Multimedia Processing Techniques for Retrieving, Extracting, and Accessing Musical Content

Techniken der Multimediaverarbeitung zur Suche, Extraktion und den Zugriff auf musikalische Inhalte

Dissertation

Der Technischen Fakultät

der Friedrich-Alexander-Universität Erlangen-Nürnberg

zur

Erlangung des Doktorgrades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

vorgelegt von

Stefan Balke

aus

Höxter

Als Dissertation genehmigt
von der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung:	21.03.2018
Vorsitzender des Promotionsorgans:	Prof. Dr.-Ing. Reinhard Lerch
1. Gutachter:	Prof. Dr. Meinard Müller
2. Gutachter:	Prof. Dr. Gerhard Widmer

Abstract

Music constitutes a challenging multimedia scenario. Besides music recordings, there exist a number of other media objects including symbolic music representations, video recordings, scanned sheet music, or textual metadata. Developing tools that allow users to retrieve information from different types of music-related data is central to the research area known as *Music Information Retrieval* (MIR). This requires techniques from various engineering fields such as digital signal processing, image processing, data management, and machine learning. In this thesis, we develop novel multimedia processing techniques and explore their capabilities and limitations within different complex music scenarios. The thesis consists of three main parts.

In the first part, we consider retrieval scenarios within a Western classical music setting. For example, given a short monophonic melodic theme in symbolic notation as a query, retrieve all corresponding documents in a collection of polyphonic music recordings. In another related retrieval scenario, we aim to link the score of musical themes, scanned from book pages, to their symbolic counterparts given in MIDI format. Both scenarios require mid-level feature representations derived from different media types, as well as robust retrieval techniques that can handle extraction errors and variations in the data.

The second part of this thesis deals with the extraction of musical parameters such as fundamental frequencies or musical pitches from audio recordings. In this context, a general goal is to reduce the variations in the degree of polyphony between monophonic queries and polyphonic music databases. In our computational approach, we propose a data-driven method based on *Deep Neural Networks* (DNNs) which aims at enhancing salient parts from jazz music recordings. As an example application, we employ the learned model in a retrieval scenario where we take a jazz solo transcription as a query to identify the corresponding music recording.

In the third part, we explore the potential of web-based user interfaces for researchers as well as music lovers. We present several prototypical interfaces that offer various functionalities for enabling access and navigation in musical content. Furthermore, these interfaces allow researchers to show their results in an interactive fashion and reduce technical barriers when cooperating with other researchers from related fields such as musicology.

Zusammenfassung

Musik stellt ein anspruchsvolles Multimediaszenario dar. Neben den Musikaufnahmen existieren eine Vielzahl von weiteren Medienobjekten (z. B. symbolisch kodierter Notentext, Videoaufnahmen, gescannte Notentextseiten und weitere textbasierte Metadaten). Die Entwicklung von Werkzeugen, die es einem Benutzer erlauben, unterschiedliche, musikalische Inhalte aufzufinden und darauf zuzugreifen, ist eine zentrale Aufgabenstellung im Bereich des *Music Information Retrieval* (MIR). Dies erfordert den Einsatz von Techniken aus unterschiedlichen Ingenieursfachrichtung, wie beispielsweise der digitalen Signal- und Bildverarbeitung, dem Datenmanagement oder dem maschinellen Lernen. In dieser Arbeit entwickeln wir neuartige Techniken der Multimediaverarbeitung und untersuchen das Leistungsvermögen als auch die Grenzen dieser Techniken in verschiedenen, komplexen Musikszenarien. Die Arbeit besteht aus drei Teilen.

Der erste Teil der Arbeit beschäftigt sich mit der automatisierten Musiksuche im Kontext von klassischer Musik. Ein Ziel besteht darin, die zu einem kurzen, monophonen Musikthema zugehörigen Musikaufnahmen in einem polyphonen Musikdatenbestand zu identifizieren. In einem ähnlichen Suchszenario verknüpfen wir Musikthemen mit den zugehörigen Buchseiten, die in gescannter Form vorliegen. Für beide Szenarien werden geeignete Merkmalsdarstellungen benötigt, die zum einen die unterschiedlichen Medientypen zusammenführen können und zum anderen robust gegen Extraktionsfehler und Variabilitäten in den Daten sind.

Der zweite Teil handelt von der Extraktion musikalischer Parameter aus den Audioaufnahmen (z. B. der Grundfrequenz oder Tonhöhe). Im Zuge dieser Arbeit besteht ein Ziel darin, die unterschiedlichen Polyphoniegrade monophoner Anfragen und polyphoner Musikdatenbestände anzugleichen. In unserem datengetriebenen Ansatz trainieren wir dazu Neuronale Netzwerke, die dominante Melodien in Jazzaufnahmen verstärken. In einer Beispielapplikation benutzen wir das trainierte Modell in einem Suchszenario, bei dem Jazzsolotranskription als Anfragen verwendet werden, um die dazugehörigen Musikaufnahmen zu identifizieren.

Im dritten Part untersuchen wir das Potenzial webbasierter Benutzerschnittstellen für die Verwendung von Wissenschaftlern sowie Musikliebhabern. Dafür stellen wir prototypische Schnittstellen vor, die eine Vielzahl an Funktionalitäten für den Zugriff auf und die Navigation in musikalischen Inhalten ermöglichen. Diese Benutzerschnittstellen erlauben es, wissenschaftliche Ergebnisse interaktiv darzustellen und zudem einen einfachen Zugriff zu ermöglichen, beispielsweise bei interdisziplinären Kooperationen mit den Musikwissenschaften.

Contents

Chapter 1

Introduction

Nowadays, anyone can record a musical performance at the press of a button. During the last decades, a tremendous amount of professionally produced audio and video recordings has been made available. Besides the actual music recordings, there exist a number of other media objects which can be associated to a recorded performance, such as symbolic music representations, scanned sheet music, or textual metadata. Tay Vaughan defines multimedia as a "*woven* combination of digitally manipulated text, photographs, graphic art, sound, animation, and video elements" [192, p. 1]. In this sense, music—with its many different facets—constitutes a very rich multimedia scenario.

A central research question in the area known as *Music Information Retrieval* (MIR) is to develop tools that create links between different, music-related media objects. Many approaches follow the so-called *query-by-example* paradigm to link such objects: given a fragment of a visual, symbolic, or acoustic music representation used as a query, the objective is to retrieve all documents from a music database which contain aspects and parts that are similar to the query [37, 130, 187]. A popular example for an audio-based retrieval system is *Shazam*, where, given an audio snippet of a song as a query, the objective is to find the *exact* matches in the music database. This task—also known as *audio fingerprinting*—is considered as basically solved. However, when the query is only given as a short, monophonic melody, whether hummed or in symbolic form, solving the task requires more flexible techniques. The system needs to deal with a number of variations including tempo and tuning deviations, key transpositions, or differences in the degree of polyphony between the query and the audio recordings in the reference database. In particular, the differences in the degree of polyphony between the monophonic query and the polyphonic music mixtures highly increase the task's complexity.

In this context, an important research task is the extraction of the predominant melody from music recordings. Traditional approaches often make use of manually designed features to enhance the predominant voice in a music signal [164]. With the advent of data-driven methods,

many MIR tasks are now approached using *Deep Neural Networks* (DNNs). In the case of predominant melody extraction, DNNs can be used to infer the relevant properties directly from the data, without the need for handcrafted features. Such a trained DNN model can serve as a preprocessing step in a retrieval scenario as described above to cope with the differences in the degree of polyphony between query and database documents.

As a central application, retrieval techniques can be used to establish links between related media objects. This is of particular interest in scenarios with distributed data sources such as the Internet. In the case of musical content, a wealth of metadata (e. g., artist biography on Wikipedia) is publicly available—sometimes even the music recordings themselves. Bringing all these resources together opens up new possibilities for users to benefit from the richness of the available data and can improve the listening experience. For example, web-based user interfaces allow unified access to a variety of media objects from several resources. Furthermore, such interfaces allow researchers to present results in an interactive fashion and reduce technical barriers when cooperating with other researchers from related fields.

1.1 Structure

The thesis is divided into three main parts, as shown in Figure 1.1. In Part I, we consider the book "A Dictionary of Musical Themes" [16] as an example for a complex music retrieval scenario. The book contains roughly 10 000 musical themes that are about 4 bars long. The themes are supposed to represent memorable excerpts from famous musical works, e. g., the "Fate-motif" from the beginning of Beethoven's 5th symphony. Additionally, there is a website with symbolic MIDI (Musical Instrument Digital Interface) versions of these themes and a large music collection with Western classical music that contains music recordings for many of those 10 000 musical themes. In a first scenario, we use the MIDI version of the monophonic musical themes to retrieve corresponding music recordings from this large music collection. Here, one main challenge stems from the difference in the degree of polyphony between the monophonic query and the polyphonic sound mixtures contained in the music recordings. Furthermore, the themes only have a duration of a few measures and may deviate in tuning and tempo from the music recordings in the collection. In a related retrieval scenario, we combine the symbolic MIDI representations of the themes with scanned images of the book pages. This constitutes a challenging cross-modal retrieval scenario that requires additional steps such as image segmentation and *Optical Music Recognition* (OMR). Both scenarios require cross-modal feature representations that can be derived from different media types as well as robust retrieval techniques that can tolerate errors introduced by the conversion from scanned sheet music to a symbolic music representation.

In Part II, we present data-driven approaches for predominant melody enhancement. Many conventional systems approach such retrieval tasks by first extracting the predominant melody

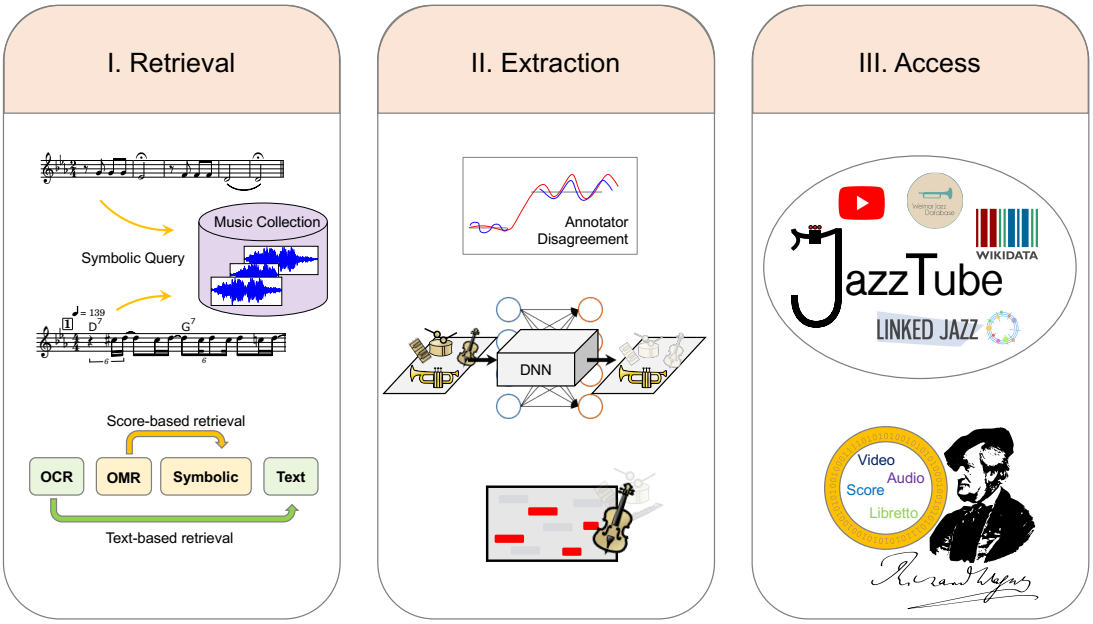


Figure 1.1: Structure of the thesis. The thesis is divided into the three main parts: *Retrieval*, *Extraction*, and *Access*. In each part, we develop novel processing techniques and explore their capabilities and limitations within different complex music scenarios.

from the recording, then quantizing the extracted trajectory to musical pitches, and finally comparing the resulting pitch sequence to the monophonic query [166]. In a first step, we evaluate current state-of-the-art algorithms and their performance within a jazz music scenario. Typically, the output of an algorithm is compared to a single reference annotation. However, annotating music recordings almost always introduces subjective decisions made by the annotator, i. e., different annotators create different annotations which may be equally valid. In our evaluation approach, we analyze the influence of this annotator disagreement and its implications on the results. Then, in our own approach to extract the predominant melody, we propose a data-driven method. In particular, we use DNNs to extract and enhance salient parts from jazz music recordings. The learned transformation can be considered as a kind of solo voice enhancement. We then apply the learned model in a retrieval scenario where we take a jazz solo transcription as a query to identify the corresponding music recording. Since DNN-based methods require adjusting a large number of parameters, we explore the influence of different hyperparameter settings in a jazz solo and walking bass transcription scenario.

In Part III, we indicate the potential of web-based user interfaces which allow easy access to linked media objects. In a first scenario, we consider a research corpus called the *Weimar Jazz Database* (WJD) as an example. As is the case with all research corpora that utilize commercial music recordings, the annotations can only be released without the audio data, and are therefore not fully usable by other researchers. However, there are publicly available videos on the Internet, featuring many of the musical pieces contained in the WJD. In our approach, we link the

WJD's music recordings with versions that are available on video platforms such as YouTube. Furthermore, we introduce a user interface that allows users to explore and interact with the annotations contained in the WJD. Finally, we integrate additional metadata from the *Semantic Web* [19], including discographic metadata, artist biographies, and artist relationships. In another multimedia scenario, we consider Richard Wagner's opera *Die Walküre* as a complex multimedia scenario. Based on suitable data structures and multimedia processing techniques, we develop a cross-modal user interface that allows a music lover to access a particular video recording (e. g., on Youtube), which is automatically aligned to an available sheet music representation, enriched with the libretto, and to other available resources about the musical work or the composer. With these example scenarios, we illustrate the potential of modern web-based technologies to share datasets and offer scientists in the digital humanities novel ways to access and interact with digitized multimedia content.

1.2 Contributions

The main contributions of this thesis can be summarized as follows.

- Systematic study of a cross-modal retrieval application within a Western classical music scenario. A special focus is set on measuring the influence that different factors—such as tempo deviations or the difference in the degree of polyphony between query and database documents—have on the retrieval performance (Chapter 2).
- A late-fusion approach that incorporates the results from text-based and score-based retrieval approaches to improve the retrieval performance (Chapter 3).
- A systematic study on the influence of annotator disagreement in fundamental frequency annotations (Chapter 4).
- A novel salience representation based on a data-driven approach (Chapter 5).
- An innovative approach to link scientific music databases including metadata, transcriptions and further annotations to the corresponding audio recordings that are publicly available on the Internet (Chapter 6).
- Modeling an opera as a multimedia scenario including a web-based user interface for cross-model and interactive access to the media objects (Chapter 7).
- A detailed literature overview of DNN-based approaches for central MIR tasks with a special focus on input representations and DNN architectures (Appendix B).

1.3 Main Publications

The main contributions of this thesis are based on the following publications, which were presented at conferences in the field of audio signal processing and Music Information Retrieval.

[8] Stefan Balke, Sanu Pulimootil Achankunju, and Meinard Müller. Matching musical themes based on noisy OCR and OMR input. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 703–707, Brisbane, Australia, 2015.

[11] Stefan Balke, Vlora Arifi-Müller, Lukas Lamprecht, and Meinard Müller. Retrieving audio recordings using musical themes. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 281–285, Shanghai, China, 2016.

[4] Stefan Balke and Meinard Müller. A graphical user interface for understanding audio retrieval results. In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Malága, Spain, 2015.

[13] Stefan Balke, Christian Dittmar, Jakob Abeßer, and Meinard Müller. Data-driven solo voice enhancement for jazz music retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 196–200, 2017.

[1] Jakob Abeßer, Stefan Balke, Klaus Frieler, Martin Pfleiderer, and Meinard Müller. Deep learning for jazz walking bass transcription. In *Proceedings of the AES International Conference on Semantic Audio*, pages 202–209, 2017.

[10] Stefan Balke, Jakob Abeßer, Jonathan Driedger, Christian Dittmar, and Meinard Müller. Towards evaluating multiple predominant melody annotations in jazz recordings. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 246–252, New York City, USA, 2016.

[15] Stefan Balke, Christian Dittmar, Jakob Abeßer, Klaus Frieler, Martin Pfleiderer, and Meinard Müller. Bridging the Gap: Enriching YouTube videos with jazz music annotations. *submitted: Frontiers in Digital Humanities*, 2018.

[5] Stefan Balke and Meinard Müller. JazzTube: Linking the Weimar Jazz Database with YouTube. In Martin Pfleiderer, Klaus Frieler, Jakob Abeßer, Wolf-Georg Zaddach, and Benjamin Burkhart, editors, *Inside the Jazzomat. New perspectives for jazz research*, pages 315–317. Schott Campus, Mainz, Germany, 2017.

[14] Stefan Balke, Manuel Hiemer, Peter Schwab, Vlora Arifi-Müller, Klaus Meyer-Wegener, and Meinard Müller. Die Oper als Multimedia Szenario: Wagners Walküren gehen online. In *Proceedings of the GI Jahrestagung*, pages 75–86, 2017. doi: 10.18420/in2017_04.

1.4 Additional Publications

The following publications are also related to music signal processing, but are not considered in this thesis.

[49] Christian Dittmar, Martin Pfleiderer, Stefan Balke, and Meinard Müller. A swingogram representation for tracking micro-rhythmic variation in jazz performances. *Journal of New Music Research*, 0(0):1–17, 2017. doi: 10.1080/09298215.2017.1367405.

[12] Stefan Balke, Paul Bießmann, Sebastian Trump, and Meinard Müller. Konzeption und Umsetzung web-basierter Werkzeuge für das Erlernen von Jazz-Piano. In *Proceedings of the GI Jahrestagung*, pages 61–73, 2017. doi: 10.18420/in2017_03.

[197] Nils Werner, Stefan Balke, Fabian-Robert Stöter, Meinard Müller, and Bernd Edler. trackswitch.js: A versatile web-based audio player for presenting scientific results. In *Proceedings of the Web Audio Conference (WAC)*, 2017.

[55] Jonathan Driedger, Stefan Balke, Sebastian Ewert, and Meinard Müller. Template-based vibrato analysis of music signals. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 239–245, New York City, USA, 2016.

[9] Stefan Balke, Lukas Lamprecht, Vlora Arifi-Müller, Thomas Prätzlich, and Meinard Müller. Automatisierte Identifikation von Audioaufnahmen anhand symbolisch codierter musikalischer Themen. In *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)*, Nürnberg, Germany, 2015.

1.5 Acknowledgments

Looking back at the past four years at the International Audio Laboratories Erlangen, it is now time to say thanks to a number of persons. At first, I want to express my deep gratitude to my supervisor and mentor Meinard Müller. After some time in his group, I more and more realized the luck I had when meeting him in late 2013—with almost no background in MIR but with a passion for music and technology. He gave me the chance to grow as a researcher and the freedom to pursue my research goals with full devotion. Thanks to the German Research Foundation for financing my research over the years in several projects (MU 2686/6-1, MU 2686/7-1, MU 2686/11-1, MU 2686/12-1).

Over the years, I was lucky to work with many talented and inspiring researchers. To start with, thanks to Meinard's current and former research group members: Vlora Arifi-Müller, Nanzhu Jiang, Patricio López-Serrano, Christian Dittmar, Jonathan Driedger, Sebastian Ewert, Harald G. Grohganz, Thomas Prätzlich, Sebastian Rosenzweig, Hendrik Schreiber, Christof Weiß, Frank Zalkow, and Julia Zalkow. Thanks to all the students I supervised. I hope you learned as much from me as I did from supervising you: Sanu Pulimootil Achankunju, Moritz Berendes, Paul Bießmann, Bharadwaj Desikan, Lukas Dietz, Manuel Hiemer, Johannes Knödtel, Lena Krauß, Lukas Lamprecht, and Julian Reck.

Furthermore, thanks to my fellow colleagues at the AudioLabs (in alphabetical order): Alexander Adami, Stefan Bayer, Sebastian Braun, Tom Bäckström, Soumitro Chakrabarty, Bernd Edler, Youssef El Baba, Esther Feichtner, Johannes Fischer, Emanuël Habets, Tracy Harris, Jürgen

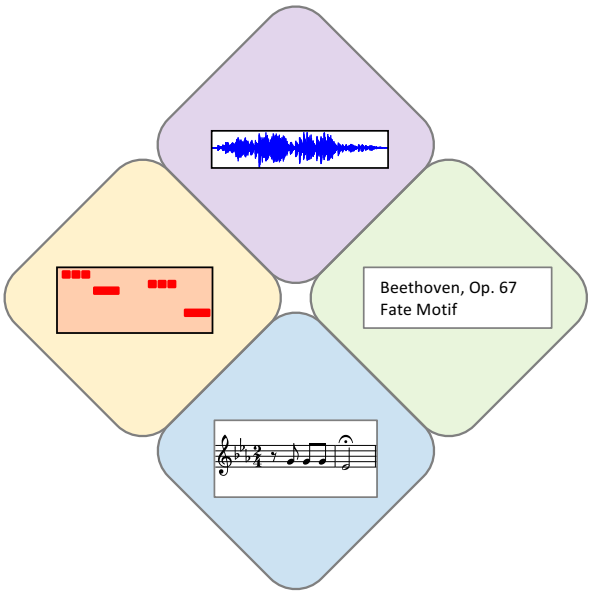
Herre, Goran Markovic, Day-See Riechmann, Konstantin Schmidt, Michael Schöffler, Fabian-Robert Stöter (thank you for paving the way from Hannover to Erlangen), Armin Taghipour, Stefan Turowski, Maja Taseska, María Luis Valero, Elke Weiland, and Nils Werner. In the same line, thanks to the colleagues at Fraunhofer IDMT in Ilmenau with whom I had lots of interesting discussions: Jakob Abeßer, Estefania Cano Cerón, Daniel Gärtner, Sascha Grollmisch, Anna Kruspe, Alexander Loos, Hanna Lukashevich, and Stylianos Ioannis Mimilakis. Thanks to the Jazzomat research team for fruitful cooperations and nice jams: Klaus Frieler, Martin Pfleiderer, Jakob Abeßer, and Wolf-Georg Zaddach. To the crew at the FAU's high performance computer center, in particular Thomas Zeiser, thank you for giving straightforward support.

Presenting at conferences resulted in getting to know a number of nice people from all over the world! Thanks to the ISMIR community for honest and inspiring discussions and the fun we had aside from the official program: Andreas Arzt, Juan P. Bello, Rachel Bittner, Sebastian Böck, Roger Dannenberg, Matthias Dorfer, Dan P. W. Ellis, Masataka Goto, Cynthia Liem, Brian McFee, Oriol Nieto, Jordi Pons, Colin Raffel, Justin Salamon, and Sebastian Stober. Special thanks to Gerhard Widmer for taking the time to review my thesis.

As important as having nice colleagues is to have good friends—thanks for your constant support—you know who you are.
To all the the musicians I have played with and hopefully continue from time to time.
To my family, for always being there.
To my wife Heike and our daughter Ella, for everything.

Part I

Retrieval of Musical Themes



Chapter 2

Retrieving Audio Recordings Using Musical Themes

In this chapter, we report on a systematic study considering a cross-modal retrieval scenario which was originally published in [11, 4]. Using a musical theme from the book "A Dictionary of Musical Themes" as a query, the objective is to identify all related music recordings from a given audio collection of Western classical music. By adapting well-known retrieval techniques, our main goal is to get a better understanding of the various challenges including tempo deviations, musical tunings, key transpositions, and differences in the degree of polyphony between the symbolic query and the audio recordings to be retrieved. In particular, we present an oracle fusion approach that indicates upper performance limits achievable by a combination of current retrieval techniques.

2.1 Introduction

There has been a rapid growth of digitally available music data including audio recordings, digitized images of scanned sheet music, album covers, and an increasing number of video clips. The huge amount of readily available music requires retrieval strategies that allow users to explore large music collections in a convenient and enjoyable way. In the last decades, many systems for content-based audio retrieval scenarios that follow the query-by-example paradigm have been suggested. Given a fragment of a symbolic or acoustic music representation used as a query, the task is to automatically retrieve documents from a music database containing parts or aspects that are similar to the query [37, 80, 155, 187]. One such retrieval scenario is known as *query-by-humming* [162, 166], where the user specifies a query by singing or humming a part of a melody. The objective is then to identify all audio recordings (or other music

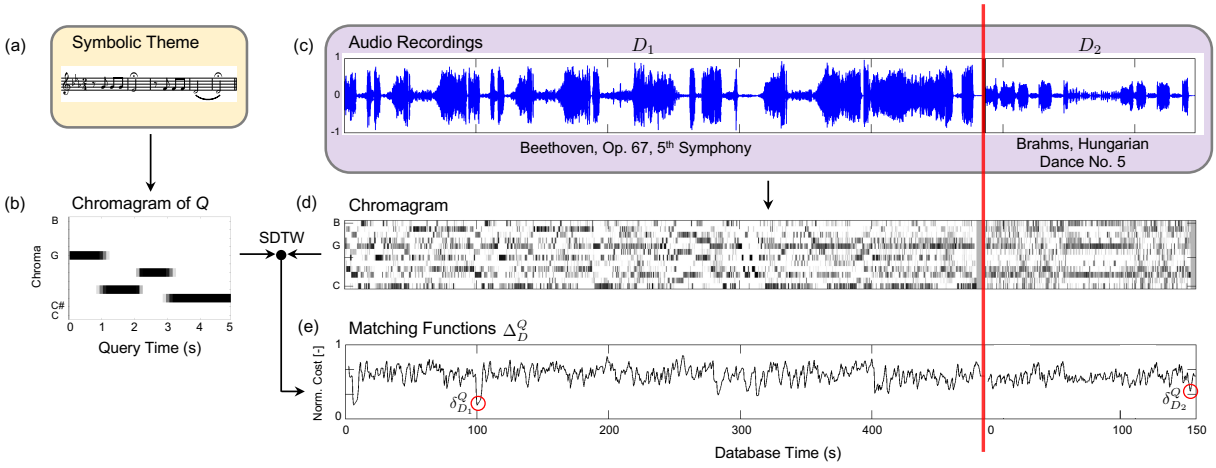


Figure 2.1: Illustration of the matching procedure. **(a)** Sheet music representations of a musical theme. **(b)** Chromagram of the query. **(c)** Music collection as a concatenated waveform. **(d)** Chroma representation of the recordings in the music collection. **(e)** Matching function Δ.

representations) that contain a melody similar to the specified query. Similarly, the user may specify a query by playing a characteristic phrase of a piece of music on an instrument [3, 120]. In a related retrieval scenario, the task is to identify an audio recording by means of a short symbolic query, e.g., taken from a musical score [65, 148, 184]. In the context of digital music libraries, content-based retrieval techniques are used to identify pieces in large archives which have not yet been systematically annotated [44, 129].

The retrieval scenario considered in this chapter is inspired by the book "A Dictionary of Musical Themes" by Barlow and Morgenstern [16], which contains roughly 10000 musical themes of instrumental Western classical music. Published in the year 1949, this dictionary is an early example of indexing music by its prominent themes. It was designed as a reference book for trained musicians and professional performers to identify musical pieces by a short query fragment. Most of the 10000 themes listed in the book [16] are also available as machine-readable versions (MIDI) on the internet [172]. Further details can be found in Appendix A.

In this chapter, we consider a cross-modal retrieval scenario, where the queries are symbolic encodings of musical themes and the database documents are audio recordings of musical performances. Then, given a musical theme used as a query, the task is to identify the audio recording of the musical work containing the theme. The retrieved documents may be displayed by means of a ranked list. This retrieval scenario offers several challenges.

- **Cross-modality.** On the one hand, we deal with symbolic sheet music (or MIDI), and with acoustic audio recordings on the other.
- **Tuning.** The tuning of the instruments, ensembles, and orchestras may differ from the

standard tuning.

- **Transposition.** The key of a recorded performance may differ from the original key notated in the sheet music (e.g., transposed versions adapted to instruments or voices).
- **Tempo differences.** Musicians do not play mechanically, but speed up at some passages and slow down at others in order to shape a piece of music. This leads to global and local tempo deviations between the query fragments and the performed database recordings.
- **Polyphony.** The symbolic themes are monophonic. However, in the database recording they may appear in a polyphonic context, where the themes are often superimposed with other voices, countermelodies, harmonies, and rhythms.

Additionally, there can be variations in instrumentation, timbre, or dynamics. Finally, the audio quality of the recorded performances may be quite low, especially for old and noisy recordings.

The main motivation of this approach is to demonstrate the performance of standard music retrieval techniques that were originally designed for audio matching and version identification [131, Chapter 7]. By successively adjusting the retrieval pipeline, we perform an error analysis, gain a deeper understanding of the data to be matched, and indicate potential and limitations of current retrieval strategies. We think that this kind of error analysis using a baseline retrieval system is essential before approaching the retrieval problem by introducing more sophisticated and computationally expensive audio processing techniques, such as [120]. The remainder of the chapter is structured as follows. In Section 2.2, we summarize the matching techniques and formalize the retrieval task. Then, in Section 2.3, we conduct extensive experiments and discuss our results. Further related work is discussed in the respective sections.

2.2 Matching Procedure

In this section, we summarize the retrieval procedure used here, following [131]. Similar procedures for synchronizing polyphonic sheet music and audio recordings were described in the literature [65, 184].

2.2.1 Chroma Features

Chroma features have been successfully used in solving different music-related search and analysis tasks [71, 131]. These features strongly correlate with tonal (harmonic, melodic) components for music whose pitches can be meaningfully categorized (often into 12 chromatic pitch classes) and whose tuning approximates to the equal-tempered scale [105]. In particular, chroma features are

suited to serve as a mid-level feature representation for comparing and relating acoustical and symbolic music, see Figure 2.1b and Figure 2.1d.

In our experiments (Section 2.3), we use the *Chroma Toolbox* [133] which uses a filterbank to decompose the audio signal in the aforementioned pitch classes. In particular, we use a chroma feature variant called CENS features. Starting with a feature rate of 10 Hz, we apply a temporal smoothing over nine frames and a downsampling by a factor of two. This results in chroma features at a rate of 5 Hz, as used in our experiments (Section 2.3).

2.2.2 Matching Technique

To compare a symbolic query to an audio recording contained in a music collection, we convert the query and recording into chroma sequences, say $X := (x_1, x_2, \ldots, x_N)$ and $Y := (y_1, y_2, \ldots, y_M)$. Typically, the length $M \in \mathbb{N}$ of Y is much larger than the length $N \in \mathbb{N}$ of the query X. Then, we use a standard technique known as *Subsequence Dynamic Time Warping* (SDTW) to compare X with subsequences of Y, see [130, Chapter 4]. In particular, we use the cosine distance (for comparing normalized chroma feature vectors) and the step size condition $\Sigma_1 := \{(1,0),(0,1),(1,1)\}$ in the SDTW. Furthermore, for the three possible step sizes, one may use additional weights w_v, w_h, and w_d, respectively. In the standard procedure, the weights are set to $w_v = w_h = w_d = 1$. In our later experiments, we use the weights to further penalize certain steps. As the results of SDTW, one obtains a matching function $\Delta : [1 : M] \to \mathbb{R}$. Local minima of Δ point to locations with a good match between the query X and a subsequence of Y, as indicated by the red circle in Figure 2.1e. For the details of this procedure and its parameters, we refer to [130, Chapter 4].

2.2.3 Retrieval Task

In the following, we formalize our retrieval task. Let $\mathcal{Q}$ be a collection of musical themes, where each element $Q \in \mathcal{Q}$ is regarded as a *query*. Furthermore, let $\mathcal{D}$ be a set of audio recordings, which we regard as a database collection consisting of *documents* $D \in \mathcal{D}$. Given a query $Q \in \mathcal{Q}$, the retrieval task is to identify the semantically corresponding documents $D \in \mathcal{D}$. In this setting, we are only interested in the associated audio recording of a given theme and not in its exact position within the recording. Therefore, we compute a matching function Δ_D^Q for Q and each of the documents $D \in \mathcal{D}$. Then, we define $\delta_D^Q = \min_m \Delta_D^Q(m)$ to be the distance between Q and D. Finally, we sort the database documents $D \in \mathcal{D}$ in ascending order according to the values δ_D^Q. The position of a document D in this ordered list is called the *rank* of D.

Figure 2.1 illustrates the matching procedure by using Beethoven's "Fate-Motif" as query. First, the given sheet music is transformed into a sequence of chroma features (see Figure 2.1a-b). In

Queries	#Themes	Database	#Recordings	Duration
$\mathcal{Q}_1$	177	$\mathcal{D}_1$	100	∼11 h
$\mathcal{Q}_2$	2046	$\mathcal{D}_2$	1113	∼120 h

Table 2.1: Overview of the datasets used for our experiments. Further details are given in Appendix A.

this example, our database consists of two audio recordings (see Figure 2.1c), which are also converted into chroma-based feature sequences (see Figure 2.1d). The matching functions Δ_D^Q are shown in Figure 2.1e. Red circles indicate the positions of the minima δ_D^Q for each document D. In this example, the matching function yields two distinct minima in the first document (Beethoven) at the beginning and after roughly 100 s. This is due to the fact that the motif, which is used as query, occurs several times in this work. In our document level scenario, both minima are considered to be correct matches as we are only interested in the entire recording and not in the exact position of the queried theme.

2.3 Experiments

We now report on our experiments using queries from the book by Barlow and Morgenstern, where we successively adapt the described matching procedure. Our main motivation is to gain a better understanding of the challenges regarding musical tuning, key transpositions, tempo deviation, and the degree of polyphony.

2.3.1 Test Datasets

The symbolic queries as given in the book by Barlow and Morgenstern [16] are available on the internet as MIDI files [172] in the "Electronic Dictionary of Musical Themes" (in the following referred to as EDM). We denote the 9803 themes from EDM by $\mathcal{Q}$. Furthermore, let $\mathcal{D}$ be a collection of audio recordings $D \in \mathcal{D}$.

We created two query test datasets, as shown by Table 2.1. The first dataset $\mathcal{Q}_1$ consists of 177 queries and serves as a development testset. The second test dataset $\mathcal{Q}_2$ contains 2046 queries and is used to investigate the scalability of the matching technique. In both test datasets, the durations of the queries ranges roughly between 1 s and 19 s with a mean of 7.5 s.

Additionally, we design two collections $\mathcal{D}_1$ and $\mathcal{D}_2$, which contain exactly one audio recording representing a true match of the queries contained in $\mathcal{Q}_1$ and $\mathcal{Q}_2$, respectively. Note that the number of queries is higher than the number of recordings because for a given musical piece, several themes may be listed in the book by Barlow and Morgenstern; e.g., there are six musical themes listed for the first movement of Beethoven's 5^{th} Symphony. A detailed overview about

the datasets can be obtained from Appendix A.

2.3.2 Evaluation Measures

In our evaluations, we compare a query $Q \in \mathcal{Q}$ with each of the documents $D \in \mathcal{D}$. This results in a ranked list of the documents $D \in \mathcal{D}$, where (due to the design of our test datasets $\mathcal{D}_1$ and $\mathcal{D}_2$) one of these documents is considered relevant. Inspired by a search-engine-like retrieval scenario, where a user typically looks at the top match and then may also check the first five, ten or twenty matches, we evaluate the top K matches for $K \in \{1, 5, 10, 20\}$. For a given K, the query is considered to be correct if its retrieved rank is at most K. Considering all queries at question, we then compute the proportion of correct queries (w.r.t. K). This results in a number $\rho_K \in [0 : 100]$ (given in percent), which we refer to as Top-K matching rate. Considering different values for K gives us insights in the distribution of the ranks and the system's retrieval performance.

2.3.3 Experiments using $\mathcal{Q}_1$ and D_1

We start with a first series of experiments based on $\mathcal{Q}_1$ and $\mathcal{D}_1$, where we systematically adapt various parameter settings while reducing the retrieval task's complexity by exploiting additional knowledge. We then aggregate the obtained results by means of an oracle fusion. This result indicates the upper limit for the performance that is achievable when using the suggested matching pipeline. Table 2.2 gives an overview of the results, which we now discuss in detail by exemplarily considering the results for ρ_1 and ρ_{10}.

Baseline. As a preliminary experiment, we use Σ_1 for the step size condition and $w_v = w_h = w_d = 1$ as weights. This yields Top-K matching rates of $\rho_1 = 38.4\%$ and $\rho_{10} = 62.7\%$. To increase the system's robustness, we restrict the SDTW procedure by using a different step size condition Σ. In general, using the set Σ_1 may lead to alignment paths that are highly deteriorated. In the extreme case, the query X may be assigned to a single element of Y. Therefore, it may be beneficial to replace Σ_1 with the set $\Sigma_2 = \{(2, 1), (1, 2), (1, 1)\}$, which yields a compromise between a strict diagonal matching (without any warping, $\Sigma_0 = \{(1, 1)\}$) and the DTW-based matching with full flexibility (using Σ_1). Furthermore, to avoid the query X being matched against a very short subsequence of Y, we set the weights to $w_v = 2$, $w_h = 1$, and $w_d = 1$. Similar settings have been used, e. g., in [132]. With these settings, we slightly improve the Top-K matching rates to $\rho_1 = 45.2\%$ and $\rho_{10} = 70.1\%$ (see also "Baseline" in Table 2.2). In general, using the set Σ_1 may lead to alignment paths that are highly deteriorated. In the extreme case, the sequence X may be assigned to a single element of Y. Therefore, it may be beneficial to replace Σ_1 with the set $\Sigma_2 = \{(2, 1), (1, 2), (1, 1)\}$, which yields a compromise between a strict diagonal matching (without any warping, $\Sigma_0 = \{(1, 1)\}$) and the DTW-based matching with full

Top-K	1	5	10	20
Baseline	45.2	62.1	70.1	76.8
Tu	46.9	64.4	72.9	81.9
Tr	52.0	68.9	79.1	87.6
Tu+Tr	53.7	72.3	83.1	91.0
Tu+Tr+Ql	68.4	79.1	88.1	93.2
Tu+Tr+Ql+Df	37.3	57.6	67.8	74.6
Oracle Fusion	72.3	84.7	92.1	97.7

Table 2.2: Top-K matching rate for music collection $\mathcal{D}_1$ with corresponding musical themes $\mathcal{Q}_1$ used as queries. The following settings are considered: Tu = Tuning estimation, Tr = Annotated transposition, Ql = Annotated query length, Df = Dominant feature band.

flexibility (using Σ_1). For details on the implementation and initialization of $\mathbf{D}$ when using Σ_2, we refer to [130, Chapter 7]. In the following, we continue using Σ_2 and the weights $w_v = 2$, $w_h = 1$, and $w_d = 1$.

Tuning (Tu) and Transposition (Tr). Deviations from the standard tuning in the actual music recording can lead to misinterpretations of the measured pitch. Estimating the tuning used in the music recording beforehand can reduce these artifacts [71]. Instead of using a dedicated tuning estimator, we simply test three different tunings by detuning the filterbank by $\pm 1/3$ semitones used to compute the chroma features (see Section 2.2.1). We then pick the tuning which yields the smallest minimum δ_D^Q. For a detailed description of a similar procedure, we refer to [71, 135]. This further improves the matching rates to $\rho_1 = 46.9\%$ and $\rho_{10} = 72.9\%$. As the musical key of the audio recording may differ from the key specified in the MIDI, we manually annotated the required transposition. Using this information in the matching procedure (by applying suitable chroma shifts [72]), the results improve to $\rho_1 = 52.0\%$ and $\rho_{10} = 79.1\%$. Combining both, the tuning estimation and the correct transposition, we get Top-K matching rates of $\rho_1 = 53.7\%$ and $\rho_{10} = 83.1\%$.

Query Length (Ql). We observed that the tempo events in some of our MIDI queries are set to an extreme parameter, which results in a query duration that strongly deviates from the corresponding passage in the audio recording. When the tempo information deviates too much from the audio recording, SDTW based on Σ_2 is unable to warp the query to the corresponding audio section. Furthermore, the features may lose important characteristics. For instance, the beginning theme of Beethoven's Pathétique has a MIDI duration of 3.5 s, whereas the corresponding section in the audio recording has a duration of 21 s. To even out tempo differences, we manually annotated the durations of the audio sections corresponding to queries and used this information to adapt the duration of the query before calculating the chroma features. This further increases the matching rate to $\rho_1 = 68.4\%$ and $\rho_{10} = 88.1\%$.

Dominant Feature Band (Df). In the next experiment, we want to compensate for the

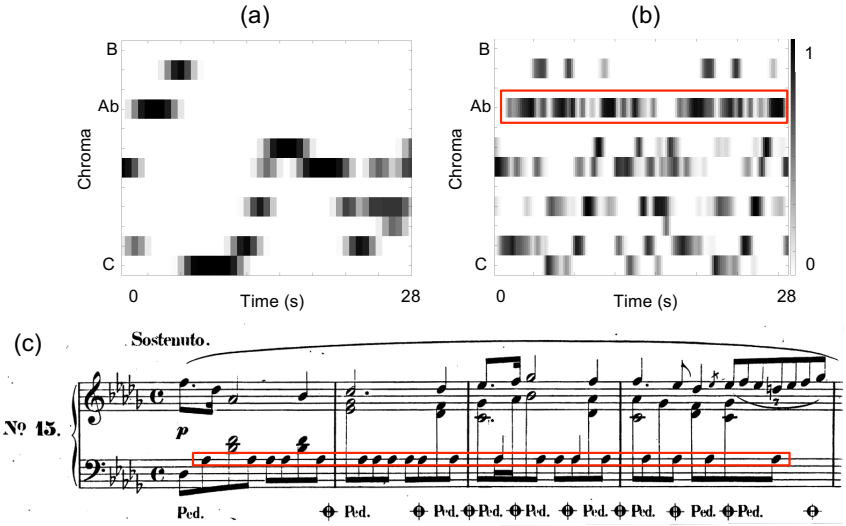


Figure 2.2: Example of Chopin's Prélude Op. 28, No. 15 ("Raindrop"). **(a)** Chromagram of monophonic query. **(b)** Chromagram of the corresponding section in the audio recording. **(c)** Sheet music representation of the corresponding measures.

different degrees of polyphony. Looking at the chromagram of the monophonic musical theme in Figure 2.1b reveals that only one chroma band is active at a time. For database documents as shown in Figure 2.1d, however, the energy is spread across several chroma bands due to the instruments' partials and accompaniments. A first method to reduce the polyphony on the audio side is to only take the dominant chroma band (the band with the largest value) for each time frame. This can be thought of as "monofying" the database document in the mid-level feature representation. Using this monofied chroma representation results in a matching rate of $\rho_1 = 37.3\%$ and $\rho_{10} = 67.8\%$. Even though this procedure works for some cases, for others it may pick the "wrong" chroma band, thus deteriorating the overall retrieval result. Further experiments showed that more refined methods (by extracting the predominant melody as described in [164]), may lead to slightly better results. However, Figure 2.2a shows a typical example where the advanced methods still fail, since the salient energy is located in the A$^\flat$-band (see Figure 2.2b), which is the accompaniment played with the left hand (see Figure 2.2c) and not the part we would perceive as being the main melody.

Oracle Fusion. In this experiment we assume having an oracle which can tell us, for each query, which setting performs best (in the sense that the relevant document is ranked better). The results obtained from oracle fusion yield a kind of upper limit which can be reached by using the suggested matching pipeline. Performing the oracle fusion for all queries leads to matching rates of $\rho_1 = 72.3\%$ and $\rho_{10} = 92.1\%$ (see Table 2.2). Oracle fusion shows that our matching pipeline may yield good retrieval results. However, a good prior estimate of transposition and tempo is important. Also, as we see in our next experiment, the results do not scale well when considering much larger datasets.

Top-K	**1**	**5**	**10**	**20**	**50**	**100**	**200**	**500**
Tu+Tr+5 s	14.9	21.8	25.8	29.2	35.5	43.0	54.1	76.1
Tu+Tr+10 s	18.3	25.1	28.3	32.6	38.7	46.1	56.1	76.2
Tu+Tr+15 s	13.6	19.5	22.7	26.1	31.6	38.9	49.7	72.4
Oracle Fusion	25.0	34.1	39.0	43.5	51.0	59.6	70.2	86.9

Table 2.3: Top-K matching rate for music collection $\mathcal{D}_2$ with corresponding musical themes $\mathcal{Q}_2$ used as queries. The following settings are considered: Tu = Tuning estimation, Tr = Transposition offset $[-2:2]$, $\{5, 10, 15\}$ s = Fixed query length.

2.3.4 Experiments using $\mathcal{Q}_2$ and $\mathcal{D}_2$

We now expand the experiments using the larger datasets $\mathcal{Q}_2$ (consisting of 2046 musical themes) and $\mathcal{D}_2$ (consisting of 1113 audio recordings). In this case, we do not have any knowledge of transposition and tempo information. One strategy to cope with different transpositions is to simply try out all 12 possibilities by suitably shifting the queries' chromagrams [72]. This, however, also increases the chance of obtaining false positive matches. Analyzing the annotations from $\mathcal{D}_1$, it turns out that most of the transpositions lie within $[-2:2]$ semitones. Therefore, in subsequent experiments, we only use these five transpositions, instead of all twelve possible chroma shifts. As for the query length, the durations of the annotated sections in $\mathcal{D}_1$ are within 3 s and 30 s. To cover this range, the duration of each query (EDM MIDI) is set to 5 s, 10 s, and 15 s, respectively. The results of the Top-K matching rates are shown in Table 2.3. For example, when using a query length of 5 s, the the matching rates are $\rho_1 = 14.9\%$ and $\rho_{10} = 25.8\%$. Using different query lengths (10 s and 15 s) does not substantially improve the retrieval results. However, using an oracle fusion over the different query lengths, the retrieval results substantially improve, leading to matching rates of $\rho_1 = 25.0\%$ and $\rho_{10} = 39.0\%$. In other words, even when using alignment methods to compensate for local tempo differences, a good initial estimate for the query duration is an essential step to improve the matching results.

Concluding these experiments, one can say that the retrieval of audio recordings by means of short monophonic musical themes is a challenging problem due to the challenges listed in the introduction (Section 2.1). We have seen that a direct application of a standard chroma-based matching procedure yields reasonable results for roughly half of the queries. However, the compensation of tuning issues and tempo differences is of major importance. The used matching procedure is simple to implement and has the potential for applying indexing techniques to speed up computations [79].

Differences in the degree of polyphony remain one main problem when matching monophonic themes against music recordings. In this context, simply taking the dominant feature band, as in our experiment in Section 2.3.3, turned out to even worsen the matching quality. (This was also the reason why we did not used this strategy in our experiment of Section 2.3.4.) One promising

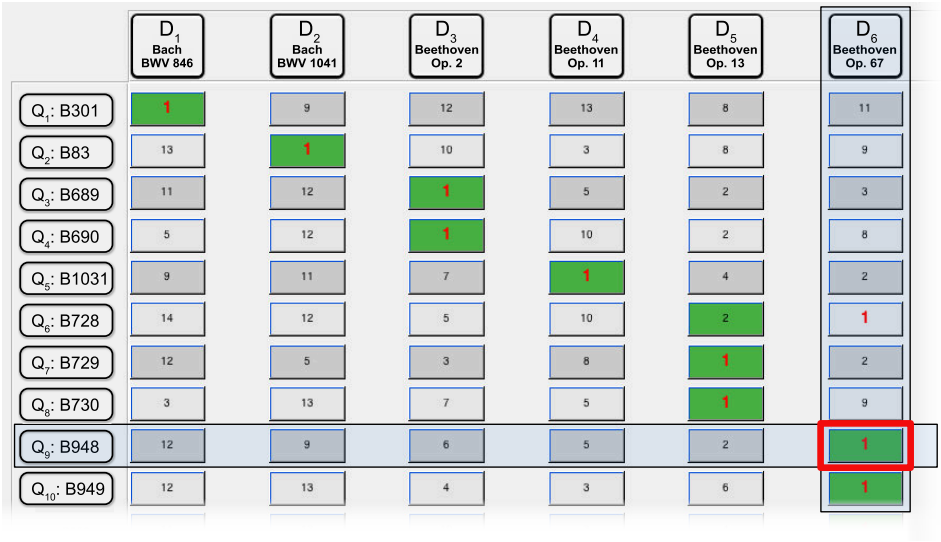


Figure 2.3: Main GUI window. The retrieval results in form of ranking values are mapped to a grid of boxes. The columns represent the audio recordings from the database and the rows the musical themes which were used as query. A green background is used to indicate ground truth annotations (the most relevant document).

approach, as suggested in [120], is to use NMF-based techniques to decompose the audio recording into monophonic-like components. These techniques, however, are computationally expensive and do not easily scale to recordings of long duration and large datasets. The development of scalable techniques to match monophonic and polyphonic music representations remain a research direction with many challenging problems.

2.4 Graphical User Interface

In this section, we present a graphical user interface (GUI) which we developed to systematically evaluate the matching results. The purpose of this GUI is to identify the challenges of this particular retrieval scenario and gain more insights into the used data. Figure 2.3 shows the main window of the GUI. The top row shows the audio recordings $\mathcal{D}$ contained in the database and the first left column lists the used queries $\mathcal{Q}$. By pushing one of the oval rectangles, one can inspect the calculated feature representation and listen to the audio or to a sonified version of the musical theme, respectively. For example, Figure 2.4a shows the chroma feature representation of query $\mathcal{Q}_9$ and the blue bar indicates the current position of the playback.

In the middle of Figure 2.3, we show all retrieval results as a grid of boxes. Additionally, the green background indicates the most relevant match as obtained from manual annotations. By pushing one of the boxes, the cost matrix of the corresponding best matching segment is visualized, see Figure 2.4b. Additionally, the warping path between the query and this segment is shown as

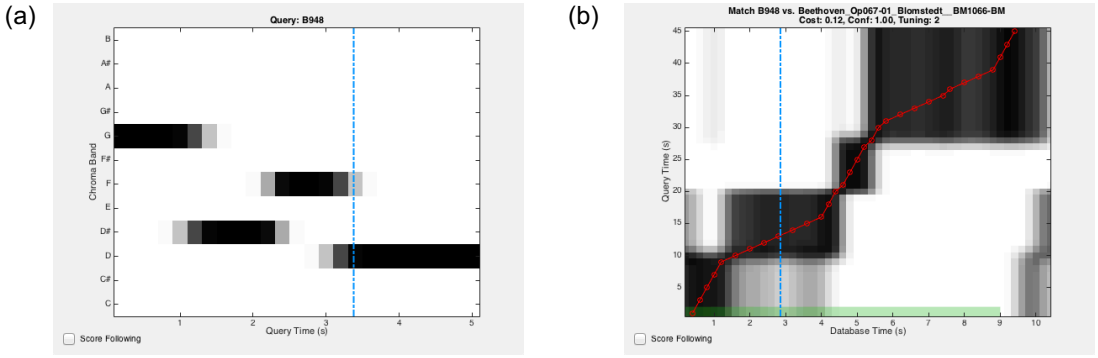


Figure 2.4: (a) Chroma feature representation of the monophonic theme. The blue bar indicates the playing position. (b) Visualization of the match in the audio recording. The plot shows the cost matrix with the actual warping path obtained from the SDTW. The green bar indicates annotations from the ground truth.

a red line. The green bar at the bottom incorporates the exact position of the query from the manual annotations. In the case of the shown example, the retrieval result is correct as the relevant database document is identified as the first element of the ranked list.

Furthermore, by sonifying the retrieval results, we get a feeling for the problems and challenges the algorithm faces when dealing with this kind of music. We do this by playing back the audio recording at the position of the estimated match and additionally acoustically overlay this recording with a sonified version of the time-aligned query. In this way, the GUI can make results from a retrieval system more accessible and also audible. Poorly performing matches can be analyzed and the gained knowledge can possibly be integrated into future versions of the retrieval algorithm.

2.5 Conclusion and Future Work

In this chapter, we have presented some baseline experiments for identifying audio recordings by means of musical themes. Due to musical and acoustic variations in the data as well as the typically short duration of the query, the matching task turned out to be quite challenging. Besides gaining some deeper insights into the challenges and underlying data, we still see potential of the considered retrieval techniques—in particular within a cross-modal search context. For example, in the case of the Barlow–Morgenstern scenario, the book contains textual specifications of the themes besides the visual score representations of the notes. Similarly, structured websites (e.g., Wikipedia websites) often contain information of various types including text, score, images, and audio. By exploiting multiple types of information sources, fusion strategies may help to better cope with uncertainty and inconsistency in heterogeneous data collections (see [136]). In the next chapter, we present a fusion approach for identifying musical themes (given in MIDI

format) based on corrupted OMR and OCR input. The further investigation of such cross-modal fusion approaches, including audio, image, and text-based cues, constitutes a promising research direction.

Chapter 3

Matching Musical Themes based on Noisy OCR and OMR Input

In this chapter, we deal with the problem of automatically matching the the themes from the book "A Dictionary of Musical Themes" to other digitally available sources. We hereby closely follow our original contribution presented in [8].

In 1949, Barlow and Morgenstern released the book "A Dictionary of Musical Themes" which contains 9803 themes of well-known instrumental pieces from the corpus of Western Classical music [16]. These monophonic themes (usually four bars long) are often the most memorable parts of a piece of music. To this end, we introduce a processing pipeline that automatically extracts from the scanned pages of the printed book textual metadata using Optical Character Recognition (OCR) as well as symbolic note information using Optical Music Recognition (OMR). Due to the poor printing quality of the book, the OCR and OMR results are quite noisy containing numerous extraction errors. As one main contribution, we adjust alignment techniques for matching musical themes based on the OCR and OMR input. In particular, we show how the matching quality can be substantially improved by fusing the OCR- and OMR-based matching results. Finally, we report on our experiments within the challenging Barlow and Morgenstern scenario, which also indicates the potential of our techniques when considering other sources of musical themes such as digital music archives and the world wide web.

3.1 Introduction

There has been a rapid growth of digitally available music data including audio recordings, digitized images of scanned sheet music, album covers and an increasing number of video clips. The huge amount of readily available music requires retrieval strategies that allow users to

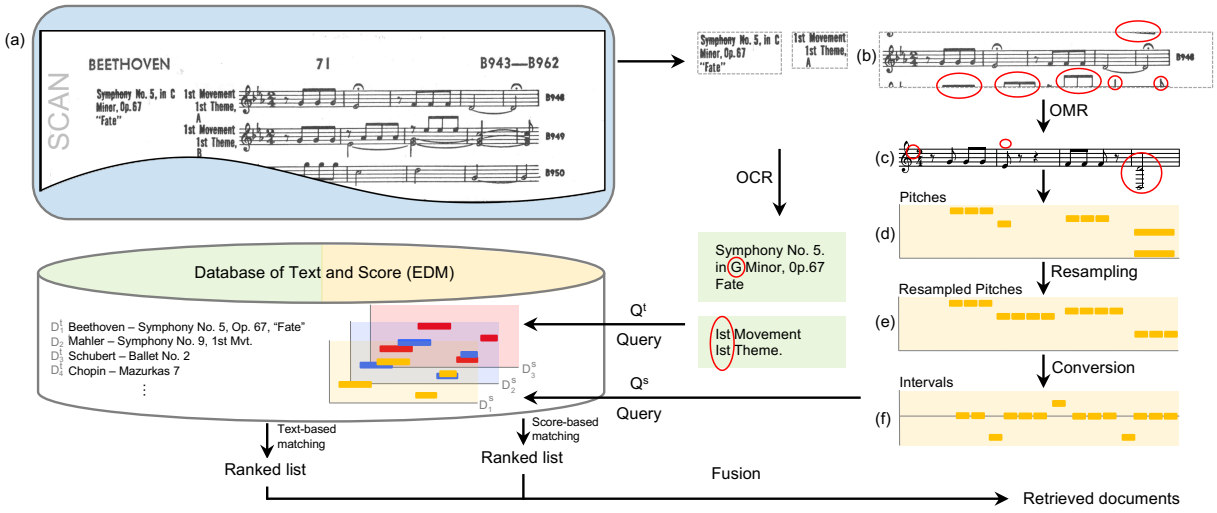


Figure 3.1: Overview of the processing pipeline. Each page is segmented into text and sheet music parts. The cropped images are transformed into computer readable representations using OCR and OMR (typical extraction errors are highlighted by a red circle). The results are used to query against a database consisting of music documents. Using a fusion strategy based on text-based and score-based matching results, the retrieval system outputs a ranked list of documents.

explore large music collections in a convenient and enjoyable way [52, 136, 141, 148, 173]. In this chapter, we focus on Western classical music, where a piece of music is typically specified by the composer, some work identifier such as a catalogue or opus number, and other types of metadata. For example, the musical work number Op. 67 by Ludwig van Beethoven specifies his Symphony No. 5 in C minor, the symphony with the famous fate motif. Besides such textual descriptions, Western classical music is given in form of printed sheet music, which visually encodes the notes to be played by musicians. Thanks to massive digitization efforts like the International Music Score Library Project[1] (IMSLP), millions of digitized pages of sheet music are publicly available on the world wide web.

Handling music collections of this size, one requires analysis and retrieval techniques for the various kinds of representations and formats. One important step consists in extracting the textual metadata as well as the note information from the digitized images. To this end, techniques such as Optical Character Recognition (OCR) to extract text-based metadata and Optical Music Recognition (OMR) to extract symbolic representations from the digital scans of printed sheet music are needed [18, 33, 66, 159, 158]. Besides of inconsistencies in the metadata that describes a musical work, the OCR and OMR may contain a significant number of extraction errors. This particularly holds for books of poor printing quality and scans of low resolution.

In this chapter, we deal with a challenging matching scenario by considering the book "A

[1] http://www.imslp.org/

Dictionary of Musical Themes" by Barlow and Morgenstern [16]. This book yields an overview of the most important musical themes from the Western classical music literature, thus covering many of the pieces contained in IMSLP. The contributions of this chapter are as follows. First, we describe a fully automated processing pipeline that matches the music themes from the book by Barlow and Morgenstern to other digitally available sources. This pipeline involves segmentation, OCR, OMR, and alignment techniques (see Section 3.2 and Figure 3.1). Then, we report on extensive experiments that indicate the retrieval quality based on inconsistent and erroneous OCR and OMR input (see Section 3.3). In particular, we show how the quality can be significantly improved by fusing the OCR-based and OMR-based matching results. Finally, we discuss how our processing pipeline may be applied to automatically identify, retrieve, and annotate musical sources that are distributed in digital music archives and the world wide web.

3.2 Processing Pipeline

3.2.1 Text and Score Recognition

As starting point for our matching scenario, we use the book by Barlow and Morgenstern [16], which contains 9803 musical themes from the most important compositions of the Western classical music literature. The book includes orchestral music, chamber music, and works for solo instruments. Each theme is specified by a textual specification as well as a visual score representation of the notes. In particular, the respective composer, the underlying musical work, and the movement are listed. Within the book, the themes are systematically organized and suitably indexed.

An example for a scanned page of the book is shown in Figure 3.1a. The excerpt shows text-based metadata as well as score information. The composer is written on the top of each page (e. g., "Beethoven"), whereas the title of each musical work (e. g., "Symphony No. 5 in C Minor") is specified in a text box aligned to the left. Furthermore, each theme is further specified by a movement and theme description (e. g., "1st Movement, 1st Theme, A") followed by a score representation of the theme. Finally, an additional identifier (e. g., "B948"), which is used for indexing purposes, is printed at the end of each theme.

As this example shows, the book is structured in a systematic fashion, even though the positions of the various text boxes may slightly vary from theme to theme and page to page. Using heuristics on the layout of the book, we first automatically segment each page by determining for each of the themes the bounding boxes of the various text elements and the image containing the score information. In particular, we exploit the knowledge on the rough position of the elements as well as the characteristic horizontal staff lines of the score. This yields a segmentation result as indicated in Figure 3.1b. Because of the regular structure of the pages, the bounding boxes

computed by our algorithm are correct for more than 99% of the themes. One problem is that the bounding boxes for the score representations may intersect with previous and subsequent bounding boxes, which often results in unwanted score fragments as highlighted in Figure 3.1b.

The text boxes are further processed by feeding in the cropped images into an OCR engine. In our processing pipeline, we have used the freely available OCR engine Tesseract [179]. As indicated by Figure 3.1, the recognition results are of good overall quality with occasional errors on the character level. In our example, the string "1st" has been recognized as "Ist" and "C Minor" was transcribed as "G Minor". Because of its prominent placement, the larger font size, and the capitalization, the extraction of the composers' names (e. g., "BEETHOVEN") works particularly well.

The score information is processed by feeding in the cropped images into an OMR engine. For this task, we use the freely available OMR software Audiveris [20]. As can be seen by our example, the score conversion is more problematic than in the case of text. On the one hand, many extraction errors occur on the note level. In our example, some of the note lengths were not detected correctly, the fermata is missing, and an additional note has been added in the last measure. Some of these errors come from score fragments due to the above mentioned intersection problem of the bounding boxes. On the other hand, there are recognition errors that have a global impact on the interpretation of the pitch parameters of the notes. In particular, the recognition of the key and time signatures as well as the kind of clef (e.g. G-clef, C-clef or F-clef) has turned out to be problematic. In the example of Figure 3.1c, the OMR engine could not detect the three flats of the key signature, which affects the interpretation of the fourth note (the E flat becomes an E). Most of the errors are due to the poor printing quality of the book by Barlow and Morgenstern. Experiments with different scan resolutions and other OMR engines (e. g., PhotoScore, SharpEye or SmartScore) have not resolved these problems. As we will show in the next section, the influence of the extraction errors can be attenuated by designing suitable cost functions and matching procedures.

3.2.2 Matching Procedures

As a result of the previously described recognition process, we obtain a textual representation of the metadata (containing the composer, work identifier, and other metadata) and a symbolic score representation for each of the 9803 themes of the book by Barlow and Morgenstern (in the following referred to as BM). The goal is to use this information for identifying other digital sources that belong or relate to the musical themes. In our experiments, we consider a scenario that allows us to study various matching procedures and to systematically evaluate matching results. To this end, we consider the "Electronic Dictionary of Musical Themes" (in the following referred to as EDM), which is publicly available at [172]. The EDM collection contains standard

MIDI files for the musical themes, which are linked to textual metadata similar to the original book by Barlow and Morgenstern. While the EDM themes more or less agree with the BM themes, there are inconsistencies with regard to the number of themes, the metadata and the score representations. Using the printed BM book as a reference, we have manually linked the BM themes to corresponding EDM themes. These correspondences serve as ground truth in the subsequent experiments.

In the following, we formulate our setting as a retrieval task. We denote the set of BM themes by $\mathcal{Q}$, where each element $Q \in \mathcal{Q}$ is regarded as a *query*. Furthermore, let $\mathcal{D}$ be the set of EDM themes, which we regard as a database collection consisting of *documents* $D \in \mathcal{D}$. Given a query $Q \in \mathcal{Q}$, the retrieval task is to identify the semantically corresponding document $D \in \mathcal{D}$.

3.2.3 Text-based Matching

Let us consider a fixed query $Q \in \mathcal{Q}$. In a first matching procedure, we only consider the textual representation, denoted by Q^{t}, which was obtained from the OCR step. Similarly, let D^{t} denote the text information for a document $D \in \mathcal{D}$. Both Q^{t} as well as D^{t} are represented as character strings. To compare these strings, one can use standard string alignment techniques such as the edit distance [42]. In our scenario, the two strings to be compared both contain the name of the composer, some work descriptor as well as a movement and theme identifier. However, the strings may also differ substantially due to additional information, segmentation errors, and OCR errors. Therefore, to compare strings, we use the longest common subsequence (LCS), which is a variant of the edit distance that is more robust to noise and outliers. For a description of this standard similarity measure, we refer to [42]. We convert the LCS-based similarity value into a normalized cost value by defining

$$c^{\mathrm{t}}(Q, D) := 1 - \frac{\mathrm{LCS}(Q^{\mathrm{t}}, D^{\mathrm{t}})}{|Q^{\mathrm{t}}|} \in [0, 1], \tag{3.1}$$

where $|Q^{\mathrm{t}}|$ denotes the length of the string Q^{t}. The performance of this matching procedure is discussed in Section 3.3.

3.2.4 Score-based Matching

Next, we define a matching procedure that only considers the score representation of the query $Q \in \mathcal{Q}$ resulting from the OMR step. In a first step, we convert the OMR result into a piano-roll like representation as indicated by Figure 3.1d. Dealing with monophonic themes (a property that may be corrupted by the OMR step), we consider the upper pitch contour of the OMR result. Since OMR often fails at detecting the correct note durations but tends to correctly recognize the bar lines, we do not use the note durations but locally resample the pitch sequence

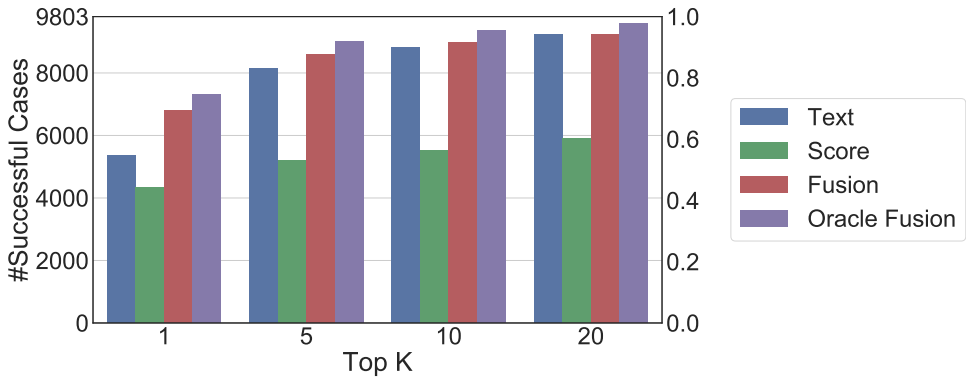


Figure 3.2: Comparison of the number of top K matches for the different procedures.

to match the bar line constraints, see Figure 3.1e. This results in a sequence of pitch values. Furthermore, since OMR often misinterprets the global clef, we convert the pitch sequence into a sequence of intervals (differences of subsequent pitches), see Figure 3.1f. The interval sequence, denoted by Q^{s}, is used for the matching step. Similarly, we process a document $D \in \mathcal{D}$, this time starting with a MIDI representation. The resulting interval sequence is denoted by D^{s}.

The OMR also often fails in detecting accidentals of notes, so that a pitch may be changed by one semitone. Using the edit distance would punish a deviation of one semitone to the same extent as larger deviations. Therefore, we use a local cost measure that takes the amount of the deviations into account. For two given intervals, say $a, b \in \mathbb{N}_0$, we define the distance by

$$\delta(a, b) = \frac{\min\{|a - b|, 12\}}{12} \in [0, 1]. \tag{3.2}$$

In this definition, we cap the value by 12 (an octave) to be robust to extreme outliers and then normalize the value. Based on this distance, we use standard dynamic time warping (DTW)—as described in [130, Chapter 4]—to obtain

$$c^{\mathrm{s}}(Q, D) := \frac{\mathrm{DTW}(Q^{\mathrm{s}}, D^{\mathrm{s}})}{|Q^{\mathrm{s}}|}. \tag{3.3}$$

Again we normalize by the length $|Q^{\mathrm{s}}|$. In the next section, we discuss the performance of the OCR-based and OMR-based matching procedures and show how they can be combined to further improve the results.

3.3 Retrieval Experiments

We now evaluate the proposed matching procedures within a retrieval setting. In this scenario, we consider the set $\mathcal{D}$ of EDM themes as a database collection of unknown musical themes.

Procedure	OCR	OMR	Fusion	Oracle Fusion
Mean rank	7.04	1186.26	6.24	3.34
Mean rank (capped)	3.64	9.63	2.96	2.24

Table 3.1: Mean ranks for the four different matching procedures. The capped mean ranks are computed by replacing the ranks above $K = 20$ to the value 21.

Using a BM theme $Q \in \mathcal{Q}$ as query, the task is to identify the database document that musically corresponds to the query. Note that in this retrieval scenario there is exactly one relevant document for each query.

In our evaluation, we compare the query Q with each of the documents $D \in \mathcal{D}$ and consider the top K matches for some number $K \in \mathbb{N}$. In a search-engine-like retrieval scenario, a user typically first looks at the top match and then may also check the first five, ten or twenty matches at most. Therefore, in the following, we consider the values $K \in \{1, 5, 10, 20\}$. In the case that the top K matches contain the relevant document, we say that the retrieval process has been *successful*. Conducting the retrieval process for all 9803 queries $Q \in \mathcal{Q}$, we then count the number of successful cases. Figure 3.2 shows the matching results for $K \in \{1, 5, 10, 20\}$ using four different matching procedures based on the text-based procedure from Section 3.2.3, the score-based procedure from Section 3.2.4, and two fusion procedures to be explained.

Let us start with a discussion of the text-based matching result. Considering the top match ($K = 1$), the retrieval system has been successful for 5354 of the 9803 queries, i. e., in 54.6% of all cases. Considering the top five matches ($K = 5$), the number of successful cases increases to 8156 queries (83.2%). This improvement can be explained by the fact that the specifications of the musical themes from the same work often differ in only a few characters, e. g. "1st Movement, 1st Theme, A" versus "2nd Movement, 1st Theme, B". Such small differences may lead to confusion among the top matches in the presence of OCR errors. Considering $K = 20$, one obtains 9225 successful cases (94.1%), which indicates that the text-based retrieval alone already yields a good overall retrieval quality.

Next, let us have a look at the score-based matching. In the case $K = 1$, the score-based retrieval has been successful for 4342 of the 9803 queries (44.3%). This much lower number (compared to the text-based procedure) reflects the fact that the OMR step introduces a large number of substantial errors. For example, an inspection showed that, for 1794 queries, the OMR engine was not able to produce a usable score representation. In these cases, the matching procedure was regarded as not successful. Increasing K, the results naturally improve reaching 5889 successful cases for $K = 20$ (60.1%). To get a better picture on the overall quality of the matching procedures, we have also analyzed the ranking positions of the relevant documents. Recall that we obtain for each query a ranked list of the documents $D \in \mathcal{D}$, where one of these documents is considered relevant. We determine the rank of this document for each query and

then compute a *mean rank* by averaging these ranks over all possible $Q \in \mathcal{Q}$. The mean ranks for all four considered matching procedures are shown in Table 3.1. The text-based procedure yields a mean rank of 7.04, whereas the score-based procedure results in a mean rank of 1186.26. The poor mean rank in the score-based case is the result of the unavailability of any score information for 1794 queries as mentioned above, where we set the rank to the value 4901 (half the size of $\mathcal{Q}$). Reducing the effect of outliers, we capped the rank by the value 21 (meaning that the rank is beyond $K = 20$). The mean rank of the capped values is 3.64 for the text-based and 9.63 for the score-based case. This again demonstrates that the text-based result is in average much more reliable than the score-based one.

Still, the score-based matching yields qualitatively different results than the text-based matching. We demonstrate this by fusing the matching results obtained by the two types of information. In a first experiment, we assume to have an oracle that tells us for each query which of the matching procedures performs better (in the sense that the relevant document is ranked better). The results obtained from this oracle fusion procedure yield a kind of upper limit for the joint performance of the text-based and score-based matching procedures. The results for the different values K are shown in Figure 3.2, while the mean rank can be found in Table 3.1. For example, one obtains 7315 (74.6%) successful cases for $K = 1$, increasing to 9592 (97.8%) for $K = 20$. This shows that the text-based matching can be significantly improved when including the score-based information.

We now present a fusion strategy that does not exploit any oracle knowledge. The text-based matching result is taken as the basis and then refined using the score-based information. The first assumption is that the top match is particularly reliable in the case that both, the text-based and score-based matching procedures, yield the same top match. The second (weaker) assumption is that the score-based top match is somewhat reliable when it is contained in the text-based $K = 20$ top matches. The third assumption is that the score-based result is particularly reliable in the case that the cost measure defined in (3.3) of the score-based first (top) match is significantly lower than the cost of the subsequent second match. Based on these assumptions, we use the ranked list of the text-based matching procedure and possibly replace the top match when the condition of the second or third assumption holds whereas the conditions of the first assumption does not hold. This simple fusion strategy yields matching results as indicated by Figure 3.2 and Table 3.1. In particular, for $K = 1$, the fusion strategy yields 6809 (69.5%) successful cases which is close to the upper limit 7315 (74.6%) obtained by oracle fusion.

Instead of presenting the exact details at this point, we only wanted to indicate the potential of fusing matching results. Using more refined fusion procedures could lead to results which are even closer to the upper limit indicated by oracle fusion.

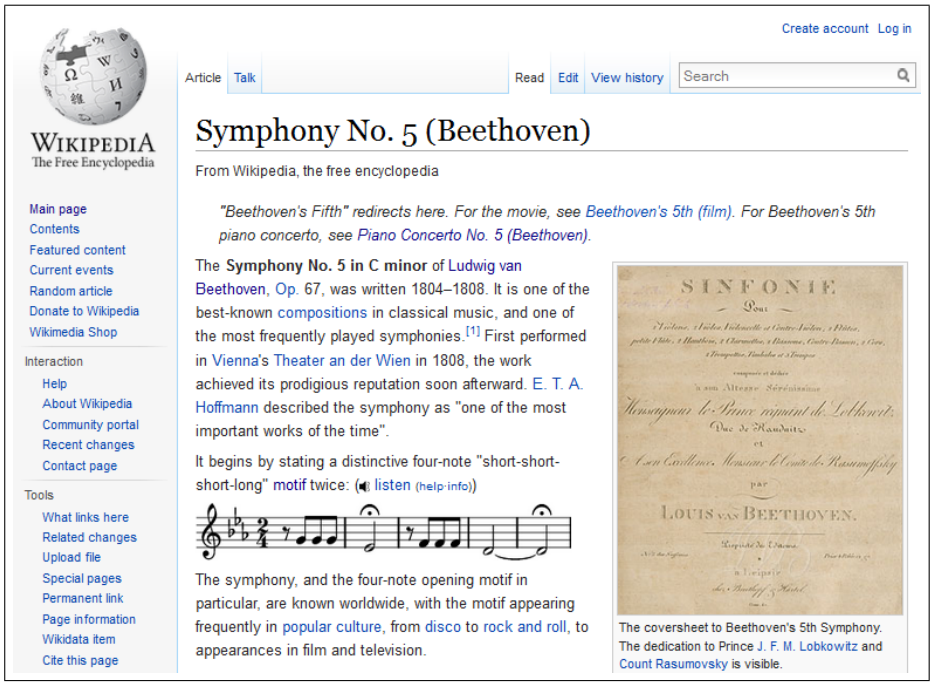


Figure 3.3: Example for a typical Wikipedia website contain various types of information (text, score, image, audio).

3.4 Applications and Conclusions

In this chapter, we have presented techniques for matching text-based and score-based musical information. As a case study, we used the sources from the book by Barlow and Morgenstern to serve as query input, while the EDM collection was used for evaluation purposes to serve as an example collection of digitally available musical items.

Going beyond the described (somehow controlled) scenario, we see potential of music information retrieval techniques for a much wider range of application scenarios. As mentioned in the introduction, there are millions of digitized pages of sheet music publicly available on the world wide web. Furthermore, music website as available at Wikipedia often contain information of various types including text, score, images, and audio, as shown in Figure 3.3. Often the description of musical works is enriched with audio examples and score fragments of musical themes. Using similar techniques as described in this chapter, one can use such structured websites to automatically derive text-based and score-based queries (and queries of other types of information such as audio or video) to look for musically related documents on the world wide web. For example, using the work specification (Beethoven, Symphony No. 5) and the score excerpt from Figure 3.3, one may want to retrieve sheet music representations from IMSLP or resources from less structured websites.

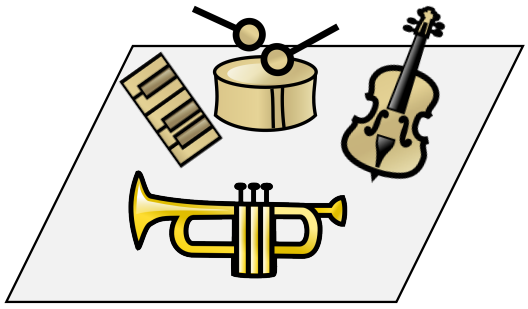
One main contribution of this chapter was to show that matching procedures based on possibly corrupted score input (e. g., coming from OMR) may still be a valuable component, in particular within a fusion scenario where an existing classifier should be further improved.

Fusion strategies that exploit multiple types of information sources will play an important role to better cope with uncertainty and inconsistency in heterogeneous data collections, see [136]. In this context, audio-related information has been studied extensively, see, e. g., [148, 173, 107].

Future work will be concerned with integrating all available sources that describe a musical work in order to identify, retrieve, and annotate musical sources that are distributed on the world wide web.

Part II

Extraction of Predominant Musical Voices



Chapter 4

Towards Evaluating Multiple Predominant Melody Annotations in Jazz Recordings

In this chapter, we analyze inter-annotator disagreement in predominant melody annotations, closely following the results presented in [10].

Melody estimation algorithms are typically evaluated by separately assessing the task of voice activity detection and fundamental frequency estimation. For both subtasks, computed results are typically compared to a single human reference annotation. This is problematic since different human experts may differ in how they specify a predominant melody, thus leading to a pool of equally valid reference annotations. In this chapter, we address the problem of evaluating melody extraction algorithms within a jazz music scenario. Using four human and two automatically computed annotations, we discuss the limitations of standard evaluation measures and introduce an adaptation of Fleiss' kappa that can better account for multiple reference annotations. Our experiments not only highlight the behavior of the different evaluation measures, but also give deeper insights into the melody extraction task.

4.1 Introduction

Predominant melody extraction is the task of estimating an audio recording's fundamental frequency trajectory values (F0) over time which correspond to the melody. For example in classical jazz recordings, the predominant melody is typically played by a soloist who is accompanied by a rhythm section (e. g., consisting of piano, drums, and bass). When estimating the soloist's F0-trajectory by means of an automated method, one needs to deal with two issues:

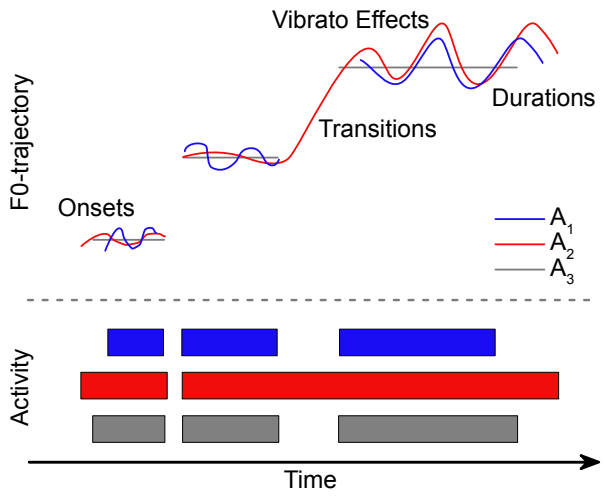


Figure 4.1: Illustration of different annotations and possible disagreements. A_1 and A_2 are based on a fine frequency resolution. Annotation A_3 is based on a coarser grid of musical pitches.

First, to determine the time instances when the soloist is active. Second, to estimate the course of the soloist's F0 values at active time instances.

A common way to evaluate such an automated approach—as also used in the Music Information Retrieval Evaluation eXchange (MIREX) [53]—is to split the evaluation into the two subtasks of activity detection and F0 estimation. These subtasks are then evaluated by comparing the computed results to a single manually created reference annotation. Such an evaluation, however, is problematic since it assumes the existence of a single ground-truth. In practice, different humans may annotate the same recording in different ways thus leading to a low inter-annotator agreement. Possible reasons are the lack of an exact task specification, the differences in the annotators' experiences, or the usage of different annotation tools [165, 167]. Figure 4.1 exemplarily illustrates such variations on the basis of three annotations $A_1, ..., A_3$ of the same audio recording, where a soloist plays three consecutive notes. A first observation is that A_1 and A_2 have a fine frequency resolution which can capture fluctuations over time (e. g., vibrato effects). In contrast, A_3 is specified on the basis of semitones which is common when considering tasks such as music transcription. Furthermore, one can see that note onsets, note transitions, and durations are annotated inconsistently. Reasons for this might be differences in annotators' familiarity with a given instrument, genre, or a particular playing style. In particular, annotation deviations are likely to occur when notes are connected by slurs or glissandi.

Inter-annotator disagreement is a generally known problem and has previously been discussed in the contexts of audio music similarity [95, 64], music structure analysis [178, 138, 144], and melody extraction [29]. In general, a single reference annotation can only reflect a subset of the musically or perceptually valid interpretations for a given music recording, thus rendering the

SoloID	Performer	Title	Instr.	Dur.
Bech-ST	Sidney Bechet	Summertime	Sopr. Sax	197
Brow-JO	Clifford Brown	Jordu	Trumpet	118
Brow-JS	Clifford Brown	Joy Spring	Trumpet	100
Brow-SD	Clifford Brown	Sandu	Trumpet	048
Colt-BT	John Coltrane	Blue Train	Ten. Sax	168
Full-BT	Curtis Fuller	Blue Train	Trombone	112
Getz-IP	Stan Getz	The Girl from Ipanema	Ten. Sax	081
Shor-FP	Wayne Shorter	Footprints	Ten. Sax	139

Table 4.1: List of solo excerpts taken from the WJD. The table indicates the performing artist, the title, the solo instrument, and the duration of the solo (given in seconds).

common practice of evaluating against a single annotation questionable.

The contributions of this chapter are as follows. First, we report on experiments, where several humans annotate the predominant F0-trajectory for eight jazz recordings. These human annotations are then compared with computed annotations obtained by automated procedures (MELODIA [164] and pYIN [121]) (Section 4.2). In particular, we consider the scenario of soloist activity detection for jazz recordings (Section 4.3.1). Afterwards, we adapt and apply an existing measure (Fleiss' Kappa [63]) to our scenario which can account for jointly evaluating multiple annotations (Section 4.3.2). Note that this chapter has an accompanying website at [7] where one can find the annotations which we use in the experiments.

4.2 Experimental Setup

In this work, we use a selection of eight jazz recordings from the *Weimar Jazz Database* (WJD) [68, 147]. For each of these eight recordings (see Table 4.1), we have a pool of seven annotations $\mathcal{A} = \{A_1, \ldots, A_7\}$ which all represent different estimates of the predominant solo instruments' F0-trajectories. In the following, we model an annotation as a discrete-time function $A : [1 : N] \to \mathbb{R} \cup \{*\}$ which assigns to each time index $n \in [1 : N]$ either the solo's F0 at that time instance (given in Hertz), or the symbol '$*$'. The meaning of $A(n) = *$ is that the soloist is inactive at that time instance.

In Table 4.2, we list the seven annotations. For this work, we manually created three annotations $A_1, \ldots, A_3$ by using a custom graphical user interface as shown in Figure 4.2 (see also [54]). In addition to standard audio player functionalities, the interface's central element is a salience spectrogram [164]—an enhanced time–frequency representation with a logarithmically-spaced frequency axis. An annotator can indicate the approximate location of F0-trajectories in the salience spectrogram by drawing *constraint regions* (blue rectangles). The tool then automatically uses techniques based on *dynamic programming* [131] to find a plausible trajectory through the

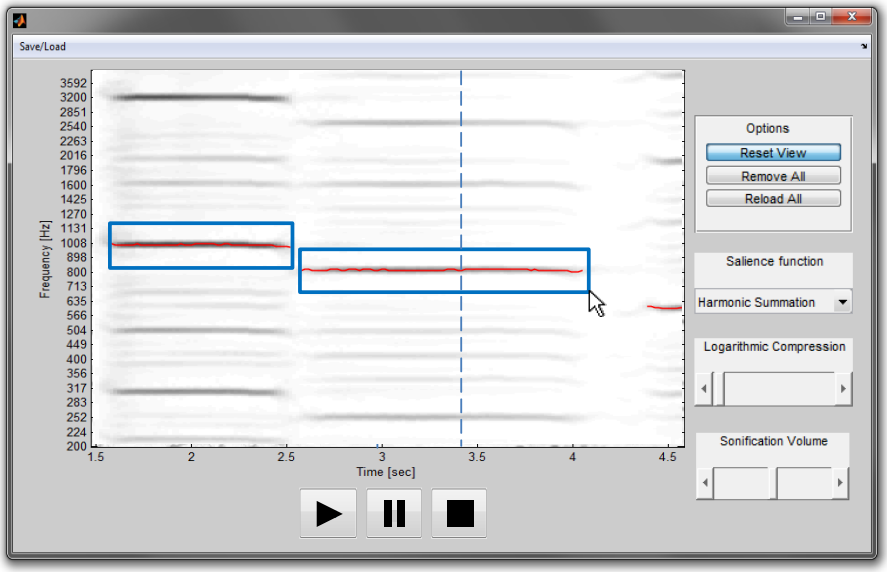


Figure 4.2: Screenshot of the tool used for the manual annotation of the F0 trajectories.

specified region. The annotator can then check the annotation by listening to the solo recording, along with a synchronized sonification of the F0-trajectory.

In addition to the audio recordings, the WJD also includes manually annotated solo transcriptions on the semitone level. These were created and cross-checked by trained jazz musicians using the *Sonic Visualiser* [35]. We use these solo transcriptions to derive A_4 by interpreting the given musical pitches as F0 values by using the pitches' center frequencies.

A_5 and A_6 are created by means of automated methods. A_5 is extracted by using the MELODIA [164] algorithm as implemented in Essentia [28] using the default settings (sample rate = 22050 Hz, hop size = 3 ms, window size = 46 ms). For obtaining A_6, we use the tool Tony [122] (which is based on the pYIN algorithm [121]) with default settings and without any corrections of the F0-trajectory.

As a final annotation, we also consider a baseline $A_7(n) = 1$ kHz for all $n \in [1 : N]$. Intuitively, this baseline assumes the soloist to be always active. All of these annotations are available on this chapter's accompanying website [7].

4.3 Soloist Activity Detection

In this section, we focus on the evaluation of the *soloist activity detection* task. This activity is derived from the annotations of the F0-trajectories $A_1, \ldots, A_7$ by only considering active time instances, i. e., $A(n) \neq *$. Figure 4.3 shows a typical excerpt from the soloist activity annotations

Annotation	Description
A_1	Human 1, F0-Annotation-Tool
A_2	Human 2, F0-Annotation-Tool
A_3	Human 3, F0-Annotation-Tool
A_4	Human 4, WJD, Sonic Visualiser
A_5	Computed, MELODIA [164, 28]
A_6	Computed, pYIN [121]
A_7	Baseline, all time instances active at 1 kHz

Table 4.2: Set $\mathcal{A}$ of all annotations with information about their origins.

for the recording `Brow-JO`. Each row of this matrix shows the annotated activity for one of our annotations from Table 4.2. Black denotes regions where the soloist is annotated as active and white where the soloist is annotated as inactive. Especially note onsets and durations strongly vary among the annotation, see e. g., the different durations of the note event at second 7.8. Furthermore, a missing note event is noticeable in the annotations A_1 and A_6 at second 7.6. At second 8.2, A_6 found an additional note event which is not visible in the other annotations. This example indicates that the inter-annotator agreement may be low. To further understand the inter-annotator agreement in our dataset, we first use standard evaluation measures (e. g., as used by MIREX for the task of *audio melody extraction* [127]) and discuss the results. Afterwards, we introduce Fleiss' Kappa, an evaluation measure known from psychology, which can account for multiple annotations.

4.3.1 Standard Evaluation Measures

As discussed in the previous section, an estimated annotation A_e is typically evaluated by comparing it to a reference annotation A_r. For the pair (A_r, A_e), one can count the number of time instances that are *true positives* #TP (A_r and A_e both label the soloist as being active), the number of *false positives* #FP (only A_e labels the soloist as being active), the number of *true negatives* #TN (A_r and A_e both label the soloist as being inactive), and the number *false negatives* #FN (only A_e labels the soloist as being inactive).

In previous MIREX campaigns, these numbers are used to derive two evaluation measures for the task of activity detection. *Voicing Detection* (VD) is identical to *Recall* and describes the ratio that a time instance which is annotated as being active is truly active according to the reference annotation:

$$\mathrm{VD} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}} \,. \tag{4.1}$$

The second measure is the *Voicing False Alarm* (VFA) and relates the ratio of time instances

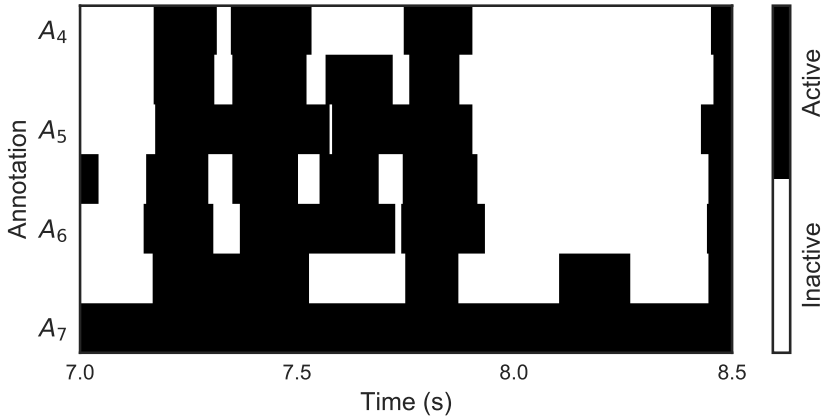


Figure 4.3: Excerpt from `Brow-JO`. $A_1, \ldots, A_4$ show the human annotations. A_5 and A_6 are results from automated approaches. A_7 is the baseline annotation which considers all frames as being active.

Ref. \ Est.	A_1	A_2	A_3	A_4	A_5	A_6	A_7	∅
A_1	–	0.93	0.98	0.92	0.74	0.79	1.00	0.89
A_2	0.92	–	0.97	0.92	0.74	0.79	1.00	0.89
A_3	0.84	0.84	–	0.88	0.69	0.74	1.00	0.83
A_4	0.85	0.86	0.94	–	0.70	0.75	1.00	0.85
A_5	0.84	0.84	0.90	0.85	–	0.77	1.00	0.87
A_6	0.75	0.76	0.81	0.77	0.65	–	1.00	0.79
A_7	0.62	0.62	0.71	0.67	0.55	0.65	–	0.64
∅	0.80	0.81	0.89	0.83	0.68	0.75	1.00	0.82

Table 4.3: Pairwise evaluation: *Voicing Detection* (VD). The values are obtained by calculating the VD for all possible annotation pairs (Table 4.2) and all solo recordings (Table 4.1). These values are then aggregated by using the arithmetic mean.

which are inactive according to the reference annotation but are estimated as being active:

$$\mathrm{VFA} = \frac{\#\mathrm{FP}}{\#\mathrm{TN} + \#\mathrm{FP}} \,. \tag{4.2}$$

In the following experiments, we assume that all annotations $A_1, \ldots, A_7 \in \mathcal{A}$ have the same status, i. e., each annotation may be regarded as either reference or estimate. Then, we apply the standard measures in a pairwise fashion. For all pairs $(A_r, A_e) \in \mathcal{A} \times \mathcal{A}$ with $A_r \neq A_e$, we extract VD and VFA (using the `MIR_EVAL` [156] toolbox) for each of the solo recordings listed in Table 4.1. The mean values over the eight recordings are presented in Table 4.3 for the VD-measure and in Table 4.4 for the VFA-measure.

As for the Voicing Detection (Table 4.3), the values within the human annotators $A_1, \ldots, A_4$ range from 0.84 for the pair (A_3, A_2) to 0.98 for the pair (A_1, A_3). This high variation in

Ref. \ Est.	A_1	A_2	A_3	A_4	A_5	A_6	A_7	∅
A_1	–	0.13	0.30	0.27	0.22	0.44	1.00	0.39
A_2	0.12	–	0.29	0.26	0.22	0.43	1.00	0.39
A_3	0.05	0.07	–	0.14	0.18	0.43	1.00	0.31
A_4	0.16	0.16	0.27	–	0.24	0.46	1.00	0.38
A_5	0.34	0.35	0.48	0.44	–	0.49	1.00	0.52
A_6	0.38	0.38	0.54	0.49	0.35	–	1.00	0.52
A_7	0.00	0.00	0.00	0.00	0.00	0.00	–	0.00
∅	0.17	0.18	0.31	0.27	0.20	0.38	1.00	0.36

Table 4.4: Pairwise evaluation: *Voicing False Alarm* (VFA). The values are obtained by calculating the VFA for all possible annotation pairs (Table 4.2) and all solo recordings (Table 4.1). These values are then aggregated by using the arithmetic mean.

VD already shows that the inter-annotator disagreement even within the human annotators is substantial. By taking the human annotators as reference to evaluate the automatic approach A_5, the VD lies in the range of 0.69 for (A_3, A_5) to 0.74 for (A_2, A_5). Analogously, for A_6, we observe values from 0.74 for (A_3, A_6) to 0.79 for (A_1, A_6).

As for the Voicing False Alarm (see Table 4.4), the values among the human annotations range from 0.05 for (A_3, A_1) to 0.30 for (A_1, A_3). Especially annotation A_3 deviates from the other human annotations, resulting in a very high VFA (having many time instances being set as active).

In conclusion, depending on which human annotation we take as the reference, the evaluated performances of the automated methods vary substantially. Having multiple potential reference annotations, the standard measures are not generalizable to take these into account (only by considering a mean over all pairs). Furthermore, although the presented evaluation measures are by design limited to yield values in $[0, 1]$, they can usually not be interpreted without some kind of baseline. For example, considering VD, the pair (A_2, A_3) yields a VD-value of 0.97, suggesting that A_3 can be considered as an "excellent" estimate. However, considering that our uninformed baseline A_7 yields a VD of 1.0, shows that it is meaningless to look at the VD alone. Similarly, an agreement with the trivial annotation A_7 only reflects the statistics on the active and inactive frames, thus being rather uninformative. Next, we introduce an evaluation measure that can overcome some of these problems.

4.3.2 Fleiss' Kappa

Having to deal with multiple human annotations is common in fields such as medicine or psychology. In these disciplines, measures that can account for multiple annotations have been developed. Furthermore, to compensate for chance-based agreement, a general concept referred

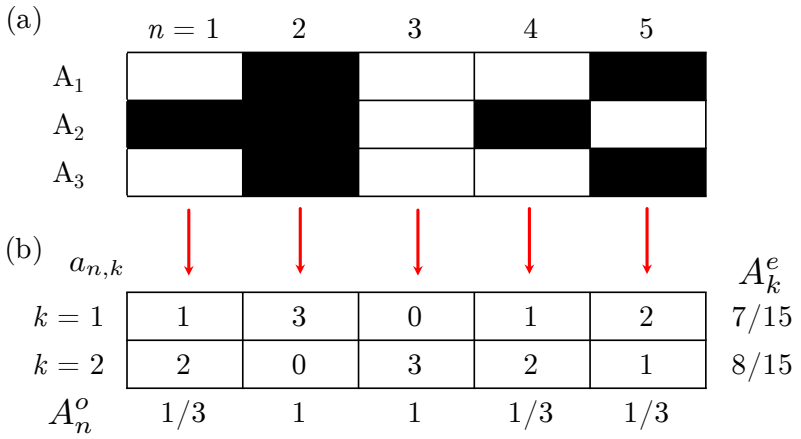


Figure 4.4: Example of evaluating Fleiss' κ for $K = 2$ categories, $N = 5$ frames, and three different annotations. (a) Annotations. (b) Number of annotations per category and time instance. Combining $A^o = 0.6$ and $A^e = 0.5$ leads to $\kappa = 0.2$.

< 0	$0 - 0.2$	$0.21 - 0.4$	$0.41 - 0.6$	$0.61 - 0.8$	$0.81 - 1$
poor	slight	fair	moderate	substantial	almost perfect

Table 4.5: Scale for interpreting κ as given by [109].

to as *Kappa Statistic* [63] is used. In general, a kappa value lies in the range of $[-1, 1]$, where the value 1 means complete agreement among the raters, the value 0 means that the agreement is purely based on chance, and a value below 0 means that agreement is even below chance.

We now adapt *Fleiss' Kappa* to calculate the chance-corrected inter-annotator agreement for the soloist activity detection task. Following [63, 109], Fleiss' Kappa is defined as:

$$\kappa := \frac{A^o - A^e}{1 - A^e} \,. \tag{4.3}$$

In general, κ compares the mean observed agreement $A^o \in [0, 1]$ to the mean expected agreement $A^e \in [0, 1]$ which is solely based on chance. Table 4.5 shows a scale for the agreement of annotations with the corresponding range of κ.

To give a better feeling for how κ works, we exemplarily calculate κ for the example given in Figure 4.4(a). In this example, we have $R = 3$ different annotations $A_1, \ldots, A_3$ for $N = 5$ time instances. For each time instance, the annotations belong to either of $K = 2$ categories (*active* or *inactive*). As a first step, for each time instance, we add up the annotations for each category. This yields the number of annotations per category $a_{n,k} \in \mathbb{N}$, $n \in [1 : N]$, $k \in [1 : K]$ which is shown in Figure 4.4(b). Based on these distributions, we calculate the observed agreement A^o_n

SoloID \ Comb.	κ_H	$\kappa_{H,5}$	$\kappa_{H,6}$	ρ_5	ρ_6
Bech-ST	0.74	0.60	0.55	0.82	0.75
Brow-JO	0.68	0.56	0.59	0.82	0.87
Brow-JS	0.61	0.47	0.43	0.78	0.71
Brow-SD	0.70	0.61	0.51	0.87	0.73
Colt-BT	0.66	0.55	0.49	0.84	0.74
Full-BT	0.74	0.66	0.61	0.89	0.83
Getz-IP	0.72	0.69	0.64	0.96	0.90
Shor-FP	0.82	0.65	0.58	0.80	0.70
∅	0.71	0.60	0.55	0.85	0.78

Table 4.6: κ for all songs and different pools of annotations. κ_H denotes the pool of human annotations $A_1, \ldots, A_4$. These values are then aggregated by using the arithmetic mean.

for a single time instance $n \in [1 : N]$ as:

$$A_n^o := \frac{1}{R(R-1)} \sum_{k=1}^{K} a_{n,k}(a_{n,k} - 1) \ , \tag{4.4}$$

which is the fraction of agreeing annotations normalized by the number of possible annotator pairs $R(R-1)$, e. g., for the time instance $n = 2$ in the example, all annotators agree for the frame to be active, thus $A_2^o = 1$. Taking the arithmetic mean of all observed agreements leads to the mean observed agreement

$$A^o := \frac{1}{N} \sum_{n=1}^{N} A_n^o \ , \tag{4.5}$$

in our example $A^o = 0.6$. The remaining part for calculating κ is the expected agreement A^e. First, we calculate the distribution of agreements within each category $k \in [1 : K]$, normalized by the number of possible ratings NR:

$$A_k^e := \frac{1}{NR} \sum_{n=1}^{N} a_{n,k} \ , \tag{4.6}$$

e. g., in our example for $k = 1$ (active) results in $A_1^e = 7/15$. The expected agreement A_e is defined as [63]

$$A^e := \sum_{k=1}^{K} (A_k^e)^2 \tag{4.7}$$

which leads to $\kappa = 0.2$ for our example. According to the scale given in Table 4.5, this is a "slight" agreement.

In Table 4.6, we show the results for κ calculated for different pools of annotations. First, we calculate κ for the pool of human annotations $H := \{1, 2, 3, 4\}$, denoted as κ_H. κ_H yields values ranging from 0.61 to 0.82 which is considered as "substantial" to "almost perfect" agreement

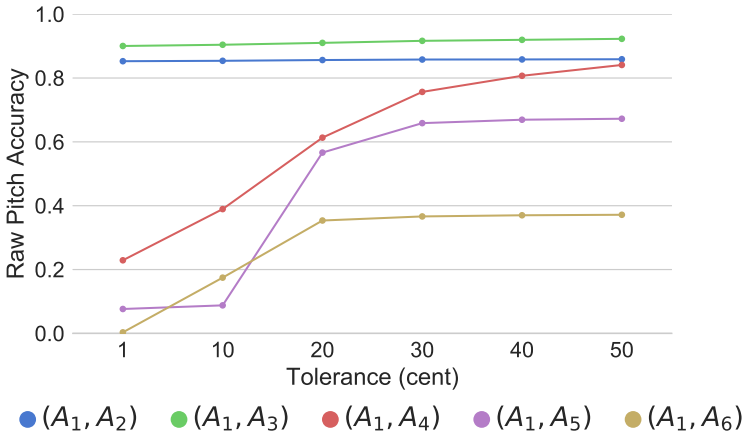


Figure 4.5: *Raw Pitch Accuracy* (RPA) for different pairs of annotations based on the annotations of the solo recording `Brow-JO`, evaluated on all active frames according to the reference annotation.

according to Table 4.5.

Now, reverting to our initial task of evaluating an automatically obtained annotation, the idea is to see how the κ-value changes when adding this annotation to the pool of all human annotations. A given automated procedure could then be considered to work correctly if it produces results that are just about as variable as the human annotations. Only if an automated procedure behaves fundamentally different than the human annotations, it will be considered to work incorrectly. In our case, calculating κ for the annotation pool $H \cup \{5\}$ yields values ranging from 0.47 to 0.69, as shown in column $\kappa_{H,5}$ of Table 4.6. Considering the annotation pool $H \cup \{6\}$, $\kappa_{H,6}$ results in κ-values ranging from 0.43 to 0.64. Considering the average over all individual recordings, we get mean κ-values of 0.60 and 0.55 for $\kappa_{H,5}$ and $\kappa_{H,6}$, respectively. Comparing these mean κ-values for the automated approaches to the respective κ_H, we can consider the method producing the annotation A_5 to be more consistent with the human annotations than A_6.

In order to quantify the agreement of an automatically generated annotation and the human annotations in a single value, we define the proportion $\rho \in \mathbb{R}$ as

$$\rho_5 := \frac{\kappa_{\mathrm{H},5}}{\kappa_{\mathrm{H}}}, \rho_6 := \frac{\kappa_{\mathrm{H},6}}{\kappa_{\mathrm{H}}}. \tag{4.8}$$

One can interpret ρ as some kind of "normalization" according to the inter-annotator agreement of the humans. For example, solo recording `Brow-JS` obtains the lowest agreement of $\kappa_H = 0.61$ in our test set. The algorithms perform "moderate" with $\kappa_{H,5} = 0.47$ and $\kappa_{H,6} = 0.43$. This moderate performance is partly alleviated when normalizing with the relatively low human agreement, leading to $\rho_5 = 0.78$ and $\rho_6 = 0.71$. On the other hand, for the solo recording `Shor-FP`, the human annotators had an "almost perfect" agreement of $\kappa_{H,6} = 0.82$. While the

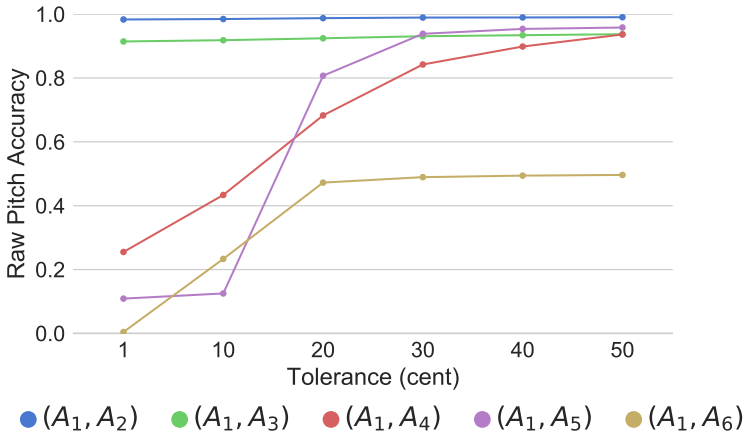


Figure 4.6: *Modified Raw Pitch Accuracy* for different pairs of annotations based on the annotations of the solo recording `Brow-JO`, evaluated on all active frames according to the *union* of reference and estimate annotation.

automated method's approaches were "substantial" with $\kappa_{H,5} = 0.65$ and "moderate" with $\kappa_{H,6} = 0.58$. However, although the automated method's κ-values are higher than for `Brow-JS`, investigating the proportions ρ_5 and ρ_6 reveal that the automated method's relative agreement with the human annotations is actually the same ($\rho_5 = 0.78$ and $\rho_5 = 0.71$ for `Brow-JS` compared to $\rho_5 = 0.80$ and $\rho_5 = 0.70$ for `Shor-FP`). This indicates the ρ-value's potential as an evaluation measure that can account for multiple human reference annotations in a meaningful way.

4.4 F0 Estimation

One of the used standard measures for the evaluation of the F0 estimation in MIREX is the *Raw Pitch Accuracy* (RPA) which is computed for a pair of annotations (A_r, A_e) consisting of a reference A_r and an estimate annotation A_e. The core concept of this measure is to label an F0 estimate $A_e(n)$ to be correct, if its F0-value deviates from $A_r(n)$ by at most a fixed tolerance $\tau \in \mathbb{R}$ (usually $\tau = 50$ cent). Figure 4.5 shows the RPA for different annotation pairs and different tolerances $\tau \in \{1, 10, 20, 30, 40, 50\}$ (given in cent) for the solo recording `Brow-JO`, as computed by `MIR_EVAL`. For example, looking at the pair (A_1, A_4), we see that the RPA ascends with increasing value of τ. The reason for this becomes obvious when looking at Figure 4.7. While A_1 was created with the goal of having fine grained F0-trajectories, annotations A_4 was created with a transcription scenario in mind. Therefore, the RPA is low for very small τ but becomes almost perfect when considering a tolerance of half a semitone ($\tau = 50$ cent).

Another interesting observation in Figure 4.5 is that the annotation pairs (A_1, A_2) and (A_1, A_3) yield almost constant high RPA-values. This is the case since both annotations were created using

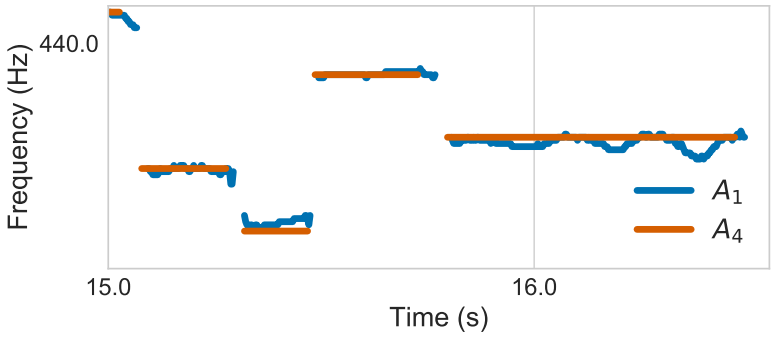


Figure 4.7: Excerpt from the annotations of the solo `Brow-JO` of A_1 and A_4.

the same annotation tool—yielding very similar F0-trajectories. However, it is noteworthy that there seems to be a "glass ceiling" that cannot be exceeded even for high τ-values. The reason for this lies in the exact definition of the RPA as used for MIREX. Let $\mu(A) := \{n \in [1:N] : A(n) \neq *\}$ be the set of all active time instances of some annotation in $\mathcal{A}$. By definition, the RPA is only evaluated on the reference annotation's active time instances $\mu(A_r)$, where each $n \in \mu(A_r) \setminus \mu(A_e)$ is regarded as an incorrect time instance (for any τ). In other words, although the term "Raw Pitch Accuracy" suggests that this measure purely reflects correct F0-estimates, it is implicitly biased by the activity detection of the reference annotation. Figure 4.8 shows an excerpt of the human annotations A_1 and A_2 for the solo recording `Brow-JO`. While the F0-trajectories are quite similar, they differ in the annotated activity. In A_1, we see that transitions between consecutive notes are often annotated continuously—reflecting glissandi or slurs. This is not the case in A_2, where the annotation rather reflects individual note events. A musically motivated explanation could be that A_1's annotator had a performance analysis scenario in mind where note transitions are an interesting aspect, whereas A_2's annotator could have been more focused on a transcription task. Although both annotations are musically meaningful, when calculating the RPA for (A_1, A_2), all time instances where A_1 is active and A_2 not, are counted as incorrect (independent of τ)—causing the glass ceiling.

As an alternative approach that decouples the activity detection from the F0 estimation, one could evaluate the RPA only on those time instances, where reference *and* estimate annotation are active, i. e., $\mu(A_r) \cup \mu(A_e)$. This leads to the modified RPA-values as shown in Figure 4.6. Compared to Figure 4.5, all curves are shifted towards higher RPA-values. In particular, the pair (A_1, A_2) yields modified RPA-values close to one, irrespective of the tolerance τ—now indicating that A_1 and A_2 coincide perfectly in terms of F0 estimation.

However, it is important to note that the modified RPA evaluation measure may not be an expressive measure on its own. For example, in the case that two annotations are almost disjoint in terms of activity, the modified RPA would only be computed on the basis of a very small number of time instances, thus being statistically meaningless. Therefore, to rate a computational

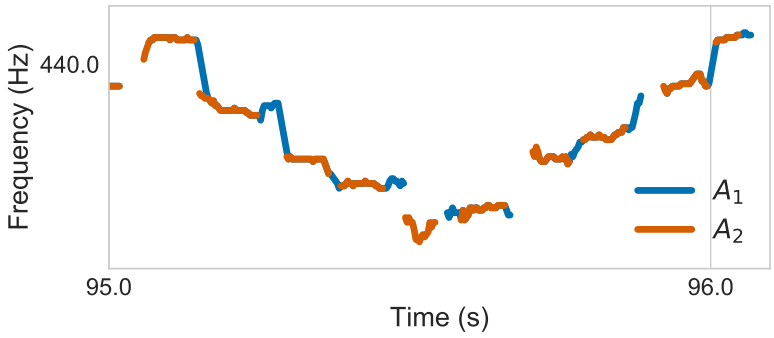


Figure 4.8: Excerpt from the annotations of the solo `Brow-JO` of A_1 and A_2.

approach's performance, it is necessary to consider both, the evaluation of the activity detection as well as the F0 estimation, simultaneously but independent of each other. Both evaluations give valuable perspectives on the computational approach's performance for the task of predominant melody estimation and therefore help to get a better understanding of the underlying problems.

4.5 Conclusion

In this chapter, we investigated the evaluation of automatic approaches for the task of predominant melody estimation—a task that can be subdivided into the subtask of soloist activity detection and F0 estimation. The evaluation of this task is not straightforward since the existence of a single "ground-truth" reference annotation is questionable. After having reviewed standard evaluation measures used in the field, one of our main contributions was to adapt Fleiss' Kappa—a measure which accounts for multiple reference annotations. We then explicitly defined and discussed Fleiss' Kappa for the task of the soloist activity detection.

The core motivation for using Fleiss' Kappa as an evaluation measure was to consider an automatic approach to work correctly, if its results were just about as variable as the human annotations. We therefore extended this kappa measure by normalizing it by the variability of the human annotations. The resulting ρ-values allow for quantifying the agreement of an automatically generated annotation and the human annotations in a single value.

For the task of F0 estimation, we showed that the standard evaluation measures are biased by the activity detection task. This is problematic, since mixing both subtasks can obfuscate insights into advantages and drawbacks of a tested predominant melody estimation procedure. We therefore proposed an alternative formulation for RPA which decoupled the two tasks.

Chapter 5

Data-Driven Solo Voice Enhancement for Jazz Music Retrieval

This chapter is based on our contributions presented in [13]. The results presented in Section 5.5, which were obtained by a close cooperation with Jakob Abeßer, were originally published in [1].

Retrieving short monophonic queries in music recordings is a challenging research problem in Music Information Retrieval (MIR). In jazz music, given a solo transcription, one retrieval task is to find the corresponding (potentially polyphonic) recording in a music collection. Many conventional systems approach such retrieval tasks by first extracting the predominant F0-trajectory from the recording, then quantizing the extracted trajectory to musical pitches and finally comparing the resulting pitch sequence to the monophonic query. In this chapter, we introduce a data-driven approach that avoids the hard decisions involved in conventional approaches: Given pairs of time–frequency (TF) representations of full music recordings and TF representations of solo transcriptions, we use a DNN-based approach to learn a mapping for transforming a "polyphonic" TF representation into a "monophonic" TF representation. This transform can be considered as a kind of solo voice enhancement. We evaluate our approach within a jazz solo retrieval scenario and compare it to a state-of-the-art method for predominant melody extraction.

5.1 Introduction

The internet offers a large amount of digital multimedia content—including audio recordings, digitized images of scanned sheet music, album covers, and an increasing number of video clips. The huge amount of readily available music requires retrieval strategies that allow users to

explore large music collections in a convenient and enjoyable way [113]. In this chapter, we consider the retrieval scenario of identifying jazz solo transcriptions in a collection of music recordings, see Figure 5.1. When presented in a musical theme retrieval scenario for classical music [11], this task offers various challenges, e. g., local and global tempo changes, tuning deviations, or key transpositions. Jazz solos usually consist of a predominant solo instrument (e. g., trumpet, saxophone, clarinet, trombone) playing simultaneously with the accompaniment of the rhythm group (e. g., piano, bass, drums). This typical interaction between the musicians leads to a complex mixture of melodic and percussive sources in the music recording. Consequently, retrieving monophonic pitch sequences of a transcribed solo can be very difficult due to the influence of the additional instruments in the accompaniment.

In this approach, we propose a data-driven approach for enhancing the solo voice in jazz recordings with the goal to improve the retrieval results. As our main technical contribution, we adapt a DNN architecture originally intended for music source separation [188] to train a model for enhancing the solo voice in jazz music recordings. Given the time–frequency (TF) representation of an audio recording as input for the DNN and a jazz solo transcription similar to a piano roll as the target TF representation, the training goal is to learn a mapping between both representations which enhances the solo voice and attenuates the accompaniment.

Throughout this work, we use the jazz solo transcriptions and music recordings provided by the Weimar Jazz Database (WJD). The WJD consists of 299 (as of August 2016) transcriptions of instrumental solos in jazz recordings performed by a wide range of renowned jazz musicians. The solos have been manually annotated and verified by musicology and jazz students at the Liszt School of Music Weimar as part of the Jazzomat Research Project [147]. Furthermore, the database contains more musical annotations (e. g., beats, boundaries, etc.) including basic meta-data of the jazz recording itself (i. e., artist, record name, etc.). A motivation for improving the considered retrieval scenario is to connect the WJD with other resources available online, e. g., YouTube. This way, the user could benefit from the additional annotations provided by the WJD while exploring jazz music.

The remainder of this chapter is structured as follows. In Section 5.2, we discuss related works for cross-modal retrieval and solo voice enhancement approaches. In Section 5.3, we introduce our DNN-based approach for solo voice enhancement. In particular, we explain the chosen DNN architecture, specify our training strategy, and report on the DNN's performance using the WJD. In Section 5.4, we evaluate our approach within the aforementioned retrieval scenario and compare it against a baseline and a conventional state-of-the-art system. In our experiments, we show that our DNN-based approach improves the retrieval quality over the baseline and performs comparably to the state-of-the-art approach. Finally, in Section 5.5, we apply similar techniques to the task of bass line transcription which indicates the scalability to non-salient instruments.

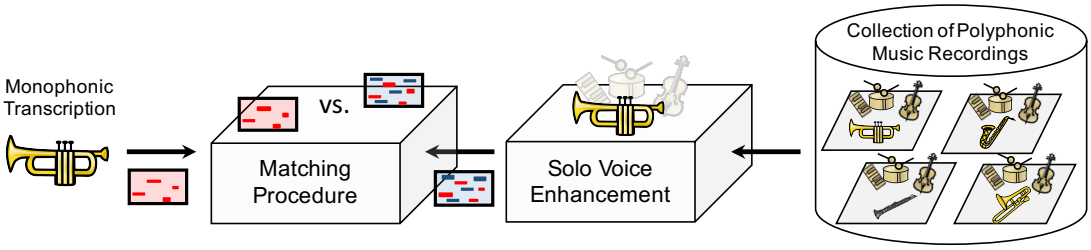


Figure 5.1: Illustration of the retrieval scenario. Given a jazz solo transcription used as a query, the task is to identify the music recording containing the solo. By enhancing the solo voice, we reduce the influence of the accompaniment in order to increase the retrieval results.

5.2 Related Work

Many systems for content-based audio retrieval that follow the query-by-example paradigm have been suggested [65, 148, 184, 37, 80, 187]. One such retrieval scenario is known as *query-by-humming* [162, 166], where the user specifies a query by singing or humming a part of a melody. Similarly, the user may specify a query by playing a musical phrase of a piece of music on an instrument [3, 120]. In a related retrieval scenario, the task is to identify a short symbolic query (e. g., taken from a musical score) in a music recording [155, 65, 148, 184, 11]. Conventional retrieval systems approach this task by first extracting the F0-trajectory from the recording, quantizing the extracted trajectory to musical pitches and finally mapping it to a TF representation to perform the matching (see [166]).

Many works in the MIR literature are concerned with extracting the predominant melody in polyphonic music recordings—a widely used example is Melodia [164]. More recent studies adapted techniques to work better with different musical styles, e. g., in [30], a combination of estimation methods is used to improve the performance on symphonic music. In [96], the authors use a source-filter model to better incorporate timbral information from the predominant melody source. A data-driven approach is described in [21], where a trained classifier is used to select the output for the predominant melody instead of using heuristics.

5.3 DNN-Based Solo Voice Enhancement

Our data-driven solo voice enhancement approach is inspired by the procedure proposed in [188], where the authors use a DNN for source separation. We will now explain how we adapt this DNN architecture to our jazz music scenario.

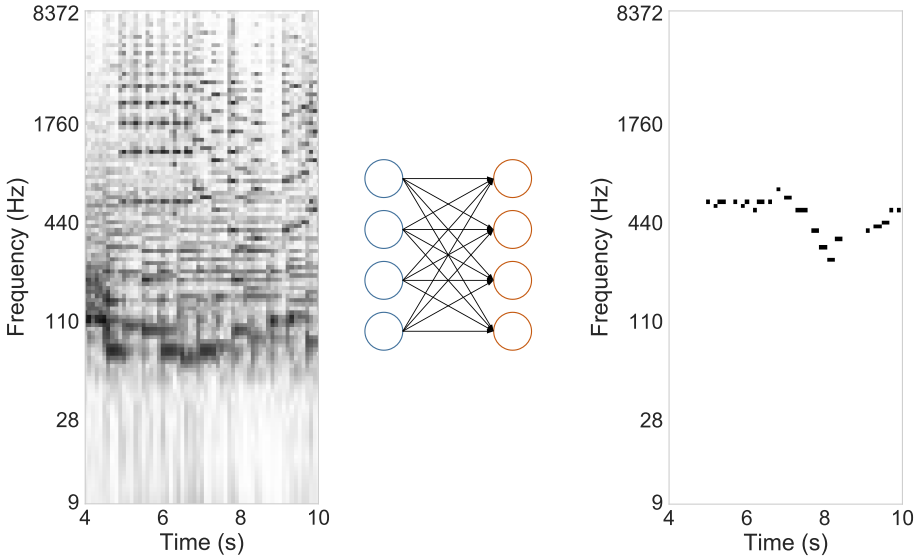


Figure 5.2: Input TF representation obtained from a music recording (left) and target TF obtained from the related WJD's solo transcription (right).

5.3.1 Deep Neural Network

Our DNN architecture closely follows [188], where the authors describe a DNN architecture and training protocol for source separation of monophonic instrument melodies from polyphonic mixtures. In principle, the network is similar to Stacked Denoising Autoencoders (SDA) [193], i. e., it consists of a sequence of conventional neural network layers that map input vectors to target output vectors by multiplying with a weight matrix, adding a bias term and applying a non-linearity (rectified linear units). In the setting described by the authors of the original work, the initial DNN consists of 3591 input units, a hidden layer, and 513 output units. The input vectors stem from a concatenation of 7 neighboring frames (513 dimensions each) obtained from a Short Time Fourier Transform (STFT) [131]. The target output vector is a magnitude spectrogram frame (513 dimensions) of the desired ground-truth. The training procedure uses the mean squared error between input and output to adjust the internal weights and biases via Stochastic Gradient Descent (SGD) until 600 epochs of training are reached. Afterwards, the next layer is stacked onto the first one and the output of the first is interpreted as an input vector. This way, the network is gradually built up and trained to a depth of five hidden layers. The originality of the approach in [188] lies in the least-squares initialization of the weights and biases of each layer prior to the SGD training [60].

In our approach, we do not try to map mixture spectra to solo instrument spectra, but rather to activation vectors for musical pitches. Our input vectors stem from an STFT (frame size = 4096 samples, hop size = 2048 samples) provided by the `librosa` Python package [123]. We then map the spectral coefficients to a logarithmically spaced frequency axis with 12 semitones per octave

	Training Set	Validation Set	Test Set
Duration (h)	5.575 (0.003)	2.389 (0.001)	0.885 (0.004)
Active Frames (%)	61.9 (0.2)	62.0 (0.3)	61.9 (1.8)
No. of Solos	269.1 (5.2)	—	29.9 (5.2)
No. of Full Rec.	204.3 (3.8)	—	22.7 (3.8)

Table 5.1: Mean duration and mean ratio of active frames aggregated over all folds (standard deviation is enclosed by brackets).

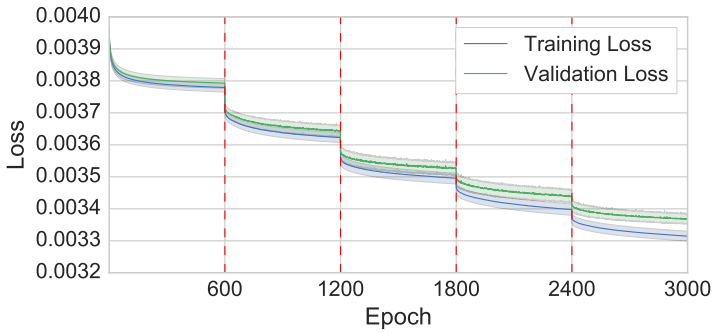


Figure 5.3: Training and validation loss during training epochs. For both losses, we show the mean values and the 95 % confidence intervals. The red lines indicate when the next layer is added to the DNN.

and 10 octaves in total which forms the TF representation for the music recordings [131]. The TF representations for the solo transcriptions are directly obtained from the WJD. In these first experiments, we want a simple DNN architecture and do not consider temporal context to keep the number of DNN parameters low. Therefore, our initial DNN consists of 120 input units, one hidden layer with 120 units, and 120 output units. Figure 5.2 shows the input TF representation of the music recording and the corresponding target output TF representation from the WJD's solo transcription.

5.3.2 Training

To train our DNNs, we consider the solo sections of the tracks provided by the WJD, i. e., where a solo transcription in a representation similar to a piano-roll is available. This selection leads to a corpus of around 9.5 hours of annotated music recordings. To perform our experiments, we sample 10 folds from these music recordings for training and testing using `scikit-learn` [146]. By using the record identifier provided by the WJD, we avoid using solos from the same record simultaneously in the training and test sets. Furthermore, we randomly split 30 % of the training set to be used as validation data during the training epochs. Table 5.1 lists the mean durations and standard deviations for the different folds and the portion of the recordings that consists of an actively playing soloist. The low standard deviations in the duration, as well as in the

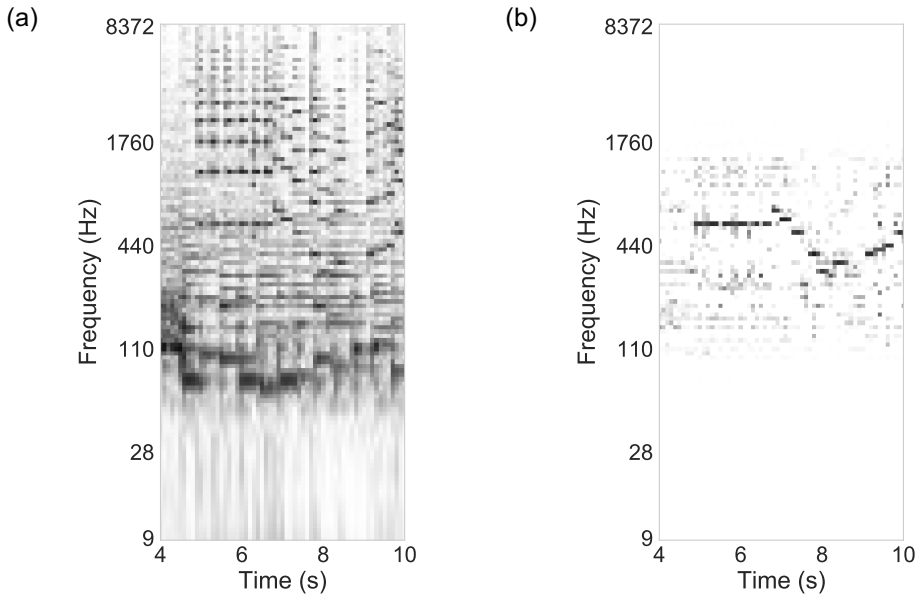


Figure 5.4: Typical example for the polyphony reduction using our DNN for an excerpt from Clifford Brown's solo on Jordu. **(a)** Input TF representation. **(b)** Output TF representation after processing with the DNN.

portion of active frames indicate that we created comparable folds. Note that a full recording can contain more than one solo transcription which explains the higher number of solo transcriptions compared to the number of full recordings. In order to reproduce the experiments, we offer the calculated features for all folds, as well as the exact details of the network architecture, on our accompanying website [6].

We start the training with our initial DNN with one hidden layer. We use SGD (momentum = 0.9, batch size = 100) with mean squared error as our loss function. After multiples of 600 epochs, we add the next layer with 120 units to the network until a depth of five hidden layers is reached. All the DNNs have been trained using the Python package `keras` [40]. The resulting mean training and mean validation loss considering all 10 folds are shown in Figure 5.3. After multiples of 600 epochs, we see that the loss improves as we introduce the next hidden layer to the network. With more added layers, we see that the validation loss diverges from the training loss as a sign that we are slowly getting into overfitting and can thus end the training.

5.3.3 Qualitative Evaluation

To get an intuition about the output results of the network, we process short passages from solo excerpts with the trained DNNs. Figure 5.4a shows the TF representation of an excerpt from a trumpet solo. Processing this with the DNN yields the output TF representation as shown in Figure 5.4b. Note that the magnitudes of the TF representations are logarithmically compressed

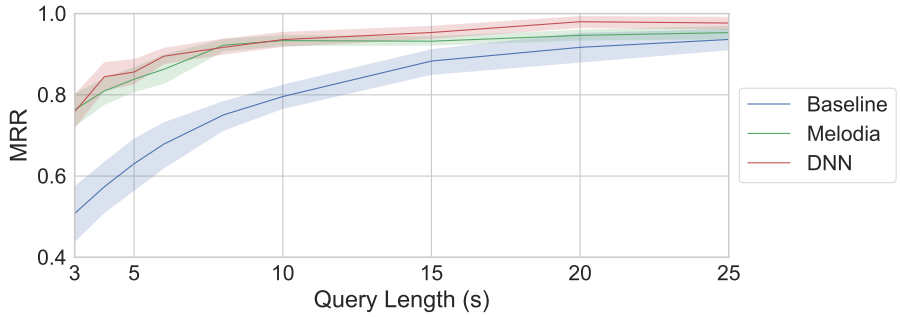


Figure 5.5: Mean reciprocal rank (MRR) for all three methods performed on all folds and with varying the query length. For all methods, we show the 95 % confidence intervals.

for visualization purposes. In the output, we can notice a clear attenuation of frequencies below 110 Hz and above 1760 Hz. An explanation for this phenomenon is that no pitch activations in those frequency bands are apparent in our training data. Thus, the DNN quickly learns to attenuate these frequency areas since they do not contribute to the target pitch activations at the output. In the region between these two frequencies, a clear enhancement of the solo voice can be seen, together with some additional noise. As seen in the input TF representation, the fundamental frequency (around 500 Hz) contains less energy than the first harmonic (around 1000 Hz), which is typical for the trumpet. However, the DNN correctly identifies the fundamental frequency. Further examples, as well as sonifications of the DNN's output, can be found at the accompanying website [6].

5.4 Retrieval Application

In this section, we first summarize our retrieval procedure and then describe our experiments. We intentionally constrain the retrieval problem to a very controlled scenario where we know that the monophonic queries correspond almost perfectly to the soloist's melody in the recording. We can rely on this assumption, since we use the manual transcriptions of the soloist as provided in the WJD.

5.4.1 Retrieval Task and Evaluation Measure

In the this section, we formalize our retrieval task following [131]. Let $\mathcal{Q}$ be a collection of jazz solo transcriptions, where each element $Q \in \mathcal{Q}$ is regarded as a *query*. Furthermore, let $\mathcal{D}$ be a set of music recordings, which we regard as a database collection consisting of *documents* $D \in \mathcal{D}$. Given a query $Q \in \mathcal{Q}$, the retrieval task is to identify the semantically corresponding documents $D \in \mathcal{D}$. In our experiments, we use a standard matching approach which is based on chroma

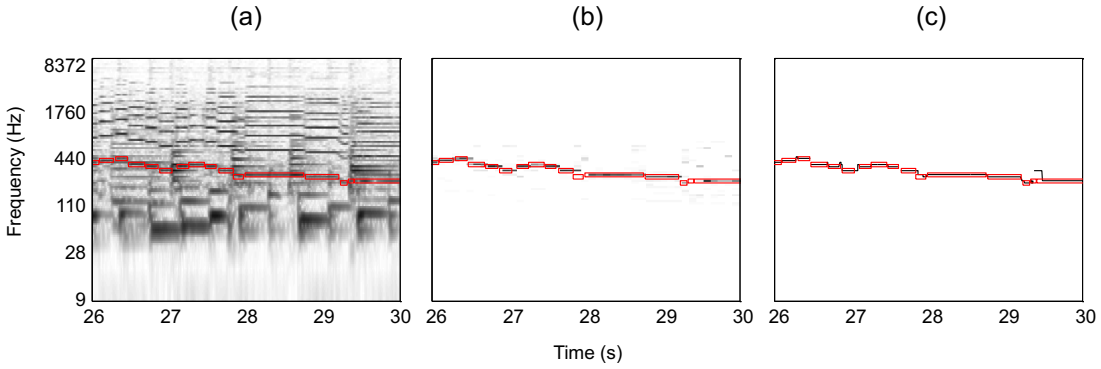


Figure 5.6: Typical example for the effect of both solo voice enhancement techniques. **(a)** Log-frequency magnitude spectrogram of a short jazz excerpt from our data. There is a clearly predominant solo melody, but also strong components from the accompaniment, such as bass and drums. **(b)** The same excerpt after running through a trained DNN as described in Section 5.3. We can see strongly attenuated influence of the accompaniment. **(c)** The same excerpt after extracting the predominant melody using the salience-based approach [164]. We can see that the trajectory of the solo melody has been tracked with only very few spurious frequencies.

features and a variant of Subsequence Dynamic Time Warping (SDTW). In particular, we use a chroma variant called CENS features with a smoothing of 9 time frames and a downsampling factor of 2 [134]. Comparing a query $Q \in \mathcal{Q}$ with each of the documents $D \in \mathcal{D}$ using SDTW yields a distance value for each pair (Q, D). We then rank the documents according to the these distance values of the documents $D \in \mathcal{D}$, where (due to the design of our datasets) one of these documents is considered relevant. In the following, we use the mean reciprocal rank (MRR) of the relevant document $D \in \mathcal{D}$ as our main evaluation measure. For the details of this procedure, we refer to the literature, e. g., [131, Chapter 7.2.2].

5.4.2 Experiments

We now report our retrieval experiments which follow the retrieval pipeline illustrated in Figure 5.1. In general, for our retrieval experiments, the queries are TF representations of the solo transcriptions from the WJD and the database elements are the TF representations of the corresponding full recordings containing the solos. We perform the retrieval for all 10 training folds separately. As listed in Table 5.1, the retrieval task consists in average for each fold of 30 solo transcriptions as queries to 23 music recordings in the database. Assuming we have a system that retrieves the relevant document randomly following a uniform distribution, for 30 queries and 23 database elements this would lead to a mean reciprocal rank of 0.13. This value serves as a lower bound of the expected performance of more intelligent retrieval systems. To further study the retrieval robustness, we consider query lengths starting from using the first 25 s of the solo transcription and then successively going down to 3 s.

In our baseline approach, we reduce the TF representations of the query and database documents (without using the DNN) to chroma sequences and apply the retrieval technique introduced earlier. The results of the baseline approach in terms of MRR for different query lengths are shown in Figure 5.5, indicated by the blue line. For a query length of 25 s, the baseline approach yields an MRR of 0.94. Reducing the query length to 5 s leads to a significant drop of the MRR down to 0.63. Now we consider our proposed DNN-based solo voice enhancement approach. The queries stay the same as in the baseline approach, but the TF representations of the database recordings are processed with our DNN before we reduce them to chroma sequences. For a query length of 25 s, this yields an MRR of 0.98; for a query length of 5 s, the MRR only slightly decreases to 0.86 which is much less than in the baseline approach. A reason for this is that the queries lose their specificity the shorter they become. This leads to wrong retrieval results especially when using the unprocessed recordings as in the baseline approach. The DNN-based approach compensates this by enhancing the solo voice and therefore makes it easier for the retrieval technique to identify the relevant recording.

Lastly, we consider a salience-based approach described in [164] for processing the music recording's TF representation. In short, this method extracts the predominant melody's F0-trajectory from the full recording, which is then quantized and mapped to a TF representation. The conceptional difference to our DNN-based approach is illustrated in Figure 5.6. For a query length of 25 s, this method yields a slightly lower MRR than the DNN-based approach of 0.96. Reducing the query to a length of 5 s, we achieve an MRR of 0.84. All three methods perform well when considering query lengths of more than 20 s. When the query length is shortened, all methods show a decrease in performance, whereas the DNN-based and salience-based methods significantly outperform the baseline approach.

5.5 Towards Non-Salient Instruments: Jazz Walking Bass Transcription

So far, we wanted to enhance the solo instrument's voice, which is typically the most salient voice in a jazz piece. Of course, there are additional voices such as such as the bass or the piano which are less salient yet play an important role for giving harmonic context. In this section, we report on experiments where we adapted our DNN-based approach to learn a salience representation for the bass instead for the solo voice. Our procedure is illustrated by Figure 5.7a: Given a jazz recording as input to a DNN, the objective is to learn the pitches played by the bass player. We evaluate the performance of the learned models in a transcription scenario where we assume that a (monophonic) walking bass line is present in the music, i. e., a bass pitch is present at each beat position. The estimated walking bass line is represented as a sequence of *beat-wise pitch values* and is estimated in a two-step way: we first extract the bass salience representation and

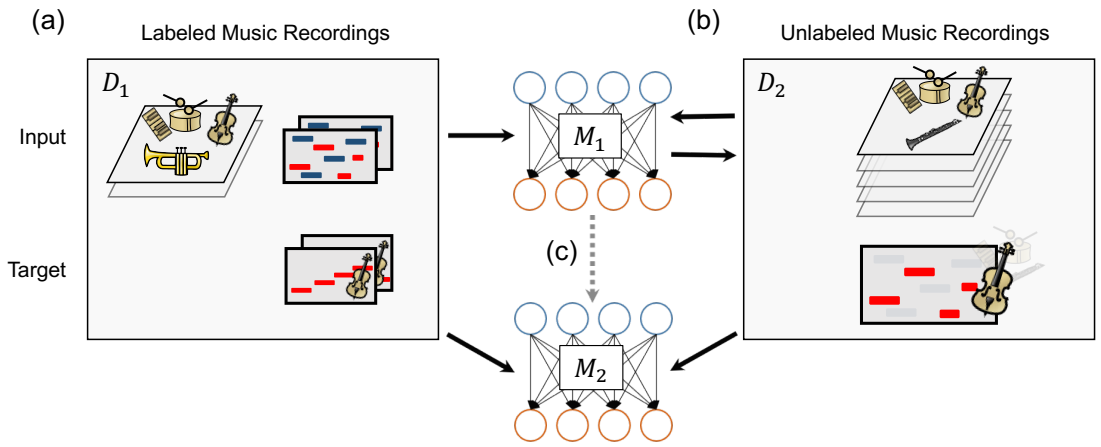


Figure 5.7: Visualization of the proposed system. (a) The labelled dataset D_1 is used for training the DNN to derive model M_1. (b) M_1 is used to create labels for the unlabelled music recordings in dataset D_2. (c) D_1 and D_2 are used as a combined training set to derive the DNN model M_2.

then aggregate the frames by using the beat annotations obtained from the WJD to obtain a beat-wise bass salience representation. Extracting the walking bass line is done by picking the highest value for the current beat estimate.

5.5.1 Input Features and Targets

In our approach, we resample each audio signal to a sampling rate of 22.05 kHz and compute the constant-Q magnitude spectrogram using the `librosa` python library [124] with a hopsize of 1024 (46.4 ms) and a frequency resolution of 12 bins per octave as input features. We consider the pitch range of a double bass ranging from MIDI pitch 28 and 67 (f_0 values from 41.2 Hz to 392 Hz). Pitch annotations are converted to binary pitch saliency vectors, which serve as target representation for multi-label classification. Both the input features and target values have the same dimensionality of 40.

In order to enlarge the datasets D_1 and D_2, we perform *data augmentation* and created two additional versions of each audio recording by applying pitch shifting[1] one semitone upwards and downwards, respectively. As a side effect, this procedure balances the overall pitch distribution in the training set. The enlarged datasets are denoted as D_1^+ (duration = 1.3 h) and D_2^+ (duration = 21.49 h).

[1]For pitch shifting we used the `sox` audio library `http://sox.sourceforge.net/`

Alg.	Frame-wise		Beat-wise		Sparseness
	A	$A_{\mathbf{PC}}$	A	$A_{\mathbf{PC}}$	[88]
SG	0.28 (0.14)	0.39 (0.15)	0.68 (0.22)	0.75 (0.21)	-
RK	0.12 (0.13)	0.18 (0.14)	0.60 (0.27)	0.64 (0.26)	-
D	0.37 (0.20)	0.41 (0.19)	0.72 (0.16)	0.75 (0.15)	-
M_1	0.31 (0.09)	0.43 (0.10)	0.71 (0.17)	0.78 (0.14)	0.684 (0.035)
M_1^+	**0.57** (0.13)	0.70 (0.11)	**0.83** (0.13)	0.89 (0.11)	0.761 (0.018)
$M_2^{0,+}$	0.54 (0.12)	0.68 (0.11)	0.81 (0.14)	0.88 (0.12)	**0.954** (0.010)
$M_2^{1,+}$	0.54 (0.13)	0.70 (0.11)	0.81 (0.14)	0.89 (0.11)	0.935 (0.015)
$M_2^{2,+}$	0.55 (0.12)	0.71 (0.11)	0.82 (0.14)	0.89 (0.12)	0.922 (0.019)
$M_2^{3,+}$	0.56 (0.12)	0.70 (0.11)	0.82 (0.14)	0.88 (0.12)	0.862 (0.030)

Table 5.2: Mean pitch detection accuracy values A, chroma accuracy A_{PC}, and mean frame-wise sparseness values, averaged over all test files (standard deviation values given in brackets). Both accuracy measures are computed frame-wise and beat-wise. Highest accuracy values A and sparseness values are denoted in bold print.

5.5.2 Training

The training was done in a similar fashion as described in Section 5.3.2. For the optimization, we used the ADADELTA algorithm [199], a mini-batch size of 500 (samples per gradient update), 500 epochs (gradient updates) for the training of each layer, and the mean squared error as the loss function.

Our experiments showed that a network with 4 layers, 5 context frames, 25% dropout, and no weight regularization showed the best performance on the dataset D_1^+. The optimal number of layers is close to the 5 layers used for melody pitch salience estimation. The incorporation of temporal context (frame stacking) seems beneficial for our application scenario. One possible reason could be that most bass notes in the walking bass style are relatively long (quarter notes) and have a stable pitch contour.

5.5.3 Evaluation

For our evaluation, a separate dataset D_3 (duration = 0.12 h) is used as test set. We obtain bass salience predictions from the six models trained on different combinations of datasets. The state-of-the-art bass transcription algorithms by Ryynänen & Klapuri (RK) [163] and Dittmar et al. (D) [48] output a list of note events (score). The algorithm from Salamon & Gomez (SG) [166] outputs a frame-wise f_0 contour of the bass line.[2] The quantitative results in terms of frame-wise transcription accuracy are shown in Table 5.2. Summarizing the results reveals that using data augmentation was beneficial in our scenario. Transcription accuracy values are consistently higher around 5–10% when using the DNN-based methods when disregarding octave

[2]It must be noted that the algorithm SG is limited to a two-octave pitch range between the MIDI pitch values 21 and 45(f_0 values between 27.5–110 Hz).

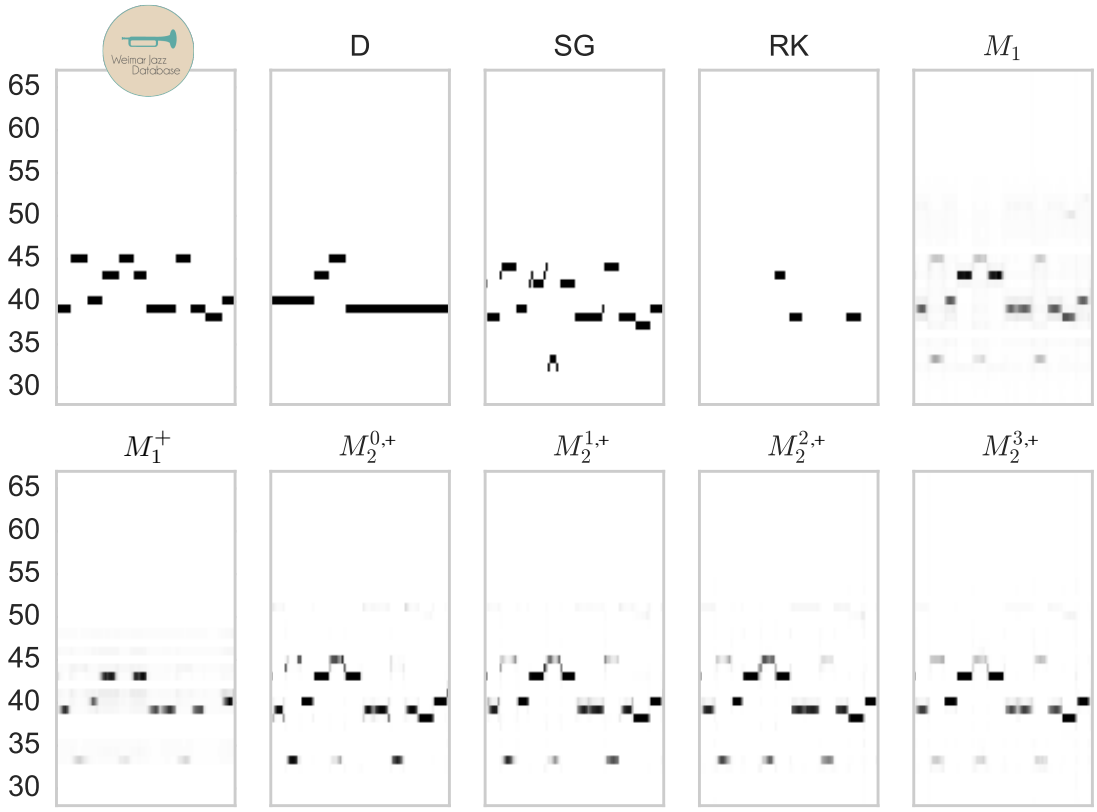


Figure 5.8: Excerpts from bass pitch salience representations for excerpt from 0:04 to 0:09 of Chet Baker's Solo on *Let's Get Lost*. Pitch salience matrices for deep learning methods are squared for better visibility. The MIDI pitch is shown on the vertical axis.

errors. However, one important note is that the state-of-the-art algorithms are not at all tailored towards jazz music while the proposed models are trained on music recordings with similar music style as the test data. For a more detailed evaluation, we refer to [1].

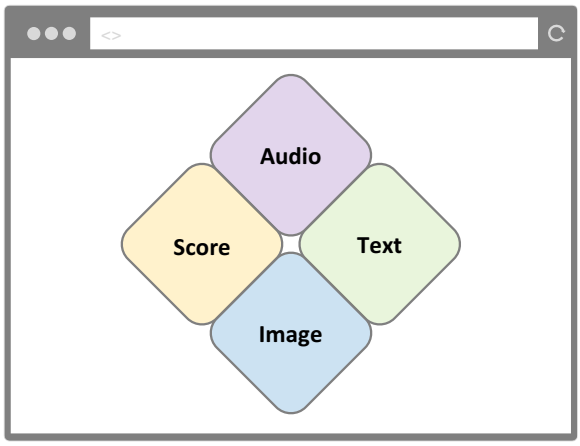
Figure 5.8 shows the predicted bass salience representations from all compared algorithms for the bass line excerpt from 0:04 to 0:09 of Chet Baker's Solo on *Let's Get Lost*. With respect to the proposed methods, adding additional training data using the semi-supervised training procedure does not significantly improve the accuracy values for the given transcription task. However, we found that the predictions obtained from all variants of model M_2, which incorporate additional training data are sparser than the predictions obtained from the initial model M_1. We interpret this as an indicator that these models are more confident in predicting the pitch salience of monophonic bass lines and show less confusion to other pitches. While this is not directly visible in the evaluation within the given transcription task, we belief that it has potential to improve source separation algorithms.

5.6 Conclusion

In this chapter, we described a data-driven approach for solo voice enhancement by adapting a DNN-based method originally used for source separation. As a case study, we used this enhancement strategy to improve the performance of a cross-modal retrieval scenario and compared it to a baseline and a conventional method for predominant melody estimation. From the experiments we conclude that in the case of jazz recordings, solo voice enhancement improves the retrieval results. Furthermore, the DNN-based and salience-based approaches perform on par in this scenario of jazz music and can be seen as two alternative approaches. In a related experiments, we used a similar network architecture and training procedure to estimate a bass salience representation. In this scenario, only few training data is available which made the use of data augmentation techniques such as pitch shifting important. In future work, we would like to investigate if we can further improve the results by enhancing the current data-driven approaches, e. g., by incorporating more temporal context through recurrent architectures and testing different unsupervised training methods.

Part III

Web-Based Technologies for Accessing Musical Content



Chapter 6

Enriching YouTube Videos with Jazz Music Annotations

In this chapter, we present an approach to enrich publicly available videos with musical annotations. We hereby closely follow our original contributions presented in [15, 5].

Web services allow permanent access to music from all over the world. Especially in the case of web services with user-supplied content, e. g., YouTube^{TM}, the available metadata is often incomplete or erroneous. On the other hand, a vast amount of high-quality and musically relevant metadata has been annotated in research areas such as Music Information Retrieval (MIR). Although they have great potential, these musical annotations are ofter inaccessible to users outside the academic world. With our contribution, we want to bridge this gap by enriching publicly available multimedia content with musical annotations available in research corpora, while maintaining easy access to the underlying data. Our web-based tools offer researchers and music lovers novel possibilities to interact with and navigate through the content. In this chapter, we consider a research corpus called the Weimar Jazz Database (WJD) as an illustrating example scenario. The WJD contains various annotations related to famous jazz solos. First, we establish a link between the WJD annotations and corresponding YouTube videos employing existing retrieval techniques. With these techniques, we were able to identify 988 corresponding YouTube videos for 329 solos out of 456 solos contained in the WJD. We then embed the retrieved videos in a recently developed web-based platform and enrich the videos with solo transcriptions that are part of the WJD. Furthermore, we integrate publicly available data resources from the Semantic Web in order to extend the presented information, for example, with a detailed discography or artists-related information. Our contribution illustrates the potential of modern web-based technologies for the digital humanities, and novel ways for improving access and interaction with digitized multimedia content.

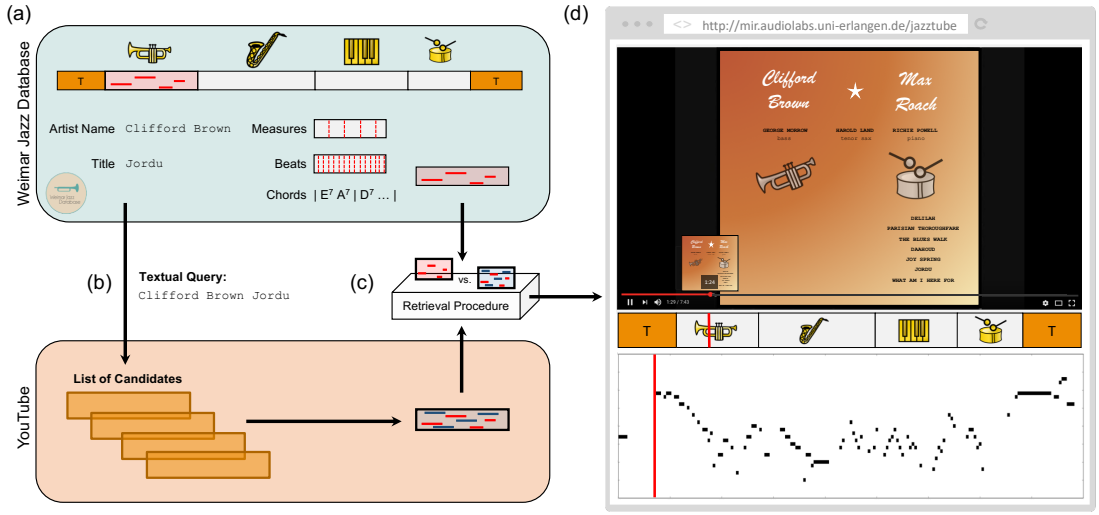


Figure 6.1: Illustration of the two-stage retrieval scenario applied to retrieve videos from YouTube. (a) Overview of the various annotations for Clifford Brown's solo *Jordu* contained in the Weimar Jazz Database. (b) First retrieval stage: Text-based retrieval on YouTube resulting in a list of candidates. (c) Second retrieval stage: Content-based retrieval using the solo recording from the WJD as query. (d) Identified video embedded in a web-based demonstrator and enriched with the annotations obtained from the WJD. *Figure (d) has been created by the authors and, therefore, no permission is required for its use in this manuscript.*

6.1 Introduction

Online video platforms, such as YouTube, make billions of videos available to users from all over the world. Many of these videos contain recordings of music performances. Often, these performances are tagged with basic metadata—mainly the artist and the title of the song. However, since this metadata is not curated, it might be incomplete or incorrect. The lack of reliable metadata makes it hard to identify particular recordings, especially for music genres where many renditions of the same musical work exist (e. g., symphonies in Western classical music, ragas in Indian music, or standards in jazz music). Imagine a jazz student who is practicing a jazz solo played by a famous musician and is now interested in the original recording. In the case that the student searches for a musician whose name is not mentioned in the metadata (e. g., because the musician was "only" a sideman in the band), a textual search may not be successful or may result in too many irrelevant results. Assuming that the student has already a partial or even a complete transcription of the solo available, content-based retrieval techniques could help to resolve this problem. Here, *content-based* means that, in the comparison of music data, the system makes use of the raw music data itself (e. g., from the music recording or the YouTube video), rather than relying on manually generated keywords referring to the artists' names, the song's title or lyrics [130].

Jazz musicians, musicologists, and publishers have made many jazz solo transcriptions publicly available during the last decades, e. g., Hal Leonard's *Omnibook* series.[1] One comprehensive corpus of solo transcriptions is the Weimar Jazz Database (WJD), which consists of 456 (as of May 2017) transcriptions of instrumental solos in jazz recordings performed by a wide range of renowned musicians [147]. The solos have been manually transcribed by musicology and jazz students. In addition, the database offers various music-related annotations such as chord sequences or beat positions. We believe that these annotations are a great resource that could help musicians and other researchers in gaining a deeper understanding of jazz music. However, these annotations and the underlying audio material are not directly accessible, mainly for two reasons. First, the audio files originate from commercial music recordings which are protected by copyright and ancillary copyright laws. Therefore, they cannot be made publicly available by scientific institutions. This restricts the usefulness of the dataset for scientific research, where both the annotations and the corresponding audio material are required. Second, the annotations are encoded in a database format which is not easily accessible for users without technical skills. Both problems apply to many scientific datasets which offer musical annotations for commercial music recordings. Simply switching to music recordings that are released under public domain licenses is not an option for research questions which rely on specific music recordings. In our approach, we try to bypass some of these copyright restrictions by using music recordings that are publicly available via YouTube. However, there is no doubt that both musicians and composers should be gratified financially for the music they create according to national and international copyright and ancillary copyright laws. YouTube seems to guarantee this financial entitlement through agreements with national copyright collecting societies. In contrast, for scientific institutions offering music databases it is very difficult or impossible to handle these legal claims. As a case study, we focus on the recordings which have corresponding annotations in the WJD.

As the main contribution of this work, we introduce various retrieval methods based on metadata and content-based descriptors and show how these techniques can be applied for identifying and enriching YouTube videos. In the following, we sketch a typical two-stage retrieval scenario which is then described in more detail in the subsequent sections (Figure 6.1 provides an overview). In this example, we are interested in the song *Jordu*, recorded by Clifford Brown in 1954 (Figure 6.1a). In the first step, we use the title and the name of the soloist as provided by the WJD to perform a metadata-based search on YouTube (Figure 6.1b). This search results in a list of candidates. Besides relevant music recordings, this list may also contain other recordings by the same artist or cover versions by other artists. Using the recording associated to the WJD's annotations, we apply an audio-based retrieval approach to identify the relevant music recordings in this list of candidates (Figure 6.1c). The result of this matching procedure is a list of relevant documents that can be used to link the WJD's annotations to the YouTube videos. The retrieved video is

[1] `https://www.halleonard.com/search/search.action?seriesfeature=OMNIBK`

then embedded in a web-based application (Figure 6.1d). Additionally, we use the annotations provided by the WJD to further enrich the video, e. g., by offering new navigation possibilities based on the song structure or transcriptions of the song's solo. As a result, the user is able to follow the soloist's improvisation in a piano-roll-like representation. For intuition and hands-on experience with this concept, our web-based application can be accessed under the following address:

`http://mir.audiolabs.uni-erlangen.de/jazztube`

The remainder of this chapter is structured as follows. We start by giving a brief overview of the literature and related projects (Section 6.2). Then, we introduce the different data resources used in this study (Section 6.3). Subsequently, we describe the various retrieval procedures which are used to link the WJD to the YouTube videos (Section 6.4). Finally, we present a web-based service which integrates the introduced data resources in a unifying user interface (Section 6.5).

6.2 Related Work

Similar web-based services which aim to enhance the listening experience have been proposed in the past. *Songle*,[2] for instance, lets users explore music from different perspectives [75]. In this web-based service, computational approaches are used to annotate music recordings (including beats, melodic lines, or chords). Afterwards, these generated annotations are presented in a web-based interface. Since the automatically generated annotations may contain errors, the users can correct them or add new ones. The annotations contained in *Songle* can then be used in third-party applications or research projects (e. g., for singing-voice analysis). Another service called *Songrium*,[3] allows users to add lyrics to publicly available videos (e. g., obtained from YouTube). In addition, the lyrics can be visualized and played back along with the linked video similar to karaoke applications. For an overview of other systems by Goto and colleagues, we refer to the literature, see [74, 73].

Another project which aims at enhancing the listening experience, especially for classical music, is called *PHENICX* (Performances as Highly Enriched aNd Interactive Concert eXperiences) [115, 69, 125, 116]. As one main functionality, suitable visualizations are generated in real-time and displayed during the live performance of an orchestra. Such visualizations may be a rendition of a musical score (score-following applications) or an animation controlled by the baton movements of the orchestra's conductor. Furthermore, as in our scenario, the project offers a web-based service, which allows the playback of enriched videos.[4] In the research project *Freischütz Digital*,

[2] `http://songle.jp`
[3] `http://songrium.jp`
[4] `http://phenicx.prototype.videodock.com`

[5] user interfaces for dealing with critical editions in an opera scenario were developed [161, 151]. In this scenario, an essential step is to link the different sheet music editions with the various existing music recordings. These alignments are then used in special user interfaces which may support musicologists in their work on critical editions.

Besides publicly available music recordings or videos, the internet offers additional information (metadata or textual annotations) for music recordings. Many services offer metadata in a structured way, often following standardized data formats as defined in the *Semantic Web* [19]. The Semantic Web contains standardized schemas, called *ontologies*, for exchanging different kinds of data. A way to exchange musical annotations is defined in the *Music Ontology* [157]. One of the most frequently used services in the Semantic Web is *DBpedia*[6] which offers information from Wikipedia in a structured data format. Popular services for music metadata in general are *MusicBrainz*[7] or *Discogs*.[8] In particular for jazz music, the *JDISC*[9] project aims to provide complete discographies for a number of selected artists. Another related project is called *Linked Jazz*,[10] which offers relationships between jazz musicians in a structured way [143]. Beside sharing metadata, researchers have used YouTube as a way of specifying datasets which were used in their experiments [171]. In particular for audio applications, Google released *AudioSet*, a dataset consisting of over two million 10-second sound clips obtained from YouTube which have then been labeled by human annotators [70].

This work follows similar concepts as used in the *SyncPlayer* [108, 44, 186]. The SyncPlayer offers various ways of interacting and navigating with a large, multi-modal corpus of music recordings, sheet music, and lyrics. Furthermore, users are able to search within this corpus by specifying a short melodic phrase or an excerpt from the lyrics. The results are then presented in an interactive graphical user interface which allows auditioning the results. In previous works, we studied the use of interfaces for two different music scenarios. In [12], a web-based user interface motivated by applications in jazz piano education is presented. In particular, a video recording, a piano-roll representation and additional annotations are incorporated in a unifying interface which allows the user to simultaneously play back the different media objects. A related approach focusses on the opera *Die Walküre (The Valkyrie)* from Richard Wagner's cycle *Der Ring des Nibelungen (The Ring of the Nibelung)*. The goal of the interface is to supply intuitive functions that allow a user to easily access and explore all available data (including different recordings, videos, lyrics, sheet music) associated to a large-scale work such as an opera.

[5] `http://www.freischuetz-digital.de`
[6] `http://www.dbpedia.org`
[7] `https://www.musicbrainz.org`
[8] `https://www.discogs.com`
[9] `http://jdisc.columbia.edu`
[10] `https://www.linkedjazz.org`

Abbr.	Instrument	#Solos
cl	Clarinet	15
bcl	Bass clarinet	2
ss	Soprano saxophone	23
as	Alto saxophone	80
ts	Tenor saxophone	157
ts-c	Tenor saxophone in C	1
bs	Baritone saxophone	11
tp	Trumpet	102
cor	Cornet	15
tb	Trombone	26
g	Guitar	6
p	Piano	6
vib	Vibraphone	12
13		$\sum$ 456

Table 6.1: Solo instruments occurring in the WJD. The first column introduces an abbreviation, whereas the last column indicates the number of solos of the respective instrument.

6.3 Data Resources

In this chapter, we consider jazz-related data of different modality stemming from different resources. We now introduce the Weimar Jazz Database (WJD), the relevant jazz recordings, the streaming platform YouTube from which we obtain videos, and the used web resources for additional metadata.

6.3.1 Weimar Jazz Database (WJD)

The WJD is part of the Jazzomat Research Project,[11] which aims at a better understanding of creative processes in improvisations using computational methods [147]. The WJD comprises 456 (as of July 2017) high-quality solo transcriptions (similar to a piano-roll representation), extracted from 343 tracks taken from 197 different records. The solos are performed by a wide range of renowned jazz musicians in the period from 1925 to 2009 (e. g., Louis Armstrong, Don Byas, or Chris Potter). All solos were manually annotated by musicology and jazz students at the University of Music Franz Liszt Weimar using the *SonicVisualiser* [34]. The annotators had different musical backgrounds but a general familiarity with jazz music, mostly through listening and playing. The produced transcriptions were then inspected with an automated verification procedure which primarily searched for syntactical errors and suspicious annotations, such as beat outliers. In a final step, the transcription were cross-checked by an experienced supervisor

[11] `http://jazzomat.hfm-weimar.de`

and added to the database. Table 6.1 lists the number of solo transcriptions grouped by the 13 different occurring solo instruments. As one might expect for jazz music, the database is biased towards tenor saxophone and trumpet solos, which represent about 56% of the currently available solo transcriptions.

Figure 6.2a shows the distribution of the solos with respect to their durations and recording years. The solos have a minimum duration of 19 s (Steve Coleman's second solo on *Cross-Fade*), a maximum duration of 818 s (John Coltrane's solo on *Impressions*), and an average duration of 107 s. Similarly, Figure 6.2b indicates the distribution of the whole tracks (which usually contain more than a single solo part), with a minimum duration of 128 s, a maximum duration of 1620 s, and an average duration of 354 s. From all 343 tracks, there are 247 tracks with one annotated solo part, 80 tracks with two, 15 with three, and a single track with four annotated solo parts. Summing over the number of annotated note events in all solo transcriptions results in over 200 000 elements.[12]

Figure 6.3 displays the beginning of *Clifford Brown's* solo on *Jordu* as an example for the data contained in the WJD. Figure 6.3a shows a time–frequency representation (see Section 6.3.2 for details) of this excerpt superimposed by the available solo transcriptions (each note is represented by a red rectangle) and measure positions (represented as blue vertical lines). Figure 6.3b shows a sheet music representation derived from the solo annotations. Note that deriving sheet music from the transcriptions requires algorithms which are able to quantize the onsets and durations of the annotated note events into musically meaningful notes, see [67]. [13]

6.3.2 Jazz Recordings

A typical jazz recording consists of a soloist who is accompanied by a rhythm section (e. g., double bass, piano, and drums). From an engineering perspective, such a recording is a sequence of amplitude values sampled from a microphone signal (or a mixture of multiple signals). By applying digital signal processing methods, one can analyze and manipulate such signals. A common way to analyze music signals is to transform them into a time–frequency representation, e. g., a spectrogram. For example in Figure 6.2a, we show an excerpt of a spectrogram from *Clifford Brown's* solo on *Jordu*. There exist different approaches to obtain such a time–frequency representation.[14] In particular, we use a logarithmically-spaced frequency axis with a bandwidth of a single semitone per frequency band (row in the spectrogram)—motivated by human's logarithmic perception of frequency and the equal-tempered scale underlying the music. In this representation, one can locate note onsets and durations, as well as harmonic partials generated

[12] Additional statistics: `http://mir.audiolabs.uni-erlangen.de/wjd_web/statistics/`

[13] The sheet music representation was generated by using the *LilyPond* (`http://www.lilypond.org/`) export which can be obtained from the WJD by using the *MeloSpyGUI* (`http://jazzomat.hfm-weimar.de/download/download.html#download-melospygui`).

[14] For the example in Figure 6.2a, we use the semitone filterbank described in [130, 133].

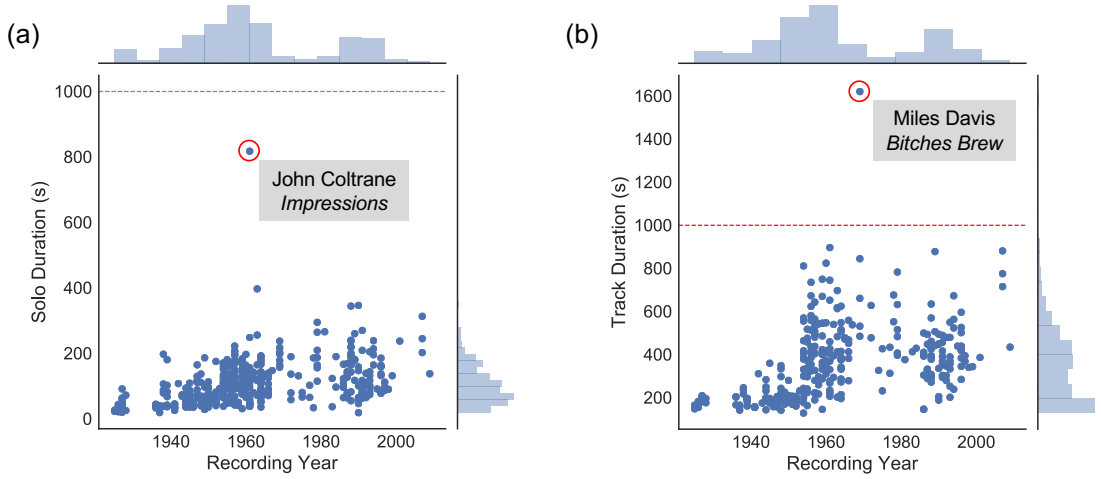


Figure 6.2: (a) 456 solos and (b) 343 tracks considered in the WJD, represented according to their duration and respective recording years. The dashed lines indicate the limits for the maximum duration of the considered YouTube videos.

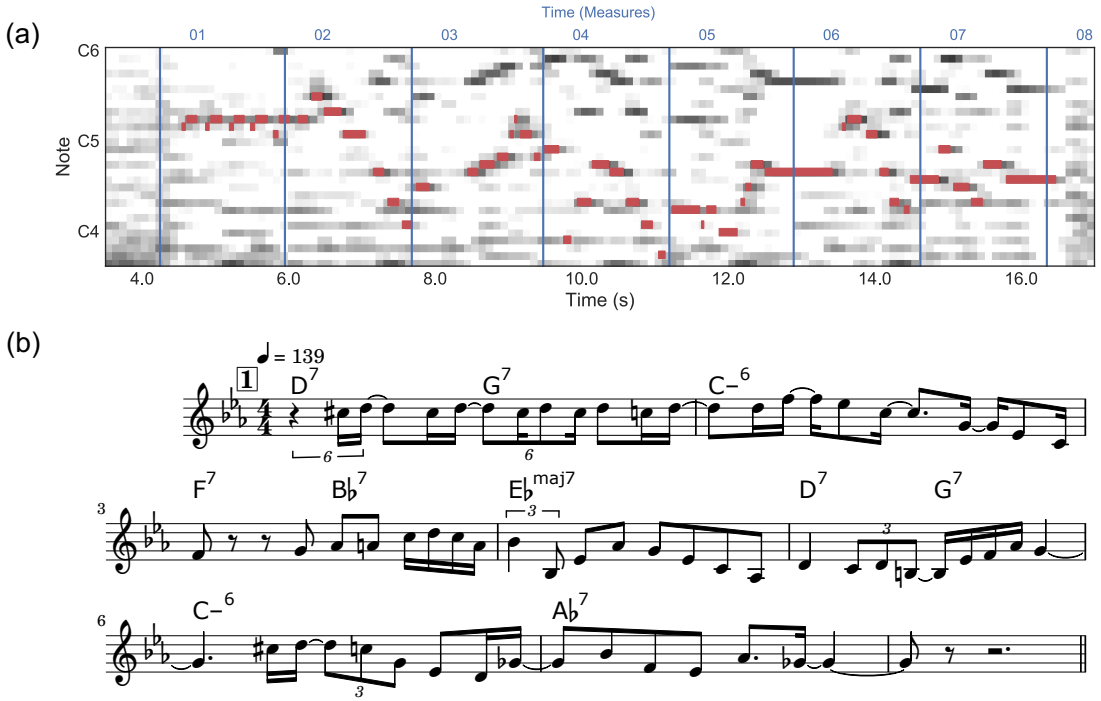


Figure 6.3: Beginning s of *Clifford Brown's* solo on *Jordu*. (a) Log-frequency spectrogram, where rectangles (red) indicate the solo transcription and vertical lines (blue) the annotated measure positions. (b) Sheet music representation of the transcribed solo.

by the sounding instruments. For an overview of computational approaches and music processing in general, we refer to the literature, e. g., [131, 195, 100].

6.3.3 Videos

There exist many different web services which offer users to publish videos. Among these services, YouTube[15] is without doubt the largest and most famous platform for video sharing. For our scenario, we are particularly interested in YouTube videos that contain music—especially the music that underlies the WJD. Some of the offered music videos are official releases by record labels, but the majority are videos uploaded by private platform users. Especially the music videos uploaded by the private users often contain only a static image (cover art) or a slideshow while the audio track is a digitized version of the commercially available record. By embedding YouTube videos in a web service, one relies on the availability of these videos. Due to user deletions, copyright infringements, or legal constraints in some countries, videos may not be available. However, YouTube has a lot of redundancy, i. e., the same music recording may be available in more than one version.

6.3.4 Additional Metadata

In addition to the solo transcriptions, the WJD contains basic metadata for the music recordings (e. g., artist and record name), as well as a special identifier for the MusicBrainz[16] platform. MusicBrainz is a community-driven platform, which collects music metadata and makes it publicly available. With the identifier available in the WJD, one is able to request a comprehensive list of available metadata from the MusicBrainz platform (e. g., participating musicians, producer's name, and so on). Furthermore, MusicBrainz can serve as a gateway to other web services which offer different kinds of metadata or even other multimedia objects (e. g., pictures of the artist). This "web of data" is often referred to as the Semantic Web [19]. By using the web service DBpedia,[17] we can furthermore obtain and integrate content published on Wikipedia (e. g., bibliographic information about the artist).

6.4 Retrieval and Linking Strategies

In this section, we report on experiments where we systematically created links between the annotations contained in the WJD and corresponding YouTube videos. The retrieval task is as follows: Given a specific music recording or a solo annotation provided by the WJD as query, identify the relevant videos in the pool of YouTube videos. Since the number of YouTube videos is very large, we follow a two-step retrieval strategy which we describe in the following.

[15] `http://www.youtube.com`
[16] `http://www.musicbrainz.org/`
[17] `http://wiki.dbpedia.org`

6.4.1 Retrieval Scenario

We started by formalizing our retrieval task following [131]. Let $\mathcal{D}$ be the set of all documents available on YouTube. A YouTube document $D \in \mathcal{D}$ consists of the video and the available metadata. Let $\mathcal{Q}$ be a collection of documents available within the WJD. A WJD document $Q \in \mathcal{Q}$ consists of a solo annotation, the underlying music excerpts, as well as metadata. In our scenario, the document Q served as query, whereas $\mathcal{D}$ was the database to search in. Given a query Q, the retrieval task was to identify the corresponding documents D. In our scenario, we followed a two-step retrieval strategy. First, we performed a metadata-based retrieval using the YouTube search engine. For a query Q, the result of the text-based retrieval is denoted as $\mathcal{D}_Q^{\text{Text}} \subset \mathcal{D}$. In the second step, we performed content-based retrieval only based on $\mathcal{D}_Q^{\text{Text}}$ to identify the relevant documents, denoted as $\mathcal{D}_Q^{\text{Rel}} \subseteq \mathcal{D}_Q^{\text{Text}}$.

6.4.2 Text-Based Retrieval

In the first step of our retrieval strategy, we extracted a subset of possible video candidates from YouTube. These were retrieved by performing two text-based queries (per solo) using the standard YouTube search engine (using YouTube's default settings). The first text-based query term consisted of the name of the soloist and the song title (e. g., `John Coltrane Kind of Blue`). Since the soloist is not always the artist who released the record, we performed a second text-based query which consisted of the artist's name under which the record was released, followed by the song title (in our example: `Miles Davis Kind of Blue`). From each retrieval result, we took the top 20 candidates (or less, depending on the number of YouTube search results). Furthermore, we only considered videos which are shorter or equal to 1000 s to avoid videos where users uploaded, for instance, complete records to YouTube (rather than individual songs).

In our experiments, using the first text-based query terms for all 456 solos considered in the WJD led to a pool of 4114 video candidates. The second text-based query resulted in a pool of 4069. In a next step, we fused the two candidate pools together, where we removed duplicates by using the video identifiers attached to every YouTube video. Our final candidate pool comprised 5199 video candidates—resulting in approximately 12 candidates per query (solo). Note that $\mathcal{D}_Q^{\text{Meta}}$ may still contain cover versions and other irrelevant documents. The following audio-based retrieval step is intended to resolve this issue.

6.4.3 Audio-Based Retrieval

In a second step, we used the audio recordings from the WJD to refine the list of candidates $\mathcal{D}_Q^{\text{Meta}}$ obtained from the text-based retrieval. This task is also known as *audio identification* and

can be approached in many different ways, see e. g., [36, 131, 2]. Our method is based on chroma features and diagonal matching which is easily extendable to retrieval scenarios with different query and database modalities (e. g., matching solo transcription against audio recordings or matching audio excerpt against sheet music representations). Furthermore, since we performed our retrieval only on the small subsets $\mathcal{D}_Q^{\text{Meta}}$, we do not consider efficiency issues here. In particular, we used a chroma variant called CENS with a feature rate of 5 Hz [134, 133].[18] We compared a query Q with each of the documents $D \in \mathcal{D}_Q^{\text{Meta}}$ by using diagonal matching. This comparison yields a distance value $\delta_{Q,D} \in [0:1]$ for each pair (Q, D), where $\delta_{Q,D} = 0$ refers to a perfect match and $\delta_{Q,D} = 1.0$ to a poor match. By sorting the documents $D \in \mathcal{D}_Q^{\text{Meta}}$ by $\delta_{Q,D}$ in an ascending order, one receives a ranked list. In this ranked list, the most similar documents (w.r.t. to the used distance function) are listed on top. In the case of extracting the relevant documents, one has to further process this ranked list. For instance, one may mark a document as relevant if $\delta_{Q,D}$ is smaller than a threshold $\tau \in [0:1]$. All relevant documents that fulfill this condition are then collected in the subset $\mathcal{D}_Q^{\text{Rel}} \subseteq \mathcal{D}_Q^{\text{Meta}}$.

Using this retrieval approach with a threshold $\tau = 0.1$, we were able to identify 988 relevant videos for 329 solos on YouTube (on average 3 relevant videos per solo, min = 1, max = 9). For 92 queries, we retrieved 1 relevant document, for 67 queries 2, for 60 queries 3, and for 110 queries more than 3 documents. However, for 124 queries, we were not able to find any relevant videos. We found different reasons for this from manually inspecting some candidate lists. One obvious reason is that the metadata-based retrieval step did not return any relevant documents in $\mathcal{D}_Q^{\text{Meta}}$ (e. g., for the textual `David Murray Ask me Now`). Sometimes, only other versions of the same song are available on YouTube, for instance, the textual query `Art Pepper Anthropology` yields mainly results for the version of this song from the record *Art Pepper + Eleven: Modern Jazz Classics* instead of the relevant version from the record *The Intimate Art Pepper*. Furthermore, in many instances, we found that relevant documents were present in $\mathcal{D}_Q^{\text{Meta}}$, but not recognized since the distance value $\delta_{Q,D}$ surpassed the chosen threshold $\tau = 0.1$ by a small margin.

6.4.4 Solo-Based Retrieval

The previous experiments were based on the assumption that we have access to the music recordings underlying the WJD annotations. However, in certain scenarios this might not be the case, for instance, when only a score representation of the piece or the solo is available. In this case, audio identification is no longer possible and one needs more general retrieval strategies. In the following experiment, we simulate a retrieval scenario by using the WJD's solo transcriptions (see Figure 6.3a) as query and convert them to chroma features. This constitutes a challenging retrieval task, where one needs to compare monophonic queries (the solo transcriptions) against polyphonic audio mixtures (music recordings contained in the YouTube videos).

[18] All computations can be done by using the implementations provided by the Python library *librosa* [124].

K	1	3	5	10	15	20
Top-K	0.85	0.97	0.99	1.00	1.00	1.00

Table 6.2: Top-K matching rate for the solo-based retrieval. The Top-K matching rate is calculated by dividing the Top-K matches by the 329 retrieved solos from the audio-based retrieval.

In a first experiment for this advanced retrieval task, we took the same list of candidates $\mathcal{D}_Q^{\text{Meta}}$ and parameters as used in Section 6.4.3 and only exchanged the audio-based queries against solo-based queries. In order to evaluate the results, we took the results $\mathcal{D}_Q^{\text{Rel}}$ from the audio-based retrieval as reference. The solo-based retrieval is considered as correct if among the top K documents in the ranked list, there is at least one relevant document. The results for this Top-K evaluation measure are shown in Table 6.2.

We retrieve for 85% of the queries a relevant document at rank 1 (Top-1). For 99% of the queries, the first relevant document is within the Top-5 matches. The mean reciprocal rank for the first matches for all queries is 0.91 ($\sigma = 0.22$) Although the solos and audio recordings vary in their degree of polyphony, we reach respectable results. The main reason is that the solo transcriptions are relatively long and perfectly aligned, leading to a high "discriminative power". Furthermore, the queries are very unique, since they stem from an improvisation. When the queries get shorter, usually the discriminative power decreases rapidly, as they may represent more frequently used patterns.

6.4.5 Perspectives

So far, our approach for retrieving videos from YouTube relies on either audio recordings or very clean solo transcriptions taken as queries. This is exactly the situation we had in our WJD scenario. In other scenarios, one may have to deal with imperfect or less specific queries. For instance, a query might be a person humming the solo, which then requires an extra step for extracting the fundamental frequency from the hummed melody (query-by-humming), see e. g., [162, 145, 166]. Furthermore, the query may have a different tuning, may be transposed to another key, or played with rhythmic variations. A possible solution could be to use multiple queries, e. g., transposing the query in all possible 12 keys and performing a separate retrieval for each resulting version. Another way of handling some of these issues is to use different feature representations, e. g., features that robust against temporal deformations, see [2, 181, 180].

In a related retrieval scenario, described in [11], audio recordings were retrieved from a database containing Western classical music recordings by using monophonic queries with a duration of only a few measures. Besides the discrepancy in the degree of polyphony between query and database documents, tuning, key, and tempo deviations, which frequently occur in Western classical music

performances, make this retrieval task very challenging. A common preprocessing step, which targets the "polyphony gap" between query and database document, is to enhance the predominant melody in audio recordings. In [166], the authors used a so-called *salience representation* in a query-by-humming system which led to a substantial increase in performance [164]. In [13], a data-driven approach is used to estimate a salience representation for jazz music recordings which showed a similar performance as the aforementioned, salience-based method.

Another untapped resource for jazz music retrieval are the many publicly available solo transcriptions. However, these transcriptions are typically not available in a machine-readable format. In this case, one could use Optical Music Recognition (OMR) systems to convert sheet music images to symbolic music representations. This conversion, however, may introduce severe errors, such as missing notes, wrongly detected clefs, key signatures, or accidentals, see [18, 33, 66, 159, 158, 11]. A recently proposed approach for score following tries to circumvent the difficult OMR step by directly working on the scanned images of the sheet music [50]. Two Convolutional Neural Networks (CNN)—one applied to the sheet music, a second one to the audio recordings—are used for feature extraction. In an extra layer, these features are then combined to retrieve temporal relationships between the two modalities, for instance with a learned embedding space [153, 51]. Currently, new OMR approaches based on deep neural networks show promising results and may lead to a significant increase in conversion quality [81].

6.5 Application

In this section, we present the functionalities of our web-based application, called *JazzTube*, which allows users to easily access the WJD's annotations, as well as the corresponding YouTube videos, in different interactive ways. The application offers various ways to access the WJD. First, tables of the compositions, soloists, and transcribed solos contained in the WJD are in given in form of suitable tables. Furthermore, one can access the information on the record, the track, and at the solo level.

6.5.1 Solo View

Figure 6.4 shows a screenshot of the core functionality of our interactive, web-based user interface. In the top panel, some general information about the solo (Figure 6.4a) is shown. Many of these entries are hyperlinks and lead to the artist's overview page or the corresponding track. Furthermore, several possibilities of exporting the solo transcription, either as comma-separated values (CSV), or as sheet music, are offered. The conversion from the annotations to the sheet music is obtained by using the algorithm described in [147]. Below this basic information, all available YouTube videos are listed (Figure 6.4b). Having more than one match gives alternatives

to the user. Note that YouTube videos may have different recording qualities or may disappear from YouTube. After pressing the play button, the corresponding YouTube video is automatically retrieved and embedded in the website (Figure 6.4c). Below the YouTube player, a piano-roll representation of the solo transcription is presented running synchronously with the video playback (Figure 6.4d). Finally at the bottom, additional statistics about the solo (e. g., pitch histograms) are provided (Figure 6.4e).

6.5.2 Soloist View

Starting from an overview table of available soloists, the user can navigate to the soloist view containing additional details about the artist. Here, one can also find the available solo transcriptions for the given soloist. Furthermore, *Semantic Web* technologies are used to perform a search query on *DBpedia* to retrieve further details. Usually the received response is very rich in information. Currently, a short biography and a link to the corresponding *Wikipedia* entry for further reading are included. In addition to the biographical data, further relationships to other artists, obtained from the *LinkedJazz* project, are embedded.

6.5.3 Technical Details

Our web-based demonstrator is a typical client-server application. The client uses the Hypertext Transfer Protocol (HTTP) to perform requests to the server (e. g., by entering an URL through a web browser). These requests are then processed by the server and the response is displayed in the user's web browser. For setting the layout, we use the open-source framework *Bootstrap*[19]. This framework allows for designing a website for different devices (e. g., laptops, tablets, or smartphones). Interactions and animations within the client are realized with *JavaScript*.[20] In particular, a framework called *D3 (Data-Driven Documents)*[21] for visualizing the piano-roll is employed. For the server backend, the Python framework *Flask* is used.[22]

6.5.4 Possible Advancements for JazzTube

In the case of the Weimar Jazz Database, looking at the scrolling piano roll visualization of a jazz improvisation while simultaneously listening to the recording could be both of high educational value and a great pleasure. To relate sounding music to moving pitch contours, rhythms, changing event densities, and recurring or contrasting motifs and patterns, which are easily recognizable from a piano roll visualization, can enrich and deepen the understanding of the tonal, rhythmical,

[19] `http://www.getbootstrap.com`
[20] `https://www.javascript.com`
[21] `https://www.d3js.org`
[22] `http://flask.pocoo.org`

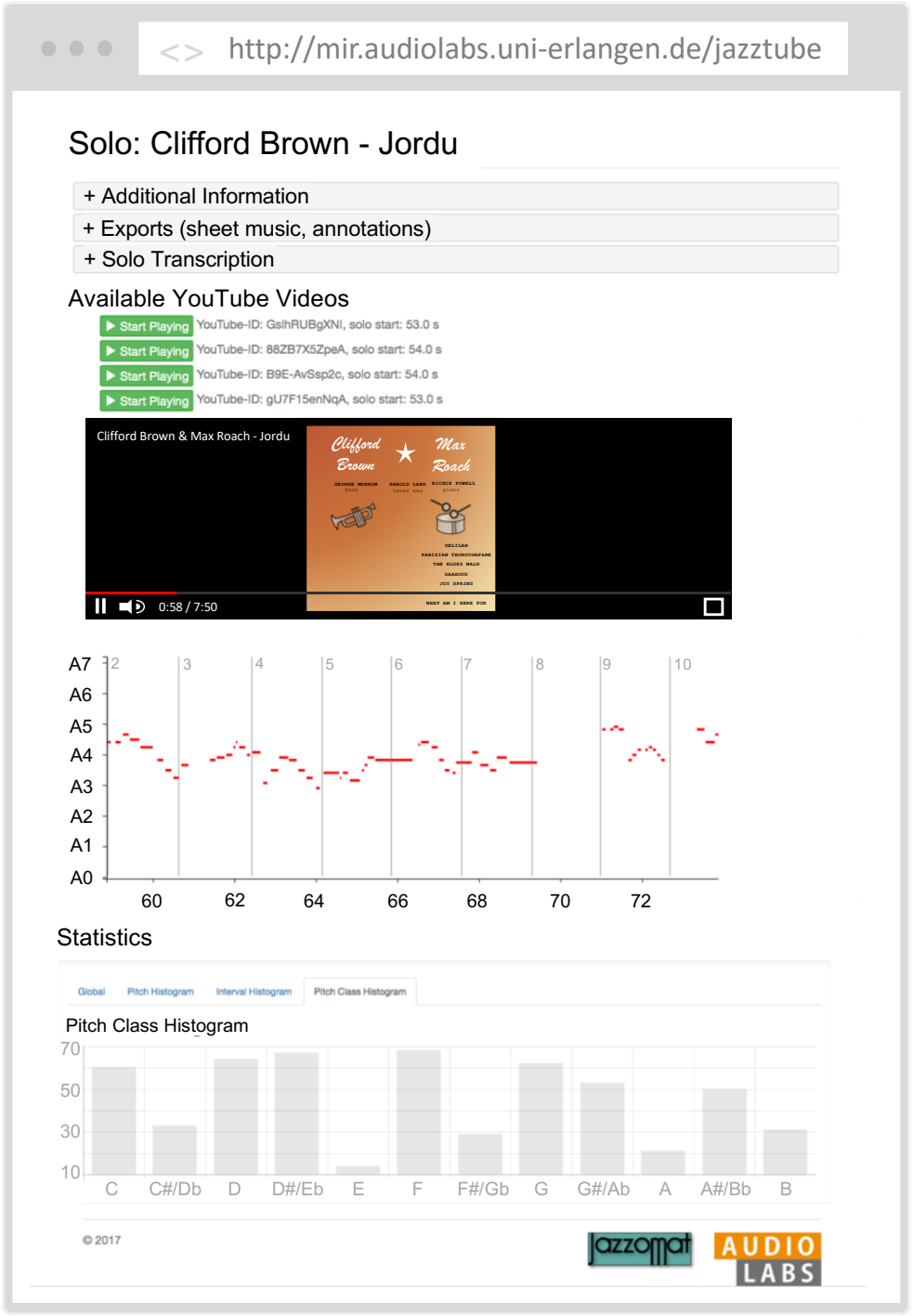
and formal dimensions of the music in an inimitable way. Moreover, recognizing musical passages visually immediately before listening to the sounds can contribute to the play with musical expectancies (or "sweet anticipation", [93]) which lie at the heart of the pleasures of listening to music.

For the future, several extensions to the current form of *JazzTube* are desirable. The piano roll representation could be extended with different layers of annotations, such as phrases, midlevel units, chords, choruses, form part, or tone formation, which are already available in the WJD. Coloring or annotating events with respect to different functions, e. g., roots of underlying chords, passing tones, or melodic accents, would give an even deeper insight in the inner structure of an improvisation. In the case of jazz, automated identification and annotation of patterns and licks would provide options for analysis not easily achievable with traditional paper and pencil tools. Furthermore, retrieving patterns or motifs from the database would be of great value, for example, by selecting a few tones in a solo and finding and displaying all cross-references in the corpus. Finally, adding options of score-following would be of great help, since music notation is still the standard communication and representation tools of musicians and musicologists. On a different footing, the vast educational implications could be further exploited by adding specialized display options or specifically designed course materials and tutorials (e. g., on jazz history) based on the contents and possibilities of *JazzTube*.

6.6 Conclusions

With *JazzTube*, we offer researchers and music lovers novel possibilities to interact with and navigate through the content of the WJD. With JazzTube's innovative approach to link scientific music databases including metadata, transcriptions and further annotations to the corresponding audio recordings that are publicly available via YouTube, copyright restrictions can be bypassed in an elegant way. However, there is no doubt that both musicians and composers should be gratified financially for the music they create according to national and international copyright and ancillary copyright laws. YouTube seems to guarantee this financial entitlement through agreements with national copyright and performance protection associations. In contrast, for scientific institutions offering music databases it is very difficult or impossible to handle these legal claims. Therefore, the approach of *JazzTube* could open up a way for music projects to connect metadata and annotations with audio recordings that can not be freely provided on the internet but can be used for searching for the corresponding audio recordings at YouTube. This could be a way to easily link, e. g., the recording metadata provided within the JDISC[23] project with YouTube recordings. Furthermore, we envision that *JazzTube* is a source for inspiration and fosters the necessary dialogue between musicologists and computer scientists.

[23] `http://jdisc.columbia.edu`



(a) (b) (c) (d) (e)

Figure 6.4: Screenshot of our web-based interface called *JazzTube*. (a) Metadata and export functionalities. (b) List of linked YouTube videos. (c) Embedded YouTube video. (d) Piano-roll representation of the solo transcription synchronized with the YouTube video. (e) Additional statistics.

Chapter 7

Opera as a Multimedia Scenario: Wagner's Valkyries Go Online

This chapter closely follows the results presented in [14].

Music—with its many representations—can be seen as a multimedia scenario: There are a number of media objects (e. g., video recordings, lyrics, or sheet music) beside the actual music recording, which describe the music in different ways. Through digitization efforts, many of these media objects are now publicly available on the Internet. However, the media objects are usually accessed individually, without using their musical relationships. Harnessing these relationships could open up new ways of navigating and interacting with the music, or extending existing media objects with additional metadata. In this chapter, we model these relationships with a suitable database model by taking Richard Wagner's opera *Die Walküre* as a case study. Based on this model, we present a web-based demonstrator which combines the interconnections between the media objects and allows users to access the data through a graphical user interface.

7.1 Introduction

Operas are an essential part of concert programs worldwide. They combine elements from theater, singing, and orchestral music to form a complete artwork. Many famous composers wrote operas such as Giuseppe Verdi, Giacomo Puccini, Georges Bizet, Wolfgang Amadeus Mozart, or Richard Wagner. First and foremost, Wagner is well-known for his opera cycle *Der Ring des Nibelungen*. This cycle consists of the four operas *Das Rheingold*, *Die Walküre*, *Siegfried*, and *Götterdämmerung*. Depending on the performance, the total duration of all operas combined is around 16 hours. The corresponding libretto (text of an opera) and the sheet music fill hundreds of pages. Today, much of this multimedia data is available on the Internet; for

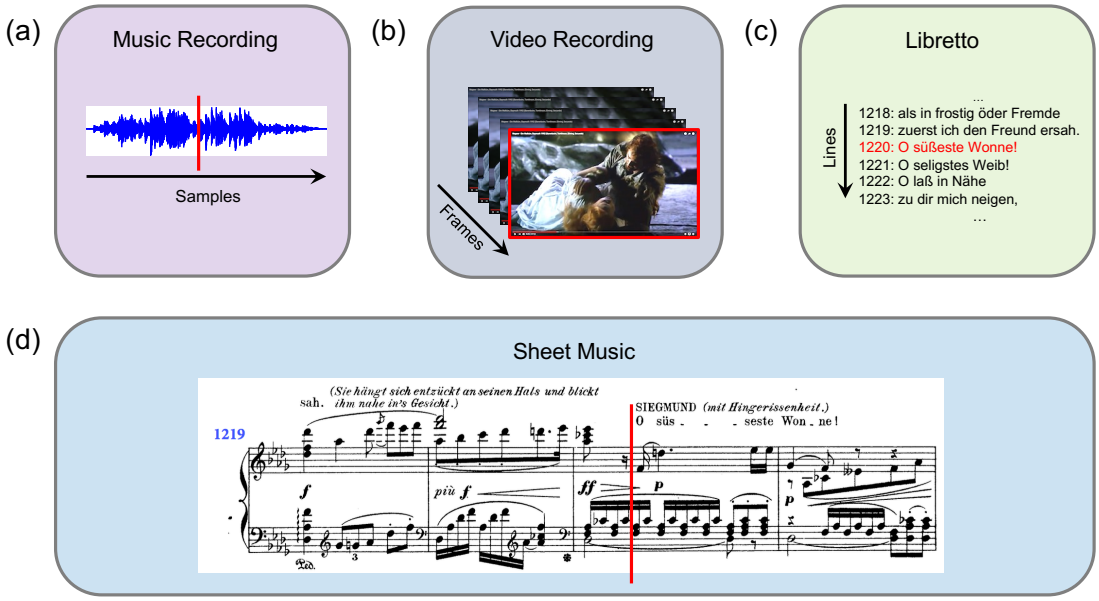


Figure 7.1: Overview of the different media objects in the context of an opera. (a) Music recording. (b) Video recording. (c) Libretto. (d) Scanned sheet music. The red annotations indicate the relationships of the individual media objects. Once these annotations exist, they allow simultaneous navigation across the media objects.

instance, recordings of the ring cycle are publicly available and can be found on video platforms such as YouTube. The sheet music, as well as the libretto, are distributed on platforms such as the International Music Score Library Project (IMSLP).[1] A central task is to unveil the existing musical relationships between media objects with the help of computational approaches, and make the results accessible for the user. Our vision comprises a user interface which allows access to a particular video recording (e. g., on Youtube), which is then automatically aligned to the available sheet music, enriched with the libretto, and linked to other available resources about the musical work or the composer.

The main contributions in this chapter are the modeling of the opera as a multimedia scenario and the enrichment with publicly available video recordings by using a web-based user interface. Flawlessly implementing our design for all genres is a very challenging task which is not yet solved. To keep the system's complexity feasible, we focus in the following on the opera scenario, which comprises media objects of many types, see Figure 7.1 for an example which is taken from Wagner's opera *Die Walküre*. All different media objects—audio and video recording, the opera's libretto, and the sheet music—describe the same opera from different perspectives. A trivial linking of these media objects is impossible due to their different types.

The remainder of this chapter is structured as follows: In Section 7.2, we first present related concepts and existing systems from the literature. In Section 7.3, we take the opera *Die Walküre*

[1] `http://www.imslp.org`

as an example to explain the available media objects and a way to link them. Based on this linking concept, we introduce and discuss a relational database schema. In Section 7.4, we present a prototypical web-based demonstrator which allows for a simultaneous access and navigation within the linked media objects.

7.2 Related Work

Many research contributions deal with similar approaches for navigation across different media objects. In the following, we present a number of central works from this research area. This work is closely related to the concepts of the *SyncPlayer* [44]. In this system, the user can search a music database by using various query types (e. g., a short melody or a line from the lyrics), which are then processed by a server. The results from this search are presented and played back in a graphical user interface. The possibility to interact and navigate within the sheet music representation is especially useful for musicologists. Similar user interfaces were developed in the project *Freischütz Digital* for dealing with critical editions in an opera scenario [161, 198]. In this project, a core task is to link the existing music recordings with the different editions of the musical score. Based on the established links, several user interfaces were developed to assist musicologists in creating a critical edition.

Another project which aims at enhancing the listening experience, especially for classical music, is called *PHENICX* (Performances as Highly Enriched aNd Interactive Concert eXperiences) [115, 69]. As one main functionality, suitable visualizations are generated in real-time and displayed during the live performance of an orchestra. Such visualizations may be a rendition of a musical score (score-following applications) or an animation controlled by the baton movements of the orchestra's conductor.

Computational approaches for synchronizing music recordings with the corresponding sheet music is an important task in Music Information Retrieval (MIR). Besides the music synchronization task, there exist other tasks (e. g., content-based retrieval, music structure analysis, or audio source separation). For an overview, we refer to the literature (e. g., [131, 195, 114, 129, 196, 100, 136]).

7.3 Case Study: *Die Walküre* by Richard Wagner

In this section, we take the opera *Die Walküre* by Richard Wagner as an example to explain different media objects and discuss their interactions. Inspired by [126], we first describe the media objects which play a role in our scenario. With the help of a relational database schema, we present the properties of the media objects and their interconnections. Finally, based on this database schema, we develop a web-based demonstrator which shows possibilities for flexible

Media Object	Media type	Elements	Divison of Time
Music recording	Audio	Sequence of audio samples	Time axis (samples/seconds)
Video recording	Video	Sequence of images	Time axis (frames/seconds)
Libretto	Text	Sequence of strings	Lines
Sheet music	Image	Sequence of images	Pages

Table 7.1: Overview of media objects in an opera scenario. In addition, the media type is given, as well as the elements of the media objects and their division of time.

navigation and audiovisual representations for the various media objects that are related to the musical work at hand.

7.3.1 Presenting the Media Objects

An opera consists of different media objects. In this chapter, we consider typical media objects which are shown in Figure 7.1. These are either specific to a particular performance (e. g., music and video recordings) or specific to a particular musical work (e. g., libretto and sheet music). In our model, we assume that each media object is a sequence of elements which are naturally ordered w.r.t. time (i. e., incrementing physical time corresponds to the next element in the sequence). In Table 7.1, we present these media objects together with a description of their elements and the respective division of time.

Music Recording: Microphones can be used to capture sound waves during musical performances. In the case of digital music recordings, the amplitudes of these sound waves are sampled periodically and saved as a sequence. These amplitude values are also known as *samples*. A typical sampling rate for compact disc (CD) recordings is 44 100 samples per second. Navigation within a music recording is done via a physical time axis (given in seconds).

Video Recording: Video recordings consist of a sequence of images, called *frames*. A common frame rate is 30 frames per second. The navigation within video recordings is performed—similarly to music recordings—using a physical time axis (given in seconds).

Libretto: The libretto is the spoken and sung text in an opera. In our scenario, the libretto consists of a sequence of text lines which can be accessed through a line number.

Sheet Music: Sheet music is, in our scenario, a sequence of scanned sheet music pages. A single page usually consists of 30 measures and can be accessed with its page number.

Bringing these different media objects together in a unified framework is a challenging multimedia task. Besides the different media types, a central requirement for our demonstrator's functionality is to establish a link across the various time axes.

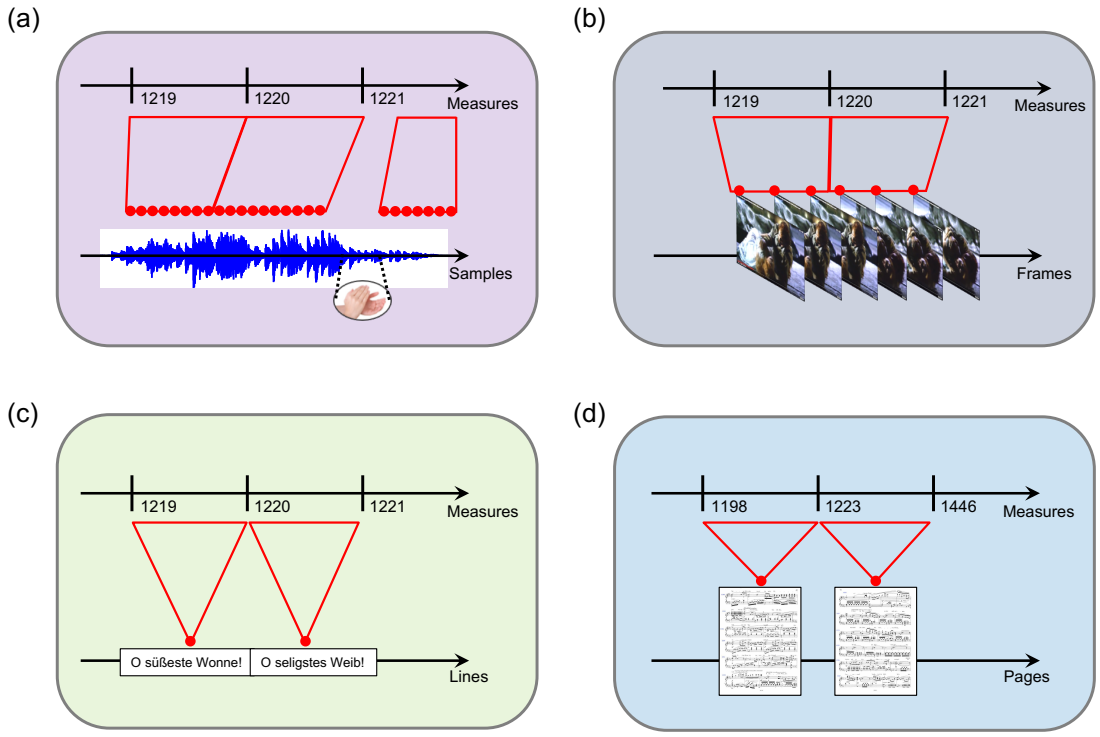


Figure 7.2: The media objects are linked through region annotations on a reference axis (given in measures). Music (a) and video recording (b) allow for a precise projection onto the reference axis due to their fine temporal resolutions. The libretto (c) is projected line-wise and the scanned sheet music (d) is projected page-wise onto the reference axis.

7.3.2 Linking the Multimedia Objects

The media objects consist of sequences of discrete elements (samples, frames, text lines, pages of sheet music) which have a temporal relationship with each other. This relationship can be modeled with an abstract reference axis. In contrast to the discrete time axis of the existing elements, we consider a continuous time axis. Points in time $t \in \mathbb{R}$ on this axis are given in musical measures. Measure beginnings are encoded with whole numbers $\mathbb{N} \subseteq \mathbb{R}$ (e. g., $t = 3$ encodes the beginning of measure three). All positions within measures are encoded with decimals (e. g., $t = 3.5$ references to the middle of the third measure). In practice, we round the positions on the continuous reference axis to a fixed number of three digits. The beginning of measure 3 is encoded as 3.000, the middle as 3.500 and the end of this measure as 3.999. Furthermore, we assume that a temporal resolution on the measure level is sufficient for our later applications. In the case that a measure position is requested at a finer level than the temporal resolution (e. g., measure position 3.750), we use linear interpolation to calculate the corresponding time positions.

Figure 7.2 gives a schematic overview of the media objects and their projections onto the reference

axis in our opera scenario. In general, we map subsequences of elements to suitable regions on the reference axis. Figure 7.2a illustrates this concept for a music recording. A subsequence of audio samples is assigned to the measure region 1219.000–1219.999 on the reference axis. A big advantage of the region mapping is that we can skip unnecessary elements within the media objects (e. g., applause or silence at the beginning or end of a music recording). Figure 7.2b shows the annotations of a video recording where subsequences of frames are mapped to musically corresponding regions on the reference axis. The libretto in Figure 7.2c is mapped line-wise onto the reference axis. In contrast to the music and video recordings, the libretto consists of significantly fewer elements—a line in the libretto contains the text for several measures. In our scenario, the scanned music has the coarsest temporal resolution (approx. 30 measure per page), see Figure 7.2d.

This concept of projecting subsequences of discrete elements on a shared, continuous reference axis allows us to handle different kinds of requests based on measure regions for all media objects. These requests can be efficiently processed with a relational database schema, which we will present in the following section.

7.3.3 Database Schema

A common starting point when designing a database schema is to outline the scenario which should be covered by the database (the so-called *miniworld* [58]). In our miniworld, we want to model the media objects in an opera scenario and their corresponding relationships. Typically, operas are addressed by a work title and a composer (e. g., *Die Walküre* by Richard Wagner). Many operas are subdivided into smaller parts, for instance *Die Walküre* is divided into three acts. Opera performances are usually identified with a combination of year, place, and conductor (e. g., 1992, Bayreuth, Daniel Barenboim). A performance has many media objects which can be linked with a reference axis (given in measures).

Figure 7.3 shows one possible model for our miniworld as an entity-relationship diagram (ER-diagram in Chen-Notation [58]). Our model starts with the description of a musical work by defining the entity type `WORK`. For the sake of simplicity, we consider the acts of an opera as separate works (similar to the movements of a symphony or dances in a suite). The act of an opera is therefore an entity of the type `WORK` which is identified by the attribute `WorkID`. The `WorkID` for the first act in the opera *Die Walküre* is encoded as *WWV086B-1*, whereas WWV stands for *Wagner-Werk-Verzeichnis* (catalogue of Wagner's works). In this catalogue, *86B-1* describes the first act of the opera *Die Walküre*. In this scenario, we furthermore assume that additional metadata can be obtained from an existing *Metadata DB* which offers structured data using existing schemas. For instance, a well-known service which offers metadata for music recordings

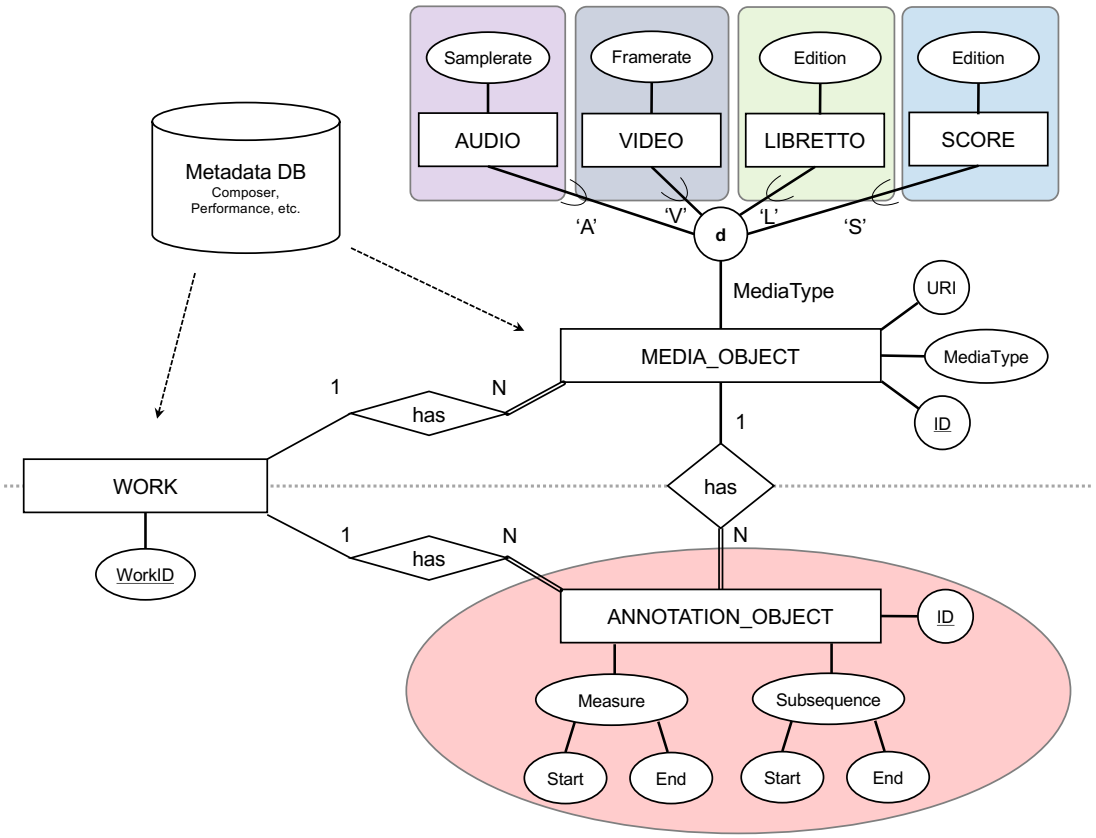


Figure 7.3: Entity-Relationship diagram of the conceptual database schema (notation follows [58]).

is the open music encyclopedia MusicBrainz[2]. Another resource for additional metadata is the Semantic Web[3] which offers special schemas for music such as the *Music Ontology* [157]. The *Music Ontology* can be used to obtain structured data from web services, but at the same time allows users to share their data with other services.

Our opera scenario consists of various interconnected media objects. We model these media objects with the abstract entity type `MEDIA_OBJECT`. Each media object constitutes a realization of an abstract type: `AUDIO`, `VIDEO`, `LIBRETTO`, or `SCORE`. This allows for an easy extension of the database schema with new media types. The attribute `URI` (Uniform Resource Identifier) is used to link to the source data (e. g., a video file on a local hard drive or a website). Each entity is usually connected to several entities of type `ANNOTATION_OBJECT`. The composite attributes `Subsequence` and `Measure` model the mapping of a media object to the reference axis. `Start` and `End` are used to map a sequence of elements to a region on the reference axis; e. g., Figure 7.2a maps the measure region {1219.000, 1219.999} to the region {3408.65, 3411.03} in the music recording. Finally, we add a 1:N relationship between `WORK` and `ANNOTATION_OBJECT` to stress

[2] `https://www.musicbrainz.org`
[3] `https://www.w3.org/standards/semanticweb/`

the connection between works and their annotations.

7.3.4 Adding Musical Works

Adding new musical works requires a mixture of manually creating annotations and using computational methods. The most time-consuming step lies in annotating the start positions of the measures in the music recordings (e. g., by using the publicly available tool *Sonic Visualiser* [35]). In the case of the opera *Die Walküre*, we only created the measure annotations of a single reference recording. By using computational approaches for music synchronization we were able to transfer these annotations to other performances [152]. All further work-specific media objects such as libretto and sheet music were annotated manually and added to the database.

7.4 Web-based Demonstrator

We developed a web-based demonstrator based on the database schema. The demonstrator is meant to present possibilities for interaction and navigation in opera recordings. The core idea of the demonstrator is to use publicly available video recordings (e. g., from the online video platform YouTube) and enrich them with additional information about the musical work. We used the grid-based HTML framework `Bootstrap`[4] to design the user interface. `JavaScript`[5] is used for the user interaction in the browser and `Flask`[6] for the logic on the server side. In the following, we further explain the functionalities of our demonstrator with different recordings from the opera *Die Walküre*.

Figure 7.4 presents a screenshot of the demonstrator where each media object has a corresponding element in the user interface. Figure 7.4a shows the available versions (YouTube videos) for the opera's first act. The accessible measure regions are highlighted in green: *Barenboim* and *Sawallisch* are complete recordings of the first act (measures 0–1524), whereas, for instance, *Levine* only offers a part of the first act (measures 1214–1524). The reference axis and the current measure are shown above the versions. The selected version is indicated with a note icon at the beginning of the version. The corresponding video recording is shown in Figure 7.4b. Figure 7.4c highlights the current line in the libretto. For better readability, we also print the previous and next text line. The sheet music is depicted page-wise, see Figure 7.4d.

Navigating within the opera is possible in various ways. The user can either pick a measure on the reference axis, directly enter a measure in a text box, or jump to a measure in particular version by clicking on the highlighted regions. Switching between versions allows a comparison

[4] `http://www.getbootstrap.com`
[5] `https://www.javascript.com`
[6] `http://flask.pocoo.org`

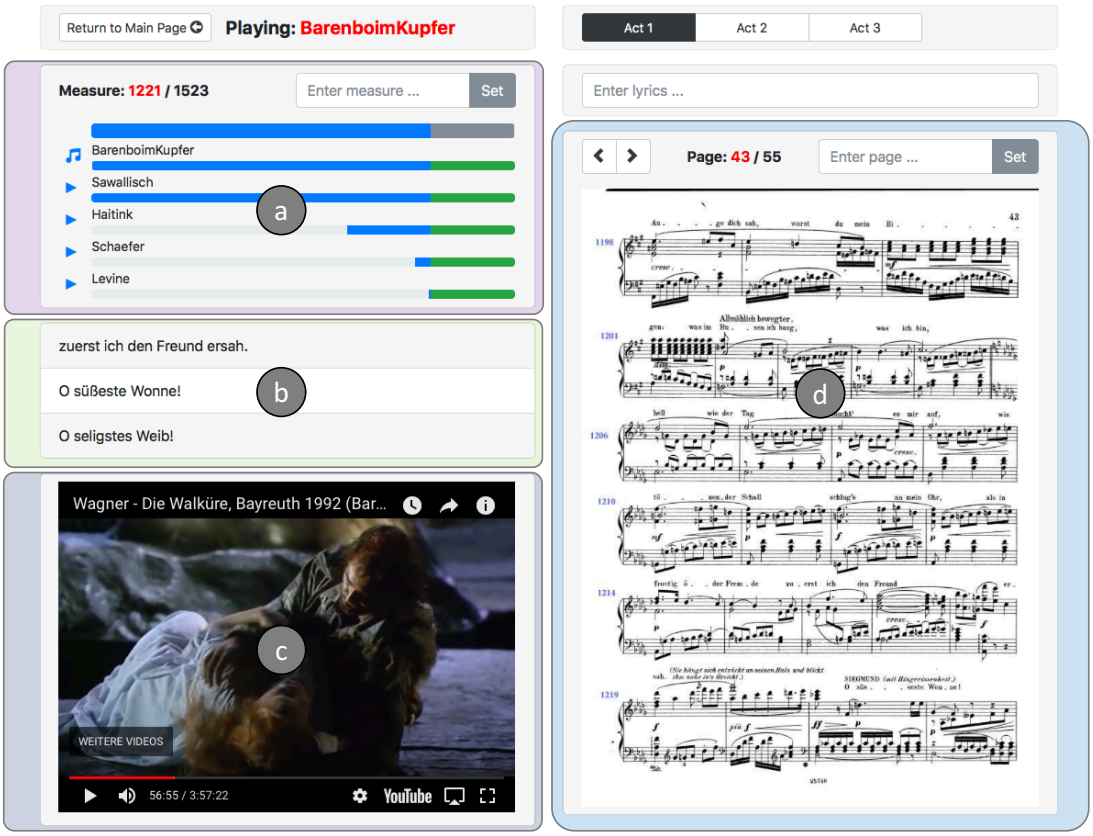


Figure 7.4: Screenshot of the web-based interface which consists of five main elements. (a) Overview of the available music recordings. The available regions within the respective version are color-coded. (b) Integration of the video recording from the online platform YouTube. (c) Textual visualization of the previous, current, and subsequent libretto lines. (d) Scanned sheet music with buttons for page-wise navigation.

of different performances on a measure level (e. g., compare Daniel Barenboim' in Bayreuth in 1992 with Wolfgang Sawallisch's interpretation in the *Bayerische Staatsoper* in 1989). In the sheet music, a page-wise navigation is possible, as well as directly accessing a page.

The web-based demonstrator is available under the following address:

`http://mir.audiolabs.uni-erlangen.de/2017-GI-DemoWalkuere`

7.5 Conclusions and Future Work

In this chapter, we described a way to model an opera as a multimedia scenario to enrich publicly available video recordings with additional information. The associated media objects were linked with a reference axis. These connections were then described with a relational database schema.

We presented a web-based demonstrator that uses this schema to allow simultaneous and easy access across media objects as well as the comparison of different performances of the same opera.

Our demonstrator enables a number of possible future tasks. One important aspect is the extension of further search functionalities. The text-based search in the libretto could be extended with a content-based search in the music recording. For instance, a new navigation level could be introduced on the basis of repetitive melodic, harmonic, or rhythmic patterns. Especially in Wagner's operas, leitmotifs play an important role. Leitmotifs are repeating characteristic melodic lines which are, for instance, associated to a character (e. g., the *Siegfried-Motif*) or an object (e. g., the *Schwert-Motif*). Through their regular appearance, the leitmotifs give a sort of structure to these long operas and help the listener to understand the plot.

Another extension to our demonstrator is the integration of additional metadata from the Semantic Web. For instance, services such as DBpedia[7] offer Wikipedia entries (e. g., on the musical work or the composer) as structured data. Besides the purely technical challenges, we consider the demonstrator as a presentation for a wider audience. Many publicly funded research projects (like ours) create data which hardly finds any usage besides reaching the promised project goals. Our demonstrator allows all music lovers to easily access and interact with the data.

[7] `http://wiki.dbpedia.org`

Chapter 8

Summary and Future Work

In this thesis, we dealt with fundamental tasks in Music Information Retrieval (MIR): retrieving, extracting, and accessing music-related data of various media types including audio, video, images, and text. Considering several complex music scenarios, we presented different aspects of these tasks which we now summarize.

In the first part, motivated by the book "A Dictionary of Musical Themes" [16], we introduced a cross-modal retrieval scenario with different media types including music recordings, as well as musical themes given in MIDI format and as images. Linking these media types through the musical content they describe, turned out to be a challenging problem. Audio-based retrieval (Chapter 2) requires flexible feature representations and robust retrieval techniques to cope with tempo and key deviations and, especially, with differences in the degree of polyphony between symbolic MIDI queries and music recordings. In the second retrieval scenario (Chapter 3), we had to deal with extraction errors introduced by Optical Character Recognition (OCR) and Optical Music Recognition (OMR) systems. Currently, new DNN-based OMR approaches are developed which may lead to a significant increase in extraction quality [81]. An interesting extension could be to adapt these systems to a given sheet music style—such as the one used in the Barlow–Morgenstern book—through transfer learning or active learning.

In the second part, we focused on the extraction of predominant voices from music recordings. We first analyzed the influence of annotator disagreement on the evaluation of computational approaches for predominant melody estimation within a jazz music scenario (Chapter 4). We then presented a data-driven approach for solo voice enhancement by adapting a DNN-based method originally used for source separation (Chapter 5). As a case study, we used this enhancement approach in a cross-modal retrieval scenario. Given a monophonic solo transcription as a query, the task was to retrieve the corresponding jazz music recording. As another contribution, we indicated the potential of DNN-based methods in a bass transcription scenario where the instrument is no longer salient in the music. One general problem of DNN-based methods

is that they require a large amount of labeled data which is often very time-consuming to create. Future research could deal with unsupervised or semi-supervised techniques to lower the need for labeled data. Furthermore, the training data is often unbalanced which could lead to entanglement problems, as presented for piano transcription in [97]. Although similar to the Barlow–Morgenstern scenario, the jazz scenario was slightly more controlled since the monophonic solo transcriptions were perfectly aligned to the music recordings and there were no key deviations. Applying the presented solo voice enhancement with a specialized model for classical music—similar to the salience representation for symphonic music presented in [30], but data-driven—could increase the retrieval performance in the Barlow–Morgenstern scenario.

In the final part of this thesis, we indicated the potential of web-based interfaces considering two multimedia scenarios. With *JazzTube*, we linked the Weimar Jazz Database (WJD) with publicly available music recordings on YouTube. On the one hand, this tool offers researchers a way to retrieve the music recordings corresponding to the annotations in the WJD. Granting access to the original music recordings is a crucial prerequisite for reproducing scientific experiments based on this data. On the other hand, JazzTube gives music lovers a way to access and explore the information contained in the WJD. In a second scenario, we enriched publicly available video recordings of operas with additional information. Besides the technical challenges of modeling an opera as a multimedia scenario and defining suitable data structures, we indicated how music lovers (in this case "Wagner lovers") can make direct use of existing annotations which were so far exclusively used for research purposes.

This thesis showed that music constitutes a rich and complex domain for studying a variety of challenging research questions in multimedia processing, digital signal processing, and machine learning. This is also reflected by the fact that there exists a growing number of MIR contributions that explore the potential of DNN-based methods (see Appendix B). Making progress in the field of deep learning requires different applications and experts with domain-knowledge who know the training data and can interpret the predictions. The field of MIR—as indicated in this thesis—can contribute to this progress with various interesting and challenging applications.

Part IV

Appendix

Appendix A

A Dictionary of Musical Themes

In 1949, Barlow and Morgenstern (BM) released the book "A Dictionary of Musical Themes" which contains 9803 themes of well-known instrumental pieces from the corpus of Western Classical music [16]. These monophonic themes (usually four bars long) are often the most memorable parts of a piece of music. Figure A.1 shows an example page from the book for the themes of Beethoven's *Symphony No. 5 in C Minor*. Each theme is given a unique identifier which we call *BM-ID*, e. g., `B948`. There may be several BM-themes (e. g., "1st Theme" for `B948` or "2nd Theme" for `B951`) for a single musical work.

In addition to the book, there exists a website[1] called "The Electronic Dictionary of Musical Themes" (EDM) which offers 9825 themes as MIDI files, along with additional metadata (composer and work name). There is a large overlap between BM-themes and EDM-themes, but unfortunately the direct correspondences are not given explicitly. Figure A.2 shows a screenshot taken from the website which shows the corresponding entries in the EDM to the BM-themes shown in Figure A.1. Unfortunately, at the point of writing the thesis, the website is no longer reachable.[2] We are in contact with the original authors to reactivate the website in the future.

For our cross-modal retrieval experiments, we assemble a large collection of music recordings containing Western classical music. We then manually associated the BM-themes to the recordings in the music collection. Some of the listed themes stem from nowadays relatively unknown musical works where few or no recordings exist. Thus, striving for a complete coverage of the book with music recordings is almost impossible. However, we were able to link a several thousand BM-themes to corresponding music recordings. Establishing these links is a very time-consuming and work-intense task. At this point, I want to express my gratitude to all the colleagues and students who participated in this task, in particular: Vlora Arifi-Müller, Lena Krauß, Lukas

[1] `http://www.multimedialibrary.com`

[2] A snapshot of the website can be obtained from the Internet Archive (without the MIDI files): `https://web.archive.org/web/20160330030505/http://multimedialibrary.com/barlow/all_barlow.asp`

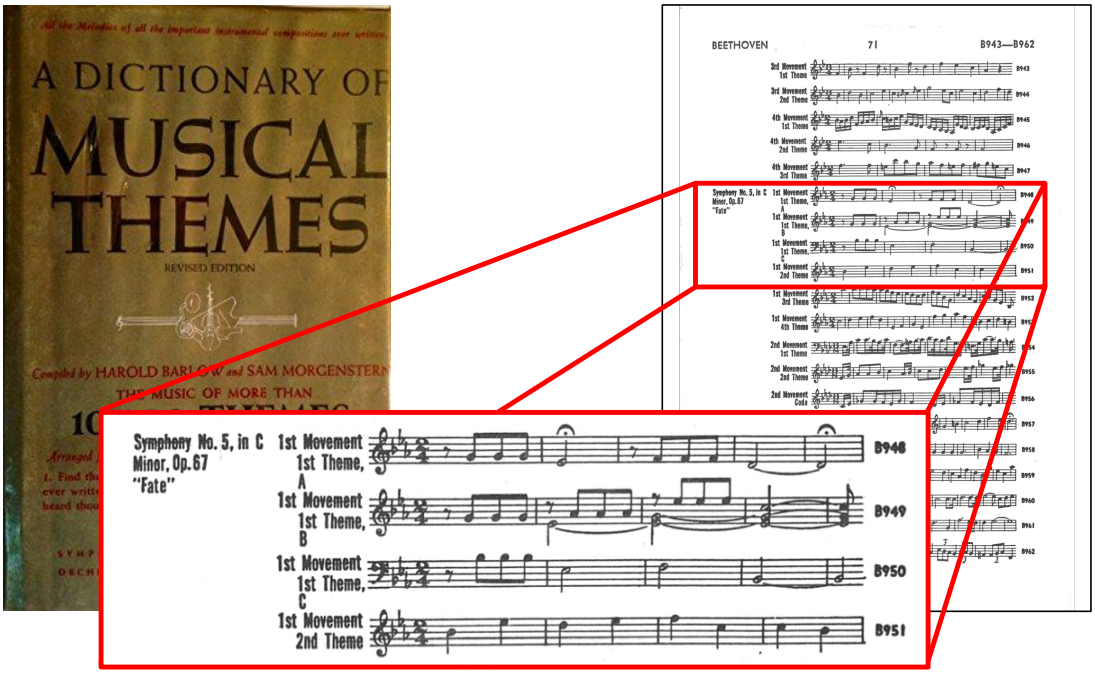


Figure A.1: An example page from the book "A Dictionary of Musical Themes." The composer is given at the top of the page. The musical themes are associated to a musical work or a movement (for symphonies). Furthermore, each theme has a unique identifier, e. g., *B948*.

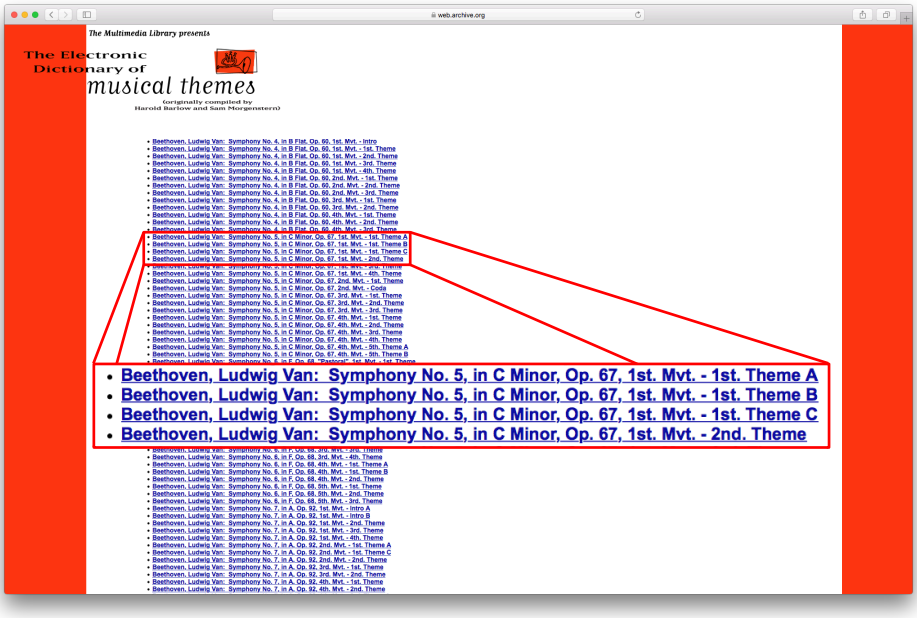


Figure A.2: A screenshot taken from the "Electronic Dictionary of Musical Themes". The themes are given as lists which link to the corresponding MIDI files.

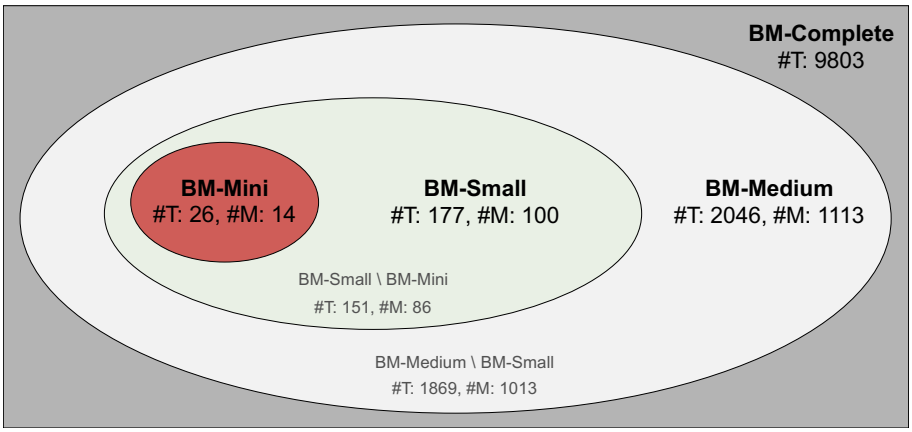


Figure A.3: Overview of the research subsets *BM-Mini*, *BM-Small*, and *BM-Medium*.

Lamprecht, Sanu Pulimootil Achankunju, Meinard Müller, Frank Zalkow, and all the other students who helped digitizing and archiving the CDs.

A.1 Research Subsets

For our experiments in Section 2, we created three subsets: `BM-Mini`, `BM-Small`, and `BM-Medium`. Figure A.3 illustrates the sizes of the three subsets. `BM-Mini` mainly serves as a "development" testset. `BM-Small` already allows for controlled experiments, whereas `BM-Medium` is a first step towards a "real–world" scenario (e. g., in a retrieval setting). The subsets were designed such that `BM-Medium` is a superset of `BM-Small`, and `BM-Small` is a superset of `BM-Mini`.

As a convention for all subsets, each theme is associated to a single corresponding music recording. On the other side, however, a single music recording may contain more than one theme (which explains the higher number of themes compared to the recordings). We use the following identifiers for the music recordings:

- **ComposerID:** Composer's last name (e. g., "Bach" for Johann Sebastian Bach).
- **WorkID:** Identifier for the musical work. If a composer's works are listed in a catalogue (e. g., "Bach-Werke-Verzeichnis (BWV)"), we use them (otherwise we use an abridged version of the work title). In the case of works with more parts (e. g., movements of symphony), the respective part number is appended with a dash.
- **PerformanceID:** Name of the performing ensemble or artist. In case of an orchestra, usually the conductor's last name is used.

Of course, this is only an excerpt of the rules we had when defining the identifiers (and there

were many ad-hoc decisions involved in cases where the rules did not apply). However, with a combination of these three IDs, we can locate the relevant recordings.

A.1.1 BM-Mini

BM-Mini contains 26 Themes (#T) and 14 music recordings (#M). Meant as a development testset, we picked famous compositions and from which we the recordings very well (e. g., Beethoven's 5th Symphony). Table A.1 lists the themes, the corresponding music recordings, their durations, and the type of ensemble.

ID	BM-ID	ComposerID	WorkID	PerformanceID	Dur. (s)	Ensemble
0169	B83	Bach	BWV1041-01	Sitkovetsky	247.933	Concerto
0389	B301	Bach	BWV0846-01	Belder	147.627	Solo
0807	B689	Beethoven	Op002No1-01	Brendel	246.267	Solo
0808	B690	Beethoven	Op002No1-01	Brendel	246.267	Solo
0846	B728	Beethoven	Op013-01	Brendel	548.000	Solo
0847	B729	Beethoven	Op013-01	Brendel	548.000	Solo
0848	B730	Beethoven	Op013-01	Brendel	548.000	Solo
1066	B948	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1067	B949	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1068	B950	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1069	B951	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1070	B952	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1071	B953	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1149	B1031	Beethoven	Op011-01	Berkes	556.080	Trio
1511	B1375	Brahms	HungarianDances-05	SchmidtIsserstedt	160.133	Orchestra
1512	B1376	Brahms	HungarianDances-05	SchmidtIsserstedt	160.133	Orchestra
2219	C232	Chopin	Op024-02	Groot	141.773	Solo
2221	C234	Chopin	Op030-02	Groot	85.920	Solo
2236	C249	Chopin	Op063-03	Groot	135.933	Solo
2276	C289	Chopin	Op028-04	Davidovich	143.093	Solo
2287	C300	Chopin	Op028-15	Davidovich	368.893	Solo
2288	C301	Chopin	Op028-15	Davidovich	368.893	Solo
7752	S533	Schubert	D0759-01	Goodman	803.507	Orchestra
7753	S534	Schubert	D0759-01	Goodman	803.507	Orchestra
7754	S535	Schubert	D0759-01	Goodman	803.507	Orchestra
7940	S713	Schumann	Op015-07	Horowitz	169.507	Solo

Table A.1: Subset *BM-Mini*.

A.1.2 BM-Small

BM-Small contains 177 Themes (#T) and 100 music recordings (#M). In setting up this subset, the goal was to have a cross section of famous composers and different musical styles. Table A.2

lists the themes, the corresponding music recordings, their durations, and the type of ensemble.

ID	BM-ID	ComposerID	WorkID	PerformanceID	Dur. (s)	Ensemble
0116	B30	Bach	BWV1046-01	Belder	234.133	Concerto
0126	B40	Bach	BWV1048-01	Belder	319.160	Concerto
0167	B81	Bach	BWV1065-01	Schornsheim	239.560	Concerto
0169	B83	Bach	BWV1041-01	Sitkovetsky	247.933	Concerto
0179	B93	Bach	BWV0543-01	Fagius	208.693	Solo
0242	B156	Bach	BWV1002-03-1	Lubotsky	172.800	Solo
0257	B171	Bach	BWV1030-01	Wentz	438.720	Duo
0261	B175	Bach	BWV1001-01	Lubotsky	295.800	Solo
0297	B209	Bach	BWV1009-05	Linden	224.360	Solo
0333	B245	Bach	BWV0808-01	Asperen	201.533	Solo
0386	B298	Bach	BWV0565-01	Fagius	169.493	Solo
0387	B299	Bach	BWV0565-02	Fagius	376.507	Solo
0389	B301	Bach	BWV0846-01	Belder	147.627	Solo
0390	B302	Bach	BWV0846-02	Belder	119.480	Solo
0391	B303	Bach	BWV0847-01	Belder	87.560	Solo
0392	B304	Bach	BWV0847-02	Belder	90.520	Solo
0557	B461	Bartok	Sz112-01	Chung	968.867	Concerto
0558	B462	Bartok	Sz112-01	Chung	968.867	Concerto
0650	B554	Beethoven	WoO059	Brendel	167.987	Solo
0688	B592	Beethoven	Op018No4-01	BudapestStringQuartet	392.133	Quartet
0689	B593	Beethoven	Op018No4-01	BudapestStringQuartet	392.133	Quartet
0804	B687	Beethoven	Op017-01	Tarjani	486.227	Duo
0807	B689	Beethoven	Op002No1-01	Brendel	246.267	Solo
0808	B690	Beethoven	Op002No1-01	Brendel	246.267	Solo
0846	B728	Beethoven	Op013-01	Brendel	548.000	Solo
0847	B729	Beethoven	Op013-01	Brendel	548.000	Solo
0848	B730	Beethoven	Op013-01	Brendel	548.000	Solo
0880	B762	Beethoven	Op027No2-01	Brendel	361.533	Solo
0881	B763	Beethoven	Op027No2-01	Brendel	361.533	Solo
1066	B948	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1067	B949	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1068	B950	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1069	B951	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1070	B952	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1071	B953	Beethoven	Op067-01	Blomstedt	485.493	Orchestra
1149	B1031	Beethoven	Op011-01	Berkes	556.080	Trio
1473	B1337	Brahms	Op015-01	Fleisher	1279.627	Concerto
1474	B1338	Brahms	Op015-01	Fleisher	1279.627	Concerto
1475	B1339	Brahms	Op015-01	Fleisher	1279.627	Concerto
1476	B1340	Brahms	Op015-01	Fleisher	1279.627	Concerto
1511	B1375	Brahms	HungarianDances...	SchmidtIsserstedt	160.133	Orchestra
1512	B1376	Brahms	HungarianDances...	SchmidtIsserstedt	160.133	Orchestra
1551	B1415	Brahms	Op025-04	Han	489.200	Quartet
1552	B1416	Brahms	Op025-04	Han	489.200	Quartet
1553	B1417	Brahms	Op025-04	Han	489.200	Quartet
1857	B1711j	Britten	Op002	EndellionStringQuartet	797.707	Quartet
1858	B1711k	Britten	Op002	EndellionStringQuartet	797.707	Quartet
1859	B1711l	Britten	Op002	EndellionStringQuartet	797.707	Quartet
1860	B1711m	Britten	Op002	EndellionStringQuartet	797.707	Quartet

Continued on next page

ID	BM-ID	ComposerID	WorkID	PerformanceID	Dur. (s)	Ensemble
2184	C198	Chopin	Op010-12	Lortie	169.173	Solo
2210	C223	Chopin	Op066	Davidovich	305.000	Solo
2211	C224	Chopin	Op066	Davidovich	305.000	Solo
2219	C232	Chopin	Op024-02	Groot	141.773	Solo
2221	C234	Chopin	Op030-02	Groot	85.920	Solo
2236	C249	Chopin	Op063-03	Groot	135.933	Solo
2245	C258	Chopin	Op009-02	Harasiewicz	296.093	Solo
2276	C289	Chopin	Op028-04	Davidovich	143.093	Solo
2287	C300	Chopin	Op028-15	Davidovich	368.893	Solo
2288	C301	Chopin	Op028-15	Davidovich	368.893	Solo
2598	D30	Debussy	L103-01	Badings	278.160	Concerto
2665	D97	Debussy	L095-01	Kocsis	252.080	Solo
2666	D98	Debussy	L095-01	Kocsis	252.080	Solo
2712	D144	Debussy	L075-03	Kocsis	352.000	Solo
2713	D145	Debussy	L075-03	Kocsis	352.000	Solo
2841	D262	Dukas	ApprentiSorcier	Fricsay	565.747	Orchestra
2842	D263	Dukas	ApprentiSorcier	Fricsay	565.747	Orchestra
2843	D264	Dukas	ApprentiSorcier	Fricsay	565.747	Orchestra
2960	D372	Dvorak	B083/8	Farrer	247.733	Orchestra
2961	D373	Dvorak	B083/8	Farrer	247.733	Orchestra
3022	D434	Dvorak	B178-01	Szell	520.760	Orchestra
3023	D435	Dvorak	B178-01	Szell	520.760	Orchestra
3024	D436	Dvorak	B178-01	Szell	520.760	Orchestra
3031	D443	Dvorak	B178-04	Szell	654.573	Orchestra
3032	D444	Dvorak	B178-04	Szell	654.573	Orchestra
3033	D445	Dvorak	B178-04	Szell	654.573	Orchestra
3440	G39	Gershwin	AmericanParis	Gershwin	947.533	Orchestra
3441	G40	Gershwin	AmericanParis	Gershwin	947.533	Orchestra
3442	G41	Gershwin	AmericanParis	Gershwin	947.533	Orchestra
3443	G42	Gershwin	AmericanParis	Gershwin	947.533	Orchestra
3617	G206	Granados	Op037-02	Bream	301.947	Solo
3618	G207	Granados	Op037-02	Bream	301.947	Solo
3722	G310	Grieg	Op036-01	CohenR	585.707	Duo
3723	G311	Grieg	Op036-01	CohenR	585.707	Duo
3790	H9	Handel	HWV287-01	Miller	157.440	Concerto
3794	H13	Handel	HWV289-01	Schmitt	314.267	Concerto
3795	H14	Handel	HWV289-01	Schmitt	314.267	Concerto
3928	H147	Handel	HWV365-01	Bosgraaf	155.440	Duo
3944	H163	Handel	HWV368a-01+HWV3...	EcoleOrphee	654.453	Quartet
3964	H183	Handel	HWV361-01+HWV36...	EcoleOrphee	439.707	Trio
4081	H300	Haydn	Hob07bNo002-01	Walevska	809.907	Concerto
4082	H301	Haydn	Hob07bNo002-01	Walevska	809.907	Concerto
4100	H320	Haydn	Hob03No006-01	BuchbergerQuartet	113.867	Quartet
4116	H336	Haydn	Hob03No031-01	BuchbergerQuartet	385.013	Quartet
4484	H703	Haydn	Hob15No027-01	VanSwietenTrio	470.893	Trio
4485	H704	Haydn	Hob15No027-01	VanSwietenTrio	470.893	Trio
4856	K30	Khachaturian	ConcertoViolinD...	Kogan	787.853	Concerto
4857	K31	Khachaturian	ConcertoViolinD...	Kogan	787.853	Concerto
4858	K32	Khachaturian	ConcertoViolinD...	Kogan	787.853	Concerto
4889	K63	Kodaly	GalantaDances	FischerI	933.533	Orchestra
4890	K64	Kodaly	GalantaDances	FischerI	933.533	Orchestra

Continued on next page

ID	BM-ID	ComposerID	WorkID	PerformanceID	Dur. (s)	Ensemble
4891	K65	Kodaly	GalantaDances	FischerI	933.533	Orchestra
4892	K66	Kodaly	GalantaDances	FischerI	933.533	Orchestra
4893	K67	Kodaly	GalantaDances	FischerI	933.533	Orchestra
4894	K68	Kodaly	GalantaDances	FischerI	933.533	Orchestra
4895	K69	Kodaly	GalantaDances	FischerI	933.533	Orchestra
4896	K70	Kodaly	GalantaDances	FischerI	933.533	Orchestra
5137	L189	Liszt	S541/3	Howard	253.600	Solo
5551	M289	Mendelssohn	Op030-03	Laar	158.867	Solo
5702	M429	Mozart	KV315	Graf	386.040	Concerto
5707	M434	Mozart	KV622-01	Boer	719.813	Concerto
5718	M445	Mozart	KV447-01	Jeurissen	400.040	Concerto
5719	M446	Mozart	KV447-01	Jeurissen	400.040	Concerto
5796	M522a	Mozart	KV219-01	Carmignola	576.573	Concerto
5797	M523	Mozart	KV219-01	Carmignola	576.573	Concerto
5798	M524	Mozart	KV219-01	Carmignola	576.573	Concerto
5831	M557	Mozart	KV186-02	Graaf	124.720	Decet
5835	M561	Mozart	KV188-01	Zon	105.853	Septet
5853	M579	Mozart	KV334-01	Nodel	398.320	Concerto
5890	M615	Mozart	KV298-01	Grauwels	355.707	Quartet
5894	M619	Mozart	KV370-01	KochL	436.627	Quintet
5957	M682	Mozart	KV581-01	Leister	543.000	Quintet
5962	M687	Mozart	KV452-01	Wurtz	568.587	Quintet
5963	M688	Mozart	KV452-01	Wurtz	568.587	Quintet
6011	M736	Mozart	KV361-01	Schneider	597.707	Orchestra
6097	M822	Mozart	KV145	Matousek	175.080	Concerto
6115	M840	Mozart	KV378-01	Accardo	753.507	Duo
6247	M972	Mozart	KV550-01	Linden	448.907	Orchestra
6248	M973	Mozart	KV550-01	Linden	448.907	Orchestra
6419	P37	Paganini	MS025-24	Accardo	262.507	Solo
6500	P114	Piston	IncredibleFluti...	Mariano	69.200	Orchestra
6820	R53	Rachmaninoff	Op019-01	Shafran	781.693	Duo
6821	R54	Rachmaninoff	Op019-01	Shafran	781.693	Duo
6892	R125	Ravel	MR081	Steinberg	930.120	Orchestra
6893	R126	Ravel	MR081	Steinberg	930.120	Orchestra
7233	S14	Saint	CarnavalAnimaux...	Licata	87.667	Orchestra
7236	S17	Saint	Op033-01	Fournier	328.733	Concerto
7237	S18	Saint	Op033-01	Fournier	328.733	Concerto
7266	S47	Saint	Op040	Ansermet	448.027	Orchestra
7267	S48	Saint	Op040	Ansermet	448.027	Orchestra
7282	S63	Saint	Op028	Milstein	522.067	Concerto
7283	S64	Saint	Op028	Milstein	522.067	Concerto
7284	S65	Saint	Op028	Milstein	522.067	Concerto
7285	S66	Saint	Op028	Milstein	522.067	Concerto
7308	S89	Saint	Op065-01	Pople	288.533	Septet
7433	S214	Scarlatti	K009	Belder	202.600	Solo
7520	S301	Schubert	D0899-03	Hoek	338.200	Solo
7538	S319	Schubert	D0780-03	Nauta	128.067	Solo
7589	S370	Schubert	D0810-01	Brandis	972.000	Quartet
7590	S371	Schubert	D0810-01	Brandis	972.000	Quartet
7606	S387	Schubert	D0667-04	Sharon	478.520	Quintet
7752	S533	Schubert	D0759-01	Goodman	803.507	Orchestra

Continued on next page

ID	BM-ID	ComposerID	WorkID	PerformanceID	Dur. (s)	Ensemble
7753	S534	Schubert	D0759-01	Goodman	803.507	Orchestra
7754	S535	Schubert	D0759-01	Goodman	803.507	Orchestra
7940	S713	Schumann	Op015-07	Horowitz	169.507	Solo
8029	S802	Schumann	Op094-01	EnsembleIncanto	208.787	Duo
8098	S870a	Shostakovich	Op040-01	Rosler	682.547	Duo
8099	S870b	Shostakovich	Op040-01	Rosler	682.547	Duo
8352	S1114	Smetana	MyCountry-02	Talich	690.147	Orchestra
8353	S1115	Smetana	MyCountry-02	Talich	690.147	Orchestra
8354	S1116	Smetana	MyCountry-02	Talich	690.147	Orchestra
8355	S1117	Smetana	MyCountry-02	Talich	690.147	Orchestra
8780	S1534a	Stravinsky	DumbartonOaks-0...	Stravinsky	261.800	Orchestra
8781	S1534b	Stravinsky	DumbartonOaks-0...	Stravinsky	261.800	Orchestra
8782	S1534c	Stravinsky	DumbartonOaks-0...	Stravinsky	261.800	Orchestra
8793	S1541	Stravinsky	OctetWinds-01	Stravinsky	236.573	Orchestra
8794	S1542	Stravinsky	OctetWinds-01	Stravinsky	236.573	Orchestra
8795	S1543	Stravinsky	OctetWinds-01	Stravinsky	236.573	Orchestra
8988	T58	Telemann	TWV043-e4-01	MusicaRhenum	307.693	Quintet
8989	T59	Telemann	TWV043-e4-01	MusicaRhenum	307.693	Quintet
8990	T60	Telemann	TWV043-e4-01	MusicaRhenum	307.693	Quintet
9167	T198	Tschaikovsky	Op071-13	Ansermet	386.360	Orchestra
9168	T199	Tschaikovsky	Op071-13	Ansermet	386.360	Orchestra
9169	T200	Tschaikovsky	Op071-13	Ansermet	386.360	Orchestra
9170	T201	Tschaikovsky	Op071-13	Ansermet	386.360	Orchestra
9270	T301	Turina	Op036	Bream	313.467	Solo
9271	T302	Turina	Op036	Bream	313.467	Solo
9412	V123	Visee	SuiteDMinor-01	Bream	51.400	Solo

Table A.2: Subset *BM-Small*.

A.1.3 BM-Medium

BM-Small contains 2045 Themes (#T) and 1114 music recordings (#M). In this subset, the goal was to have a representative cross section of Western classical music. The recordings in the dataset were selected in a way that complete work groups are represented, e. g., Bach's "Well-Tempered Clavier", all of Beethoven's symphonies, or Chopin's "Mazurkas".

Instead of listing all the 2045 themes, we give an overview of the dataset in the form of histograms. Figure A.4a shows the number of themes for the top 20 composers. Beethoven's quantitative dominance stems from the many symphonies (e. g., his 5th Symphony contains 13 BM-themes). Figure A.4b shows the durations of the music recordings. The mean duration is 385.01 s ($\sigma = 251.29$ s). The complete duration of all music recordings combined is about 120 h. Figure A.4c depicts the durations of the queries as occurring in the respective music recording. The mean query duration is 8.81 s ($\sigma = 5.87$ s).

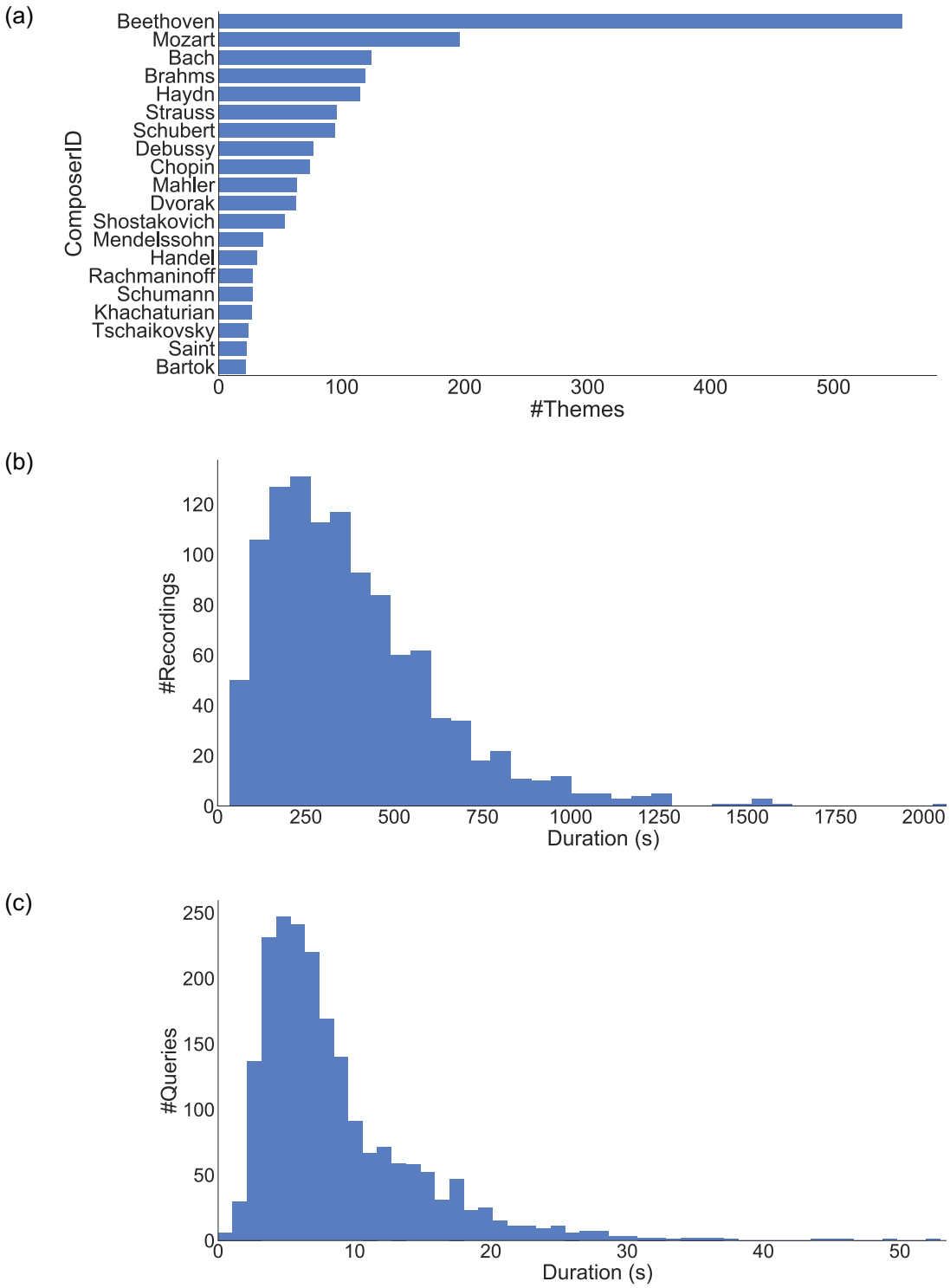


Figure A.4: Overview histograms for *BM-Medium*. (a) Top-20 composers. (b) Durations of the music recordings. (c) Theme durations in the music recordings.

Appendix B

Deep Neural Networks in MIR

In recent years, data-driven approaches such as Deep Neural Networks (DNN) enabled significant progress in many research areas and applications. In the field of Music Information Retrieval (MIR), various analysis, retrieval, and classification tasks are nowadays approached using DNN-based techniques, often leading to significant performance improvements when compared to state-of-the-art algorithms. However, succeeding with DNNs requires researchers to adapt many hyperparameters, such as the type of input representation, network architecture, and training procedure. In this appendix, current trends for several MIR tasks, including music structure analysis, beat tracking, automatic music transcription, harmony analysis, and audio source separation, are reviewed. For each task, typical approaches form the literature are considered. Each approach is then summarized with respect to its underlying neural network techniques, focusing on hyperparameters.

B.1 Sources

The considered publications mainly stem from the following sources:

- International Society for Music Information Retrieval Conferences (ISMIR) 2010–2016
- AES International Conference Semantic Audio 2017
- IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2015
- International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)
- IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
- IEEE Transactions on Audio, Speech, and Language Processing
- IEEE Transactions on Multimedia

B.2 Abbreviations

We use a number of abbreviations in the subsequent overview. Depending on the task at hand and publishing community, different "dialects" exist. However, we try to be consistent within this review. Table B.1 lists the abbreviations used for the reviewed tasks, and Table B.2 lists the abbreviations used for the methods and hyperparameters.

Abbr.	**Explanation**
FL	Feature Learning
F0	F0-Estimation
AMT	Automatic Music Transcription
BRA	Beat and Rhythm Analysis
MSA	Music Structure Analysis
CR	Chord Recognition
ASP	Audio Source Separation
VAR	Various (e. g., Singing Voice Detection, Tagging)

Table B.1: Task abbreviations used in Table B.3.

Abbr.	**Explanation**
RBM	Restricted Boltzmann Machine
DBN	Deep Belief Network
FNN	Fully-connected Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
MODEL	Some model-based extraction output
NMF	Activations from an NMF-based model
LinS	Linear-frequency Spectrogram
LogS	Logarithmic-frequency Spectrogram
MelS	Mel-frequency Spectrogram
NCMS	Normalized Cepstral Modulation Spectrum
LogLinS	Logarithmic-magnitude linear-frequency Spectrogram
LogLogS	Logarithmic-magnitude logarithmic-frequency Spectrogram
LogMelS	Logarithmic-magnitude Mel-frequency Spectrogram
SquLinS	Square-root-magnitude linear-frequency Spectrogram
CubLinS	Cubic-root-magnitude linear-frequency Spectrogram
NORM	Normalization (no further details given)
NORMTL1	Frame-wise L1-normalization
NORMTL2	Frame-wise L2-normalization
DERIV	Derivative
STD	Standardization ($\mu = 0$, $\sigma = 1$, no further details given)
STDF	Standardization per frequency-band/feature
PCA	Principal Component Analysis/Whitening
HPSS	Harmonic–Percussive Source Separation

Table B.2: Method and hyperparameter abbreviations used in Table B.3.

B.3 Overview

Task	Year	Authors	Ref.	Type	Input	Pre-proc.
AMT	2012	Böck and Schedl	[24]	RNN-BLSTM	LogLogS	DERIV
AMT	2004	Marolt	[119]	Var	HC	—
AMT	2011	Nam et al.	[137]	DBN	CubLinS	NORMFL1, PCA
AMT	2013	Boulanger-Lewandowski et al.	[32]	RNN-RBM	CubLinS	NORM
AMT	2014	Sigtia et al.	[174]	RNN	LogS	—
AMT	2017	Ewert and Sandler	[61]	RNN-LSTM	NMF	—
AMT	2017	Kelz and Widmer	[97]	CNN	LogLogS	—
AMT	2017	Abeßer et al.	[1]	FNN	LogS	NORMFL2
AMT	2016	Sigtia et al.	[176]	RNN	—	—
AMT	2016	Vogl et al.	[194]	RNN	LogLogS	—
AMT	2016	Southall et al.	[182]	RNN	LinS	—
AMT	2016	Kelz et al.	[98]	CNN	LogLogS	—
ASS	2015	Simpson et al.	[177]	CNN	LinS	—
ASS	2016	Nugraha et al.	[139]	FNN	LinS	PCA, STD
ASS	2016	Nugraha et al.	[140]	FNN	LinS	PCA, STD
ASS	2017	Chandna et al.	[38]	CNN	LinS	—
ASS	2015	Uhlich et al.	[188]	FNN	LinS	NORMFL2
ASS	2017	Miron et al.	[128]	CNN	LinS	—
ASS	2016	Grais et al.	[76]	FNN	LinS	STDF
ASS	2015	Huang et al.	[90]	RNN	LMS	DERIV
ASS	2017	Luo et al.	[118]	RNN-BLSTM	MelS	DERIV
ASS	2017	Uhlich et al.	[189]	RNN-BLSTM	LinS	—
ASS	2014	Huang et al.	[89]	RNN	LinS	—
BRA	2010	Eyben et al.	[62]	RNN-BLSTM	LogMelS	DERIV
BRA	2011	Böck and Schedl	[23]	RNN-BLSTM	LogMelS	DERIV
BRA	2012	Battenberg and Wessel	[17]	DBN	—	—
BRA	2014	Böck et al.	[25]	RNN-BLSTM	LogS	—
BRA	2016	Böck et al.	[27]	RNN-BLSTM	LogS	DERIV
BRA	2016	Elowsson	[59]	FNN	HC	—
BRA	2016	Holzapfel and Grill	[87]	CNN	LogLogS	STDF
BRA	2016	Krebs et al.	[104]	RNN-BGRU	HC	—
BRA	2016	Durand and Essid	[56]	CNN	HC	—
BRA	2017	Durand et al.	[57]	CNN	HC	—
BRA	2015	Böck et al.	[26]	RNN-BLSTM	LogMelS	DERIV
CR	2012	Humphrey et al.	[92]	CNN	LogS	NORM
CR	2013	Boulanger-Lewandowski et al.	[31]	RNN	SquLinS	NORMTL2, PCA
CR	2012	Humphrey and Bello	[91]	CNN	LogS	NORM
CR	2017	Korzeniowski and Widmer	[102]	CNN	LogLogS	—
CR	2015	Zhou and Lerch	[200]	CNN	LogS	PCA, STDF
CR	2017	Korzeniowski and Widmer	[103]	CNN	LogLogS	—
CR	2016	Deng and Kwok	[45]	RNN-BLSTM	LogS	—
CR	2015	Sigtia et al.	[175]	FNN	LogS	—
F0	2017	Bittner et al.	[22]	CNN	LogLogS	—
F0	2017	Park and Yoo	[142]	RNN-LSTM	LinS	STDF
F0	2016	Rigaud and Radenen	[160]	RNN-BLSTM	MelS	DERIV
F0	2016	Kum et al.	[106]	FNN	LogLinS	—
FL	2013	Schmidt and Kim	[170]	DBN	HC	—
FL	2010	Hamel and Eck	[82]	DBN	LinS	—

Continued on next page

Task	Year	Authors	Ref.	Type	Input	Pre-proc.
FL	2017	Dai et al.	[43]	CNN	Raw	—
FL	2012	Hamel et al.	[85]	FNN	LogMelS	PCA
FL	2016	Korzeniowski and Widmer	[101]	FNN	LogLogS	—
FL	2017	Balke et al.	[13]	FNN	LogS	—
FL	2011	Hamel et al.	[84]	FNN	MelS	PCA
FL	2014	Dieleman and Schrauwen	[46]	CNN	Raw	—
MSA	2017	Cohen-Hadria and Peeters	[41]	CNN	LogMelS, SSM	—
MSA	2014	Ullrich et al.	[190]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[77]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[78]	CNN	LogMelS	HPSS
VAR	2009	Hamel et al.	[83]	DBN	MFCC	DERIV
VAR	2011	Dieleman et al.	[47]	CNN	HC	—
VAR	2014	van den Oord et al.	[191]	FCC	MODEL	—
VAR	2017	Pons and Serra	[149]	CNN	LogMelS	STD
VAR	2015	Leglaive et al.	[110]	RNN-BLSTM	LogMelS	HPSS
VAR	2017	Pons et al.	[150]	CNN	LogMelS	STD
VAR	2015	Schlüter and Grill	[169]	CNN	LogMelS	STDF
VAR	2015	Raffel and Ellis	[154]	CNN	LogS	NORMTL2, STDF
VAR	2016	Raffel and Ellis	[153]	CNN	LogS	—
VAR	2016	Schlüter	[168]	CNN	LogLinS	STDF
VAR	2016	Choi et al.	[39]	CNN	LogMelS	—
VAR	2016	Lostanlen and Cella	[117]	CNN	LogLogS	—
VAR	2016	Jeong and Lee	[94]	FNN	LinS, CMS	—
VAR	2017	Hershey et al.	[86]	CNN	LogMelS	—
VAR	2017	Dorfer et al.	[51]	CNN	LogMelS	—
VAR	2015	Lehner et al.	[112]	RNN-LSTM	HC	STD
VAR	2016	Dorfer et al.	[50]	CNN	LogMelS	—

Table B.3: Literature Overview.

Appendix C

DNN Hyperparameter Experiments

In Chapter 5, we trained DNNs to enhance salient voices in time–frequency representations of polyphonic sound mixtures. We used a certain set of hyperparameters—including the feature type of the input feature representation the depth of the network—to train a DNN-based model. In this appendix, we present further experiments that indicate the influence of the choice of hyperparameters. We evaluate the trained models within a jazz solo transcription scenario where we assume that the solo instrument is monophonic and that it is salient in the mixture. We regard the transcription task as a multi-class classification scenario, where, given a frame of a time–frequency representation of a jazz recording as input, the objective is to classify the musical pitch of the solo instrument in this frame.

C.1 Technical Details

DNN Architecture

- Fully-connected Neural Network as shown in Figure C.1 (inspired by [98])
- **Input:** Frames with $K \in \mathbb{N}$ frequency bands obtained from a suitable time–frequency representation
- **Output:** 89 classes (88 pitches and 1 non-voicing class)
- Input feature representations were extracted using the Python package `librosa` [124]

C.2 Hyperparameters

Input Feature Type

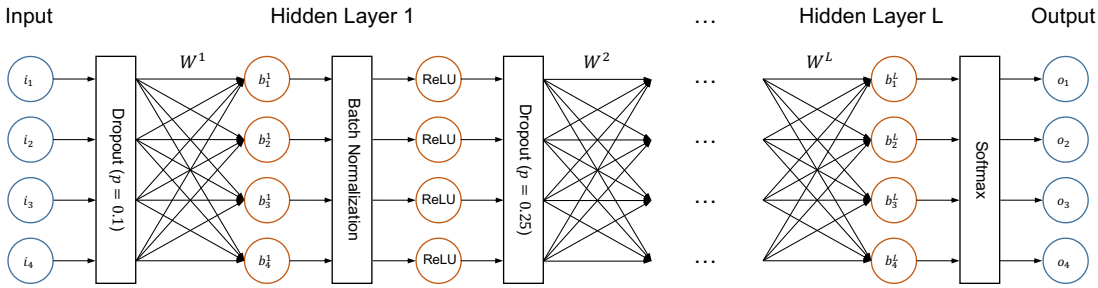


Figure C.1: Network architecture.

- `LINS`: STFT (linear frequency), $K = 4096$ ($f_s = 22.05$ kHz, frame size = 200 ms, hop size = 100 ms) [131]
- `MELS`: STFT (mel frequency), $K = 180$ (settings as `LINS`, but frequency axis consists of 180 mel-bands between 27.5 Hz and 11.025 kHz [183])
- `LOGS`: STFT (logarithmic frequency), $K = 88$ (12 semitones per octave, ranging from 27.5 Hz (`A0`) to 4186 Hz (`C8`)) [131]
- `LOGFB`: Semitone filterbank, $K = 88$ (12 semitones per octave, ranging from 27.5 Hz (`A0`) to 4186 Hz (`C8`)) [130]

$\Rightarrow$ **Input features:** $X := (x_1, x_2, \ldots, x_T)$, $x_t \in \mathbb{R}^K$ (frames), $t \in [1:T]$

Feature Compression

- Logarithmic compression $x_t^c(k) := \log(1 + x_t(k))$, $x_t \in \mathbb{R}^K$, $t \in [1:T]$, $k \in [1:K]$

Temporal Context

- Stacking $\pm\tau \in \mathbb{N}$ neighboring input frames as shown in Figure C.2
- Note that this hyperparameter increases the size of the input features to $K_\tau = K(2\tau + 1)$.

C.3 Experiments

The following set of hyperparameters were considered:

- **Input feature type:** {`LINS`, `MELS`, `LOGS`, `LOGFB`}
- **Feature compression:** {no compression, log compression}
- **Temporal context:** {0, 1, 2, 4, 8}
- **#Hidden layers:** {1, 2, 3, 4, 5}

This results in 200 different parameter settings.

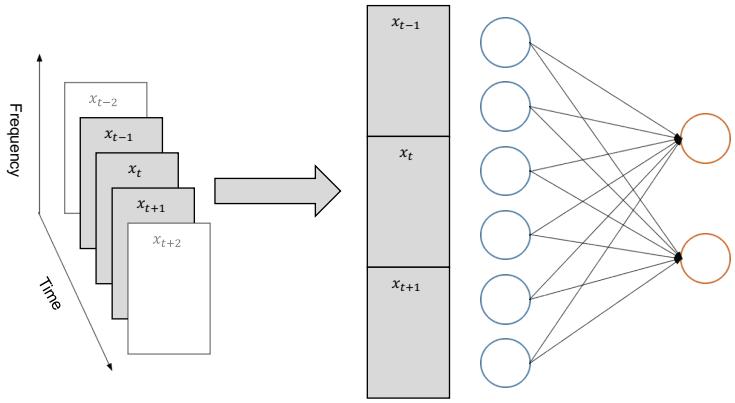


Figure C.2: Frame stacking when using a temporal context of $\tau = 1$. The center frame is combined with its $\pm\tau$ neighboring frames and served as input to the DNN.

Training

- DNNs were trained with the Python library `keras` [40] using the `Theano` backend [185].
- Training was performed at the university's local computer center[1] where we have access to a GPU pool of 14 NVIDIA GEFORCE GTX 980 and 14 NVIDIA GEFORCE GTX 1080.
- **Optimizer:** Adam (learning rate = 0.02, mini-batch size = 100) [99]
- **Epoch size:** 4096 mini-batches
- **Early stopping**: Patience of 10 epochs on the validation data (maximum 100 epochs)

Training Data

- 456 jazz solo transcriptions from the WeimarJazzDatabase [147]
- **Total Solo Duration:** 810 min.
- **Average Solo Duration:** 106.75 s ($\sigma = 68.48$)
- **Voiced Frames:** 60.67% ($\sigma = 7.82$) (soloist is active)
- **Split:**
  - Five fold cross-validation (80% training set, 20% test set)
  - 20% from training set for validation
  - Record identifier is used to reduce album effects
  - Details on the folds are listed in Table C.1

[1] `https://hpc.fau.de/systems/hpc-systems/`

	Training Set	Test Set	Validation Set
Number of solos	291.80 (0.45)	91.20 (0.45)	73.00 (0.00)
Total solo duration (h)	8.71 (0.12)	2.70 (0.19)	2.11 (0.07)
Avg. solo duration (s)	107.42 (1.46)	106.76 (7.44)	104.08 (3.64)
Voiced frames (%)	60.62 (0.32)	60.64 (0.64)	60.73 (1.25)

Table C.1: Mean values (and standard deviations) of number of solos, total solo duration, and average solo duration averaged of 5 folds for training, test and validation set. *Voiced frames* indicates the percentage frames that contain an active soloist.

Evaluation Metrics

- *Raw Pitch Accuracy (RPA)*, *Raw Chroma Accuracy (RCA)*, *Voicing Recall (VR)*, *Voicing False Alarm (VFA)*, and *Overall Accuracy (OA)* (MIREX' *Audio Melody Extraction* task, see Chapter 4 or [53] for details)
- Evaluation based on the implementations provided by the `mir-eval` package [156].

C.4 Results

- **Baseline:** Melodia [164], average (and std. dev.) over all validation folds
  RPA: 0.51 (0.14), RCA: 0.58 (0.11), VR: 0.99 (0.02), VFA: 0.98 (0.03), OA: 0.31 (0.09)
- **DNN:** Figure C.3 and Figure C.4 summarize the results on all validation folds
  - Performance of different input feature representations varies.
  - More than 3 hidden layers does not increase accuracy significantly.
  - A temporal context of $\tau = 1$ is sufficient in this scenario.
  - Applying log-compression to the input features improves the results, especially for `LOGS` and `LOGFB`.
  - A network with a single hidden layer and $\tau = 8$ shows unstable results.
- Number of parameters for models range between 10000 and 3 Million (see Table C.2)
- Model training time varied between 1–3 h (see Table C.3).

Possible Future Directions

- Consider a "fairer" model-based baseline approach, e. g., an adaptation of *Melodia* which is informed with the pitch distribution estimated on the training set.
- Comparison with other methods, e. g., Random Forests for voicing detection [111].

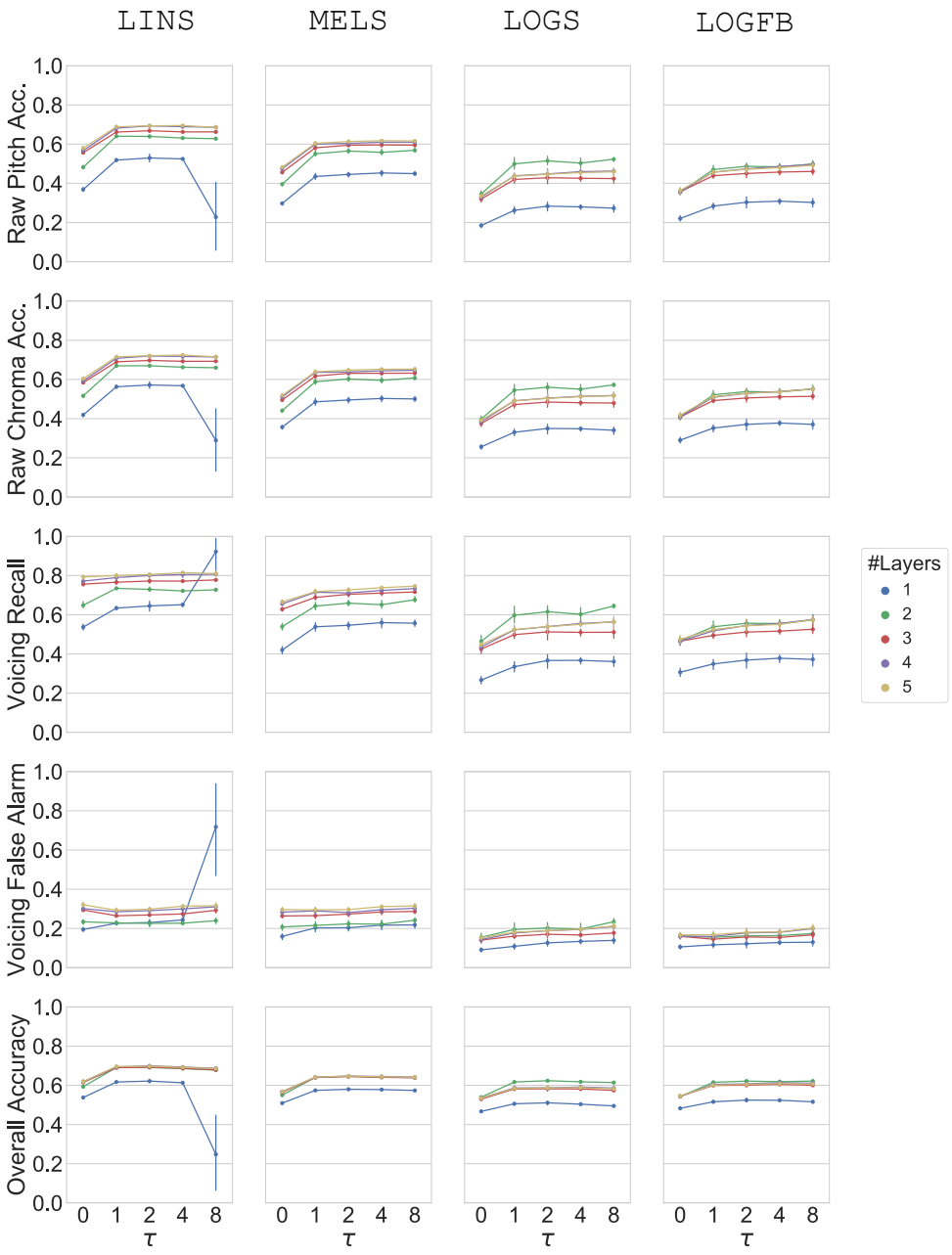


Figure C.3: Raw Pitch Accuracy and Voicing Accuracy for varying temporal context τ and number of layers using the *unprocessed* input features.

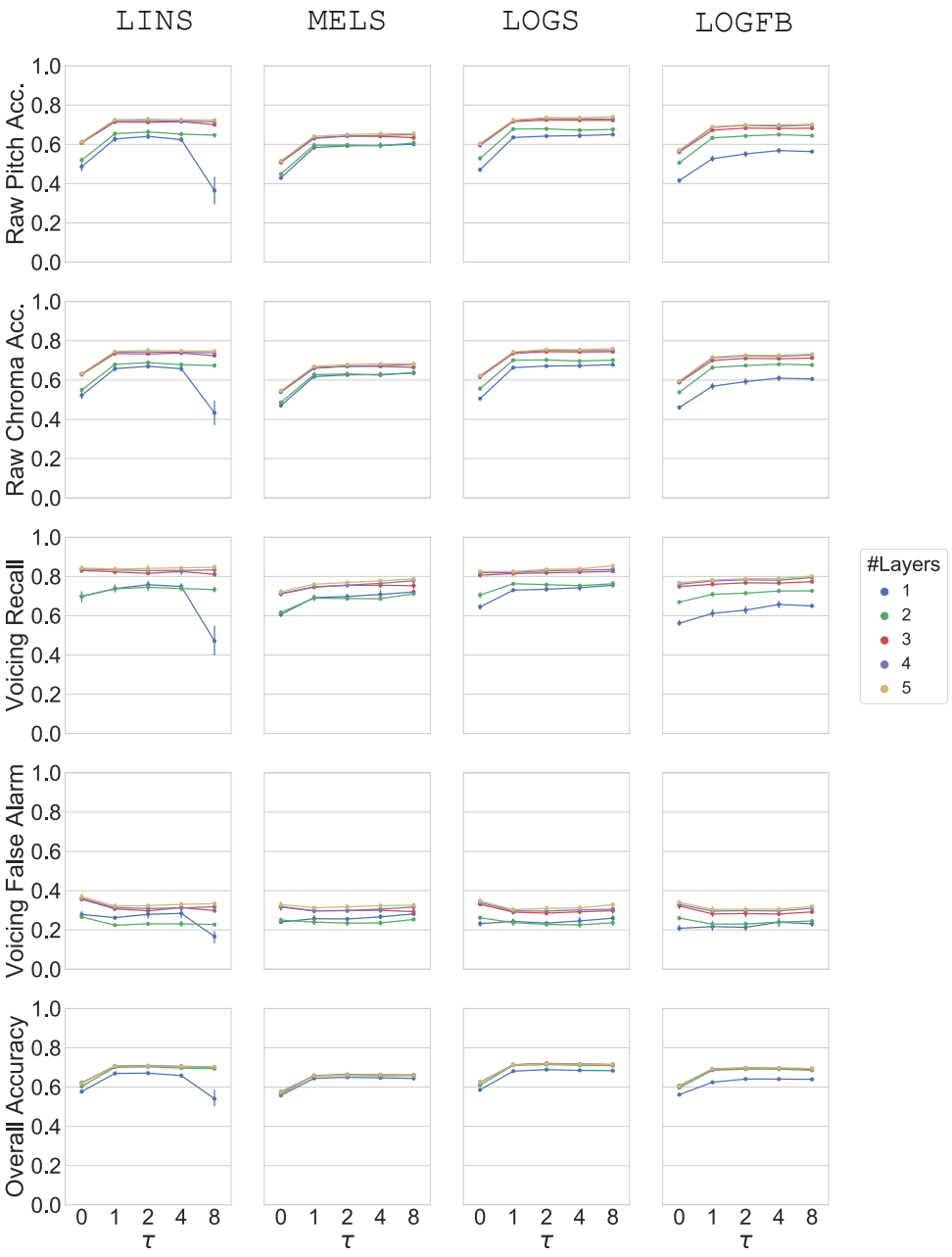


Figure C.4: Raw Pitch Accuracy and Voicing Accuracy for varying temporal context τ and number of layers using the *logarithmically compressed* input features.

	L / τ	1	2	3	4	5
LINS	0	190646	199012	207378	215744	224110
	1	571760	580126	588492	596858	605224
	2	952874	961240	969606	977972	986338
	4	1715102	1723468	1731834	1740200	1748566
	8	3239558	3247924	3256290	3264656	3273022
MELS	0	16829	25195	33561	41927	50293
	1	50309	58675	67041	75407	83773
	2	83789	92155	100521	108887	117253
	4	150749	159115	167481	175847	184213
	8	284669	293035	301401	309767	318133
LOGS	0	8273	16639	25005	33371	41737
	1	24641	33007	41373	49739	58105
	2	41009	49375	57741	66107	74473
	4	73745	82111	90477	98843	107209
	8	139217	147583	155949	164315	172681
LOGFB	0	8273	16639	25005	33371	41737
	1	24641	33007	41373	49739	58105
	2	41009	49375	57741	66107	74473
	4	73745	82111	90477	98843	107209
	8	139217	147583	155949	164315	172681

Table C.2: Number of parameters.

		#Epochs					Training Duration (s)				
	L / τ	1	2	3	4	5	1	2	3	4	5
LINS	0	17.60	17.60	29.40	30.80	29.80	63.52	65.80	113.92	120.20	115.82
	1	19.00	19.00	29.20	25.60	28.20	70.79	72.73	113.79	101.61	113.39
	2	15.20	15.20	19.80	21.60	24.00	57.27	59.60	79.00	88.00	97.51
	4	29.60	15.20	18.40	19.00	21.40	114.32	62.44	75.61	80.69	90.08
	8	20.60	14.80	15.20	17.60	19.00	84.25	64.22	67.84	79.89	83.92
MELS	0	16.40	12.00	16.80	16.80	15.20	50.86	39.36	55.29	56.63	51.68
	1	22.60	12.60	23.00	18.00	16.00	70.69	41.67	75.42	61.70	54.91
	2	20.80	13.20	16.00	16.40	17.00	65.54	43.47	54.46	56.03	59.53
	4	19.40	13.00	19.00	23.00	19.40	61.92	44.13	64.97	79.30	68.82
	8	14.80	13.20	15.40	16.40	14.00	50.19	46.19	54.63	58.98	51.61
LOGS	0	21.00	16.00	31.40	28.60	35.20	60.68	49.41	97.28	91.55	113.13
	1	21.80	40.60	51.20	40.20	54.20	65.55	127.23	161.12	129.97	178.70
	2	21.40	46.40	55.20	43.40	47.20	65.40	145.27	175.59	143.86	157.13
	4	24.40	39.00	43.80	47.60	40.80	74.93	122.30	141.37	158.25	138.04
	8	20.00	34.40	44.40	39.20	44.00	63.19	108.91	145.77	133.81	150.60
LOGFB	0	17.20	12.00	17.60	16.20	15.00	51.04	38.07	55.16	52.85	50.09
	1	26.20	12.60	25.60	20.60	18.00	79.68	40.44	80.92	68.17	59.69
	2	17.00	12.80	19.00	15.20	16.40	51.92	40.98	61.87	50.49	56.63
	4	19.60	14.00	27.20	23.00	24.00	61.08	44.03	86.25	76.81	80.49
	8	19.40	15.80	23.60	18.40	16.40	60.75	51.99	77.88	61.91	56.28

Table C.3: Number of epochs and training duration averaged over the five training folds (*uncompressed* input features).

Bibliography

[1] Jakob Abeßer, Stefan Balke, Klaus Frieler, Martin Pfleiderer, and Meinard Müller. Deep learning for jazz walking bass transcription. In *Proceedings of the AES International Conference on Semantic Audio*, pages 202–209, 2017.

[2] Andreas Arzt. *Flexible and robust music tracking.* PhD thesis, Universität Linz, 2016.

[3] Andreas Arzt, Sebastian Böck, and Gerhard Widmer. Fast identification of piece and score position via symbolic fingerprinting. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 433–438, Porto, Portugal, 2012.

[4] Stefan Balke and Meinard Müller. A graphical user interface for understanding audio retrieval results. In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Malága, Spain, 2015.

[5] Stefan Balke and Meinard Müller. JazzTube: Linking the Weimar Jazz Database with YouTube. In Martin Pfleiderer, Klaus Frieler, Jakob Abeßer, Wolf-Georg Zaddach, and Benjamin Burkhart, editors, *Inside the Jazzomat. New perspectives for jazz research*, pages 315–317. Schott Campus, Mainz, Germany, 2017.

[6] Stefan Balke, Christian Dittmar, and Meinard Müller. Accompanying website: Data-driven solo voice enhancement for jazz music retrieval. `http://www.audiolabs-erlangen.de/resources/MIR/2017-ICASSP-SoloVoiceEnhancement/`, .

[7] Stefan Balke, Jonathan Driedger, Jakob Abeßer, Christian Dittmar, and Meinard Müller. Accompanying website. `http://www.audiolabs-erlangen.de/resources/MIR/2016-ISMIR-Multiple-Annotations/`, .

[8] Stefan Balke, Sanu Pulimootil Achankunju, and Meinard Müller. Matching musical themes based on noisy OCR and OMR input. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 703–707, Brisbane, Australia, 2015.

[9] Stefan Balke, Lukas Lamprecht, Vlora Arifi-Müller, Thomas Prätzlich, and Meinard Müller. Automatisierte Identifikation von Audioaufnahmen anhand symbolisch codierter musikalischer Themen. In *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)*, Nürnberg, Germany, 2015.

[10] Stefan Balke, Jakob Abeßer, Jonathan Driedger, Christian Dittmar, and Meinard Müller. Towards evaluating multiple predominant melody annotations in jazz recordings. In *Proceedings of the*

International Conference on Music Information Retrieval (ISMIR), pages 246–252, New York City, USA, 2016.

[11] Stefan Balke, Vlora Arifi-Müller, Lukas Lamprecht, and Meinard Müller. Retrieving audio recordings using musical themes. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 281–285, Shanghai, China, 2016.

[12] Stefan Balke, Paul Bießmann, Sebastian Trump, and Meinard Müller. Konzeption und Umsetzung webbasierter Werkzeuge für das Erlernen von Jazz-Piano. In *Proceedings of the GI Jahrestagung*, pages 61–73, 2017. doi: 10.18420/in2017_03.

[13] Stefan Balke, Christian Dittmar, Jakob Abeßer, and Meinard Müller. Data-driven solo voice enhancement for jazz music retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 196–200, 2017.

[14] Stefan Balke, Manuel Hiemer, Peter Schwab, Vlora Arifi-Müller, Klaus Meyer-Wegener, and Meinard Müller. Die Oper als Multimedia Szenario: Wagners Walküren gehen online. In *Proceedings of the GI Jahrestagung*, pages 75–86, 2017. doi: 10.18420/in2017_04.

[15] Stefan Balke, Christian Dittmar, Jakob Abeßer, Klaus Frieler, Martin Pfleiderer, and Meinard Müller. Bridging the Gap: Enriching YouTube videos with jazz music annotations. *submitted: Frontiers in Digital Humanities*, 2018.

[16] Harold Barlow and Sam Morgenstern. *A Dictionary of Musical Themes*. Crown Publishers, Inc., revised edition third printing edition, 1975.

[17] Eric Battenberg and David Wessel. Analyzing drum patterns using conditional deep belief networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 37–42, Porto, Portugal, 2012.

[18] Pierfrancesco Bellini, Ivan Bruno, and Paolo Nesi. Assessing optical music recognition tools. *Computer Music Journal*, 31(1):68–93, 2007.

[19] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5): 28–37, 2001.

[20] Hervé Bitteur. Audiveris - open music scanner. Website `https://audiveris.kenai.com`, last accessed 09/29/2014, 2013.

[21] Rachel M. Bittner, Justin Salamon, Slim Essid, and Juan Pablo Bello. Melody extraction by contour classification. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 500–506, Málaga, Spain, 2015.

[22] Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan P. Bello. Deep salience representations for F0 tracking in polyphonic music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 63–70, Suzhou, China, 2017.

[23] Sebastian Böck and Markus Schedl. Enhanced beat tracking with context-aware neural networks. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 135–139, Paris, France, 2011.

[24] Sebastian Böck and Markus Schedl. Polyphonic piano note transcription with recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–124, Kyoto, Japan, March 2012.

[25] Sebastian Böck, Florian Krebs, and Gerhard Widmer. A multi-model approach to beat tracking considering heterogeneous music styles. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 603–608, Taipei, Taiwan, 2014.

[26] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Accurate tempo estimation based on recurrent neural networks and resonating comb filters. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–631, Málaga, Spain, 2015.

[27] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 255–261, New York City, USA, 2016.

[28] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R. Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–498, Curitiba, Brazil, 2013.

[29] Juan J. Bosch and Emilia Gómez. Melody extraction in symphonic classical music: a comparative study of mutual agreement between humans and algorithms. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM)*, 2014.

[30] Juan J. Bosch, Ricard Marxer, and Emilia Gómez. Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. *Journal of New Music Research*, 45(2): 101–117, 2016.

[31] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Audio chord recognition with recurrent neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 335–340, Curitiba, Brazil, 2013.

[32] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. High-dimensional sequence transduction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3178–3182, Vancouver, Canada, 2013.

[33] Donald Byrd and Megan Schindele. Prospects for improving OMR with multiple recognizers. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 41–46, Victoria, Canada, 2006.

[34] Chris Cannam, Christian Landone, Mark Sandler, and Juan Pablo Bello. The Sonic Visualiser: A visualisation platform for semantic descriptors from musical signals. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 324–327, Victoria, Canada, 2006.

[35] Chris Cannam, Christian Landone, and Mark B. Sandler. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the International Conference on Multimedia*, pages 1467–1468, Florence, Italy, 2010.

[36] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A review of audio fingerprinting. *The Journal of VLSI Signal Processing*, 41(3):271–284, November 2005. ISSN 0922-5773. doi: 10.1007/s11265-005-4151-3. URL `http://dx.doi.org/10.1007/s11265-005-4151-3`.

[37] Michael A. Casey, Remco Veltkap, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.

[38] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 258–266, Grenoble, France, 2017.

[39] Keunwoo Choi, György Fazekas, and Mark B. Sandler. Automatic tagging using deep convolutional neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 805–811, New York City, USA, 2016.

[40] François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[41] Alice Cohen-Hadria and Geoffroy Peeters. Music structure boundaries estimation using multiple self-similarity matrices as input depth of convolutional neural networks. In *Proceedings of the AES International Conference on Semantic Audio*, pages 202–209, Erlangen, Germany, 2017.

[42] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2001. ISBN 0070131511.

[43] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 421–425, New Orleans, USA, 2017.

[44] David Damm, Christian Fremerey, Verena Thomas, Michael Clausen, Frank Kurth, and Meinard Müller. A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction. *International Journal on Digital Libraries: Special Issue on Music Digital Libraries*, 12(2-3):53–71, 2012.

[45] Jun-qi Deng and Yu-Kwong Kwok. A hybrid gaussian-hmm-deep learning approach for automatic chord estimation with very large vocabulary. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 812–818, New York City, USA, 2016.

[46] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968, Florence, Italy, 2014.

[47] Sander Dieleman, Philemon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 669–674, Miami, Florida, 2011.

[48] Christian Dittmar, Karin Dressler, and Katja Rosenbauer. A toolbox for automatic transcription of polyphonic music. pages 58–65, Ilmenau, Germany, 2007.

[49] Christian Dittmar, Martin Pfleiderer, Stefan Balke, and Meinard Müller. A swingogram representation for tracking micro-rhythmic variation in jazz performances. *Journal of New Music Research*, 0 (0):1–17, 2017. doi: 10.1080/09298215.2017.1367405.

[50] Matthias Dorfer, Andreas Arzt, and Gerhard Widmer. Towards score following in sheet music images. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 789–795, New York, USA, 2016.

[51] Matthias Dorfer, Andreas Arzt, and Gerhard Widmer. Learning audio-sheet music correspondences for score identification and offline alignment. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 115–122, Suzhou, China, 2017.

[52] J. Stephen Downie. Music information retrieval. *Annual Review of Information Science and Technology (Chapter 7)*, 37:295–340, 2003.

[53] J. Stephen Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[54] Jonathan Driedger and Meinard Müller. Verfahren zur Schätzung der Grundfrequenzverläufe von Melodiestimmen in mehrstimmigen Musikaufnahmen. In Wolfgang Auhagen, Claudia Bullerjahn, and Richard von Georgi, editors, *Musikpsychologie – Anwendungsorientierte Forschung*, volume 25 of *Jahrbuch Musikpsychologie*, pages 55–71. Hogrefe-Verlag, 2015.

[55] Jonathan Driedger, Stefan Balke, Sebastian Ewert, and Meinard Müller. Template-based vibrato analysis of music signals. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 239–245, New York City, USA, 2016.

[56] Simon Durand and Slim Essid. Downbeat detection with conditional random fields and deep learned features. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 386–392, New York City, USA, 2016.

[57] Simon Durand, Juan P. Bello, Bertrand David, and Gaël Richard. Robust downbeat tracking using an ensemble of convolutional networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):76–89, 2017.

[58] Ramez Elmasri and Shamkant B. Navathe. *Fundamentals of Database Systems*. Pearson Higher Education, 7 edition, 2016. ISBN 978-0-13-397077-7.

[59] Anders Elowsson. Beat tracking with a cepstroid invariant neural network. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 351–357, New York City, USA, 2016.

[60] D. Erdogmus, O. Fontenla-Romero, J. C. Principe, A. Alonso-Betanzos, and E. Castillo. Linear-least-squares initialization of multilayer perceptrons through backpropagation of the desired response. *IEEE Transactions on Neural Networks*, 16(2):325–337, 2005.

[61] Sebastian Ewert and Mark B. Sandler. An augmented lagrangian method for piano transcription using equal loudness thresholding and LSTM-based decoding. In *Proceedings of the IEEE Workshop*

on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, New York, USA, 2017.

[62] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. Universal onset detection with bidirectional long short-term memory neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 589–594, Utrecht, The Netherlands, 2010.

[63] Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Inc., 2003. ISBN 0-471-52629-0.

[64] Arthur Flexer. On inter-rater agreement in audio music similarity. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 245–250, Taipei, Taiwan, 2014.

[65] Christian Fremerey, Michael Clausen, Sebastian Ewert, and Meinard Müller. Sheet music-audio identification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 645–650, Kobe, Japan, October 2009.

[66] Christian Fremerey, Meinard Müller, and Michael Clausen. Towards bridging the gap between sheet music and audio. In Eleanor Selfridge-Field, Frans Wiering, and Geraint A. Wiggins, editors, *Knowledge representation for intelligent music processing*, number 09051 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, January 2009. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany. URL `http://drops.dagstuhl.de/opus/volltexte/2009/1965`.

[67] Klaus Frieler and Martin Pfleiderer. Onbeat oder offbeat? Überlegungen zur symbolischen Darstellung von Musik am Beispiel der metrischen Quantisierung. In *Proceedings of the GI Jahrestagung*, pages 111–125, 2017.

[68] Klaus Frieler, Wolf-Georg Zaddach, Jakob Abeßer, and Martin Pfleiderer. Introducing the jazzomat project and the melospy library. In *Third International Workshop on Folk Music Analysis*, 2013.

[69] Martin Gasser, Andreas Arzt, Thassilo Gadermaier, Maarten Grachten, and Gerhard Widmer. Classical music on the web - user interfaces and data representations. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 571–577, 2015.

[70] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 776–780, 2017.

[71] Emilia Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[72] Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 437–440, Hong Kong, China, 2003.

[73] Masataka Goto. Music listening in the future: Augmented music-understanding interfaces and crowd music listening. In *Proceedings of the Audio Engineering Society (AES) Conference on Semantic Audio*, Ilmenau, Germany, 2011.

[74] Masataka Goto. Frontiers of music information research based on signal processing. In *Proceedings of the International Conference on Signal Processing (ICSP)*, pages 7–14, 2014.

[75] Masataka Goto, Kazuyoshi Yoshii, Hiromasa Fujihara, Matthias Mauch, and Tomoyasu Nakano. Songle: A web service for active music listening improved by user contributions. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 311–316, Miami, Florida, USA, 2011.

[76] Emad M. Grais, Gerard Roma, Andrew J. R. Simpson, and Mark D. Plumbley. Single-channel audio source separation using deep neural network ensembles. In *Proceedings of the Audio Engineering Society (AES) Convention*, Paris, France, 2016.

[77] Thomas Grill and Jan Schlüter. Music boundary detection using neural networks on spectrograms and self-similarity lag matrices. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 1296–1300, Nice, France, 2015.

[78] Thomas Grill and Jan Schlüter. Music boundary detection using neural networks on combined features and two-level annotations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 531–537, Málaga, Spain, 2015.

[79] Peter Grosche and Meinard Müller. Toward characteristic audio shingles for efficient cross-version music retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[80] Peter Grosche, Meinard Müller, and Joan Serrà. Audio content-based music retrieval. In Meinard Müller, Masataka Goto, and Markus Schedl, editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 157–174. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2012. URL `http://drops.dagstuhl.de/opus/volltexte/2012/3471`.

[81] Jan Hajič jr. and Matthias Dorfer. Prototyping full-pipeline optical music recognition with music-marker. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR): Late Breaking session*, Suzhou, China, 2017.

[82] Philippe Hamel and Douglas Eck. Learning features from music audio with deep belief networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 339–344, Utrecht, The Netherlands, 2010.

[83] Philippe Hamel, Sean Wood, and Douglas Eck. Automatic identification of instrument classes in polyphonic and poly-instrument audio. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 399–404, Kobe, Japan, 2009.

[84] Philippe Hamel, Simon Lemieux, Yoshua Bengio, and Douglas Eck. Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 729–734, Miami, Florida, 2011.

[85] Philippe Hamel, Yoshua Bengio, and Douglas Eck. Building musically-relevant audio features through multiple timescale representations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 553–558, Porto, Portugal, 2012.

[86] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 131–135, 2017.

[87] Andre Holzapfel and Thomas Grill. Bayesian meter tracking on learned signal representations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 262–268, New York City, USA, 2016.

[88] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[89] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 477–482, Taipei, Taiwan, 2014.

[90] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2136–2147, 2015.

[91] Eric J. Humphrey and Juan P. Bello. Rethinking automatic chord recognition with convolutional neural networks. In *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 357–362, Boca Raton, USA, 2012.

[92] Eric J. Humphrey, Taemin Cho, and Juan P. Bello. Learning a robust tonnetz-space transform for automatic chord recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 453–456, Kyoto, Japan, 2012.

[93] David B. Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. The MIT Press, 2006.

[94] Il-Young Jeong and Kyogu Lee. Learning temporal features using a deep neural network and its application to music genre classification. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 434–440, New York City, USA, 2016.

[95] M. Cameron Jones, J. Stephen Downie, and Andreas F. Ehmann. Human similarity judgments: Implications for the design of formal evaluations. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 539–542, Vienna, Austria, 2007.

[96] Justin Salamon Juan J. Bosch, Rachel M. Bittner and Emilia Gómez. A comparison of melody extraction methods based on source-filter modelling. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 571–577, New York City, USA, 2016.

[97] Rainer Kelz and Gerhard Widmer. An experimental analysis of the entanglement problem in neural-network-based music transcription systems. In *Proceedings of the AES International Conference on Semantic Audio*, pages 194–201, Erlangen, Germany, 2017.

[98] Rainer Kelz, Matthias Dorfer, Filip Korzeniowski, Sebastian Böck, Andreas Arzt, and Gerhard Widmer. On the potential of simple framewise approaches to piano transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 475–481, New York City, USA, 2016.

[99] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015.

[100] Peter Knees and Markus Schedl. *Music Similarity and Retrieval*. Springer Verlag, 2016. ISBN 978-3-662-49720-3.

[101] Filip Korzeniowski and Gerhard Widmer. Feature learning for chord recognition: The deep chroma extractor. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 37–43, New York City, USA, 2016.

[102] Filip Korzeniowski and Gerhard Widmer. End-to-end musical key estimation using a convolutional neural network. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, 2017.

[103] Filip Korzeniowski and Gerhard Widmer. On the futility of learning complex frame-level language models for chord recognition. In *Proceedings of the AES International Conference on Semantic Audio*, pages 179–185, Erlangen, Germany, 2017.

[104] Florian Krebs, Sebastian Böck, Matthias Dorfer, and Gerhard Widmer. Downbeat tracking using beat synchronous features with recurrent neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 129–135, New York City, USA, 2016.

[105] Carol L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford, UK, 1990.

[106] Sangeun Kum, Changheun Oh, and Juhan Nam. Melody extraction on vocal segments using multi-column deep neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 819–825, New York City, USA, 2016.

[107] Frank Kurth and Meinard Müller. Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, 2008. URL `http://dx.doi.org/10.1109/TASL.2007.911552`.

[108] Frank Kurth, Meinard Müller, David Damm, Christian Fremerey, Andreas Ribbrock, and Michael Clausen. SyncPlayer - an advanced system for multimodal music access. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 381–388, London, UK, November 2005.

[109] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[110] Simon Leglaive, Romain Hennequin, and Roland Badeau. Singing voice detection with deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 121–125, Brisbane, Australia, 2015.

[111] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner. On the reduction of false positives in singing voice detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7480–7484, Florence, Italy, 2014.

[112] Bernhard Lehner, Gerhard Widmer, and Sebastian Böck. A low-latency, real-time-capable singing voice detection method with lstm recurrent neural networks. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 21–25, Nice, France, 2015.

[113] Cynthia C. S. Liem, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic. The need for music information retrieval with user-centered and multimodal strategies. In *Proceedings of the International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM)*, pages 1–6, 2011.

[114] Cynthia C. S. Liem, Meinard Müller, Steven K. Tjoa, and George Tzanetakis. 2nd International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM). In *Proceedings of the International Conference on Multimedia (ACM Multimedia)*, pages 1509–1510, 2012.

[115] Cynthia C. S. Liem, Emilia Gómez, and Markus Schedl. PHENICX: Innovating the classical music experience. In *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–4, Torino, Italy, June–July 2015.

[116] Cynthia C. S. Liem, Emilia Gómez, and George Tzanetakis. Multimedia technologies for enriched music performance, production, and consumption. *IEEE MultiMedia*, 24(1):20–23, 2017.

[117] Vincent Lostanlen and Carmine-Emanuele Cella. Deep convolutional networks on the pitch spiral for music instrument recognition. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 612–618, New York City, USA, 2016.

[118] Yi Luo, Zhuo Chen, John R. Hershey, Jonathan Le Roux, and Nima Mesgarani. Deep clustering and conventional networks for music separation: Stronger together. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–65, New Orleans, USA, 2017.

[119] Matija Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE/ACM Transactions on Multimedia*, 6(3):439–449, 2004.

[120] Taro Masuda, Kazuyoshi Yoshii, Masataka Goto, and Shigeo Morishima. Spotting a query phrase from polyphonic music audio signals based on semi-supervised nonnegative matrix factorization. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 227–232, 2014.

[121] Matthias Mauch and Simon Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663, 2014.

[122] Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justing Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In *Proceedings of the International Conference on Technologies for Music Notation and Representation*, May 2015.

[123] Brian McFee, Matt McVicar, Colin Raffel, Dawen Liang, Oriol Nieto, Eric Battenberg, Josh Moore, Dan Ellis, Ryuichi Yamamoto, Rachel Bittner, Douglas Repetto, Petr Viktorin, João Felipe Santos, and Adrian Holovaty. librosa: 0.4.1, 2015. URL `http://dx.doi.org/10.5281/zenodo.32193`.

[124] Brian McFee, Matt McVicar, Oriol Nieto, Stefan Balke, Carl Thomé, Dawen Liang, Eric Battenberg, Josh Moore, Rachel Bittner, Ryuichi Yamamoto, Dan Ellis, Fabian-Robert Stöter, Douglas Repetto, Simon Waloschek, CJ Carr, Seth Kranzler, Keunwoo Choi, Petr Viktorin, Joao Felipe Santos, Adrian Holovaty, Waldir Pimenta, and Hojin Lee. librosa 0.5.0, 2017. URL `https://doi.org/10.5281/zenodo.293021`.

[125] Mark S Melenhorst, Ron van der Sterren, Andreas Arzt, Agustín Martorell, and Cynthia C.S. Liem. A tablet app to enrich the live and post-live experience of classical concerts. In *Proceedings of the International Workshop on Interactive Content Consumption (WSICC)*, 2015.

[126] Klaus Meyer-Wegener. *Multimedia Datenbanken*. Teubner, Wiesbaden, 2 edition, 2003.

[127] MIREX. Audio melody extraction task. Website `http://www.music-ir.org/mirex/wiki/2015:Audio_Melody_Extraction`, last accessed 01/19/2016, 2015.

[128] Marius Miron, Jordi Janer, and Emilia Gómez. Monaural score-informed source separation for classical music using convolutional neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 55–62, Suzhou, China, 2017.

[129] Nicola Montecchio, Emanuele Di Buccio, and Nicola Orio. An efficient identification methodology for improved access to music heritage collections. *Journal of Multimedia*, 7(2):145–158, 2012.

[130] Meinard Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007. ISBN 3540740473.

[131] Meinard Müller. *Fundamentals of Music Processing*. Springer Verlag, 2015. ISBN 978-3-319-21944-8.

[132] Meinard Müller and Sebastian Ewert. Towards timbre-invariant audio features for harmony-based music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):649–662, 2010. URL `http://dx.doi.org/10.1109/TASL.2010.2041394`.

[133] Meinard Müller and Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 215–220, Miami, Florida, USA, 2011.

[134] Meinard Müller, Frank Kurth, and Michael Clausen. Chroma-based statistical audio features for audio matching. In *Proceedings of the IEEE Workshop on Applications of Signal Processing (WASPAA)*, pages 275–278, New Paltz, NY, USA, October 2005.

[135] Meinard Müller, Peter Grosche, and Frans Wiering. Robust segmentation and annotation of folk song recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 735–740, Kobe, Japan, October 2009.

[136] Meinard Müller, Masataka Goto, and Markus Schedl, editors. *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, 2012. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany. ISBN 978-3-939897-37-8.

[137] Juhan Nam, Jiquan Ngiam, Honglak Lee, and Malcolm Slaney. A classification-based polyphonic piano transcription approach using learned feature representations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 175–180, Miami, Florida, 2011.

[138] Oriol Nieto, Morwaread Farbood, Tristan Jehan, and Juan Pablo Bello. Perceptual analysis of the F-measure to evaluate section boundaries in music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 265–270, Taipei, Taiwan, 2014.

[139] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel music separation with deep neural networks. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 1748–1752, Budapest, Hungary, 2016.

[140] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664, 2016.

[141] Nicloa Orio. Music retrieval: A tutorial and review. *Foundation and Trends in Information Retrieval*, 1(1):1–90, 2006.

[142] Hyunsin Park and Chang D. Yoo. Melody extraction and detection through LSTM-RNN with harmonic sum loss. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2766–2770, New Orleans, USA, 2017.

[143] M. Cristina Pattuelli. Personal name vocabularies as linked open data: A case study of jazz artist names. *Journal of Information Science*, 38(6):558–565, 2012.

[144] Jouni Paulus and Anssi P. Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17 (6):1159–1170, 2009.

[145] Steffen Pauws. CubyHum: a fully operational query by humming system. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2002.

[146] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[147] Martin Pfleiderer, Klaus Frieler, Jakob Abeßer, Wolf-Georg Zaddach, and Benjamin Burkhart, editors. *Inside the Jazzomat. New perspectives for jazz research.* Schott Campus, Mainz, Germany, 2017.

[148] Jeremy Pickens, Juan Pablo Bello, Giuliano Monti, Tim Crawford, Matthew Dovey, Mark Sandler, and Don Byrd. Polyphonic score retrieval using polyphonic audio. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2002.

[149] Jordi Pons and Xavier Serra. Designing efficient architectures for modeling temporal features with convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2472–2476, 2017.

[150] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra. Timbre analysis of music audio signals with convolutional neural networks. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, 2017.

[151] Thomas Prätzlich, Meinard Müller, Benjamin W. Bohl, and Joachim Veit. Freischütz Digital: Demos of audio-related contributions. In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Malága, Spain, 2015.

[152] Thomas Prätzlich, Jonathan Driedger, and Meinard Müller. Memory-restricted multiscale dynamic time warping. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 569–573, Shanghai, China, 2016.

[153] Colin Raffel and Dan P. W. Ellis. Pruning subsequence search with attention-based embedding. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 554–558, Shanghai, China, 2016.

[154] Colin Raffel and Daniel P. W. Ellis. Accelerating multimodal sequence retrieval with convolutional networks. In *Proceedings of the NIPS Multimodal Machine Learning Workshop*, Montréal, Canada, 2015.

[155] Colin Raffel and Daniel P. W. Ellis. Large-scale content-based matching of MIDI and audio files. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 234–240, Málaga, Spain, 2015.

[156] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. MIR_EVAL: A transparent implementation of common MIR metrics. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 367–372, Taipei, Taiwan, 2014.

[157] Yves Raimond, Samer Abdallah, Mark Sandler, and Frederick Giasson. The Music Ontology. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 417–422, Vienna, Austria, 2007.

[158] Christopher Raphael and Jingya Wang. New approaches to optical music recognition. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 305–310, Miami, Florida, USA, 2011.

[159] Ana Rebelo, Ichiro Fujinaga, Filipe Paszkiewicz, Andre R. S. Marcal, Carlos Guedes, and Jaime S. Cardoso. Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, 1(3):173–190, 2012.

[160] François Rigaud and Mathieu Radenen. Singing voice melody transcription using deep neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 737–743, New York City, USA, 2016.

[161] Daniel Röwenstrunk, Thomas Prätzlich, Thomas Betzwieser, Meinard Müller, Gerd Szwillus, and Joachim Veit. Das Gesamtkunstwerk Oper aus Datensicht – Aspekte des Umgangs mit einer heterogenen Datenlage im BMBF-Projekt "Freischütz Digital". *Datenbank-Spektrum*, 15(1):65–72, 2015. doi: 10.1007/s13222-015-0179-0.

[162] Matti Ryynänen and Anssi Klapuri. Query by humming of MIDI and audio using locality sensitive hashing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2249–2252, Las Vegas, Nevada, USA, 2008.

[163] Matti Ryynänen and Anssi P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.

[164] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6): 1759–1770, 2012.

[165] Justin Salamon and Julián Urbano. Current challenges in the evaluation of predominant melody extraction algorithms. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 289–294, Porto, Portugal, October 2012.

[166] Justin Salamon, Joan Serrà, and Emilia Gómez. Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval*, 2(1):45–58, 2013. doi: 10.1007/s13735-012-0026-0. URL `http://dx.doi.org/10.1007/s13735-012-0026-0`.

[167] Justin Salamon, Emilia Gómez, Daniel P. W. Ellis, and Gaël Richard. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 31(2):118–134, 2014. doi: 10.1109/MSP.2013.2271648. URL `http://dx.doi.org/10.1109/MSP.2013.2271648`.

[168] Jan Schlüter. Learning to pinpoint singing voice from weakly labeled examples. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 44–50, New York City, USA, 2016.

[169] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 121–126, Málaga, Spain, 2015.

[170] Erik M. Schmidt and Youngmoo Kim. Learning rhythm and melody features with deep belief networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 21–26, Curitiba, Brazil, 2013.

[171] Michael Schoeffler and Jürgen Herre. The influence of audio quality on the popularity of music videos: A YouTube case study. In *Proceedings of the International Workshop on Internet-Scale Multimedia Management*, pages 35–38, 2014.

[172] Jacob T. Schwartz and Diana Schwartz. The electronic dictionary of musical themes. Website `https://web.archive.org/web/20160330030505/http://multimedialibrary.com/barlow/all_barlow.asp`, last accessed 01/12/2015, 2008.

[173] Joan Serrà, Emilia Gómez, and Perfecto Herrera. Audio cover song identification and similarity: Background, approaches, evaluation and beyond. In Z. W. Ras and A. A. Wieczorkowska, editors, *Advances in Music Information Retrieval*, volume 274 of *Studies in Computational Intelligence*, chapter 14, pages 307–332. Springer, Berlin, Germany, 2010.

[174] Siddharth Sigtia, Emmanouil Benetos, Srikanth Cherla, Tillman Weyde, Artur S. d'Avila Garcez, and Simon Dixon. An rnn-based music language model for improving automatic music transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 53–58, Taipei, Taiwan, 2014.

[175] Siddharth Sigtia, Nicolas Boulanger-Lewandowski, and Simon Dixon. Audio chord recognition with a hybrid recurrent neural network. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 127–133, Málaga, Spain, 2015.

[176] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):927–939, 2016.

[177] Andrew J. R. Simpson, Gerard Roma, and Mark D. Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 429–436, Liberec, Czech Republic, 2015.

[178] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 555–560, Miami, Florida, USA, 2011.

[179] Ray Smith. An overview of the tesseract OCR engine. In *ICDAR*, volume 7, pages 629–633, 2007.

[180] Reinhard Sonnleitner and Gerhard Widmer. Robust quad-based audio fingerprinting. *IEEE Transactions on Audio, Speech, and Language Processing*, 24(3):409–421, 2016. doi: 10.1109/TASLP.2015.2509248.

[181] Reinhard Sonnleitner, Andreas Arzt, and Gerhard Widmer. Landmark-based audio fingerprinting for DJ mix monitoring. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 185–191, New York City, New York, USA, August 2016.

[182] Carl Southall, Ryan Stables, and Jason Hockman. Automatic drum transcription using bi-directional recurrent neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–597, New York City, USA, August 2016.

[183] Stanley Smith Stevens, John Volkmann, and Edwin B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937. doi: 10.1121/1.1915893.

[184] Iman S.H. Suyoto, Alexandra L. Uitdenbogerd, and Falk Scholer. Searching musical audio using symbolic queries. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):372–381, 2008.

[185] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, 2016. URL `http://arxiv.org/abs/1605.02688`.

[186] Verena Thomas, Christian Fremerey, David Damm, and Michael Clausen. SLAVE: a Score-Lyrics-Audio-Video-Explorer. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 717–722, Kobe, Japan, 2009.

[187] Rainer Typke, Frans Wiering, and Remco C. Veltkamp. A survey of music information retrieval systems. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 153–160, London, UK, 2005.

[188] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji. Deep neural network based instrument extraction from music. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139, Brisbane, Australia, 2015.

[189] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji. Improving music source separation based on deep neural networks through data augmentation and network blending. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 261–265, New Orleans, USA, 2017.

[190] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary detection in music structure analysis using convolutional neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 417–422, Taipei, Taiwan, 2014.

[191] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. Transfer learning by supervised pre-training for audio-based music classification. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 29–34, Taipei, Taiwan, 2014.

[192] Tay Vaughan. *Multimedia: Making It Work*. McGraw-Hill, 2011. ISBN 978-0-07-174850-6.

[193] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1096–1103, Helsinki, Finland, June 2008. doi: 10.1145/1390156.1390294.

[194] Richard Vogl, Matthias Dorfer, and Peter Knees. Recurrent neural networks for drum transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 730–736, New York City, USA, August 2016.

[195] Claus Weihs, Dietmar Jannach, Igor Vatolkin, and Guenter Rudolph. *Music Data Analysis: Foundations and Applications*. CRC Press, 2016. ISBN 978-1-498-71956-8.

[196] Christof Weiß, Rainer Kleinertz, and Meinard Müller. Möglichkeiten der computergestützten Erkennung und Visualisierung harmonischer Strukturen – eine Fallstudie zu Richard Wagners "Die Walküre". In Wolfgang Auhagen and Wolfgang Hirschmann, editors, *Bericht zur Jahrestagung der Gesellschaft für Musikforschung (GfM) 2015 in Halle/Saale*, Mainz, Germany, 2016. Schott Campus.

[197] Nils Werner, Stefan Balke, Fabian-Robert Stöter, Meinard Müller, and Bernd Edler. trackswitch.js: A versatile web-based audio player for presenting scientific results. In *Proceedings of the Web Audio Conference (WAC)*, 2017.

[198] Frans Wiering, Tim Crawford, and David Lewis. Digital critical editions of music: A multidimensional model. In Tim Crawford and Lorna Gibson, editors, *Modern Methods for Musicology*, pages 23–45. Routledge, 2009.

[199] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

[200] Xinquan Zhou and Alexander Lerch. Chord detection using deep learning. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 52–58, Málaga, Spain, 2015.