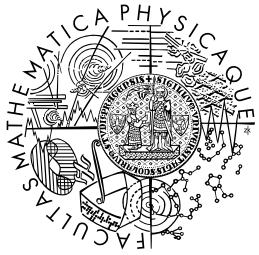

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Name Surname

Thesis title

Name of the department

Supervisor of the bachelor thesis: Supervisor's Name
Study programme: study programme
Study branch: study branch

Prague YEAR

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

Dedication.

Title: Thesis title

Author: Name Surname

Department: Name of the department

Supervisor: Supervisor's Name, department

Abstract: Abstract.

Keywords: key words

Contents

Introduction **2**

1 Related work **3**

2 Data **5**
2.1 CSV 5
2.2 Database fields 5
2.3 Volpiano 6

Conclusion **7**

Bibliography **8**

List of Figures **10**

List of Tables **11**

List of Abbreviations **12**

A Attachments **13**
A.1 First Attachment 13

Introduction

1. Related work

With the advent of computers, the field of Music Information Retrieval (MIR) emerged. Research in this interdisciplinary field focuses on extracting information from musical notation using computer science methods, such as signal processing or machine learning. Its applications vary widely, from recommender systems, to automatic audio transcription, to music generation. MIR encompasses all different kinds of music, regardless of their location, age, or function. Researchers have developed a multitude of software tools that facilitate music analysis, irrespective of what type of music it is. One of such toolkits is *music21* (Cuthbert and Ariza [2010]), a Python package able to encode musical notation as Python objects and perform analysis on large datasets.

The study of plainchant using computational methods has not been done extensively. The main research tool for musicologists in this field is the Cantus Index (Lacoste and Koláček). It is an online index of chants from several different chant databases, providing researchers with a common API for all of them.

Cornelissen et al. [2020a] developed *chant21*, a Python package able to convert two standard melodic notations, *volpiano* and *gabc* to a *music21* object, therefore making it easier to study Gregorian chant computationally. The data they used were scraped from Cantus database (Lacoste et al.) and GregoBase (Berten and contributors) and released as CantusCorpus and GregoBaseCorpus, respectively. Finally, they performed two case studies using the package. In the first one, they confirmed the melodic arch hypothesis (Huron [1996]), which had previously only been studied manually. Second, they analyzed the relation between differentiæ and antiphon openings (Shaw [2018]) and found that it differs accross modes.

Some of the computational research into plainchant has been centered on mode classification. Huron and Veltman [2006] used pitch class profiles to classify modes. They created a pitch-class distribution for each of the eight modes, and used these classes to classify previously unseen data. Cornelissen et al. [2020b] compared three approaches to mode classification: classical approach, which classifies chants based on the final pitch, range, and the initial pitch; profile approach, which was largely inspired by Huron and Veltman [2006]; and distributional approach, which focuses on the melodic aspect of mode. The authors chose various segmentations and representations of chants and used a tf-idf vector model to classify mode. The study found that we can accurately classify mode even when we discard all absolute pitch information, the melody contour contains enough information on its own.

A considerable amount of research has been done into the evaluation of melodic similarity, albeit not for Gregorian chant specifically. Wickland [2017] provides an overview of the methods. He mentions edit distance, Markov chains, and geometric measurements as the most widely used ones. Park et al. [2019] used an adapted edit distance metric to calculate the similarity of two melodic sequences by first calculating the similarity for all segments of each of the sequences and then scaling them by a weight function depending on the segment length, which yielded them what they call a multi-scale similarity stack. The overall similarity was obtained by averaging its values. Then they used the MSS stack to create a visualization that takes on the shape of a trapezoid that shows which segments

of two sequences are the most similar.

Bountouridis et al. [2017] argue that methods originally developed for bioinformatics have a great potential to be applied to music. They offer analogies for bioinformatics concepts found in musicology. For example, they liken DNA and proteins to melodic sequences, homologues (proteins that have the same ancestor) to song covers, evolution to oral transmission, etc. They claim that despite the similarities, MRI has not leveraged the full potential of bioinformatics methods. In their article, they focus on modelling melodic similarity using multiple-sequence alignment (MSA) algorithms, therefore not relying on heuristics, as opposed to previous works. Their results revealed that the MAFFT algorithm yields the best alignemnt, which can be attributed to the algorithm using gap-free segments as anchor points, therefore partitioning melodies into more meaningful segments than other algorithms.

2. Data

Our main source of data is the Cantus database (Lacoste et al.), one of the databases indexed in the Cantus Index. The database serves as a digital archive of chants, each entry containing information about its source, liturgical occasion, mode, and others. Work on the project started in the late 1980s, and to date, around 500,000 individual chants from approximately 150 manuscripts have been indexed. Each entry is transcribed manually and undergoes a thorough examination before publishing (Lacoste [2012]).

We are using a scraped version of the Cantus database released as Cantus-Corpus (**?**chant21)). Unlike the Cantus database which is continuously being updated and is therefore unsuitable for computational study, the corpus is versioned, therefore each version always contains the same data. We are using version 0.2 released in July 2020 which contains 497071 entries. The corpus is available for download in CSV format.

2.1 CSV

CSV is one of the most common formats for tabular data. The abbreviation stands for *comma-separated values*. As the name suggests, the format uses commas to separate columns (although other separators, such as a semicolon, can be used as well to allow for simpler parsing in case that the data frequently contains commas that would otherwise need to be escaped), while the individual rows are separated by a line break. The data is stored as plaintext, which makes it easily readable. Parsing CSV files becomes more complicated when the data contains column and row separators inside fields; in that case quotation marks or esape sign has to be used. There exist many well-designed parsers, one such parser is the Python module simply called *CSV*.

2.2 Database fields

The csv files in Cantus Corpus contain 21 fields (excluding its row number), of which we are only using a subset.

Each entry contains the chant's incipit, which is the first few words of the text. As chants do not have a title, incipit can substitute its role in contexts where one is needed.

The fields position, sequence, and folio represent the exact location of the original chant in a manuscript from which it was transcribed.

feastid represents the liturgical occasion when the chant was intended to be performed, or, in other words, a feast. Similarly, officeid represents the liturgical time of the day during which it was sung.

The text and melody of the chant are found in the fulltext and volpiano fields, respectively. Entries can contain both, either, or none of these fields.

2.3 Volpiano

The melodies in the volpiano fields are encoded as strings of alphanumeric characters and dashes. These can be rendered as musical notation using the volpiano font. Each character represents either a pitch, empty space, or other musical characters, such as a clef.

Volpiano was developed as a research tool optimized for databases and word processors. There are strict rules concerning the transcription, which leads to all volpiano-encoded melodies having a standardized format. Each transcription begins with a treble clef. Gaps between words are encoded as three dashes, while two dashes represent gaps between syllables (Helsen and Lacoste [2011]).

2.4 Data cleaning

Conclusion

Bibliography

Olivier Berten and contributors. Gregobase: A database of gregorian scores. URL https://gregobase.selapa.net/.

Dimitrios Bountouridis, Daniel G. Brown, Frans Wiering, and Remco C. Veltkamp. Melodic similarity and applications using biologically-inspired techniques. *Applied Sciences*, 7(12), 2017. ISSN 2076-3417. doi: 10.3390/app7121242. URL https://www.mdpi.com/2076-3417/7/12/1242.

Bas Cornelissen, Willem Zuidema, and John Ashley Burgoyne. Studying large plainchant corpora using chant21. In *7th International Conference on Digital Libraries for Musicology*, DLfM 2020, page 40–44, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450387606. doi: 10.1145/3424911.3425514. URL https://doi.org/10.1145/3424911.3425514.

Bas Cornelissen, Willen Zuidema, and John Ashley Burgoyne. Mode classification and natural units in plainchant. 2020b. URL https://program.ismir2020.org/poster_232.html.

Michael Scott Cuthbert and Christopher Ariza. music21: A toolkit for computer-aided musicology and symbolic music data, 2010.

Kate Helsen and Debra Lacoste. A report on the encoding of melodic incipits in the cantus database with the music font 'volpiano'. *Plainsong and Medieval Music*, 20(1):51–65, 2011. doi: 10.1017/S0961137110000197. URL https://doi.org/10.1017/S0961137110000197.

David Huron. The melodic arch in western folksongs. *Computing in Musicology*, pages 3–23, 1996.

David Huron and Joshua Veltman. A cognitive approach to medieval mode: Evidence for an historical antecedent to the major/minor system. *Empirical Musicology Review*, 1(1):33–55, 2006. URL https://doi.org/10.18061/1811/24072.

Debra Lacoste. The cantus database: Mining for medieval chant traditions. *Digital Medievalist*, 7, 2012. URL http://doi.org/10.16995/dm.42.

Debra Lacoste and Jan Koláček. Cantus index: Online catalog for mass and office chants. URL http://cantusindex.org/.

Debra Lacoste, Jan Koláček, Terence Bailey, and Ruth Steiner. A database for latin ecclesiastical chant - inventories of chant sources. URL https://cantus.uwaterloo.ca/.

Saebyul Park, Taegyun Kwon, Jongpil Lee, Jeounghoon Kim, and Juhan Nam. A cross-scape plot representation for visualizing symbolic melodic similarity. *Proceedings of the 20th ISMIR Conference*, pages 423–430, 2019.

Rebecca Shaw. Differentiae in the cantus manuscript database: Standardization and musicological application. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, DLfM '18, page 38–46, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450365222. doi: 10.1145/3273024.3273028. URL https://doi.org/10.1145/3273024.3273028.

David D. Wickland. Evaluating melodic similarity using pairwise sequence alignments and suffix trees. Master's thesis, The University of Guelph, 9 2017.

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment