component notes of the chord (also called *pitchclass*), instead of the chord label, as used in embedding methods in the natural language processing field. In addition, we proposed hybrid embedding forms, which combine embedding on the chord label and the novel *pitchclass2vec*. We compared different embedding models, including *pitchclass2vec* with the

state-of-the-art approach in the field of music structural segmentation. We used ChoCo, a dataset of chord annotations, for training the embeddings and Billboard, a dataset of structurally annotated tracks, for the music segmentation task. We used the different types of embeddings on a recurrent neural network (LSTM). The results obtained by using our embeddings outperform the state of the art in every case, with the best result obtained by *pitchclass2vec*, achieving a pairwise F1 score of 0.548 and an over-under-segmentation F1 score of 0.538.

*Pitchclass2vec* is effectively able to learn the harmonic relationships that ties different chords together. Even though the experiments based on *fasttext* and *word2vec* proves to be effective as well, the musical theoretical approach upon which we base *pitchclass2vec* is an essential factor that needs to be taken into account. The presented embedding model proves to be a promising method to improve results in MIR tasks that can be complemented with harmonic information. Moreover, it provides a valuable tool to better understand and analyse harmonic progressions, since it allows a richer comparison between chords and chord sequences when compared to string labels.

There are additional information that we plan to integrate on *pitchclass2vec* to obtain a richer and more accurate representation. For instance, one of the main limitations of our approach stems from the fact that we do not take temporal information into account. We plan to test this possibility by using the temporal information directly in the embedding process and further modify the LSTM model to condition the classification of the section of a chord based on its duration. As discussed in Section 3, chord labels have their own semantic as well. Since the hybrid models proposed did not directly result in more accurate results, we plan on expanding the *pitchclass2vec* method to take into account the label of a chords as well directly in its embedding model. To obtain a semantically richer representations we plan on enhancing *pitchclass2vec* by using deep contextual word embeddings [50] along with knowledge enhancement techniques [51] that combines domain-specific ontologies, such as [52], with deep contextual word embeddings.

Moreover, we plan an in-depth analysis of *pitchclass2vec* training parameters, described in section 3.1, since in [39] the authors showed that, on a product recommendation task, carefully optimized hyper-parameters nearly double the final accuracy on all the experiments.

It is worth mentioning that the LSTM model that we implemented for the structure segmentation task does not take advantage of two fundamental aspect of the task itself: the segmentation of a musical piece should be conditioned by its musical genre. In fact, the annotation guidelines provided in [42] defines some genre-specific labels and encourage their use whenever applicable. Even though a relabeling process can partially solve this issue, the need of genre-specific labels proves that the use-cases of specific structure labels, for instance *theme*, might be different in different musical genres. Finally, as already discussed in section 4, many recent techniques has shown to be effective in increasing the accuracy of different tasks in the NLP field [43, 44, 45] when using continuous word representations. We plan to address these issues in future works.

## Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004746.

## References

- [1] N. Tan, R. Aiello, T. Bever, Harmonic structure as a determinant of melodic organization, *Memory & cognition* 9 (1981) 533–9. doi:10.3758/BF03202347.
- [2] C. J. Stevens, Music perception and cognition: A review of recent cross-cultural research, *Top. Cogn. Sci.* 4 (2012) 653–667. URL: <https://doi.org/10.1111/j.1756-8765.2012.01215.x>. doi:10.1111/j.1756-8765.2012.01215.x.
- [3] P. Collins, M. Schmuckler, Phrasing influences the recognition of melodies, *Psychonomic bulletin & review* 4 (1997) 254–9. doi:10.3758/BF03209402.
- [4] C. L. Krumhansl, P. W. Juszyk, Infants' perception of phrase structure in music, *Psychological Science* 1 (1990) 70–73. URL: <http://www.jstor.org/stable/40062394>.
- [5] T. Wigram, C. Gold, Music therapy in the assessment and treatment of autistic spectrum disorder: clinical application and research evidence, *Child: care, health and development* 32 (2006) 535–542.
- [6] G. Burns, A typology of 'hooks' in popular records, *Popular Music* 6 (1987) 1–20. doi:10.1017/S0261143000006577.
- [7] D. Temperley, *The cognition of basic musical structures*, MIT press, 2004.
- [8] C. Harte, M. B. Sandler, S. A. Abdallah, E. Gómez, Symbolic representation of musical chords: A proposed syntax for text annotations, in: *ISMIR 2005, 6th International Conference on Music Information Retrieval*, London, UK, 11–15 September 2005, Proceedings, 2005, pp. 66–71. URL: <http://ismir2005.ismir.net/proceedings/1080.pdf>.
- [9] S. R. Livingstone, C. Palmer, E. Schubert, Emotional response to musical repetition., *Emotion* 12 (2012) 552.
- [10] M. Giraud, R. Groult, F. Levé, Computational analysis of musical form, in: D. Meredith (Ed.), *Computational Music Analysis*, Springer, 2016, pp. 113–136. URL: [https://doi.org/10.1007/978-3-319-25931-4\\_5](https://doi.org/10.1007/978-3-319-25931-4_5). doi:10.1007/978-3-319-25931-4\_5.
- [11] M. C. McCallum, Unsupervised learning of deep features for music segmentation, in: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019*, Brighton, United Kingdom, May 12–17, 2019, IEEE, 2019, pp. 346–350. URL: <https://doi.org/10.1109/ICASSP.2019.8683407>. doi:10.1109/ICASSP.2019.8683407.
- [12] O. Nieto, G. J. Mysore, C. Wang, J. B. L. Smith, J. Schlüter, T. Grill, B. McFee, Audio-based music structure analysis: Current trends, open challenges, and applications, *Trans. Int. Soc. Music. Inf. Retr.* 3 (2020) 246–263. URL: <https://doi.org/10.5334/tismir.54>. doi:10.5334/tismir.54.
- [13] W. B. de Haas, F. Wiering, R. C. Veltkamp, A geometrical distance measure for determining the similarity of musical harmony, *Int. J. Multim. Inf. Retr.* 2 (2013) 189–202. URL: <https://doi.org/10.1007/s13735-013-0036-6>. doi:10.1007/s13735-013-0036-6.
- [14] W. B. de Haas, A. Volk, F. Wiering, Structural segmentation of music based on repeated harmonies, in: *2013 IEEE International Symposium on Multimedia, ISM 2013*, Anaheim, CA, USA, December 9–11, 2013, IEEE Computer Society, 2013, pp. 255–258. URL: <https://doi.org/10.1109/ISM.2013.48>. doi:10.1109/ISM.2013.48.
- [15] M. Sahlgren, The distributional hypothesis, *Rivista di Linguistica (Italian Journal of Linguistics)* 20 (2008) 33–53.
- [16] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations

- in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013. URL: <http://arxiv.org/abs/1301.3781>.
- [17] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, *Trans. Assoc. Comput. Linguistics* 5 (2017) 135–146. URL: [https://doi.org/10.1162/tacl\\_a\\_00051](https://doi.org/10.1162/tacl_a_00051). doi:10.1162/tacl\_a\_00051.
- [18] K. O’Hanlon, M. B. Sandler, Fifthnet: Structured compact neural networks for automatic chord recognition, *IEEE ACM Trans. Audio Speech Lang. Process.* 29 (2021) 2671–2682. URL: <https://doi.org/10.1109/TASLP.2021.3070158>. doi:10.1109/TASLP.2021.3070158.
- [19] H. Liang, W. Lei, P. Y. Chan, Z. Yang, M. Sun, T. Chua, Pirhdy: Learning pitch-, rhythm-, and dynamics-aware embeddings for symbolic music, in: C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, R. Zimmermann (Eds.), *MM ’20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, ACM, 2020, pp. 574–582. URL: <https://doi.org/10.1145/3394171.3414032>. doi:10.1145/3394171.3414032.
- [20] R. J. Weiss, J. P. Bello, Unsupervised discovery of temporal structure in music, *IEEE J. Sel. Top. Signal Process.* 5 (2011) 1240–1251. URL: <https://doi.org/10.1109/JSTSP.2011.2145356>. doi:10.1109/JSTSP.2011.2145356.
- [21] R. J. Weiss, J. P. Bello, Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization, in: J. S. Downie, R. C. Veltkamp (Eds.), *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, International Society for Music Information Retrieval, 2010, pp. 123–128. URL: <http://ismir2010.ismir.net/proceedings/ismir2010-23.pdf>.
- [22] G. Shibata, R. Nishikimi, K. Yoshii, Music structure analysis based on an LSTM-HSMM hybrid model, in: J. Cumming, J. H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, T. de Reuse (Eds.), *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, 2020, pp. 23–29. URL: <http://archives.ismir.net/ismir2020/paper/000005.pdf>.
- [23] J. Wang, J. B. L. Smith, W. T. Lu, X. Song, Supervised metric learning for music structure features, in: J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, A. Srinivasamurthy (Eds.), *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, 2021, pp. 730–737. URL: <https://archives.ismir.net/ismir2021/paper/000091.pdf>.
- [24] A. Marmoret, J. E. Cohen, N. Bertin, F. Bimbot, Uncovering audio patterns in music with nonnegative tucker decomposition for structural segmentation, *CoRR abs/2104.08580* (2021). URL: <https://arxiv.org/abs/2104.08580>. arXiv:2104.08580.
- [25] J. Pauwels, F. Kaiser, G. Peeters, Combining harmony-based and novelty-based approaches for structural segmentation, in: A. de Souza Britto Jr., F. Gouyon, S. Dixon (Eds.), *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, November 4-8, 2013*, 2013, pp. 601–606. URL: [http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/138\\_Paper.pdf](http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/138_Paper.pdf).
- [26] T. Chen, L. Su, Harmony transformer: Incorporating chord segmentation into harmony recognition, in: A. Flexer, G. Peeters, J. Urbano, A. Volk (Eds.), *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft*,



- The Netherlands, November 4-8, 2019, 2019, pp. 259–267. URL: <http://archives.ismir.net/ismir2019/paper/000030.pdf>.
- [27] B. W. Frankland, A. J. Cohen, Parsing of melody: Quantification and testing of the local grouping rules of Ierdahl and Jackendoff's a generative theory of tonal music, *Music Perception* 21 (2004) 499–543.
- [28] E. Cambouropoulos, The local boundary detection model (LBDM) and its application in the study of expressive timing, in: *Proceedings of the 2001 International Computer Music Conference, ICMC 2001, Havana, Cuba, September 17-22, 2001*, Michigan Publishing, 2001. URL: <https://hdl.handle.net/2027/spo.bbp2372.2001.021>.
- [29] G. Velarde, T. Weyde, D. Meredith, An approach to melodic segmentation and classification based on filtering with the haar-wavelet, *Journal of New Music Research* 42 (2013) 325–345.
- [30] D. Meredith, K. Lemström, G. A. Wiggins, Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music, *Journal of New Music Research* 31 (2002) 321–345.
- [31] Q. Jiao, S. Zhang, A brief survey of word embedding and its recent development, in: *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, 2021, pp. 1697–1701. doi:10.1109/IAEAC50856.2021.9390956.
- [32] S. Meftah, N. Semmar, A neural network model for part-of-speech tagging of social media texts, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, European Language Resources Association (ELRA), 2018. URL: <http://www.lrec-conf.org/proceedings/lrec2018/summaries/913.html>.
- [33] J. P. C. Chiu, E. Nichols, Named entity recognition with bidirectional lstm-cnns, *Trans. Assoc. Comput. Linguistics* 4 (2016) 357–370. URL: [https://doi.org/10.1162/tacl\\_a\\_00104](https://doi.org/10.1162/tacl_a_00104). doi:10.1162/tacl\_a\_00104.
- [34] J. Lilleberg, Y. Zhu, Y. Zhang, Support vector machines and word2vec for text classification with semantic features, in: N. Ge, J. Lu, Y. Wang, N. Howard, P. Chen, X. Tao, B. Zhang, L. A. Zadeh (Eds.), *14th IEEE International Conference on Cognitive Informatics & Cognitive Computing, ICCI\*CC 2015, Beijing, China, July 6-8, 2015*, IEEE Computer Society, 2015, pp. 136–140. URL: <https://doi.org/10.1109/ICCI-CC.2015.7259377>. doi:10.1109/ICCI-CC.2015.7259377.
- [35] S. Madjiheurem, L. Qu, C. Walder, Chord2vec: Learning musical chord embeddings, in: *Proceedings of the constructive machine learning workshop at 30th conference on neural information processing systems (NIPS2016)*, Barcelona, Spain, 2016.
- [36] E. Anzuoni, S. Ayhan, F. Dutto, A. McLeod, admin, M. Rohrmeier, A historical analysis of harmonic progressions using chord embeddings, in: *Proceedings of the 18th Sound and Music Computing Conference, 2021*, pp. 284–291.
- [37] J. de Berardinis, A. Meroño-Peñuela, A. Poltronieri, V. Presutti, Choco: a chord corpus and a data transformation workflow for musical harmony knowledge graphs, in: *Manuscript under review*, 2022.
- [38] E. J. Humphrey, J. Salamon, O. Nieto, J. Forsyth, R. M. Bittner, J. P. Bello, JAMS: A JSON annotated music specification for reproducible MIR research, in: H. Wang, Y. Yang, J. H.



- Lee (Eds.), Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014, 2014, pp. 591–596. URL: [http://www.terasoft.com.tw/conf/ismir2014/proceedings/T106\\_355\\_Paper.pdf](http://www.terasoft.com.tw/conf/ismir2014/proceedings/T106_355_Paper.pdf).
- [39] H. Caselles-Dupré, F. Lesaint, J. Royo-Letelier, Word2vec applied to recommendation: hyperparameters matter, in: S. Pera, M. D. Ekstrand, X. Amatriain, J. O’Donovan (Eds.), Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018, ACM, 2018, pp. 352–356. URL: <https://doi.org/10.1145/3240323.3240377>. doi:10.1145/3240323.3240377.
- [40] J. A. Burgoyne, J. Wild, I. Fujinaga, An expert ground truth set for audio chord recognition and music analysis, in: A. Klapuri, C. Leider (Eds.), Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011, University of Miami, 2011, pp. 633–638. URL: <http://ismir2011.ismir.net/papers/OS8-1.pdf>.
- [41] R. M. Bittner, M. Fuentes, D. Rubinstein, A. Jansson, K. Choi, T. Kell, mirdata: Software for reproducible usage of datasets, in: A. Flexer, G. Peeters, J. Urbano, A. Volk (Eds.), Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019, 2019, pp. 99–106. URL: <http://archives.ismir.net/ismir2019/paper/000009.pdf>.
- [42] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. D. Roure, J. S. Downie, Design and creation of a large-scale database of structural annotations, in: A. Klapuri, C. Leider (Eds.), Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011, University of Miami, 2011, pp. 555–560. URL: <http://ismir2011.ismir.net/papers/PS4-14.pdf>.
- [43] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, CoRR abs/1508.01991 (2015). URL: <http://arxiv.org/abs/1508.01991>. arXiv:1508.01991.
- [44] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, W. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2015, pp. 802–810. URL: <https://proceedings.neurips.cc/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html>.
- [45] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, The Association for Computational Linguistics, 2016, pp. 606–615. URL: <https://doi.org/10.18653/v1/d16-1058>. doi:10.18653/v1/d16-1058.
- [46] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50. <http://is.muni.cz/publication/884893/en>.
- [47] M. Levy, M. B. Sandler, Structural segmentation of musical audio by constrained clustering, IEEE Trans. Speech Audio Process. 16 (2008) 318–326. URL: <https://doi.org/10.1109/TASL.2007.910781>. doi:10.1109/TASL.2007.910781.
- [48] H. M. Lukashevich, Towards quantitative measures of evaluating song segmentation, in: J. P. Bello, E. Chew, D. Turnbull (Eds.), ISMIR 2008, 9th International Conference on Music

- Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008, 2008, pp. 375–380. URL: [http://ismir2008.ismir.net/papers/ISMIR2008\\_219.pdf](http://ismir2008.ismir.net/papers/ISMIR2008_219.pdf).
- [49] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. W. Ellis, Mir\_eval: A transparent implementation of common MIR metrics, in: H. Wang, Y. Yang, J. H. Lee (Eds.), Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014, 2014, pp. 367–372. URL: [http://www.terasoft.com.tw/conf/ismir2014/proceedings/T066\\_320\\_Paper.pdf](http://www.terasoft.com.tw/conf/ismir2014/proceedings/T066_320_Paper.pdf).
- [50] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: <https://aclanthology.org/N18-1202>. doi:10.18653/v1/N18-1202.
- [51] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, N. A. Smith, Knowledge enhanced contextual word representations, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 43–54. URL: <https://aclanthology.org/D19-1005>. doi:10.18653/v1/D19-1005.
- [52] S. Kantarelis, E. Dervakos, N. Kotsani, G. Stamou, Functional harmony ontology: Musical harmony analysis with description logics, *Journal of Web Semantics* 75 (2023) 100754. URL: <https://www.sciencedirect.com/science/article/pii/S1570826822000385>. doi:<https://doi.org/10.1016/j.websem.2022.100754>.