

Master thesis on Sound and Music Computing
Universitat Pompeu Fabra

Computational Analysis of Arab-Andalusian Music Through Symbolic Pattern Discovery Algorithms

Miguel Garcia Casado

Supervisor: Xavier Serra

Co-Supervisor: Andres Ferraro

September 2020



Copyright© 2020 by Miguel G. Casado

Licensed under Creative Commons Attribution 4.0 International



Contents

1	Introduction	1
1.1	Context	1
1.2	Related Work	2
1.2.1	Arab-Andalusian Studies	2
1.2.2	Symbolic Pattern Discovery	4
1.2.3	Symbolic Pattern Similarity	6
1.3	Goals and Structure of the Thesis	8
2	Dataset	9
2.1	Description of the data	9
2.2	Accessibility of the data	11
2.3	Analysis of the data	12
3	Pattern analysis	14
3.1	Finding relevant patterns per nawba	15
3.1.1	Motivation	15
3.1.2	Methodology	15
3.1.3	Pre-processing	16
3.1.4	SIA Algorithm	16
3.1.5	Post-processing	18
3.1.6	Results	19
4	Evaluation	21

4.1	Automatic Evaluation	21
4.1.1	Nawba Classifier with pattern occurrences	21
4.1.2	Nawba classifier with pattern features	23
4.1.3	Nawba Classifier with pattern occurrences and selected features	26
4.2	Qualitative evaluation	27
5	Discussion	34
5.1	Discussion	34
5.2	Conclusions	36
5.3	Reproducibility	37
	List of Figures	38
	List of Tables	39
	Bibliography	40
A	First Appendix	45

Acknowledgement

As first, I would like to give my most sincere thank to Xavier Serra, who open the research world to me and gave me the opportunity to become a family member of the MTG, his advice and guidance during this time are priceless.

I would like to thank to Andrés Ferraro, who has been always keen to offer me his time and valuable help with no conditions, being the main core of help during this last year.

I really would like to express my gratitude to Rafael Caro, very important person of this work, being able to offer me his hands at any time and giving the musicological perspective that I totally needed for this work, it would not have been possible without his guidance.

Another special thank goes to Tom, quickly becoming more than a Master colleague, always being besides me and being the person who triggered the computational flavour of the *centos*.

I also want to give my warm thank to my former colleagues of the ASPLAB, specially Alia, being the person who I shared the Arab-Andalusian music with and always being there for any needed support. Also, I want to thank to all the amazing people I've met during this time at the MTG, as well as all my wonderful colleagues at Dolby. Thanks Sonia for sowing the seeds and encourage me at any time.

I can't close this words without thank to my family for just being unconditionally there, listening and taking care of me every day. Thanks to my friends, Pepe, Nacho and Carlos, for always being the most solid pilar. Thank you Zapa, for having the ability of being close from very far away.

As last but not least, thank you Berta for give me your hand and simply share the path.

Abstract

Arab-Andalusian music was originated in the medieval Islamic territories of the Iberian Peninsula, mixing local traditions and Arabic influences. The expert performer and researcher of the Moroccan tradition of this music, Amin Chaachoo, has developed a theory which supports that *Centonization*, a melodic composition technique used in Gregorian chant, was also utilized for the creation of this repertoire.

In this thesis, by applying symbolic pattern discovery algorithms to a music score corpus of Arab-Andalusian music of the Moroccan tradition, we aim to contribute to the development of Amin Chaachoo's theory. For this purpose, relevant patterns are discovered from the transcriptions and compared with the list of musical motives proposed by the expert. As last, results are evaluated using automatic and manual methods, finally, a visualization tool is designed and presented in order to be able to show, listen and compare the patterns in an easy way for the Arab-Andalusian researchers.

Keywords: Pattern Discovery, Arab-Andalusian, Ethnomusicology

Chapter 1

Introduction

1.1 Context

The abundant culture developed in the medieval Islamic territories of the Iberian Peninsula known as Al-Andalus originated a refined musical and literary tradition that combines local musical practices with Middle Eastern Arabic poetry and sensibilities. The forced migration of Andalusian population to North Africa brought this tradition to this area, where it survived to this date after the disappearance of Al-Andalus in the 15th century. Nowadays, it is considered the classical musical repertoire in countries such as Morocco, Algeria, and Tunisia, in each of which it developed local characteristics, and is commonly known (among other names [1]) as Arab-Andalusian music. In this thesis, the focus is on the Moroccan repertoire of this tradition, which is known as *al-Āla* [2].

This work is run under the CompMusic project¹, which challenges the current Western centered information paradigms and try to advance in our information technology research contributing to our rich multicultural society. The idea of this project comes from the desire of promoting a fresh approach to Music Information Retrieval (MIR) research, without the bias that this current research has towards western

¹CompMusic: Computational models for the discovery of the World's Music.
<https://compmusic.upf.edu/>

pop music and thus, influenced by western classical music concepts, as explained by Serra [3].

The main objective of this thesis is to help to develop a musicological theory of a Moroccan expert in the field, Amin Chaachoo, in terms of musical motives and patterns. By applying Symbolic Pattern Discovery algorithms to our Arab-Andalusian score repertoire, some useful information for experts and musicologists is provided, in terms of musical motives or patterns. This set of algorithms allow to analyse a large corpus automatically. Furthermore, the results are evaluated following a quantitative and a qualitative method. As a first use case of applying computational techniques to this musical culture, with the idea of representing and evaluating our results in the most easy way to be understood, the "Arab-Andalusian Patterns Visualization Tool" is created to help musicologists and experts on the field to further study and keep their research on this tradition.

1.2 Related Work

In order to understand the conducted work of this thesis, it is necessary to know most of the relevant work done on both musical and computational domain.

Firstly, the main musical concepts of this tradition will be described and then, some ethnomusicological studies in Arab-Andalusian music will be presented. After this, work done in the music pattern composition technique from Amin Chaachoo's theory is described for a better understanding of the general structure of the project. As last, different symbolic pattern recognition as well as symbolic pattern similarity algorithms are explained to show the state of the art of this topic and justify our selected methods.

1.2.1 Arab-Andalusian Studies

Arab-Andalusian Theory

The first important concept that has to be taken into account is the concept of *nawba* (in plural *nawabāt*). One *nawba* can be understood as type of suite which

also includes orchestral pieces and both instrumental and vocal solo improvisations [2, 4]. Each *nawba* is linked to an emotion that produces in the listener and a part of the day that describes.

The core of this tradition is the singing of *ṣanā'ī'* (plural of *ṣan'a*) or poems either by a choir accompanied by an instrumental ensemble or by a soloist. These *ṣanā'ī'* are performed along the *nawabāt*. All the *ṣanā'ī'* and other pieces in one *nawba* are composed in one single mode, known in this tradition as *ṭab'* (plural *ṭubu'*). In the specific case of the Moroccan *al-Āla* repertoire, pieces from certain *nawabāt* were lost during the process of oral transmission, so that the surviving ones were attached to other *nawabāt* according to modal similarity. In the 18th century, the scholar al-Haiek fixed the number of *nawabāt* in the *al-Āla* tradition to eleven (as described more in detail in Chapter 2) [2]. A last important concept is the *mīzān*, which is related to the rhythm and has two dimensions. The first one corresponds to the type of rhythmic patterns that a section has, and the second one is referred to the tempo ranges in which the *mīzān* rhythmic patterns are performed, which in a *nawba* are normally three: *muassa'*, *mahzūz* and *inṣirāf*.

All these concepts are represented in the diagram of Fig.1, to make a better idea of the general structure of the music. All the definitions of these concepts are

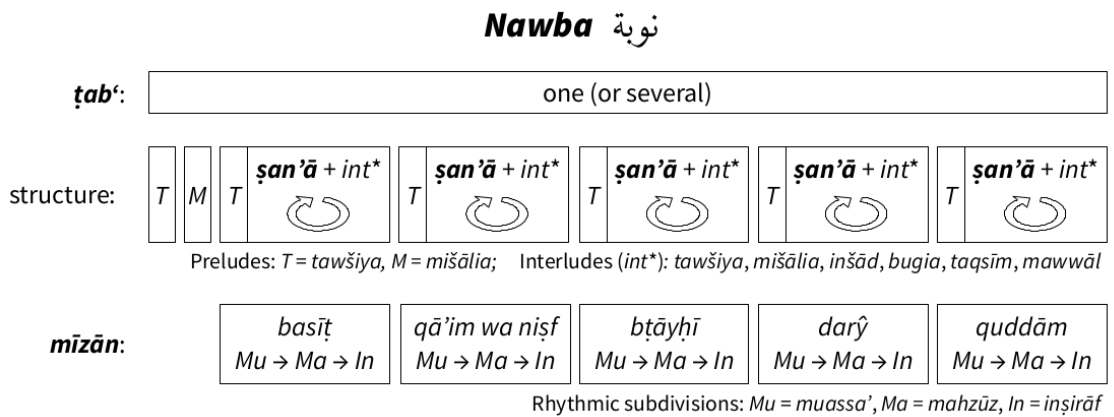


Figure 1: Diagram that summarizes the structure of a *nawba*

summarize in Appendix A.

Related musicological Work

This musical culture has not yet been explored or researched in the academic community as thoroughly as one might wish, the work carried out by Mahmoud Guettat and Christian Poché [4, 1] are well known, who did a detailed review of each component of the musical culture. However, the expert performer and researcher of the Moroccan tradition of this music, Amin Chaachoo, is developing a theory whose last formulation was recently published [5], which argues that *centonization*, a melodic composition technique used in Gregorian chant, was also utilized for the creation of this repertoire. This technique consists on the use of existing material, called *centos* to create new one. Basically, *centonization* usually uses the superposition of this melodic motives on important part of the musical phrases as beginning and end, to make new compositions. The concept of *centonization*, which comes from latin *cento* meaning patchwork, was firstly defined by Paolo Ferretti [6] as melodic composition by synthesis of pre-existing musical units known as *centos*. After that, other important authors have associated this technique to the melodic creation of Gregorian Chant [7, 8]. However, Amin Chaachoo is the first one to link this composition technique to Arab-Andalusian music and, therefore, this work is devoted to the help of this theory.

1.2.2 Symbolic Pattern Discovery

One of the most important elements in music is the recurrent repetition of structures. These repetition or musical motifs [9] provide interesting information for the analysis and collections of musical pieces. Automatic pattern discovery is an active research area in MIR that aims to discover these patterns automatically. A good example of it, is the challenge that MIREX² proposes every year regarding pattern prediction which makes the field move forward and get to know the most recent algorithms.

There exists many studies into melodic pattern discovery with symbolic scores, summaries of which have been made by Jansen et al. [10] and more recently Ren et al.

²Music Information Retrieval Evaluation eXchange. https://www.music-ir.org/mirex/wiki/MIREX_HOME

[11]. Lack of agreement on the current state-of-the-art stems from the difficulty in evaluating approaches, with expertly annotated ground truth often required for performance measurement, more often than not on a study-by-study basis. Following this idea, the work of Chaachoo [2] is used for validation of our results with a set of annotated *centos* relevant per *nawba*. In addition, it is wanted to go further in proposing/supplementing his studies with unique insights of our own. Until now, symbolic pattern discoveries algorithms have been clustered into four categories [11] which will be described in the next paragraphs.

String-based approaches

This approach takes each score as a whole string of notes and represents the patterns by counting the number of instances of re-occurring sub-sections of notes in a musical sequence and their significance computed by comparing these counts, ignoring potential *interaction* between non-consecutive notes, examples of this can be found in [12, 13, 14, 15]. The appeal of these methods is that the theory is intuitive to a non-specialist and aligns with what a musician might consider important when characterising a musical piece melodically, an important consideration when wishing to contribute to and communicate with Chaachoo and his works. Because of this, this trend was followed in our past work [16] with a similar approach by applying TF-IDF algorithm to our corpus in order to discover new potential meaningful musical patterns.

Geometric Approaches

Another important family of symbolic pattern discovery algorithms is the one that bases its approach in geometric methods. All the previously described algorithms assume that the music is represented either as a 1-dimensional string of symbols or, in the case of polyphonic music, as a set of such symbol strings. However, the geometric approaches represent the music as a multidimensional dataset [17]. In this project, one of the most rudimentary of this nature will be used, Structure Induction Algorithm (SIA) [18]. The justification of the selection is based on the interest to check if a geometric based algorithm specially designed for polyphonic scores can

also perform well with our monophonic scores and compare the results with the ones from our last work [16]. An algorithm based on 2D space such as offset and value could discover patterns of different nature than one based on a single string. In addition, the simplicity of our data, being recurrently repetitive with small intervals and linear melodic lines bring us to think that one basic algorithm might be the key to find new relevant patterns. More recent works propose improvements over the SIA algorithm, for example, the work from Collins et al. [19], called SIARCT-FP, which deals with inexactness problems. These problems come from the limited number of inexact repetition that the algorithms provides and the high number of false positives that they return.

Data mining and Machine Learning Approaches

Other two smaller families of algorithms deal with symbolic pattern discovery issues. Ren presented a new approach [20] by representing each score as a sequence of pitch-duration pairs and using the definition of 'closed patterns' to discover new patterns in a corpus. In addition, other new approaches were recently proposing the use of high-level music descriptors and deep learning models to discover patterns [21], not being more efficient than the previous ones.

1.2.3 Symbolic Pattern Similarity

The main goal of using symbolic pattern similarity algorithms in this thesis is the evaluation of our discovered patterns. Once symbolic pattern discovery algorithms have been applied, it is necessary to know if this algorithm is performing well, or just finding very similar patterns that can be clustered into more general groups. Symbolic Pattern Similarity is another active field inside MIR community, for instance they have had MIREX challenges several years as well. Given two or more sequences of notes, Symbolic Melodic Similarity aims to evaluate their degree of likeness, as human listeners are able to do. There exists just a few studies comparing the different algorithms of the field and a lack of consensus in the state of the art, good examples are the ones carried by Jansen et al. [22] and Velardo et al. [23]. Following the subdivisions from [23], four categories of algorithms can be differen-

tiated, based on the strategies they exploit: *cognition*, *music theory*, *mathematics* and *hybrid*. This thesis will be focused on the systems based on mathematics, as they correspond to the main core of algorithms in the MIR field, being the most numerous and successful.

Algorithms based on mathematics

Systems in this category use a number of mathematical approaches which serve as a basis for evaluating the degree of likeness between melodies. Many of these algorithms represent musical data as functions within an abstract space and exploit notions from geometry as a means of comparison. Other common mathematical strategies are based on statistical analysis or information-retrieval techniques.

One of the first attempt to find suitable algorithms in this category was carried out by Alopous [24]. The melody was represented as a polygonal chains within a pitch-time abstract plane and the similarity was calculated as the minimum area between polygonal chains. After this, Meredith et al. [18] presented two algorithms based on a geometric idea of the melody. The similarity between the query pattern and the target melody is defined by the number of elements that match between them, after the application of some invariants, they follow the same idea as the above described algorithm.

Finally, Urbano[25] proposed three different systems: ShapeH, ShapeTime, and Time. All three rely on a geometric model that encodes melodies as polygonal splines in a pitch/time space, and their similarity is computed as the similarity of their shape. It was implemented with a local alignment algorithm over sequences of n-grams that define spline spans. Since these algorithms had big success on the specific MIREX challenge during these years, it was decided to use them for this thesis³. In particular, ShapeH algorithm was selected, since was the one performing better and being time independent at the same time, and this requirement was needed as our patterns were time and non-time dependent.

³MelodyShape tool was used to compute the different similarity between the output patterns. <https://github.com/julian-urbano/MelodyShape>

Other symbolic similarity approaches need to be taken into account. Bohak and Marolt [26] investigated how melody-based features relate to folk-song variants. Another good example were the six proposed algorithms by Wolkowicz and Kešelj [27] that exploited text information retrieval approaches. They extracted features from an input dataset of MIDI files. As a last example, Frieler [28] proposed an approach in which melodies were represented by series of arbitrary length in an abstract space of events. N-grams were then used for measuring the similarity of two melodies.

1.3 Goals and Structure of the Thesis

Derived from the above described, a few research questions come up to the development of this work. As first goal of this thesis, it is wanted to know if this symbolic pattern algorithms are useful to support Amin Chaachoo's theory and whether they propose new and interesting material for the music experts to work on, for instance, if the discovered patterns are able to characterize the *nawba*. From this thought, other ideas will be developed on the thesis, as the legitimacy to use SIA algorithm for the same purpose and its efficiency. As last but not least, it is desired to know which pattern recognition algorithm works better with this kind of data and for the task of finding and discover new *centos* that make some sense musicologically speaking.

First, the dataset utilized is presented and described in Chapter 2. Then, the main symbolic pattern discovery experiment and the results are detailed in Chapter 3. After this, Chapter 4 is dedicated to a rigorous evaluation of the results, both quantitative and qualitative. In this chapter, the Arab-Andalusian Visualization Tool is described and presented, a tool which allows to listen to the result patterns through an easy interface. Finally, an extensive discussion is carried out in Chapter 5 in order to wrap up and get the corresponding conclusions, ensuring the fully reproducibility of the work.

Chapter 2

Dataset

In this chapter, the dataset used for the thesis is described. This data has been taken from the Dunya web page[29] of the CompMusic project, a research project that studies several world music traditions from an information technology point of view. In the next sections the content of the dataset is described, an explanation on the accessibility to it is presented and a brief analysis of the quality of the dataset is detailed.

2.1 Description of the data

The specific data selected for the thesis comes from the set of music scores from the Arab-Andalusian collection [30] [31] of Dunya. These scores are monophonic prescriptive transcriptions from heterophonic recordings from the Moroccan tradition selected and made by the expert Amin Chaachoo. Most of them were recorded in the 1960s and 1970s and they mainly come from radio programs and personal recordings. The reason to use these older recordings is because of the high musical quality of the performing orchestras at that time, which included some of the most recognised maestros of Arab-Andalusian music in Morocco [3].

The whole Arab-Andalusian dataset is formed by 156 recordings of one hour long in average, which provides an extensive quantity of material to be studied. However,

the symbolic scores are reduced to 149, because not all the recordings were able to be transcribed, in any case, this consists on a long enough group of symbolic data that allows the study of the music tradition in detail. The scores come from performance interpreted by four Moroccan orchestras: Tetouan Orchestra, Orchestra of the Conservatory of Tetouan, Brihi Orchestra and RTM Orchestra. The scores are stored in *MusicXML* format, standard score machine readable format that can be read by the majority of softwares. The distribution of numbers of scores per *nawba* can be seen in Table 1.

Every recording is accompanied by the lyric of each *sana'a* as well having arabic and transliterated lyrics. This data hasn't been used in the context of the thesis, but it is a big part of the Arab-Andalusian dataset and it could be utilized for the purpose of lyric-to-audio alignment, or for future works using *Natural Language Processing* (NLP) techniques. Indeed, they are split by sections as well as the audio is.

<i>Nawba</i> number	<i>Nawba</i> transliterated name	Number of Scores
1	<i>raml al-māya</i>	19
2	<i>al-isbahān</i>	13
3	<i>al-māya</i>	13
4	<i>rasd al-dāyl</i>	18
5	<i>al-istihlāl</i>	24
6	<i>al-rasd</i>	10
7	<i>garibat al-ḥusayn</i>	13
8	<i>al-ḥiḡāz al-kabir</i>	10
9	<i>al-ḥiḡāz al-māšriqi</i>	15
10	<i>‘irāq al-‘aḡam</i>	7
11	<i>al-‘uššāq</i>	7

Table 1: Distribution of Scores across *nawabāt*

The music of this tradition is given to numerous ornamentations and improvisations by the musicians, depending on his emotion or state of mind. For this reason, two performances of the same music piece by the same orchestra might have substantial melodic and structural differences. Hence, it is more appropriate to have transcriptions at a recording level instead of a composition level. In many cases, due to the large amount of ornamentations, it might be possible that the original melody is lost. Therefore, the core of the Arab-Andalusian music is the performance, since inter-

preters do not normally follow a score. This might suppose a weakness in our study, since our symbolic data could have a lack of melodic richness and ornamentations coming from the transcription of old audio recordings.

Apart from the presented data, each score contains valuable metadata that has been used along this thesis. This metadata has information about the interpreter orchestra, sections (*mizān*) in the recording and *ṭāb*′. With the intention of make a good use of the metadata, it has been decided to group the *ṭubu*′ from the same *nawba* together, since a single score could contain fragments of different *ṭāb*′s and this annotation is left, using the *nawba* level, there is not room for confusion.

As last, annotations following musicologist’s guidance of the sections (*mizān*) for a subset of 129 scores of the collections were used in the visualization tool. The distribution of the annotated scores can be seen in Table 2, differentiating by solo sections (*mišālia* ,*tawāšī*) and choral/orchestral (*muassa*′, *mahzūz*, *inṣirāf*).

<i>Mizān</i>	Number of sections
<i>mišālia</i>	138
<i>tawāšī</i>	87
<i>muassa</i> ′	121
<i>mahzūz</i>	125
<i>inṣirāf</i>	134

Table 2: Distribution of annotated sections

2.2 Accessibility of the data

As all the data stored in Dunya, the Arab-Andalusian dataset is centred on *Music Brainz*, open music encyclopedia that collects music metadata and makes it available to the public. All the metadata has been stored in this platform and each recording is available through an unique ID called "Music Brainz ID". This platform supports the open source initiatives, and it is a good way of promoting this kind of alternatives, having this valuable data open for everyone. At the same time, Dunya follows the same philosophy, providing open research material of five music traditions such as Carnatic, Hindustani, Makam, Jingju and Arab-Andalusian. This research material consists on recordings, scores, lyrics, metadata and other useful data such as

pitch contours and pitch histograms. In order to facilitate the access to both score and metadata, in the context of the thesis, some useful functions were added to the Arab-Andalusian section of the Dunya API, called `pycompmusic`¹. Being registered in the Dunya web page, the Arab-Andalusian API allows to download any kind of data from the whole dataset and allows the integration of the API in any possible software. In addition, there exists another set of Jupyter Notebooks that facilitate the download and accessibility of the whole dataset designed by Niccolò Pretto [32].

Finally, the sections annotations are available in the next repository². They are planned to be accesible through the Arab-Andalusian API, but it is a work in progress, since not all the annotations have been verified with the musicologist and finished yet.

2.3 Analysis of the data

As a general overview to measure the quality of the dataset, it has been analysed it following the criteria adopted in [3] through five parameters: Purpose, Coverage, Completeness, Quality, and Reusability.

- *Purpose*: As the main purpose of the project is to support *Centonization* theory and discover relevant pattern of notes, having a full symbolic set of scores for all the *nawbas*, fullfills this criteria.
- *Coverage*: Having a variety and distributed number of scores per *nawba* as detailed in Table 1 fullfills this criteria.
- *Completeness*: Since every score is accompanied with a complete set of metadata, the completeness of the dataset is ensure.
- *Quality*: Since all the scores have been transcribed by the expert whose theory is desired to support, this could generate biased scores, trying to guide the data towards the consolidation of the theory. In addition, since the scores are

¹<https://github.com/MTG/pycompmusic>

²<https://github.com/MTG/arab-andalusian-music>

monophonic transcriptions of a whole choir and orchestra singing and playing at the same time, some lack of information could be left coming from the difficulty of transcribing such a complex performance. This criteria could be the one that can generate further discussion along the thesis.

- *Reusability*: Through Jupyter Notebooks, described in section 6, and *pycomp-music* API the reusability is easily guaranteed to obtain the data and reproduce the results of the thesis.

Chapter 3

Pattern analysis

Arab-Andalusic music's tradition research has not been detailed computationally explored in terms of patterns. The expert and performer Amin Chaachoo, has been the first one to theorize the patterns of the Moroccan tradition and that's why this needs to be taken as a preliminary and work in progress. Our main purpose is to support his theory from a computational point of view, and to offer himself and the community, useful resources to keep the interest on the tradition alive, discovering interesting parts on the research. Since this is an ongoing study, this has suppose a complex task, having to collaborate and communicate closely with the musicologist.

As explained before, our main concern is to help with the *Centonization* theory using symbolic pattern discovery algorithms. Once this is understood, another difficulty needs to be taken into account, which is the complexity of defining what a *cento* is. The concept of *cento* for Arab-Andalusian music is not clear enough, even for Amin Chaachoo, just being considered as any musical motive relevant and recurrently repeated in a specific mode. Apart from that, it is not possible to know anything else from this musical pattern and, therefore, results need to be accompanied with a lot of interpretation from someone formed in the music tradition. For the comparison of our results, the latest revision of Amin Chaachoo's list of *centos* from his last Arabic version of his book [5] has been used.

From the above explained, this work needs to be understood as a novel approach,

that merges MIR techniques and musicological knowledge still under construction. Indeed, this has been the most complex part of the work, being able to understand and consolidate a workflow that would fit a musicological theory, rather than the technical complexity.

3.1 Finding relevant patterns per *nawba*

3.1.1 Motivation

The main goal is to discover new symbolic patterns that can be relevant per *nawba*. For this purpose, after applying SIA to the symbolic dataset, these results and results from other approaches [16] are compared with the ground truth of pattern listed by Amin Chaachoo. The main research question is focused on how meaningful these patterns can be for the *nawba* and how different they are in compared with other approaches and the ground truth. On the other hand, a more qualitative approach is taken by the design of a visualization tool that allows the expert to listen to output patterns from the different algorithms. Therefore, the main motivation for this experiment is to try to resolve the hypothesis of whether the nature of the algorithm can provide different meaningful output that can help to musicologist experts. In addition, the impact and variability of discovered patterns that consider the rhythmic dimension is required.

3.1.2 Methodology

In contrast with the previous approach [16], that was using an string-based algorithm as TF-IDF, an important characteristic of SIA is to accept patterns that contain silences as valid ones, with the intention of obtaining more diverse patterns. The selected algorithm in this approach is *Structure Induction Algorithm (SIA)*. The implementation was taken from an existing Github repository¹ [33]. Our analysis is implemented in Python using the music21 library [34] for processing scores and adopt the same convention for naming notes and accidentals.

¹<https://github.com/andrebola/patterns-genres>

SIA is an algorithm that might return patterns with non-consecutive notes, so that rests in between can probably be of importance and this is the reason why it was decided to add them to this work. In contrast, it is dismissed to accept all octave information from our data, being justified by Chaachoo's decision when not adding it when defining the list of *centos*. Actually, when analysing the music scores, one can observe that the melodic line does not go beyond or under a whole octave, so it is reasonable to not add the octave information in our study. The set of scores used for this experiment is composed by the entire dataset described in Chapter 2.

3.1.3 Pre-processing

In order to give the proper input to the SIA algorithm, which is implemented in Java, all the scores were converted from the original format *musicxml* to MIDI format. Apart from that, not more pre-processing is required, since the input for the algorithm is a raw MIDI file.

3.1.4 SIA Algorithm

As explained in section 1.2.2, SIA bases its approach on representing the music as a multidimensional dataset. In this case, it will be represented each score as a 2-dimensional dataset, taking onset time and morphetic pitch as the two corresponding dimensions. So then, each note of the score is represented as a data point d_1, d_2, \dots, d_n in a dataset D . Therefore, a pattern P is translatable by a vector v in a dataset D if and only if P can be translated by v to give a pattern that is a subset of D . The most important concept that needs to be understood for this algorithm is the concept of *maximal translatable pattern* (*MTP*). Formally, the *MTP* for a vector v in a dataset D , denoted by $MTP(v, D)$, is the largest pattern translatable by v in D . Mathematically:

$$MTP(v, D) = \{d | d \in D \wedge d + v \in D\}. \quad (3.1)$$

In music, *MTPs* often correspond to the patterns involved in perceptually significant

repetitions. The main goal of SIA is to compute all the non-empty *MTPs* in a dataset. The *MTP* for a vector v in a dataset D is defined as non-empty, if and only if there exists at least two datapoints d_1 and d_2 in D such that $v = d_2 - d_1$. For more information about the algorithm you can refer to the main SIA publication [18].

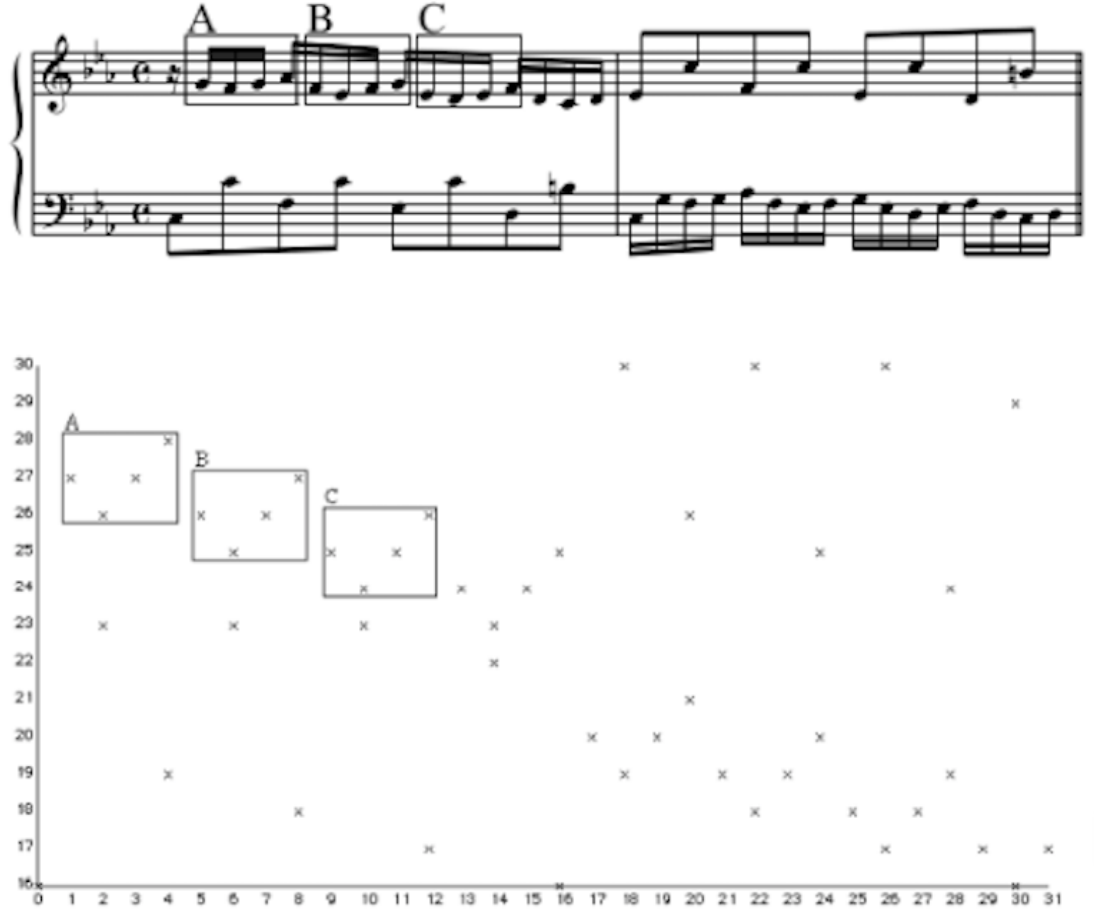


Figure 2: A two-dimensional orthogonal projection of the score excerpt onto the plane defined by the onset time and morphetic pitch dimensions taken from [18]

SIA offers two parameters that try to help with the high number of output patterns that the algorithm returns. These parameters are the *Density* and the *Compactness*. *Density* is defined as the relation between the number of notes and the temporal density and *Compactness* is defined as the relative length of the pattern with respect to the length of the whole piece of music. Depending on the tuning of these parameters, the output may vary.

In addition, it is desirable to add that it was decided to not use the improved *SIATEC* algorithm which considers transposition of patterns as equal. This decision is based on the definition of *centos* by Amin Chaachoo and the characteristic of them. The core of *Centonization's* theory lives in specific sequence of notes relevant per *nawba*. However, Amin considers as different patterns the ones that have different notes but the same intervals between the notes, this would suppose that, for example, a pattern with the notes CBAG would be considered as different from FEDC. This fact increments the complexity of the task because, a priori, there is not any common feature or characteristic that can cluster the patterns of a same *nawba* together. Amin justifies the lists of *centos* through his experience and knowledge about the music. This is what this work tries to solve and extract the most possible ideas of this theory from an objective and computational point of view.

3.1.5 Post-processing

As recurrently happening with pattern discovery algorithms, a filter of the large number of output patterns is needed in order to retrieve the most meaningful ones. Therefore, one of the useful parameters provided by SIA algorithm is used, which is the *Compactness*. This parameter is chosen to be the value that discovers the highest percentage of patterns containing *centos*.

Furthermore, patterns with length between 3 and 10 events are only considered (defining event as note or rest). This decision is based on the minimum and maximum length from the *centos* defined by Amin Chaachoo. Finally, only the first 10 most occurring patterns are considered per *nawba*, to restrict the number of helpful patterns that can be analysed by a human and based on the average number of *centos* per *nawba* in Chaachoo's list. This is done by counting the occurrences of every output pattern in all the scores of the same *nawba*, and then ordering and selecting the top 10 ones. In Fig 3, there is an example of the excerpt of a specific score with its discovered patterns. The patterns plot with the same colour correspond to the same output pattern. There might be the case that patterns with the same colour are different, this means that SIA discovered a set of non-consecutive notes

that were often repeated, but as in the example, they could have different notes in between the discovered pattern, which generates variations of the same motive. This fact gives the approach a new perspective and it is wished to check if these obtained patterns may have sense musicologically speaking.

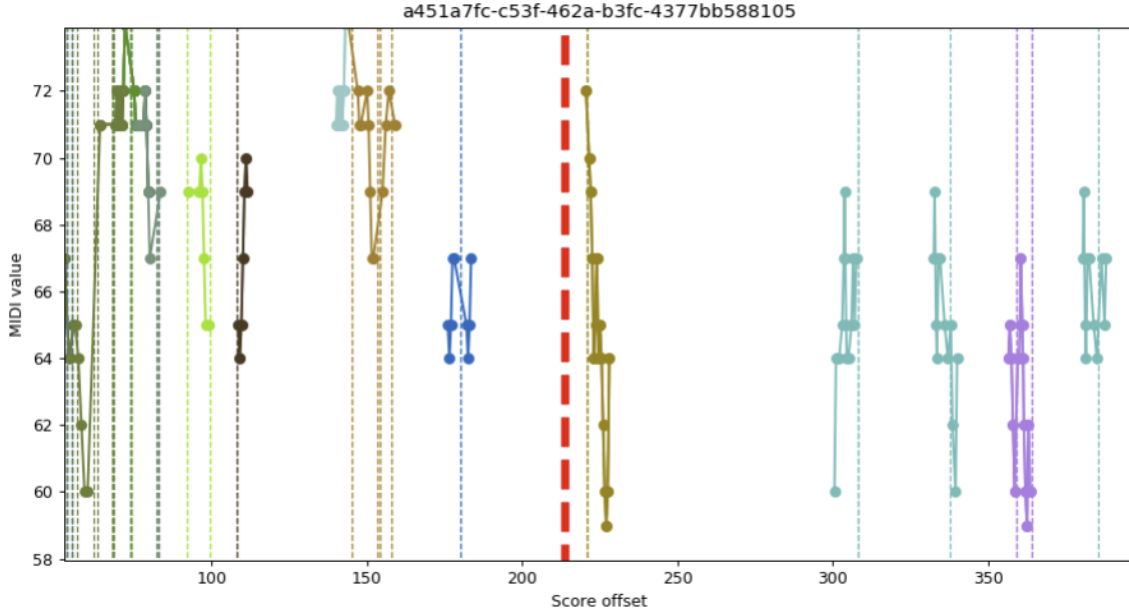


Figure 3: Excerpt of an example score with the pattern offset discovered, red width line marks the beginning of a new section

3.1.6 Results

In order to restrict our results and take some meaningful conclusions, number of relevant patterns discovered has been restricted to 10. This number is set taking the average number of centos that Amin Chaachoo defines per *nawba*. It is difficult to show clear results on how the new discovered patterns can support *centonization's* theory. However, two valuable parameters has been extracted which can be beneficial and help the support of this theory.

From all our output patterns, it's been observed that **54.55%** of them contain at least one of the centos defined in the required *nawba*. This might suppose that, at least, more than the half of the patterns that SIA is discovering are *centos* or variations of them. The rest of the patterns should be studied one by one, in order to indentify if they suppose new potential *centos* or just modifications of the known

ones. In any case, this result seems to confirm *Centonization* and show that *centos* are recurrent patterns in this music tradition.

In addition, the symbolic similarity measure [25] described in section 1.2.3 has also been applied to these results. This has shown that **42.73%** of the output patterns perform to be similar as the *centos* of the same *nawba* defined by Amin. This fact reinforces again the *Centonization's* theory, but leave a lot of room for improvement in terms of the similarity measure.

As observed, these parameters may help to interpret better the obtained results. However, they don't show evidences to extract clear conclusions. Another evaluation of the results is needed to assess the results and to be able to compare with TF-IDF algorithm, which is the other only algorithm that has been used for the same purpose.

Chapter 4

Evaluation

As explained, due to the nature of the task, an exhaustive evaluation of the results is needed. This evaluation should be both quantitative and qualitative, in order to make some sense of the results from the pattern recognition algorithm and to get the musicological sense of them as well.

4.1 Automatic Evaluation

For the automatic evaluation, firstly, the same process as the one from [16] will be taken, this will allow the comparison of the results with another algorithm. Then, a new way of evaluation through symbolic features is proposed.

4.1.1 Nawba Classifier with pattern occurrences

As a first measure to evaluate the results, a *nawba* classifier model is trained using as features the occurrences of each output pattern in every score of the dataset. Through this classifier it is possible to evaluate, for each algorithm, how relevant the output patterns are to characterize the *nawba*. For this purpose, a simple logistic regression model is used to build the *nawba classifier*, using 60/40 part of the dataset for train/split. In this case, 110 features (10 per *nawba*) were used. The accuracy of the *nawba* classification of this model may give an idea on how important these

patterns are related to the task of *nawba* prediction and thus, how meaningful they are. This may result in a good way of comparing with the pre-defined centos, since Amin Chaachoo elaborated the list of centos claiming that they are representative of the *ṭāb'* or the *nawba*. The model was built for the three pattern categories that has been describing: Centos, TF-IDF and SIA. It was applied to the corpus of 149 scores, using a 60/40 train/test split. The obtained accuracy for each model can be seen in Table 3.

Pattern Set	μ	σ
Chaachoo	70.80	4.90
TF-IDF	69.71	5.88
SIA	65.27	5.58

Table 3: Bootstrapped accuracy (n=100) when classifying *nawba* using three pattern categories

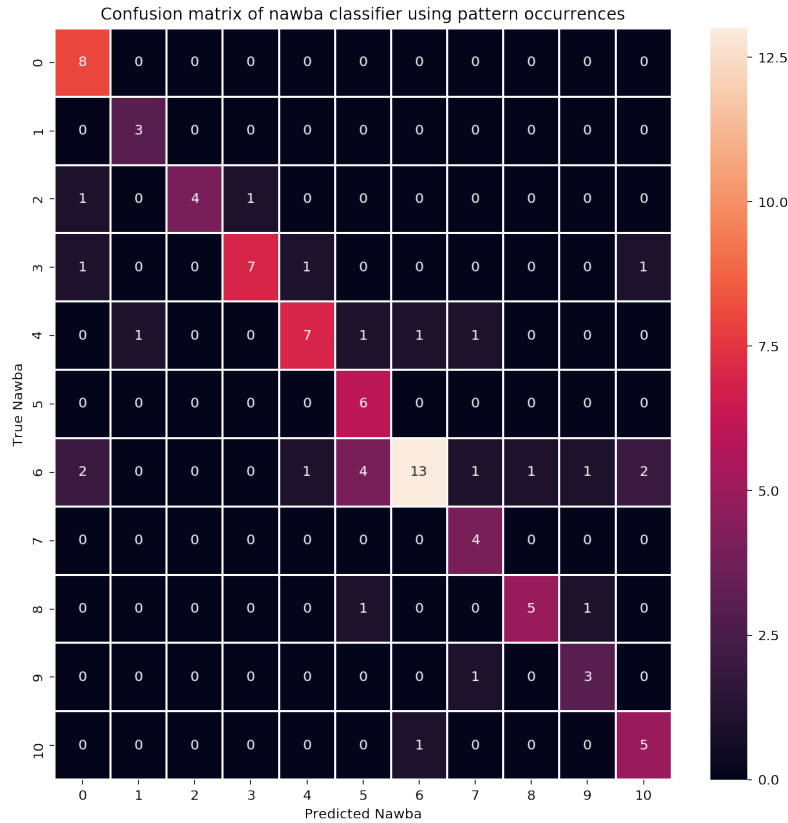


Figure 4: Confusion matrix for the *nawba* classifier using SIA pattern occurrences

The accuracy obtained for the three models is very similar. As expected, and also explained in the work from Nuttall et al. [16], the model that uses the centos's

occurrences as features is the one that gets the highest accuracy, not being far from the TF-IDF model. However, the model that was using the SIA patterns doesn't perform that well. This can be explained by different factors: SIA algorithm is run at score level, this process is different to TF-IDF that computes its output patterns globally per *nawba*. Furthermore, SIA is taking into account the rhythmic dimension of the patterns, and this results in more specific patterns that can be less relevant for the *nawba* level. Nevertheless, this result can bring to think that SIA algorithm is still producing patterns almost as important as the pre-defined centos, which is a valuable result. In Figure 4, the confusion matrix of the *nawba* classifier is shown, this matrix explains that there is a consistency in the classification. There exists a few false positives that would require a more exhaustive analysis of the patterns of that *nawba*, but this is not the main goal of this section and need to be studied more in detail in a new work.

4.1.2 Nawba classifier with pattern features

As a second step for evaluate the results, a similar *nawba* classifier model was trained. For this case, the features of the model were a set of symbolic musical extracted from the patterns, see Table 4. The features are selected and designed to catch different characteristic of the music as melody or rhythm. Using this classifier, it is discovered how important the music characteristic of the patterns are in terms of characterizing the *nawba*.

This set of features were chosen to get the main components of the *nawba* dimension. Therefore, there has been included relevant features to obtain the melodic characteristic of a *nawba* through the patterns that are part of it, as the predominant note or the persistence note, with getting the most repeated note in the pattern or the last one. Other important essence of a *nawba* are the notes of the scale, hence, features related to patterns with accidentals or with the different intervals contained in the patterns were added. A few more features in terms of rhythm were used, even though, normally the rhythmic dimension is not a predominant component of a *nawba*, seeing that tempo and rhythmic patterns are independent of the mode.

Feature name	Explanation
average_rest_length	Average duration of the rests contained in the pattern in quarter length
number_notes	Number of notes contained in the pattern
average_note_length	Average duration of the notes contained in the pattern in quarter length
most_repeated_note	MIDI value of the most repeated note contained in the pattern
first_note	MIDI value of the first note of the pattern
last_note	MIDI value of the last note of the pattern
longest_note	MIDI value of the note with more duration in the pattern
range_rhythmic	Difference between the longest and the shortest note of the pattern in quarter length
mean_midi_value	Average MIDI value of the notes contained in the pattern
contain_b	Bool feature to check if the pattern contains a flat (b)
contain_x	Bool feature to check if the pattern contains a sharp (#)
note_density	Number of notes per quarter length
is_continuous	Bool feature to check if the pattern contains correlative notes
contains_dot	Bool feature to check if the pattern contains any note with a dot
over_first_octave	Bool feature to check if the patterns goes over the first octave C4
under_first_octave	Bool feature to check if the patterns goes over the first octave C3
direction	Direction of the pattern (1 is ascendent, -1 is descendent, 0 is flat)
int_first_last_note	Interval between the first and last note of the pattern in semitones
int_two_last_notes	Interval between the last two notes of the pattern in semitones
most_repeat_int	Most repeated interval in the pattern in semitones
number_M_intervals	Number of major intervals contained in the pattern
number_m_intervals	Number of minor intervals contained in the pattern

Table 4: List of symbolic features used in the nawba prediction model

For this model, the symbolic feature values with the number of occurrences of every pattern in the score has been averaged. This allowed to represent each score as a 1-dimension array of length 21, which is the number of used features. The general workflow for building this logistic regression model is described in Figure 5.

This model was applied to the same corpus of 149 scores, using a 60/40 train/test

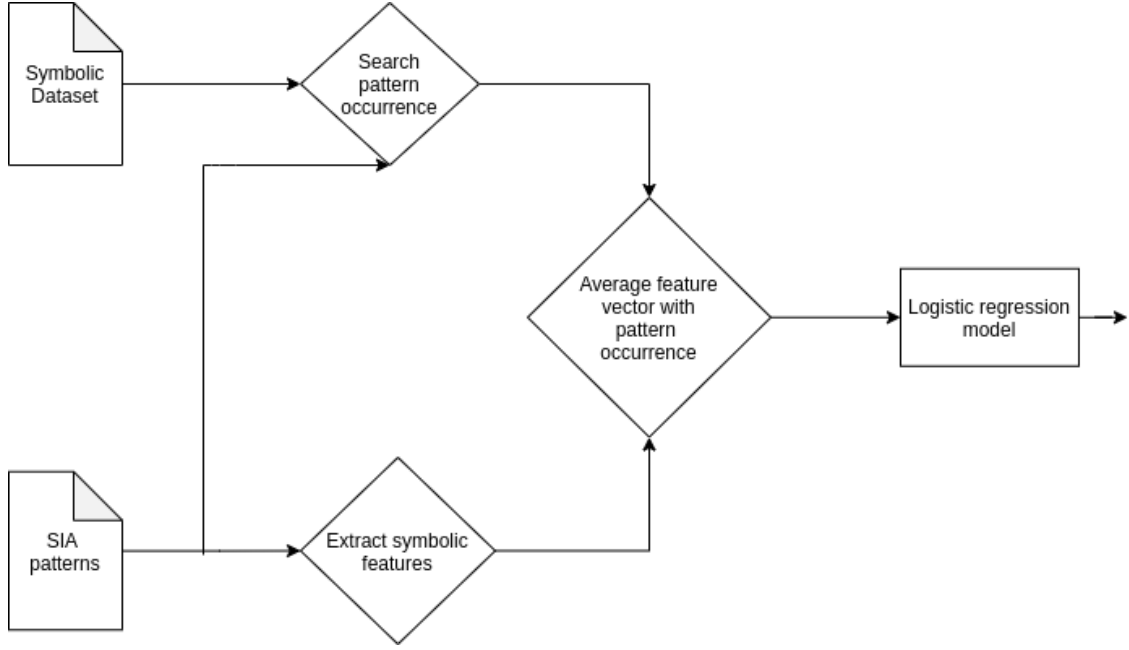


Figure 5: Workflow for building logistic model with features

split as well. The obtained accuracy was $\mu = 29.6, \sigma = 5.4$. This deduces that symbolic features are not as relevant as occurrences, in order to predict the *nawba* of a score. This result may serve to think whether all the selected features are needed, but it shouldn't be considered as a bad result, since conclusions can be taken from it.

In order to understand the impact of each feature on the model, an analysis of the importance of each feature can be conducted. This is made through another logistic regression model and retrieving the property that contains the coefficients found for each input variable. These coefficients can provide the basis for a crude feature importance score. This assumes that the input variables have the same scale or have been scaled prior to fitting the model. The result of this analysis can be seen in Figure 6.

Some ideas can be extracted from Figure 6. As explained before, the rhythmic features do not play a role in the *nawba* classification. In contrast, the features related to melody and important notes of the patterns are the ones that achieve a higher importance score. After discussing about this issue with the expert Amin Chaachoo, as described in section 4.2, it can be affirmed that this fact is corroborated

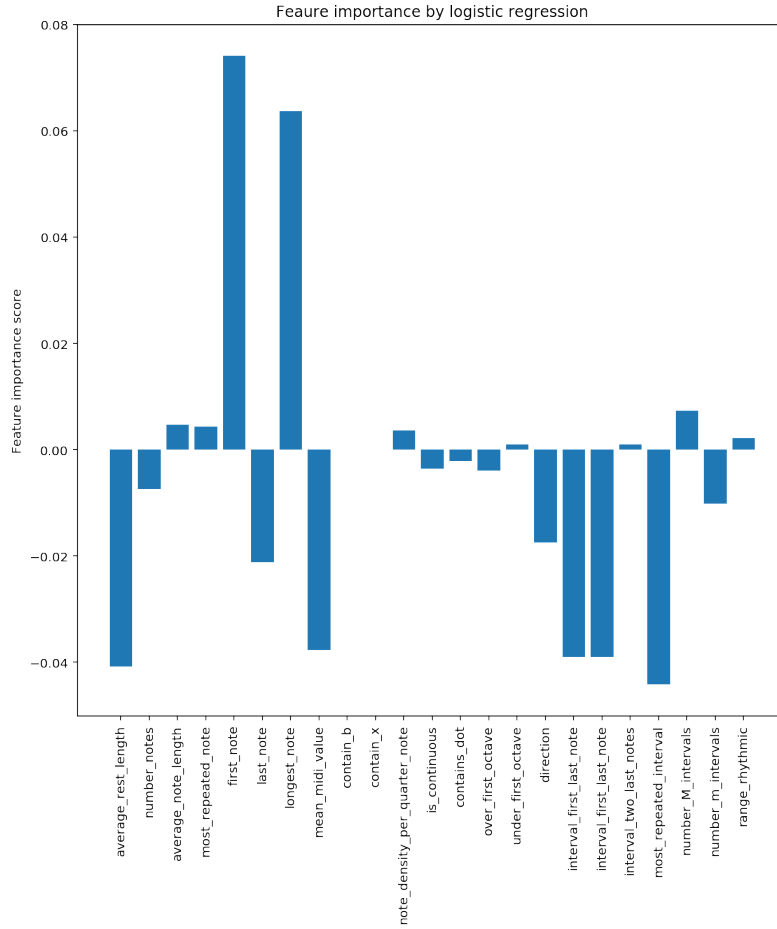


Figure 6: Feature importance for the logistic regression models that uses symbolic features

by him. Being the fundamental note and the persistence note, one of the most important properties of a *nawba*.

4.1.3 Nawba Classifier with pattern occurrences and selected features

In order to finalize with the automatic evaluations of the results, and based on the feature importance described in last section, a final a new *nawba* classifier model is designed. This model uses the features from the models described in sections 4.1.1 and 4.1.2 in the same classifier. This means, uses the SIA pattern occurrences as features but also uses a set of symbolic features. This symbolic features are part of a subset of the ones described in Table 4. Only features related to the melody are cho-

sen: `most_repeated_note`, `first_note`, `last_note`, `longest_note`, `mean_midi_value`, `contains_b` and `contains_x`. The goal of this classifier is to check whether the symbolic features can improve the occurrence classifier, so that this would mean that the musical characteristic of the patterns of the same *nawba* share some similarities. This could give some hints to introduce some improvements on the symbolic pattern discovery algorithm which would include some melodic feature in the implementation.

Same process was followed to make the *nawba* classifier. This time, the accuracy increased notably with: $\mu = 70.14, \sigma = 6.64$, see Table 5. This result ends up explaining the high importance that melodic features have in the patterns, in terms of the *nawba*. Additionally, this accuracy is the same as the one achieved with the occurrences of the *centos*. This is one very interesting fact, that means that the obtained SIA patterns need to be taken in consideration when defining the patterns of a *nawba*, and leave a big amount of material for the study of musicologists in the field. In Figure 7, the corresponding confusion matrix is shown, displaying that the classification is still consistent, having more true positives for this scenario. This matrix reinforce the idea of the necessity of adding musical constraint to the symbolic algorithm to successfully discover more characteristic pattern for the *nawba*.

Pattern Set	μ	σ
Pattern occurrence	65.27	5.58
Pattern occurrence + features	70.14	6.64

Table 5: Bootstrapped accuracy (n=100) when classifying *nawba* using pattern occurrences and symbolic features

4.2 Qualitative evaluation

A visualization tool was designed in order to evaluate the results from a musicological point of view and with an user-friendly interface. This tool is aimed at any Arab-Andalusian researcher who would like to have a general glance to a large range of patterns and their relationships while listening to them.

The main motivation for the creation of the tool was the idea of having a summary

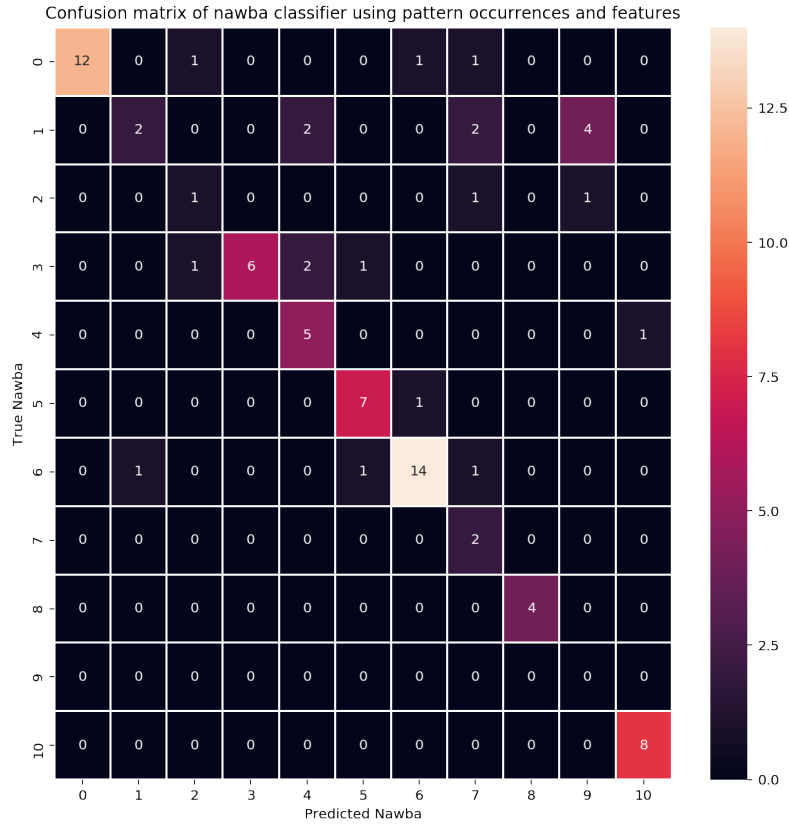


Figure 7: Confusion matrix for the *nawba* classifier using SIA pattern occurrences and selected symbolic features

of the high number of output patterns that symbolic pattern discovery algorithms normally give. Thus, results are manageable by a human and can be evaluated in a more careful way rather than diving in an extensive number of patterns, being complex to analyse and to listen to all of them. In addition, while adding the dimension of similarity in the tool, experts can compare between different kind of patterns and check plausibly the main differences they have.

Description of the visualization tool

The main view of the visualization tool is divided in three blocks with different purposes, that can be visible in Figure 8: parameter selector, main graph and score selector.

As first, a general explanation on how to use the interface is given, in order to guide the user to make a good use of the GUI. After it, the selector shows the different

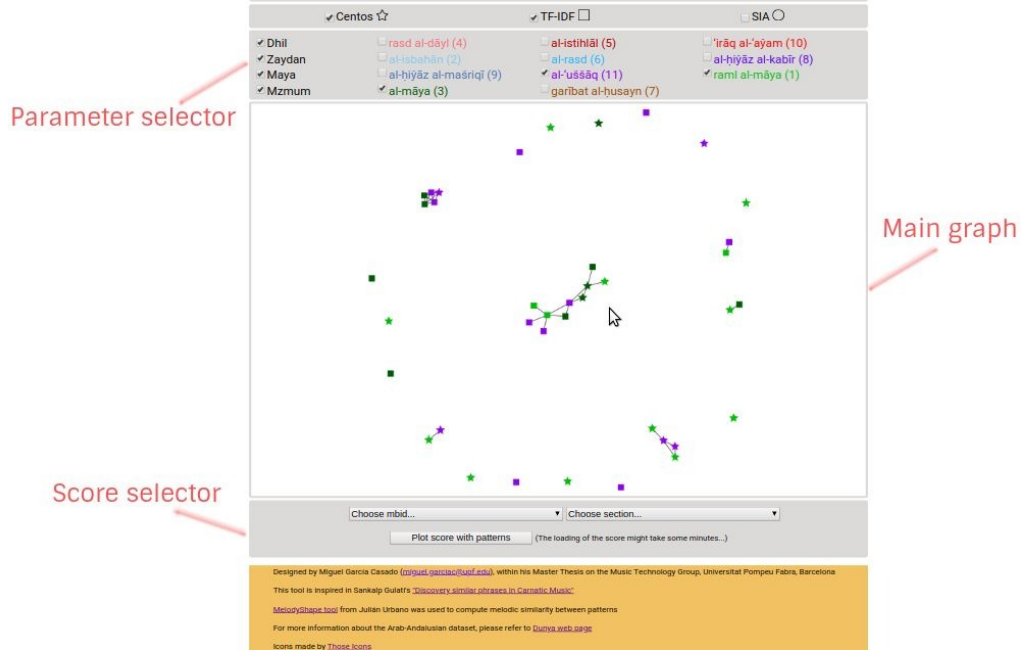


Figure 8: Example of the main parts of the visualization tool

parameters that can be chosen to make a new plot in the graph. These parameters are divided in three categories or kind of filters.

- **Pattern category:** Related to the group of patterns that have been discovered with each presented algorithm and the ground truth of *centos* defined by the expert.
 - *Centos*: Relevant patterns proposed by Amin Chaachoo representative of a *nawba*. They are represented with a star shape in the graph.
 - *TF-IDF*: Patterns discovered using the Term frequency – Inverse document frequency algorithm. They are represented with a square shape in the graph.
 - *SIA*: Patterns discovered using the Structure Induction Algorithm. They are represented with a circle shape in the graph.
- **Nawba family:** Related to the *nawba* grouping based on the well known *nawba* tree in [2]. Each family is represented with a range of hues, with the intention of visually separate the *nawabāt* of the same family in the graph

[35]. This separation comes from the *nawba* tree concept, example that can be visualize in Fig 9.

- **Nawba:** Patterns discovered or contained in the different *nawabāt*. Each *nawba* has its own color in the graph.



Figure 9: Nawba tree representation taken from [35]

Once the patterns are represented, the user should be able to explore and analyse the selection. Through the symbolic similarity algorithm, similar patterns have been linked between them, being able to see all the possible relations between the selected ones. The graph allows to zoom in and out, to better understand the different pattern clusters built in the plot. If you click on each pattern, you can listen to them. Furthermore, when hovering over a pattern, the user can see a number linked to the pattern, which will allow to recognise it when plotting the score.

At the bottom of the tool, the user can find the score selector, that contains two

The main purpose of the tool is to evaluate how relevant are the found patterns for both SIA and TF-IDF algorithm. Since this thesis is new and beginning work, first hypothesis is that the output patterns are not proposing anything new and could be variations of already known patterns, e.g. *centos*. To evaluate this, the visualization tool has been showed to the Arab-Andalusian expert Amin Chaachoo. He was asked to interact with the tool, and answer to some questions. As part of this interview, Amin Chaachoo was able to evaluate the quality of the results and analyse more in detail each pattern one by one to finally corroborate that this work was able to support *Centonization* theory, which was the main purpose.

During this analysis, the musicologist could give some value to this work. He claimed that each discovered pattern needed to be interpreted independently, and each case should be studied separately. Even though, almost all the discovered patterns were *centos* or variations of them, he said that "*this work would help for the addition of new patterns to Centonization theory that could not be found manually*". For the specific case of SIA patterns, Amin Chaachoo deduced that these patterns were catching a lot of half phrases, which would be solved by adding some constraint to the algorithm to separate pattern by the tonic of the *nawba*. However, SIA output patterns were also emphasizing a lot the importance of the persistence note of each *nawba*, which is something interesting to propose characteristic patterns of that mode. For the case of TF-IDF, Amin highlighted the power of differentiate between genres, that would be the characteristic patterns of the whole Arab-Andalusian music, and *centos*, patterns characteristic of the *nawba*. For this algorithm, the patterns seemed to be closer to the *centos* but probably they would not contain any new or unexplored material.

As last important part of the evaluation, Amin Chaachoo found out that the similarity measure between patterns was performing decently, considering it as a good first approach to see relations between *centos* and patterns. One of his main proposes was to add more musical conditions to the symbolic algorithms, in order to be working closer to the characteristic of Arab-Andalusian music. This would be done by adding some post-processing to the SIA algorithm, that would emphasize patterns

containing the persistence note of the *nawba* and that would separate patterns by its tonic, resulting in more complete patterns. Through this improvements, Amin Chaachoo considered that the work would be ideal for the further study of the music tradition.

As a general overview, it can be said that the visualization tool is a valuable resource for Arab-Andalusian researchers, what Amin pointed out, *"this tool can be very helpful for the better understanding of the music, being able to see similarities and differences between nawabāt through the patterns. Indeed, I would be keen to use this mean for my research as I consider that can be a very useful resource"*

As last, this evaluation tool allows to explore the Moroccan Arab-Andalusian repertoire further than for only the support of *Centonization* theory. The tool has been designed to enable to explore the concept of *nawba* and family of *nawabāt* through their most characteristic patterns. Furthermore, it is presented as one of the first piece of technology to visualize patterns of this tradition in the context of a score. In addition, it allows to study the dataset by sections, something that was not possible before. To sum up, apart from the main purpose of support the evaluation of the main line of research of this thesis, this visualization tool could be part of the beginning of a new computational ethnomusicology research in the Arab-Andalusian tradition.

Chapter 5

Discussion

Even though in every section of Chapter 3 a detailed explanation of the obtained results and a brief discussion has been presented, this chapter tries to wrap everything up and come with the most relevant points to be discussed. After that, some conclusions from the carried work are drawn in order to close the thesis.

5.1 Discussion

One of the first ideas that has been recurrently repeated along this thesis, is the question of what a *cento* is. The answer is not clear yet and this is what makes the proposed task very complex. Normally, symbolic pattern recognition algorithms are evaluated against a ground truth of annotated patterns, but, in this case, these patterns were not a closed and rigorous list. Therefore, the results need to be analysed from someone formed in the music tradition, being able to interpret every output pattern and give a general evaluation metric of the algorithm. This is what has been done with Amin Chaachoo. Under his point of view, the chosen algorithm helps to discover repeated patterns that almost always contain a *cento* or a variation of it. However, the necessity of adding musical constraints to the algorithm is highly needed. This is meant and planned for the future work of the thesis. Having some post-processing that would filter out the output patterns with some conditions related to persistence note and tonic note of the *nawba*, would give much more

value to the work, as explained by Amin. Indeed, the analysis through symbolic features seems to confirm this hypothesis, being the melodic symbolic features the core characteristics of the patterns regarding the definition of the *nawba*.

Another good topic to be discussed is the performance of the result in compare with the previous work. By applying SIA, at the end, lower *nawba* recognition accuracy was obtained, but this doesn't mean that this method should be discarded. Being TF-IDF and SIA completely different algorithms, it was desired to know if a geometric method algorithm could offer interesting results. This has confirmed that *centos* are indeed an important part of the music, and has left more space for the musicologist to continue supporting this theory, which was the main purpose. In addition, has given more new patterns to be studied as potential *centos*. This is an important difference with TF-IDF, which was not offering much more new material because they were not having into account the rhythmic dimension on the results. Furthermore, it would be useful as well to try out other different algorithms with the same data, to keep validating the *Centonization* theory and complement SIA and TF-IDF. One of the main limitations on this process is the nature of the heterophonic transcriptions. The scores of the dataset could have a lack of ornamentations coming from the difficulty of transcribing and capturing all the essence of the music. Furthermore, the same expert has been the one who did the transcription and the one that has evaluated the results, which could generate a bit of bias in the process.

As last, from the beginning of tackling this work, the necessity of manage a large number of patterns has been always felt. At the end, a logic way of navigating through them has been found by the design of the visualization tool. This tool, allows the user to learn more about this music tradition, being able to visualize, compare and listen to a large number of patterns at a glance. However, the tool has some limitations. In fact, not having a relationship between the pattern and its spatial position on the graph can make confusion on the interpretation of the data. Furthermore, while the multiple selection of the pattern categories is a helpful feature on the tool, it is responsibility of the user to choose a reasonable amount of *nawba* to obtain an interesting graph which doesn't have overlapping between the

patterns. This is planned to be solved in the future as well as the addition of new features that make a better understanding of the data.

5.2 Conclusions

There exists a few conclusions that can be extracted from this thesis. This work is considered the continuation of the previously started work started [16], being the first attempt to work with Arab-Andalusian music and its symbolic patterns from a computational point of view. All the obtained results lead to think that they consolidate the main research hypothesis. This new geometric family of symbolic pattern discovery algorithms is helpful for the *Centonization* theory, and all the results corroborate what Amin Chaachoo was claiming from his expert opinion. A new set of potential *centos* is proposed for the musicologist to be studied, which are, at least, as important as the proposed *centos*. This work has confirmed again the *Centonization* as a key concept in the study of the music tradition. Furthermore, from the study of symbolic features of the obtained patterns, it has given visibility to the melodic characteristic of the *nawba*, leaving apart the rhythmic dimension of the patterns, for the task of discovering new *centos*.

With the intention of having an easy way to evaluate the results, a visualization tool has been designed, which aims to be a valuable resource for the Arab-Andalusian research community, as well as for the Arab-Andalusian students. Depending on the use of the tool, it can have research or didactic goals. This software can serve as a good way of showing standard pattern discovery results for different research studies. The visualization tool has accomplished two main goals, which are the tool for the qualitative evaluation of the thesis results but, at the same time, it's also designed to be a tool used by musicologists in the future.

However, as one of the first computational analysis on Arab-Andalusian music, we hope to have contributed to the musicological theory around *Centonization* and hope that this approach may trigger new and interesting study on the topic, establishing the principles and basis for future and helpful study for musicological theories that

can contribute to a better understanding and preservation of the musical tradition.

5.3 Reproducibility

The main code of this project is available in GitHub with the description of each file. In the repository you will find the code to implement, train and validate the presented models, as well as the code to perform most of the steps of the evaluation in Chapter 4.

https://github.com/miguelgcasado/arab_andalusian_pattern_analysis

https://github.com/miguelgcasado/nawba_visualization

List of Figures

1	Diagram that summarizes the structure of a <i>nawba</i>	3
2	A two-dimensional orthogonal projection of the score excerpt onto the plane defined by the onset time and morphetic pitch dimensions taken from [18]	17
3	Excerpt of an example score with the pattern offset discovered, red width line marks the beginning of a new section	19
4	Confusion matrix for the <i>nawba</i> classifier using SIA pattern occurrences	22
5	Workflow for building logistic model with features	25
6	Feature importance for the logistic regression models that uses symbolic features	26
7	Confusion matrix for the <i>nawba</i> classifier using SIA pattern occurrences and selected symbolic features	28
8	Example of the main parts of the visualization tool	29
9	Nawba tree representation taken from [35]	30
10	Example of a score represented with pattern in colours	31

List of Tables

1	Distribution of Scores across <i>nawabāt</i>	10
2	Distribution of annotated sections	11
3	Bootstrapped accuracy (n=100) when classifying <i>nawba</i> using three pattern categories	22
4	List of symbolic features used in the <i>nawba</i> prediction model	24
5	Bootstrapped accuracy (n=100) when classifying <i>nawba</i> using pattern occurrences and symbolic features	27

Bibliography

- [1] Poché, C. *La música arábigo-andaluza (con CD)*. Músicas del mundo (Ediciones Akal, 1997). URL <https://books.google.es/books?id=MTleLv81KpUC>.
- [2] Chaachoo, A. *La musique hispano-arabe, al-Ala*. Univers musical (Editions L'Harmattan, 2016). URL <https://books.google.es/books?id=tMncCwAAQBAJ>.
- [3] Serra, X. Creating research corpora for the computational study of music: the case of the compmusic project. In *AES 53rd International Conference on Semantic Audio*, 1–9. AES (AES, 2014). URL <http://hdl.handle.net/10230/44221>.
- [4] Guettat, M. *La musique arabo-andalouse, l’empreinte du maghreb 560* (2000).
- [5] Chaachoo, A. *La musique hispano-arabe, al-Ala* (Amendis, 2019). URL <http://www.aminchaachoo.com/2019/07/21/books-space-%d9%81%d8%b6%d8%a7%d8%a1-%d8%a7%d9%84%d9%83%d8%aa%d8%a8/>.
- [6] Ferretti, P. & Agaësse, A. *Esthétique grégorienne ou Traité des formes musicales du chant grégorien. Volume I. Traduit de l’italien par Dom A. Agaësse* (Desclée, 1938). URL <https://books.google.es/books?id=P4wRnQEACAAJ>.
- [7] Chewand, G. & McKinnon, J. W. Centonization. *Oxford Music Online* (2001). URL <http://sare.upf.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsomo&AN=edsomo.05279&lang=es&site=eds-live>.

- [8] Apel, W. *Gregorian Chant*. A Midland book (Indiana University Press, 1958). URL https://books.google.es/books?id=tH_WAAAAAAAJ.
- [9] Reti, R. *The Thematic Process in Music* (Macmillan, 1978).
- [10] Janssen, B., Haas, W. D., Volk, A. & Kranenburg, P. V. Discovering repeated patterns in music: state of knowledge, challenges, perspectives. In *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research, Marseille, France*, vol. 20, 74 (2013).
- [11] Ren, I. Y., Koop, H. V. R., Volk, A. & Swierstra, W. In search of the consensus among musical pattern discovery algorithms. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 671–678 (2017).
- [12] Cambouropoulos, E. Musical parallelism and melodic segmentation: A computational approach. In *Music Perception*, 23(3):249–268 (2006).
- [13] Karydis, I., Nanopoulos, A. & Manolopoulos, Y. Finding maximum-length repeating patterns in music databases. In *Multimedia Tools and Applications*, vol. 32, 49–71 (2006).
- [14] Conklin, D. Discovery of distinctive patterns in music. In *Intell. Data Anal., Vol 14*, 547–554 (2010).
- [15] Hsu, J., Liu, C. & Chen, A. L. P. Discovering nontrivial repeating patterns in music data. In *IEEE Transactions Multimedia*, 311–325 (2001).
- [16] Nuttall, T., García-Casado, M., Núñez-Tarifa, V., Repetto, R. C. & Serra, X. Contributing to new musicological theories with computational methods: The case of centonization in arab-andalusian music. In *20th Conference of the International Society for Music Information Retrieval*, 223–228 (Delft, The Netherlands, 2019). URL <https://repositori.upf.edu/handle/10230/42789>.
- [17] Szeto, W. M. & Wong, M. H. A graph-theoretical approach for pattern matching in post-tonal music analysis. *Journal of New Music Research* **35**, 307–321 (2006). URL <https://doi.org/10.1080/09298210701535749>. <https://doi.org/10.1080/09298210701535749>.

- [18] Meredith, D., Lemström, K. & Wiggins, G. A. Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research* **31**, 321–345 (2002). URL <https://www.tandfonline.com/doi/abs/10.1076/jnmr.31.4.321.14162>. <https://www.tandfonline.com/doi/pdf/10.1076/jnmr.31.4.321.14162>.
- [19] Collins, T., Arzt, A., Flossman, S. & Widmer, G. Siarct-cfp: Improving precision and the discovery of inexact musical patterns in point-set representations. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, 549–554 (2013).
- [20] Ren, I. Y. Closed patterns in folk music and other genres. In *6th International Workshop on Folk Music Analysis* (2016). URL <https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1003&context=fema>.
- [21] Pesek, M., Leonardis, A. & Marolt, M. A compositional hierarchical model for music information retrieval (2014). URL <http://eprints.fri.uni-lj.si/id/eprint/4257>.
- [22] Janssen, B., van Kranenburg, P. & Volk, A. Finding occurrences of melodic segments in folk songs employing symbolic similarity measures. *Journal of New Music Research* **46**, 118–134 (2017). URL <https://doi.org/10.1080/09298215.2017.1316292>. <https://doi.org/10.1080/09298215.2017.1316292>.
- [23] Velardo, V., Vallati, M. & Jan, S. Symbolic melodic similarity: State of the art and future challenges. *Computer Music Journal* **40**, 70–83 (2016).
- [24] Aloupis, G. *et al.* Algorithms for computing geometric measures of melodic similarity. *Computer Music Journal* **30**, 67–76 (2006). URL <http://www.jstor.org/stable/4617944>.
- [25] Urbano, J. MIREX 2013 Symbolic Melodic Similarity: A Geometric Model supported with Hybrid Sequence Alignment. Tech. Rep., Music Information Retrieval Evaluation eXchange (2013).

- [26] Bohak, C. & Marolt, M. Calculating similarity of folk song variants with melody-based features. In *10th International Society for Music Information Retrieval Conference*, 597–602 (2009). URL <http://ismir2009.ismir.net/proceedings/PS4-4.pdf>.
- [27] Wolkowicz, J. & Kešelj, V. A text information retrieval approach to music information retrieval. In *9th International Society for Music Information Retrieval Conference* (2011). URL https://ismir2008.ismir.net/papers/ISMIR2008_220.pdf.
- [28] Frieler, K. & Müllensiefen, D. The simile algorithm for melodic similarity (2005).
- [29] Porter, A., Sordo, M. & Serra, X. Dunya: A system for browsing audio music collections exploiting cultural context. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 101–106 (Curitiba, Brazil, 2013). URL <http://hdl.handle.net/10230/32251>.
- [30] Sordo, M., Chaachoo, A. & Serra, X. Creating corpora for computational research in arab-andalusian music. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology, DLfM '14*, 1–3 (Association for Computing Machinery, New York, NY, USA, 2014). URL <https://doi.org/10.1145/2660168.2660182>.
- [31] Caro Repetto R, C. A. B. B. S. X., Pretto N. An open corpus for the computational research of arab-andalusian music (Association for Computing Machinery, New York, NY, USA, 2018).
- [32] Pretto N, C. R. R., Bozkurt B & X., S. Nawba recognition for arab-andalusian music using templates from music scores. In *Proceedings of the 15th Sound and Music Computing Conference (SMC2018)*, p. 394–9. (Cyprus University of Technology, 2018 Jul 4-7).
- [33] Ferraro, A. & Lemström, K. On large-scale genre classification in symbolically encoded music by automatic identification of repeating patterns. In *Proceedings*

- of the 5th International Conference on Digital Libraries for Musicology*, DLfM '18, 34–37 (Association for Computing Machinery, New York, NY, USA, 2018). URL <https://doi.org/10.1145/3273024.3273035>.
- [34] Cuthbert, M., Scott, C. & Ariza, C. Music21: A toolkit for computer-aided musicology and symbolic music data. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 637–642 (International Society for Music Information Retrieval, 2010).
- [35] Cortés, M. Nuevos datos para el estudio de la música en al-andalus de dos autores granadinos: As-sustarí e ibn al-jatib. *Música coral del sur* (1995). URL <http://www.centrodedocumentacionmusicaldeandalucia.es/export/sites/default/publicaciones/pdfs/estudio-musica-andalus-autores-granadinos.pdf>.

Appendix A

First Appendix

List of useful terms:

- *Nawba* (*nawabāt*): An homogeneous set of all the melodies of a particular tab', literally turn.
- *Tab'* (*tūbu'*): Musical mode, but also the emotional state produced by the melodies of this mode in performer and listener.
- *Mizan*: Rhythmic pattern in a nawba. Each nawba is structured as a sequence of five mizán
- *Sana'a* (*ṣanā'ī'*): Sung poem.
- *Cento*: Melodic pattern representative of a specific *nawba*
- *Msalia*: Instrumental prelude of undefined mode.
- *Tawsiya*: Instrumental prelude of defined mode.
- *Muassa'*: First movement of a *nawba*, slow tempo.
- *Mahzuz*: Second movement of a *nawba*, intermediate tempo.
- *Insiraf*: Third movement of a *nawba*, fast tempo.

- *TF-IDF*: Term frequency – Inverse document frequency
- *SIA*: Structure Induction Algorithm.
- *MBID*: Music Brainz ID, unique ID linked to a particular recording.