Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Automatic Harmony Analysis of Jazz Audio Recordings

Vsevolod Eremenko

Supervisor: Xavier Serra

Co-Supervisor: Baris Bozkurt

August 2018

Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Automatic Harmony Analysis of Jazz Audio Recordings

Vsevolod Eremenko

Supervisor: Xavier Serra

Co-Supervisor: Baris Bozkurt

August 2018

Contents

Acknowledgement

First and foremost, I would like to thank Xavier Serra for introducing me to the world of Sound and Music Computing with his online course and then "live" at the MTG, suggesting a fascinating topic for the thesis and his trust and confidence. I want to thank Baris Bozkurt for his practical approach and numerous advices helping to shape this work.

Also, I want to thank Emir Demirel for his meticulous chord annotations for JAAH dataset. Without him, this thesis would not be possible. Thanks to Ceyda Ergul for helping with metadata annotations.

Big thanks to my lab and MTG colleagues for the fruitful discussions, help, and encouragement, especially to Rafael, Rong, Oriol, Alastair, Albin, Pritish, Eduardo, Olga, and Dmitry.

Warm thanks to the fellow SMC Master students with whom we have lived so many good experiences and learned a lot. Especially to Tessy, Pablo, Dani, Minz, Joe, Kushagra, Gerard, Marc, and Natalia.

Finally, great thanks to my wife Olga and sons Petr and Iván for taking courage to travel with me into the unknown, for encouraging me to study and complete this work.

Abstract

This thesis aims to develop a style specific approach to Automatic Chord Estimation and computer-aided harmony analysis for jazz audio recordings. We build a dataset of time-aligned jazz harmony transcriptions, develop an evaluation metrics which accounts for jazz specificity in assessing the quality of automatic chord transcription. Then we evaluate some existing state of the art algorithms and develop our own using beat detection, chroma features extraction, and probabilistic model as building blocks. Also, we suggest a novel way of visualizing chroma distribution based on Compositional Data Analysis techniques. The visualization allows exploring the specificity of chords rendering in general and in different jazz sub-genres. The presented work makes a step toward expanding current Music Information Retrieval (MIR) approaches for Audio Chord Estimation task, which are currently biased towards rock and pop music.

Keywords: Jazz; Harmony; Datasets; Automatic Chord Estimation; Chroma; Compositional Data Analysis

Chapter 1

Introduction

1.1 Context

This work was started in the vein of the CompMusic project described by Serra in [1], which emphasize the computational study of musical artifacts in a specific cultural context. While CompMusic includes researches on rhythm, intonation, and melody in several non-Western traditions, here we concentrate on harmony phenomenon in jazz using the similar workflow (Figure 1). We assemble a corpus of manually annotated audio recordings, consider approaches to generate chords annotations automatically and provide tools for finding typical patterns in chord sequences and in chroma features distribution.

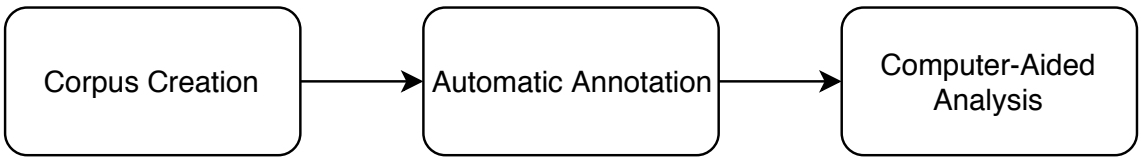

Figure 1: CompMusic-inspired workflow.

The other basis for the work is Audio Chord Estimation (ACE) task in Music Information Retrieval (MIR) field [2]. There are vast related literature and a lot of implemented algorithms for chord label extraction, but all the concepts validated mainly for mainstream rock and pop music. We aim to adopt these concepts to jazz, thus pushing the boundaries of the current MIR approach.

1.2 Motivation

Harmony is an essential concept in many jazz sub-genres. It is not only used for expressive purposes, but serves as a vehicle for improvisation and the base upon which the interaction between musicians is built. Knowing tune's harmony is necessary for any further analysis either performed by music theorist or practicing musician who studies a particular style. Audio recording is the primary source of information about jazz performance. Harmonic analysis of an audio recording requires transcription "by ear" which is a time-consuming process and requires certain experience and qualification. Reliable automatic harmony transcription tool would make possible large-scale corpus-based musicological research and improve musicians learning experience (e.g., a student could find all the recordings, where chosen artist plays over specific chord sequence).

For me as for jazz guitar student, transcribing harmony was a slow process, and I was always looking for a "second opinion" to assure me. Besides, I am fascinated with the gap between the abstraction of chord symbols and its rendering by musicians. So I am looking for a way to transcribe chord labels automatically and trying to capture a particular performer's style elusive "sonic aura."

In MIR, chord detection is also considered as a building block for many other tasks, e.g., Structural Segmentation, Cover Song Identification and Genre Classification. Jazz idiom also affects some popular styles such as funk, soul, and jazz-influenced pop. Thus, improving automatic chord transcription for jazz would also improve solution for multiple MIR tasks for many styles.

1.3 Objectives

Main objectives of this study are:

- Build a representative collection of jazz audio recordings with human-performed audio-aligned harmony annotations which could be used for research in automatic harmony analysis.

- Evaluate and develop methods of automatic harmony transcription in a way informed of jazz practice and tradition.
- Provide tools for computer-aided harmony analysis.

1.4 Structure of the Report

After the introduction in Chapter 1, Chapter 2 gives an overview of the musical, cognitional and engineering concepts used in the thesis and presents state of the art.

In Chapter 3 we justify the necessity of creating a representative, balanced jazz dataset in a new format and formulate our guidelines for building the dataset. We describe the annotated attributes, track selection principle, and transcription methodology. Then we provide and discuss the basic statistical summary of the dataset.

Chapter 4 considers existing and presents a novel way of visualizing joint chroma features distribution for chord classes based on ternary plots. It discusses general aspects of chroma sampling distribution and its application to characterization of performance style.

In Chapter 5 we discuss how to evaluate the performance of Automatic Chord Transcription on jazz recordings. Baseline evaluation scores for two state-of-the-art chord estimation algorithms are shown. We implement a basic automatic transcription pipeline which allows to evaluate performance of the algorithm components separately and make suggestions about further improvements.

Finally, in Chapter 6 we summarize the results, underline the contributions and provide suggestions for future work.

Chapter 2

Background

In the first section of this chapter, I present accounts of harmony from perspectives of:

- jazz musicians and theorists
- music perception and cognition psychologists
- music information retrieval engineers

I'll try to find more connections between them throughout the narration.

Datasets related to chords and harmony are in particular interest to all these parties, so they are reviewed in a dedicated section.

The last section reviews probabilistic and machine learning concepts used in this work.

2.1 "Three Views of a Secret"

2.1.1 Harmony in Jazz

Introduction

As Strunk laid down in [3], harmony is "the combining of notes simultaneously to produce chords and the placing of chords in succession".

Berliner states in his ethnomusicological study [4], that for the most jazz styles[1], composed tunes consisting of a melody and accompanying harmonic progression (or "changes", as performers call it) provides the structure for improvisation. Usually, musicians perform original melody in the first and the last sections of the song and improvised solos goes in between, in a cyclical form. Harmony not only contributes to a tune's mood and character but serves as a "timing cycle which guides, stimulates and limits jazz solo playing."

Whether a tune was composed by performers themselves or being chosen from the repertory of jazz standards, jazz piece must be analyzed as work which is "created by musicians at each performance event" ([4]). This is stipulated by what Berliner has called "The Malleability of Form": musicians use a range of techniques to individualize chords changes. They usually make decisions about harmony during rehearsals, and some features could be determined immediately before music events or while actually performing.

Berliner calls jazz "Ear Music," which emphasize the fact that musicians often perform without a written score, "creating much of the detail of their music in performance" and their skills are "aural knowledge", e.g.: "some artists remain ear and hand players", "some musicians conceptualize the structure of a piece primarily in aural and physical terms as a winding melodic course through successive fields of

[1] Nowadays jazz embraces many styles and approaches used by musicians to improvise and communicate with each other. But we'll refer mainly to formative years of jazz (1900-1960s) starting from blues and ragtime through New Orleans jazz, swing, bebop, hard bop and some later styles (such as Latin jazz and bossa nova). We consider only styles where chord changes played a significant role (thus, such styles as free jazz, where harmony is non-existent or modal jazz, where it's static are left aside). List of typical examples could be found at `https://mtg.github.io/JAAH`

distinctive harmonic color ...". Taking this into account, we could see that audio recordings are the primary sources of reliable information about jazz music.

Harmony Traces in the Graphemic Domain

According to Berliner [4], some musicians find it useful to reinforce their mental picture of the tune's harmony with notational and theoretical symbols. Let's consider it in historical perspective: for jazz veterans, harmony was primarily an aural knowledge. Perhaps, because at the time being (1900-1930s), sheet music score was the predominant way of the popular music notation (see [5]), and it doesn't seem to provide the necessary level of abstraction for jazz musicians. At Figure 2 you could see a typical three-staff score with full piano accompaniment and vocal melody published for the general public, who were expected to realize it as it was written.

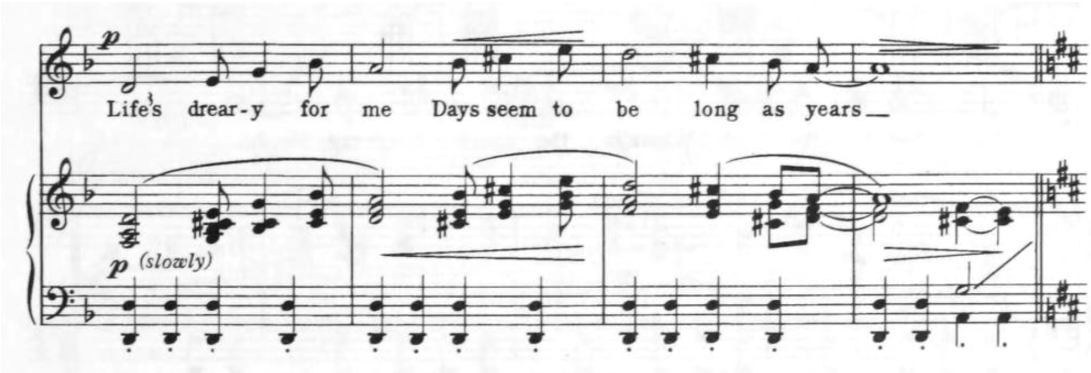

Figure 2: Excerpt from "Body and Soul" sheet music (1930s publication) [5].

According to Kernfeld [5], "craze for the ukulele" of the 1920s led to the addition of ukulele chord tablature above the vocal stave. Sometimes chord labels were added (as at Figure 3), so the same chords could be read by the guitar player.

Later, in some publications tablatures were dropped out, but chord symbols remain. According to Kernfeld, this transition from piano music to tablatures and then to chord symbols represented "an absolutely crucial move from the specific to the abstract". It gave rise to phenomena of fake books, which were widespread between pop and jazz musicians (see example at Figure 4). Fake book assumes that a performer

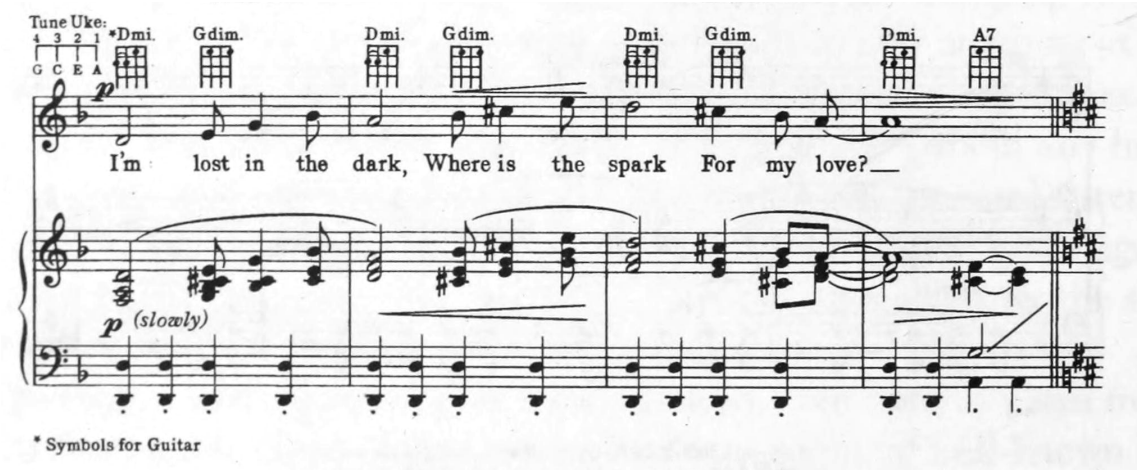

Figure 3: Excerpt from "Body and Soul" sheet music with ukulele tablature and chord symbols (1930s publication) [5].

should make his or her own version of the song, or "fake it." Kernfeld underlines, that chord symbols says nothing about how these chords have to be realized.

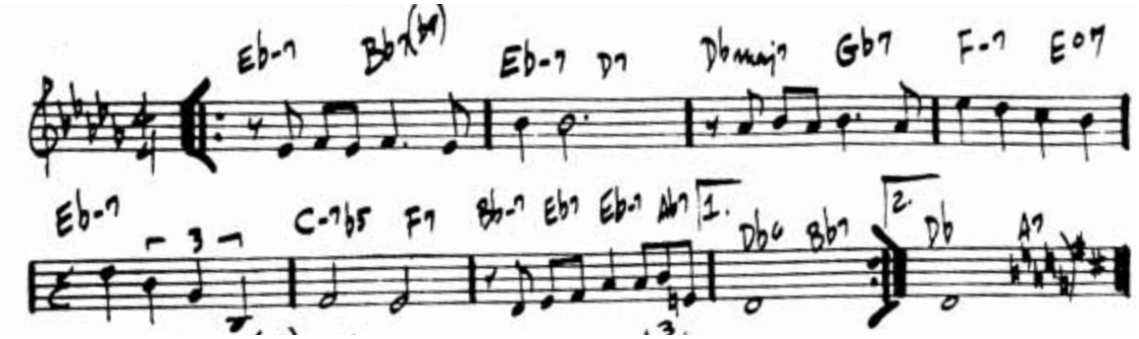

Figure 4: Excerpt from "Body and Soul" from a jazz musician's fake book.

Basic Aspects of Harmony in Jazz Theory

A detailed account of harmony is a huge topic, so I'll review only basic facts or aspects which could present an issue for or could be exploited during automatic analysis of audio recordings. Strunk gives introduction to the area in [3]. Martin presents a theoretical analytical approach in [6], [7]. For a pedagogical instructional account, one could check books by Levine [8], [9] or theory primer for popular Aebersold play-along series [10].

Navigating through chords types. Chord symbols consist of glued together root pitch class and chord type, e.g., in Figure 4 "$E\flat^{-7}$" symbol contains "$E\flat$" root

Chord Type	Abbreviation	Degrees in seventh chord rendering
Major	maj	$I - III - [V] - VII$
Minor	min	$I - III\flat - [V] - VII\flat$
Dominant	dom7	$I - III - [V] - VII\flat$
Half-diminished	hdim7	$I - III\flat - V\flat - VII\flat$
Diminished	dim	$I - III\flat - V\flat - VII\flat\flat$

Table 1: Main chord types.

and "$^{-7}$" type, which means $E\flat$ minor seventh chord. Type component could optionally contain information about added and altered degrees (e.g., "$B\flat 7^{(\flat 9}$" denotes $B\flat$ dominant seventh chord with flat nine). Also, chord inversion is sometimes encoded with the bass note after the slash (e.g., "C^7/G" means C dominant seventh with the bass G).

While preparing a lead sheet database, Pachet et al. encountered 86 types of chords in various sources [11]. But are all of them equally important for music analysis and interpretation? Most of the theoretical, e.g., [7], and pedagogical, e.g., [10], literature consider five main chord types: major, minor, dominant seventh, half-diminshed seventh, and diminished seventh (presented in Table 1).

As noted by Martin [7], sevenths are the most common chords, sixths are also found in jazz, but to a smaller degree. Though in some early styles (e.g., gypsy jazz of Django Reinhardt) triads and sixth chords were widespread. A perfect fifth degree is not required for distinguishing between major, minor and dominant seventh chord types, so it's often omitted. In pedagogical literature, such minimalistic approach is called "three-note voicing" [8]

Of course, other chord taxonomies are also used. For example, to trace changes in 20th-century jazz harmonic practice, Broze and Shanahan assembled a corpus of symbolic data covering compositions from the year 1900 to 2005 and use eight chord types for the analysis: dominant sevenths, minor sevenths, major, minor, half-diminished, diminished, sustained, augmented. But augmented chord is notated very rarely (0.1 percent in Broze and Shanahan corpus), sustained chord (1.0 percent in their corpus) used in modal and post-modal jazz which we decided to exclude from

the consideration.

When chord sequence is analyzed, roman numerals are used sometimes to denote a chord's root relation to local tonal center[3].

Chords rendering. Comping. According to Strunk[3], many experts regard the concept of chord inversion as not generally relevant to jazz. Martin [6] also considers chords as pitch class sets. Perhaps, such approach is prevalent because the bass lines are improvised and the bass player usually plays a root of a chord at least once while the chord lasts [3].

The art of improvising accompaniment in jazz is called "comping" ("a term that carries the dual connotations of accompanying and complementing" [4]). Comping is performed collectively by the rhythm section (e.g., drums, bass, piano, guitar, other non-soloing instruments). To give an idea, how varied chords rendering could be, let's draw a few examples concerning tonal aspects of comping from Berliner [4].

Bass player could interpret chords "literally or more allusively — at times, even elusively". He or she could use tones other than chord's root on the downbeat, may emphasize non-chord tones to create harmonic color and suspense. Bass players choose some pitches to represent the underlying harmony, and they select other pitches to make smooth and interesting connections between chord tones.

A piano player could decorate chord progression with embellishing chords, play counterpoint to soloist melody or use so-called "orchestral approach" to comping, which means that he or she could create wide-ranging textures, interweave simultaneous melodies, and thus go far beyond chord tones.

Soloing "outside" harmony. Unlike European tradition, where melodies are very often horizontal projections of a harmonic substructure [12], jazz allows the soloist to play with harmony in different ways. In particular:

- "vertical" approach, when chord tones are involved and the soloist pays attention to each chord in the progression.

- "horizontal" approach, when the soloist plays in certain mode related to the local tonal center and doesn't outline all the passing chords.

- playing "outside", when the soloist deliberately plays pitches "outside" from the chord played by the rhythm section [4].

E.g., one of the simplest "outside" playing devices described by Levine [9] is to play half step up or down from the original chord. Many of such "outside" moments could be qualified as bitonality. We should regard this point when transcribing the chord progression from an audio recording.

Popular patterns. Harmonic sequences derived from the circle of fifths or chromatic movement are widespread [7]. As pianist Fred Hersch stated during an interview [4]: "there were as few as ten or so different harmonic patterns."

There are works which apply Natural Language Processing (NLP) techniques to the problem of analyzing harmonic progressions of jazz standards in the symbolic domain ([13], [14]).

Harmonic structure regularity. Comparing to Western concert music, the progression structure in jazz is very regular:

- The form of the whole performance is mainly cyclic: the same chorus is repeated over and over. The number of bars in a chorus is usually thirty-two, sixteen or twelve (blues). Occasional intermissions are typically divisible-by-four bars length.

- Hypermetric regularity [15]. Chorus usually consists of four or eight bars long harmonic phrases ("hypermeasures"). Each chord in a phrase lasts for two, four, six or eight beats.

Thus, jazz tune harmonic structure reminds brickwork: there are lines of even bricks (harmonic phrases), some lines could be shifted but usually only by half-brick.

Note about harmony transcription. There's a huge jazz literature, which analyses and teaches how to render abstract chord labels into music, and I briefly reviewed some ideas in previous paragraphs. But up to my knowledge, there's no detailed account about how musicians transcribe it, what aural cues they use for that purpose and how they represent it mentally. According to Berliner [4], it's a complex and highly individual process. E.g., while some prodigies could transcribe harmony as they hear it at the age of seven, for many, the process is largely "a matter of trial and error, trying out different pitches until you get as close as you can to the quality of the chords" (from an interview with Kenny Barron). To shed some light on it, let's turn to the literature on music perception.

2.1.2 Harmony Perception

Chords and pitch. In their review [16] McDermott and Oxenham conclude, that the neural representation of chords could be an interesting direction for future research. In some cases it is clear that multiple pitches can be simultaneously represented; in others, listeners appear to perceive an aggregate sound. They formulate the intriguing hypothesis that for more than three simultaneous pitches, chord tones are not naturally individually represented, therefore chord properties must be conveyed via aggregate acoustic features that are as yet unappreciated. The following facts could be drawn to support the hypothesis:

- It's proven that humans separate tones in simultaneous two tones intervals. Though no study was performed for musical chords of three notes, there was an experiment with artificial "chords" generated from random combinations of pure tones. If there are more then three tones, for the listener they are fused together into a single sound, even when the frequencies are not harmonically related.

- It's hard for a listener to name the tones comprising a chord, it requires considerable practice in ear training, while far less experienced listener could properly identify chord quality (e.g., major or minor).

Chords and sensory consonance. According to [16], consonance is another widely studied property of chords. Western listeners consistently regard some combinations of notes as pleasant (consonant), whereas others seem unpleasant (dissonant).

The most prevalent theory of consonant and dissonance perception is usually attributed to Helmholtz, who links dissonance with roughness ("beating" or fluctuation of the amplitude of superposition of two tone's partials with adjacent frequencies). A physiological correlate of roughness have been observed in both monkeys and humans.

An alternative theory also has plausibility. Consonant pair of notes produce partials which could be generated by a single complex tone with a lower but still perceivable fundamental frequency. Partials of dissonant intervals theoretically could be produced by a single complex tone with an implausibly low F0, often below the lower limit of pitch (of around 30 Hz). As with roughness, there are physiological correlates of this periodicity: periodic responses are observed in the auditory nerve of cats for the consonant intervals, but not for the dissonant intervals [16].

Pitch. Pitch itself is a complex phenomenon. For harmonic sounds, the pitch is a perceptual correlate of periodicity [16]. It's known that humans encode absolute pitches of the notes, in particular, this supported by tonotopic representations that are observed from the cochlea to the auditory cortex [16]. It is a surprising fact, given that the relative pitch is crucial for music perception (e.g., intervals between notes and melody contour are more important than the notes absolute location). Relative pitch abilities are present even in young infants, and may thus be a feature inherent to the auditory system, although neural mechanisms of relative pitch remain poorly understood [16].

Interval of octave has a special perceptual status. As was shown by Shepard, people could perceive tones which are an octave apart as equivalent in some sense. He demonstrated that pitch perception is two dimensional and consists of "height" (or overall pitch level) and "chroma", which refers to pitch classes in music theory (or

simply, note names). Chords, in turn, often described as pitch class sets (e.g., [6]).

2.1.3 Harmony in MIR: Audio Chord Estimation Task

In the context of MIR, harmony is considered mainly in the scope of the Audio Chord Estimation (ACE) Task. ACE software systems process audio recordings and predict chord labels for time segments. Performance of the systems is evaluated by comparing their output to human annotations for some set of audio tracks. Mauch [17] and Harte [18] gave a good introduction to ACE framework.

There is certain ambiguity in the task goals mentioned by Humphrey [19]: two slightly different problems are addressed. The first, chord transcription is an abstract task which could take into account high-level concepts. The second, chord recognition "is quite literal, and is closely related to polyphonic pitch detection." E.g., Mauch in his thesis [17] regards chord labels as an abstraction, while McVicar et al in their review [20] consider chords as slow-changing sustained pitches played concurrently and described by pitch class sets. As we learned from the jazz practice review, chord symbols are highly abstract concept and should not be considered as strict pitch class sets (at least in context of jazz).

MIR community dedicated some effort to develop unified plain text chord labels, which obey strict rules, but still convenient for humans. Harte et al. [21] suggested an approach which is used since then. It describes the basic syntax and a shorthand system. The basic syntax explicitly defines a chord pitch class set. For example, `C:(3, 5, b7)` is interpreted as C, E, G, B♭. The shorthand system contains symbols which resemble chord representations on lead sheets (e.g., `C:7` stands for C dominant seventh). According to [21], `C:7` should be interpreted as `C:(3, 5, b7)`.

Performance Evaluation

Performance evaluation procedure for the ACE task is described at MIREX site [2]. The current approach to match predicted chords to ground truth was proposed by

Harte [18] and Pauwels and Peeters [22]. The main metrics is

$$Chord\ Symbol\ Recall(CSR) = \frac{summedduration\ of\ correct\ chords}{total\ duration}$$

evaluated for the whole dataset.

Before comparing, ground truth and predicted chords are mapped to equivalence classes according to a certain dictionary. For example, there's "MajMin" dictionary which reduces each chord to contained major or minor triad (or exclude it from consideration, if it doesn't contain a major or minor triad). The others dictionaries used for MIREX evaluation are: Root (only chord roots are compared), MajMinBass (major and minor triads with inversions), Sevenths (major and minor triads, dominant, major and minor seventh chords) SeventhsBass (major and minor triads, dominant, major and minor seventh chords, all with inversions). Dictionaries allow matching datasets and algorithms with different chord "resolutions" (e.g., with and without inversions), and to better see the strengths and weaknesses of the algorithms.

Humphrey and Bello criticized such evaluation approach [19], because the score is too "flat": evaluation concept doesn't support rules to penalize softly chords, which are not exactly the same, but are close to each other in some sense. They supposed that chord hierarchy based on pitch class sets inclusions could resolve some of the issues, and draw an example where `E:7` and `E:maj` could be considered as close, because `E:7` contains `E:maj`. I agree that current "flat" evaluation procedure is too restrictive, on the other hand, plausible solution must be much more complex or it will inevitably include some genre bias. E.g., in blues-rock `E:maj` and `E:7` could be often interchangeable, but in jazz, plausible chord substitutions depend on wider context, such as local tonal center or implied "stack of thirds", and in most cases just replacing `E:maj` with `E:7` would be a harsh error.

The other related issue outlined in [19], [23], [24] and [25] is harmony annotation subjectivity. Since harmonic analysis is subjective, it's not quite correct to treat the reference datasets as ground truth, unless we want to develop an algorithm which mimics the style of one particular annotator. According to estimations by Ni et al.

[24] and Koops et al. [25], top performing algorithms are close or even surpass "subjectivity ceiling" (which means that quantitatively conformance between algorithms predictions and human annotations are close or even higher than conformance between different human annotations for a part of the same dataset). Though, only the quantity of disagreements is considered, but not the quality. I suppose that disagreements between human annotators might be more perceptually plausible then disagreements with an algorithm (e.g., humans might disagree on chords which are not structurally important and their alternative choices of chords might be close musically).

Chord overlap is not the only relevant quality indicator. Harte introduces segmentation quality measure [18], which complements chord symbol recall metrics at MIREX challenge. He supposed that proper chord boundaries predictions (even without chord labels consideration) are more musically useful than wrong boundaries predictions. To tackle this, directional hamming distance is used to evaluate the precision and continuity of estimated segments [18].

To evaluate an algorithm (and to train it, if the algorithm is data-driven) annotated datasets are needed. We consider them in a separate section, because our datasets purpose and usage context is broader than ACE task.

Evolution of ACE Algorithms

McVicar et al. [20] provided a comprehensive review of the achievements in the field until 2014. First algorithms performed polyphonic note transcription and then infer chords in the symbolic domain. Then Fujishima [26] suggest to use Pitch Class Profiles (PCPs), which preserves more information and achieve more robustness than note names. It allows developing more simple and accurate chord detection systems. PCP is a 12-dimensional vector, which components represent intensities of semitone pitch classes in a sound segment. It's calculated from the spectra of a short sound segment by summing values of the spectral bins which frequencies are close enough to particular pitch classes. Later many similar features were introduced which were called "Chroma Features".

In 2008 MIREX ACE task was started, giving us the opportunity to trace the algorithms performance evolution.

For the time being, expert knowledge systems were mostly used. In 2010 the first version of Matthias and Cannam Chordino[2] algorithm showed the best performance at MIREX. Chordino is available as a vamp plugin for Sonic Visualizer, and it's quite popular even beyond MIR community due to its robust multiplatform open source implementation. It is used as a baseline benchmark in MIREX ACE challenge since then.

Then MIR community accumulates enough datasets to start development of the data-driven systems. Harmonic Progression Analyzer[3] (HPA) [27], was a typical example of such system. It was among the top performing algorithms at MIREX 2011-2012 with chord symbol recall 4% higher then Chordino (Table 1). Since then, many authors suggested improvements to various parts of the pipeline, but essentially within the same paradigm: chroma features + data-driven algorithms for pattern matching + sequence decoding.

The other interesting trend is to depart from chroma features and obtain another low-dimensional projection of time-frequency space, which represents more information about chords and is more robust than chroma. Humphrey et al. [28] used deep learning techniques to produce a transformation to 7-D Tonnetz-space, which allowed them to outperform state-of-the-art algorithm of the period. The most recent MIREX top performing ACE system madmom[4] from Korzeniowski [29] also follows this trend. It uses a fully convolutional deep auditory model to extract 128 features to feed them to chord matching pipeline. And interestingly, Korzeniowski shows [29], that extracted features not only demonstrate dependency on pitch classes but reveal explicit "knowledge" about chord type (e.g., features which have a high contribution to major chords, shows a negative connection to *all* minor chords and vice versa), which reminds human perception of chords to some degree.

[2] `http://www.isophonics.net/nnls-chroma`
[3] `https://github.com/skyloadGithub/HPA`
[4] `https://github.com/CPJKU/madmom`

Year of creation	Algorithm	CSR (Billboard2013)	CSR (Isophonics)
2010	Chordino	67.25%	75.41 %
2012	HPA	71.40 %	81.49%
2017	madmom	78.37%	86.80%

Table 2: Selected top algorithms performance at MIREX ACE task with MinMaj chord dicitionary, Billboard2013 and Isophonics2009 datasets. Chordino results are taken from 2014 re-evaluation.

As we could see from Table 2, there's a 11% CSR increase in seven years (in the table I show only the pivotal years of sharp performance increase and paradigm change). We see that accuracy for Isophonics is higher, probably because Isophonics dataset as the whole was used for algorithms training and it is more homogeneous than the Billboard. There's a good progress for both datasets, but is the task close to being completely solved? At any rate, according to Koops et al. [25], algorithms performance hit the ceiling of evaluation framework possibilities due to annotation subjectivity problem.

Another issue is the limitations of available datasets. They include mainly major and minor triad annotations, which makes difficult to evaluate algorithm performance on other chord types.

Design of ACE Algorithms

ACE pipeline consists of several components with the multitude decisions which could be made about each of them. Besides, some authors use infamous "everything but the kitchen sink" approach by adding a lot of stages to the algorithm without trying to explain, how much each step adds to the overall performance. Therefore there is an overwhelming amount of complex ACE algorithms, from which it's hard to learn the best practices. Cho and Bello [30] made a valuable effort to unfold the standardized pipeline to its components and evaluate each component contribution and interaction between them. The paper describes a good practice of development of a multistage algorithm and provides a reasonable introduction to ACE in general. My explanation is based on the paper [30] amended with several recently reported achievements.

Scheme of the basic ACE algorithm pipeline shown on Figure 5. Firstly, a transition from sound waveform to a time-frequency representation is performed. The popular choice is Short Time Fourier Transform (e.g., [31]) due to its computationally efficient implementation based on Fast Fourier Transform (FFT). The main problem here is a trade-off between time resolution and frequency resolution, which is ruled by STFT window size. Constant Q (CTQ) transform is used (e.g., [28]) which has different frequency resolution in different ranges: higher resolution in low range, which mimics the human auditory system. Rocher et al. [17] proposed STFT analysis with different windows size simultaneously. Khadkevich and Omologo [32] used time-frequency reassignment technique to improve the resolution which allows them to develop high-performing algorithm which topped MIREX scores in years 2014-2015.

For each time frame, spectrogram is reduced to chroma vector with twelve values. Sometimes, several vectors are considered for different frequency ranges ([31], [27]). Many approaches to chroma extraction have been developed (e.g., [31], [27], [33], [34], [35]). Authors of the chroma extraction algorithms pursue the following objectives:

- Chroma should be invariant of timbre
- Chroma should be invariant of loudness
- Non-harmonic sounds such as noise and transients should not affect chroma
- Chroma should be tolerant to tuning fluctuations

Design of computationally effective, robust and representative chroma features is a subject of ongoing research.

Since chroma features are not robust, usually they are smoothed, e.g., by applying moving average filter or a moving median filter [30]. Alternatively, average chroma is calculated for inter-beat intervals (beat-synchronous chromagram) [30], where beat positions are provided by a beat detection algorithm.

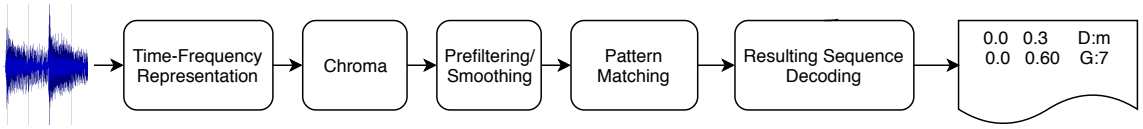

Figure 5: Basic ACE algorithm pipeline.

At the pattern matching stage, chroma features are mapped to chord labels. Soft labeling with Gaussian Mixture Models (GMM) [30] is usually involved.

At the final stage, probabilistic model such as HMM [30] or CRF [29] is used do decode the whole sequence.

We don't consider approaches involving Neural Networks, because our dataset is not big enough at the moment (114 tracks) and trained network will tend to overfit.

Now we are finishing the general review of approaches to harmony and switching to consideration of existing datasets, which could be used for training and evaluating ACE algorithms and corpus-based musicological research.

2.2 Datasets Related to Jazz Harmony

Jazz musicians use an abbreviated notation, known as a lead sheet, to represent chord progressions. Digitized collections of lead sheets are used for computer-aided corpus-based musicological research, e.g., [36, 37, 38, 15]. Lead sheets do not provide information about how specific chords are rendered by musicians [5]. To reflect this rendering, music information retrieval (MIR) and musicology communities have created several datasets of audio recordings annotated with chord progressions. Such collections are used for training and evaluating various MIR algorithms (e.g., Automatic Chord Estimation) and for corpus-based research. Here we review publicly available datasets that contain information about harmony, such as chord progressions and structural analysis including both: pure symbolic datasets and audio-aligned annotations. We consider the following aspects: content selection principle, format, annotation methodology, and actual use cases. Then we discuss some discrepancies in approaches to chord annotation and advantages and drawbacks of different formats.

Isophonics family

Isophonics[5] is one of the first time-aligned chord annotation datasets, introduced in [18]. Initially, the dataset consisted of twelve studio albums by The Beatles. Harte justified his selection by stating that it is a small but varied corpus (including various styles, recording techniques and "complex harmonic progressions" in comparison with other popular music artists). These albums are "widely available in most parts of the world" and have had enormous influence on the development of pop music. A number of related theoretical and critical works was also taken into account. Later the corpus was augmented with some transcriptions of Carole King, Queen, Michael Jackson, and Zweieck.

The corpus is organized as a directory of ".lab" files[6]. Each line describes a chord segment with a start time, end time (in seconds), and chord label in the "Harte et al." format [21]. The annotator recorded chord start times by tapping keys on a keyboard. The chords were transcribed using published analyses as a starting point, if possible. Notes from the melody line were not included in the chords. The resulting chord progression was verified by synthesizing the audio and playing it alongside the original tracks. The dataset has been used for training and testing chord evaluation algorithms (e.g., for MIREX[7]).

The same format is used for the "Robbie Williams dataset" [8] announced in [40]; for the chord annotations of the RWC and USPop datasets[9]; and for the datasets by Deng: JayChou29, CNPop20, and JazzGuitar99[10]. Deng presented this dataset in [41], and it is the only one in the family which is related to jazz. However, it uses 99 short, guitar-only pieces recorded for a study book, and thus does not reflect the variety of jazz styles and instrumentations.

[5] Available at `http://isophonics.net/content/reference-annotations`

[6] ASCII plain text files which are used by a variety of popular MIR tools, e.g., Sonic Visualizer [39].

[7] `http://www.music-ir.org/mirex/wiki/MIREX_HOME`

[8] `http://ispg.deib.polimi.it/mir-software.html`

[9] `https://github.com/tmc323/Chord-Annotations`

[10] `http://www.tangkk.net/label`

Billboard

Authors of the Billboard[11] dataset argued that both musicologists and MIR researchers require a wider range of data [42]. They selected songs randomly from the Billboard "Hot 100" chart in the United States between 1958 and 1991.

Their format is close to the traditional lead sheet: it contains meter, bars, and chord labels for each bar or for particular beats of a bar. Annotations are time-aligned with the audio by the assignment of a timestamp to the start of each "phrase" (usually 4 bars). The "Harte et al." syntax was used for the chord labels (with a few additions to the shorthand system). The authors accompanied the annotations with pre-extracted NNLS Chroma features [31]. At least three persons were involved in making and reconciling a singleton annotation for each track. The corpus is used for training and testing chord evaluation algorithms (e.g., MIREX ACE evaluation) and for musicological research [37].

Rockcorpus and subjectivity dataset

Rockcorpus[12] was announced in [23]. The corpus currently contains 200 songs selected from the "500 Greatest Songs of All Time" list, which was compiled by the writers of *Rolling Stone* magazine, based on polls of 172 "rock stars and leading authorities."

As in the Billboard dataset, the authors specify the structure segmentation and assign chords to bars (and to beats if necessary), but not directly to time segments. A timestamp is specified for each measure bar.

In contrast to the previous datasets, authors do not use "absolute" chord labels, e.g., `C:maj`. Instead, they specify tonal centers for parts of the composition and chords as Roman numerals. These show the chord quality and the relation of the chord's root to the tonic. This approach facilitates harmony analysis.

Each of the two authors provides annotations for each recording. As opposed to the

[11] `http://ddmal.music.mcgill.ca/research/billboard`
[12] `http://rockcorpus.midside.com/2011_paper.html`

aforementioned examples, the authors do not aim to produce a single "ground truth" annotation, but keep both versions. Thus it becomes possible to study subjectivity in human annotations of chord changes. The Rockcorpus is used for training and testing chord evaluation algorithms [19], and for musicological research [23].

Concerning the study of subjectivity, we should also mention the Chordify Annotator Subjectivity Dataset[13], which contains transcriptions of 50 songs from the Billboard dataset by four different annotators [25]. It uses JSON-based JAMS annotation format.

Jazz-related datasets

Here we review datasets which do not have audio-aligned chord annotations as their primary purpose, but nevertheless can be useful in the context of jazz harmony studies.

Weimar Jazz Database The main focus of the Weimar Jazz Database (WJazzD[14]) is jazz soloing. Data is disseminated as a SQLite database containing transcription and meta information about 456 instrumental jazz solos from 343 different recordings (more than 132000 beats over 12.5 hours). The database includes: meter, structure segmentation, measures, and beat onsets, along with chord labels in a custom format. However, as stated by Pfleiderer [43], the chords were taken from available lead sheets, “cloned” for all choruses of the solo, and only in some cases transcribed from what was actually played by the rhythm section.

The database's metadata includes the MusicBrainz[15] Identifier, which allows users to link the annotation to a particular audio recording and fetch meta-information about the track from the MusicBrainz server.

Although WJazzD has significant applications for research in the symbolic domain [43], our experience has shown that obtaining audio tracks for analysis and aligning them with the annotations is nontrivial: the MusicBrainz identifiers are sometimes

[13] `https://github.com/chordify/CASD`
[14] `http://jazzomat.hfm-weimar.de`
[15] A community-supported collection of music recording metadata: `https://musicbrainz.org`

wrong, and are missing for 8% of the tracks. Sometimes WJazzD contains annotations of rare or old releases. In different masterings, the tempo and therefore the beat positions, differs from modern and widely available releases. We matched 14 tracks from WJazzD to tracks in our dataset by the performer's name and the date of the recording. In three cases the MusicBrainz Release is missing, and in three cases rare compilations were used as sources. It took some time to discover that three of the tracks ("Embraceable You", "Lester Leaps In", "Work Song") are actually alternative takes, which are officially available only on extended reissues. Beat positions in the other eleven tracks must be shifted and sometimes scaled to match available audio (e.g., for "Walking Shoes"). This may be improved by using an interesting alternative introduced by Balke et al. [44]: a web-based application, "JazzTube," which matches YouTube videos with WJazzD annotations and provides interactive educational visualizations.

Symbolic datasets The iRb[16] dataset (announced in [36]) contains chord progressions for 1186 jazz standards taken from a popular internet forum for jazz musicians. It lists the composer, lyricist, and year of creation. The data are written in the Humdrum encoding system. The chord data are submitted by anonymous enthusiasts and thus provides a rather modern interpretation of jazz standards. Nevertheless, Broze and Shanahan proved it was useful for corpus-based musicology research: see [36] and [15].

"Charlie Parker's Omnibook data"[17] contains chord progressions, themes, and solo scores for 50 recordings by Charlie Parker. The dataset is stored in MusicXML and introduced in [45].

Granroth-Wilding's "JazzCorpus"[18] contains 76 chord progressions (approximately 3000 chords) annotated with harmonic analyses (i.e., tonal centers and roman numerals for the chords), with the primary goal of training and testing statistical parsing models for determining chord harmonic functions [14].

[16] https://musiccog.ohio-state.edu/home/index.php/iRb_Jazz_Corpus
[17] https://members.loria.fr/KDeguernel/omnibook/
[18] http://jazzparser.granroth-wilding.co.uk/JazzCorpus.html

Some discrepancies in chord annotation approaches in the context of jazz

An article by Harte et al. [21] de facto sets the standard for chord labels in MIR annotations. It describes the basic syntax and a shorthand system. The basic syntax explicitly defines a chord pitch class set. For example, `C:(3, 5, b7)` is interpreted as C, E, G, B♭. The shorthand system contains symbols which resemble chord representations on lead sheets (e.g., `C:7` stands for C dominant seventh). According to [21], `C:7` should be interpreted as `C:(3, 5, b7)`. However, this may not always be the case in jazz. According to theoretical research [7] and educational books, e.g., [8], the 5th degree is omitted quite often in jazz harmony.

Generally speaking, since chord labels emerged in jazz and pop music practice in the 1930s, they provide a higher level of abstraction than sheet music scores, allowing musicians to improvise their parts [5]. Similarly, a transcriber can use the single chord label `C:7` to mark the whole passage containing the walking bass line and comping piano phrase, without even noticing, "Is the 5th really played?" Thus, for jazz corpus annotation, we suggest accepting the "Harte et al." syntax for the purpose of standardization, but sticking to shorthand system and avoiding a literal interpretation of the labels.

There are two different approaches to chord annotation:

- "Lead sheet style." Contains a lead sheet [5], which has obvious meaning to musicians practicing the corresponding style (e.g., jazz or rock). It is aligned to audio with timestamps for beats or measure bars. Chords are considered in a rhythmical framework. This style is convenient because the annotation process can be split into two parts: lead sheet transcription done by a qualified musician, and beats annotation done by a less skilled person or sometimes even automatically performed.
- "Isophonics style." Chord labels are bound to absolute time segments.

We must note that musicians use chord labels for instructing and describing performance mostly within the lead sheet framework. While the lead sheet format and the

chord-beats relationship is obvious, detecting and interpreting "chord onset" times in jazz is an unclear task. The widely used comping approach to accompaniment [8] assumes playing phrases instead of long isolated chords, and a given phrase does not necessarily start with a chord tone. Furthermore, individual players in the rhythm section (e.g., bassist and guitarist) may choose different strategies: they may anticipate a new chord, play it on the downbeat, or delay. Thus, before annotating "chord onset" times, we should make sure that it makes musical and perceptual sense. All known corpus-based research is based on "lead sheet style" annotated datasets. Taking all these considerations into account, we prefer to use the lead sheet approach to chord annotations.

Now let's consider mathematical devices which would be help during the work.

2.3 Probabilistic and Machine Learning Concepts

In general, mathematical devices used in this work are quite typical for ACE (see [30]). For a more in-depth introduction to probability, Gaussian Mixture Models, Hidden Markov Models and Conditional Random Fields one could check corresponding chapters of Murphy textbook [46]. The one device which is not typical is Compositional Data analysis, and here is a short introduction.

2.3.1 Compositional Data Analysis and Ternary Plots

Chroma feature vectors are usually normalized per instance to make features independent of dynamics of a signal [30]. Musically it means, that we are not interested in loudness of the chord played, but only in the balance between its components, therefore we divide all the components on some loudness correlate. Norms such as l^1, l^2 or l^∞ are mostly used. After normalization, chroma vector components become interdependent and are distributed not in 12-dimensional space, but on some constrained 11-dimensional surface. Lets consider normalized chroma vector components $X_i \geqslant 0, i = 1..12$. Than

- for l^1: $\sum_{i=1}^{12} X_i = 1$, the figure is 11-simplex

- for l^2: $\sum_{i=1}^{12} X_i^2 = 1$, the figure is a "half-wedge" of 11-sphere
- for l^∞: $X \in \{0 \leqslant X_i \leqslant 1, \exists j : X_j = 1\}$, the figure is a quarter of the surface of 12-cube

If we exploit the knowledge about normalized chroma vector space geometry, we could:

1. develop more compact and accurate probabilistic model which doesn't give positive probabilities to impossible events [47]
2. we could obtain more compact and meaningful visualization of the distribution density

For this purpose, l^1-normalized chroma is preferable, because distribution on simplex is studied well enough (simplex has better geometric properties, then the other two: it's convex) and it's a subject of Compositional Data Analysis.

Compositional Data are vectors $X = (X_1, ...X_D)$ with all its components strictly positive and carrying only relative information. It is restricted to sum to a fixed constant, i.e. $\sum_{i=1}^{D} X_i = \kappa$ [47]. Here we assume that $\kappa = 1$.

Compositional Data Analysis is used whenever the balance of parts is studied, but overall "volume" of the composition κ is irrelevant. For example, in geochemistry (e.g., when proportions of different chemicals in rocks are studied), in economics (e.g., when proportions of different items in a state's budget over the years are studied).

The set of all possible compositions is called D-part simplex [48]:

$$\mathbb{S}^D = \{X = (X_1, ...X_D) : X_i \geqslant 0, \sum_{i=1}^{D} X_i = 1\}$$

In case of $D = 2$, the simplex is a segment between points $(1, 0)$ and $(0, 1)$, see Figure 6. In case of $D = 3$, the simplex is a triangle with vertexes at $(1, 0, 0)$, $(0, 1, 0)$

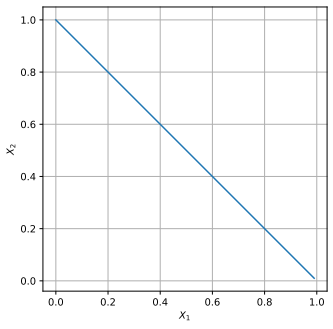

Figure 6: 2-part Simplex.

and $(0, 0, 1)$ (Figure 7 left). Since there are actually only 2 degrees of freedom, the data could be presented in two dimensions as Ternary .

The data from the 3-part simplex can be presented in 2 dimensions as Ternary Diagram. The Ternary Diagram represents set of all possible compositions as an equilateral triangle with vertexes annotated by the axis on which the corresponding vertex lies, and height equal to one. For each data point, we could obtain the components in the following way: the shortest distance from the point to the side is a value of the component corresponding to the opposite vertex. This method exploits the fact that the sum of distances from any interior point to the sides of the equilateral triangle is equal to the length of the triangle's height (Figure 7 right). Similarly, we could plot X distribution density as a ternary heat map.

4-part simplex is a solid, regular tetrahedron, where each possible 3-part composition is represented on one side of the tetrahedron. We could project content of the tetrahedron to its sides, by obtaining 3-part subcompositions. As observed in [48], we could consider higher-dimensional simplexes as hyper-tetrahedrons, obtain 3-part subcompositions and visualize them as ternary diagrams. Thus we could picture content of multidimensional simplex by projecting it to its triangle sides.

Some Fundamental Operations on Compositions. Let's review some of the operations on compositions which will use in this work:

- Closure. It's another name for normalization: $C(X) = \frac{1}{\sum_{i=1}^{D} X_i} X$. It's applied if the X is not a composition.

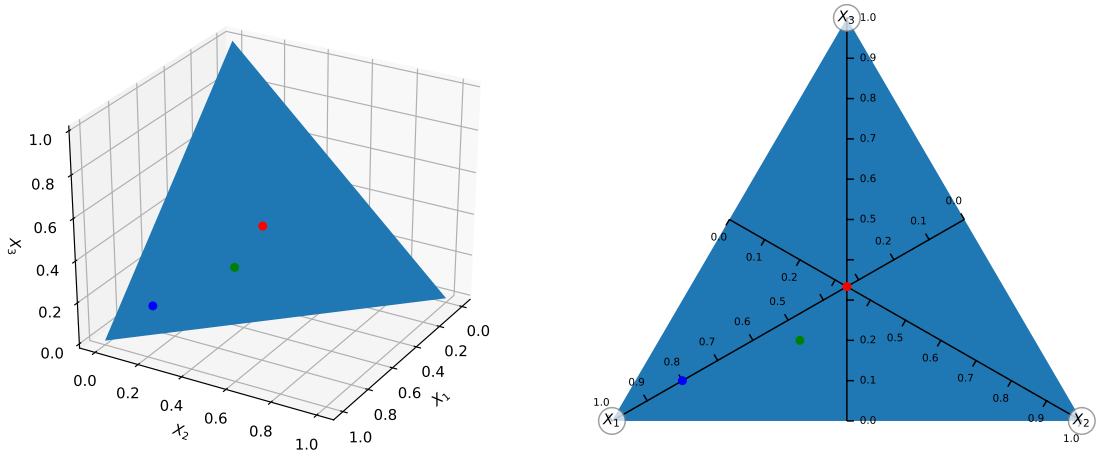

Figure 7: 3-part Simplex (left) and its' representation as Ternary Plot (right). The following compositions are shown: red $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, blue $(0.8, 0.1, 0.1)$ and green $(0.5, 0.3, 0.2)$

- Subcomposition. Used to discard (or "marginalize out" using probabilistic terminology) some of the components. Thus, it projects original multidimensional data to subspace of smaller dimensionality.

- Amalgamation. Reduce dimensionality by summing some of the components together (e.g., in case of chroma vectors in tonal content which has strong and weak degrees, we could sum chroma components for all strong degrees and weak degrees separately and obtain a new vector $Y : (Y_{strong}, Y_{weak}), Y_{strong} + Y_{weak} = 1)$.

More detail can be found in [48].

Chapter 3

JAAH Dataset

3.1 Guiding Principles

Based on given review and our own hands-on experience with chord estimation algorithm evaluation, we present our guidelines and propositions for building audio-aligned chord dataset.

1. Dataset boundaries should be clearly defined (e.g., certain music style or period). Selection of audio tracks should be proven to be representative and balanced within these boundaries.

2. Since sharing audio is restricted by copyright laws, recent releases and existing compilations should be used to facilitate access to dataset audio.

3. Use time aligned lead sheet approach and "shorthand" chord labels from [21], but avoid its literal interpretation.

4. Annotate entire tracks, but not excerpts. It makes possible to explore structure and self-similarity.

5. Provide MusicBrainz identifier to exploit meta-information from this service. If it's feasible, add meta-information to MusicBrainz instead of storing it privately within the dataset.

6. The annotation format should be stored in a machine readable format and suitable for further manual editing and verification. Relying on plain text files and specific directory structure for storing heterogeneous annotation is not practical for users. Thus, it's better to use JSON or XML, which allows to store complex structured data in a single file or a database entry or transfer through the web in a compact and unified form. JSON-based JAMS format introduced by Humphrey et al. [49] is particularly useful, but currently, it doesn't support lead sheet style for chord annotation and is verbose to be comfortably used by the human annotators and supervisors.

7. It is preferable to include pre-extracted chroma features. It will make possible to conduct some MIR experiments without accessing the audio. It would be interesting to incorporate chroma features into corpus-based research, to demonstrate, how particular chord class is rendered in a particular recording.

3.2 Proposed Dataset

3.2.1 Data Format and Annotation Attributes

Taking into consideration the discussion from the previous section, we decided to use the JSON format. An excerpt from an annotation is shown in Figure 8. We provide the track title, artist name, and MusicBrainz ID. The start time, duration of the annotated region, and tuning frequency estimated automatically by Essentia [50] are shown. The beat onsets array and chord annotations are nested into the "parts" attribute, which in turn could recursively contain "parts." This hierarchy represents the structure of the musical piece. Each part has a "name" attribute which describes the purpose of the part, such as "intro," "head," "coda," "outro," "interlude," etc. The inner form of the chorus (e.g., AABA, ABAC, blues) and predominant instrumentation (e.g., ensemble, trumpet solo, vocals female, etc.) are annotated explicitly. This structural annotation is beneficial for extracting statistical information regarding the type of chorus present in the dataset, as well as other musically important properties. We made chord annotations in the lead sheet style:

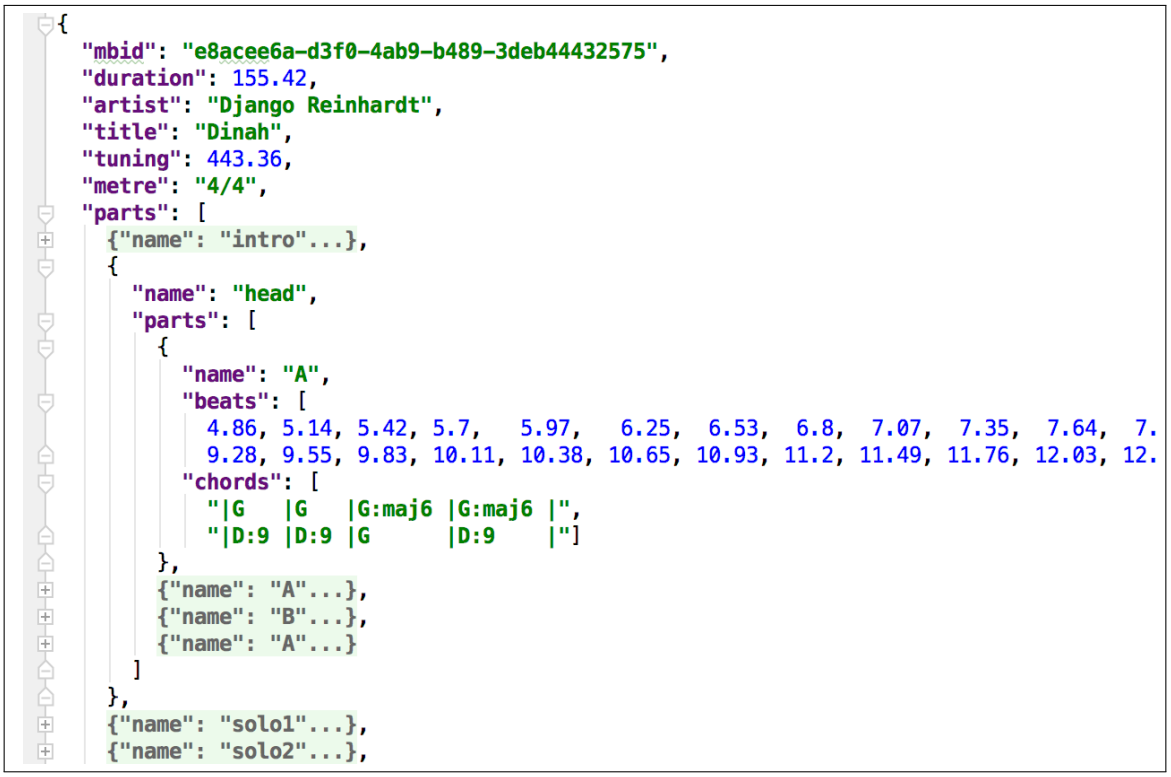

Figure 8: An annotation example.

each annotation string represents a sequence of measure bars, delimited with pipes: "|". A sequence starts and ends with a pipe as well. Chords must be specified for each beat in a bar (e.g., four chords for 4/4 meter). A simplification of this is possible: if a chord occupies the whole bar, it could be typed only once; and if chords occupy an equal number of beats in a bar (e.g., two beats in 4/4 metre), each chord could be specified only once, e.g., `|F G|` instead of `|F F G G|`.

For chord labeling, we use the Harte et al. [21] syntax for standardization reasons, but mainly use the shorthand system and do not assume the literal interpretation of labels as pitch class sets. More details on chord label interpretation will follow in 3.3.1.

3.2.2 Content Selection

The community of listeners, musicians, teachers, critics and academic scholars defines the jazz genre, so we decided to annotate a selection chosen by experts.

After considering several lists of seminal recordings compiled by authorities in jazz history and in musical education [51, 9], we decided to start with "The Smithsonian Collection of Classic Jazz" [52] and "Jazz: The Smithsonian Anthology" [53].

The "Collection" was compiled by Martin Williams and first issued in 1973. Since then, it has been widely used for jazz history education and numerous musicological research studies draw examples from it [54]. The "Anthology" contains more modern material compared to the "Collection." To obtain unbiased and representative selection, its curators used a multi-step polling and negotiation process involving more than 50 "jazz experts, educators, authors, broadcasters, and performers." Last but not least, audio recordings from these lists can be conveniently obtained: each of the collections are issued in a CD box.

We decided to limit the first version of our dataset to jazz styles developed before free jazz and modal jazz, because lead sheets with chord labels cannot be used effectively to instruct or describe performances in these latter styles. We also decided to postpone annotating compositions which include elements of modern harmonic structures (i.e., modal or quartal harmony).

3.2.3 Transcription Methodology

We use the following semi-automatic routine for beat detection: the DBNBeatTracker algorithm from the madmom package is run [55]; estimated beats are visualized and sonified with Sonic Visualizer; if needed, DBNBeatTracker is re-run with a different set of parameters; and finally beat annotations are manually corrected, which is usually necessary for ritardando or rubato sections in a performance.

After that, chords are transcribed. The annotator aims to notate which chords are played by the rhythm section. If the chords played by the rhythm section are not clearly audible during a solo, chords played in the "head" are replicated. Useful guidelines on chord transcription in jazz are given in the introduction of Henry Martin's book [54]. The annotators used existing resources as a starting point, such as published transcriptions of a particular performance or Real book chord

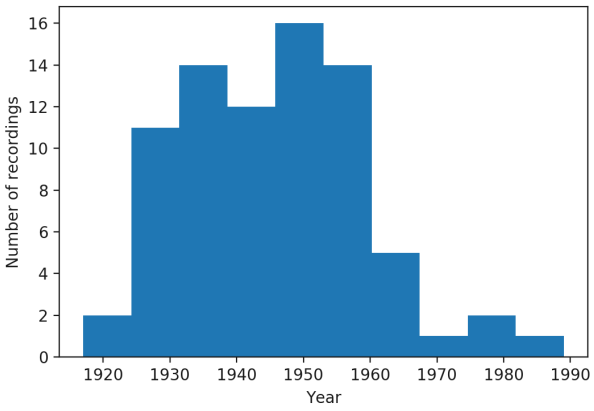

Figure 9: Distribution of recordings from the dataset by year.

progressions, but the final decisions for each recording were made by the annotator. We developed an automation tool for checking the annotation syntax and chord sonification: chord sounds are generated with Shepard tones and mixed with the original audio track, taking its volume into account. If annotation errors are found during syntax check or while listening to the sonification playback, they are corrected and the loop is repeated.

3.3 Dataset Summary and Implications for Corpus Based Research

To date, 113 tracks are annotated with an overall duration of almost 7 hours, or 68570 beats. Annotated recordings were made from music created between 1917 and 1989, with the greatest number coming from the formative years of jazz: the 1920s-1960s (see Figure 9). Styles vary from blues and ragtime to New Orleans, swing, be-bop and hard bop with a few examples of gypsy jazz, bossa nova, Afro-Cuban jazz, cool, and West Coast. Instrumentation varies from solo piano to jazz combos and to big bands.

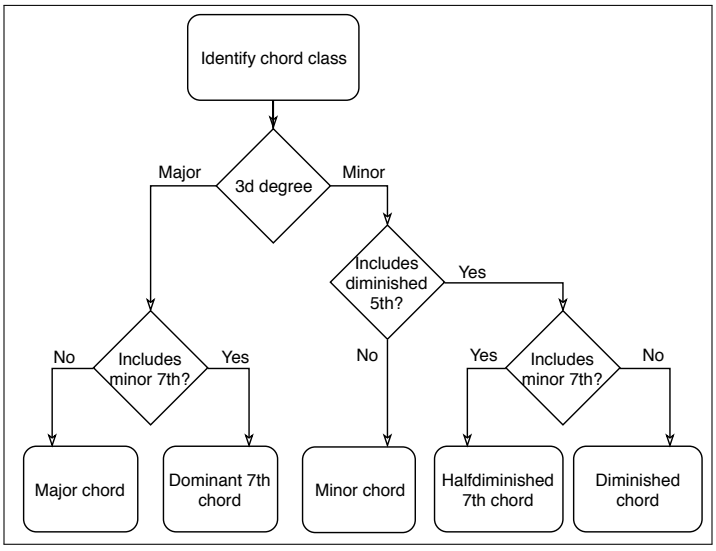

Figure 10: Flow chart: how to identify chord class by degree set.

3.3.1 Classifying Chords in the Jazz Way

In total, 59 distinct chord classes appear in the annotations (89, if we count chord inversions). To manage such a diversity of chords, we suggest classifying chords as it done in jazz pedagogical and theoretical literature. According to the article by Strunk [3], chord inversions are not important in the analysis of jazz performance, perhaps because of the improvisational nature of bass lines. Inversions are used in lead sheets mainly to emphasize the composed bass line (e.g., pedal point or chromaticism). Therefore, we ignore inversions in our analysis.

According to numerous instructional books, and to theoretical work done by Martin [7], there are only five main chord classes in jazz: major (maj), minor (min), dominant seventh (dom7), half-diminished seventh (hdim7), and diminished (dim). Seventh chords are more prevalent than triads, although sixth chords are popular in some styles (e.g., gypsy jazz). Third, fifth and seventh degrees are used to classify chords in a bit of an asymmetric manner: the unaltered fifth could be omitted in the major, minor and dominant seventh (see chapter on three note voicing in [8]); the diminished fifth is required in half-diminished and in diminished chords; and $\flat\flat 7$ is characteristic for diminished chords. We summarize this classification approach in the flow chart in Figure 10.

Chord class	Beats Number	Beats %	Duration (seconds)	Duration %
dom7	29786	43.44	10557	42.23
maj	18591	27.11	6606	26.42
min	13172	19.21	4681	18.72
dim	1677	2.45	583	2.33
hdim7	1280	1.87	511	2.04
no chord	3986	5.81	2032	8.13
unclassified	78	0.11	30	0.12

Table 3: Chord classes distribution.

The frequencies of different chord classes in our corpus are presented in Table 3. The dominant seventh is the most popular chord, followed by major, minor, diminished and half-diminished. Chord popularity ranks differ from those calculated in [36] for the iRb corpus: dom7, min, maj, hdim, and dim. This could be explained by the fact that our dataset is shifted toward the earlier years of jazz development, when major keys were more pervasive.

3.3.2 Exploring Symbolic Data

Exploring the distribution of chord transition bigrams and n-grams allows us to find regularities in chord progressions. The term bigram for two-chord transitions was defined in [36]. Similarly, we define an n-gram as a sequence of n chord transitions. The ten most frequent n-grams from our dataset are presented in Figure 11. The picture presented by the plot is what would be expected for a jazz corpus: we see the prevalence of the root movement by the cycle of fifths. The famous IIm-V7-I three-chord pattern (e.g., [7]) is ranked number 5, which is even higher than most of the shorter two-chord patterns.

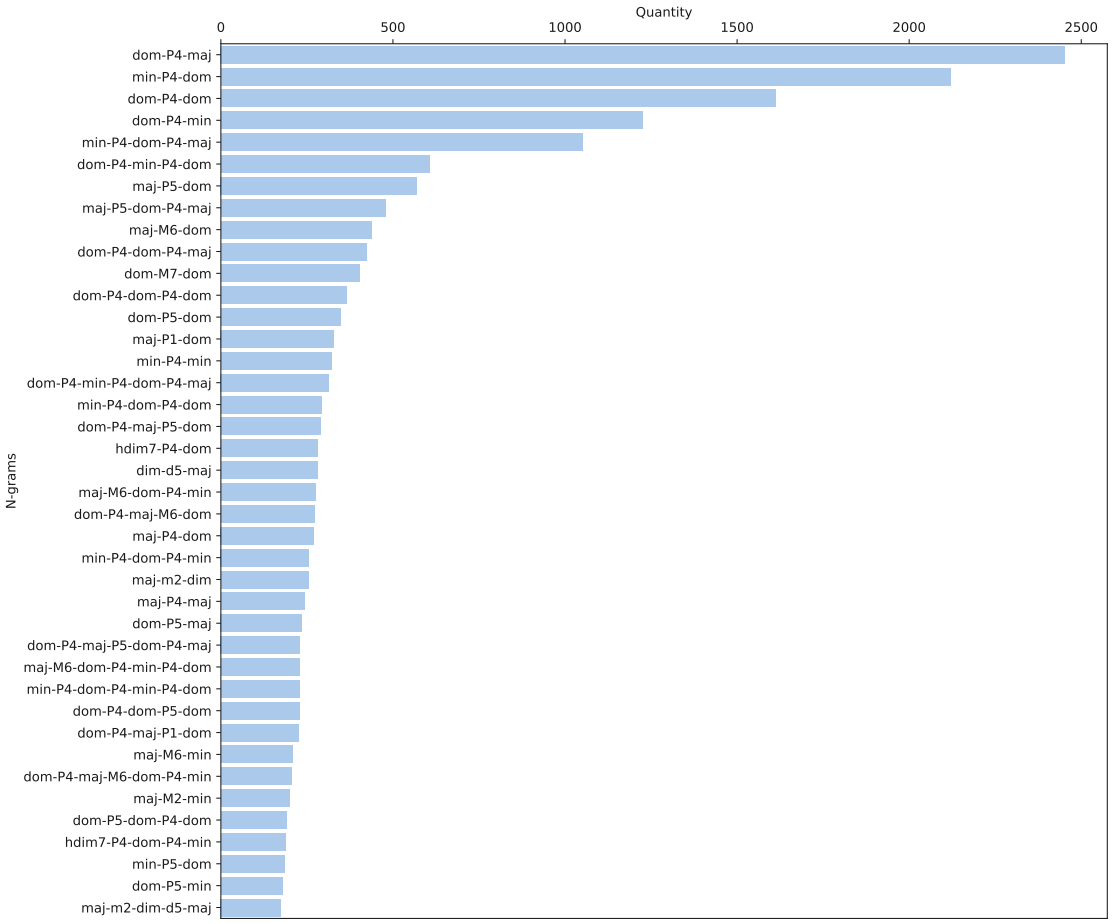

Figure 11: Top forty chord transition n-grams. Each n-gram is expressed as sequence of chord classes (dom, maj, min, hdim7, dim) alternated with intervals (e.g., P4 - perfect fourth, M6 - major sixth), separating adjacent chord roots.

Chapter 4

Visualizing Chroma Distribution

Since chord labels are an abstract representation of harmony which doesn't prescribe how exactly they should be rendered by performers, it could be interesting to study the tonal specificity of chords rendition in different sub-genre or by particular performer or even for a particular composition. In this chapter, I make an attempt to provide a visualization for joint sampling distribution of chroma features for particular chord type.

I intend to:

- build a visual profile for each chord type for particular performance or selection of performances. Then it becomes possible to visually study the tonal specificity of this performance, or compare different performances with each other.
- obtain insights about weaknesses (or strengths) of chroma-based chord detection algorithm performance.

For each track from JAAH dataset we extract NNLS chroma features[1]. Then we obtain weighted average chroma for each beat and transpose it to a common root (original root is known from the annotation). We denote chroma components by

[1] `http://www.isophonics.net/nnls-chroma`

Roman numerals showing their relation to the common root. For averaging, we use Hanning window function with the peak on the beat. As it will be shown in the "Algorithms" chapter 5, such averaging yields best result then averaging by rectangular window between the nearest beats, perhaps because chord tones are more often concentrated close to the beat, and less strong tones are played in weaker rhythmic positions. Let's call such vectors "beat centered chroma". Each component of this 12-D vector represents the intensity of a particular pitch class near the beat ("intensity" here doesn't have any proven physical or perceptual meaning). So, we will consider beat centered chroma sampling distribution.

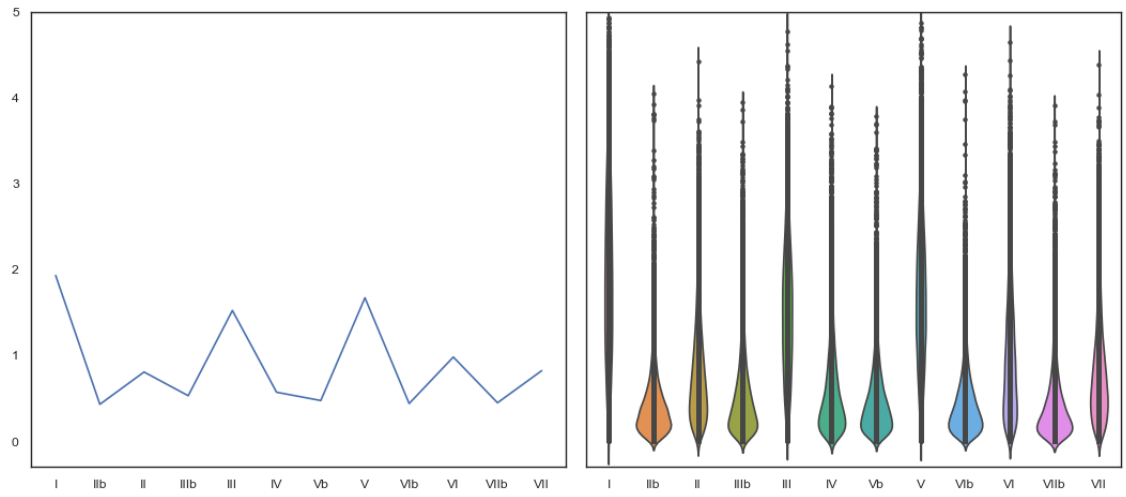

Figure 12: Beat centered chroma distribution for Major chord in JAAH dataset. Shown as pitch class profile (left), violin plots(right).

The standard way to visualize pitch classes distribution in music cognition is Pitch Class Profile [56]. Let's apply the same method to chroma vector distribution: Figure 12 left. It could be extended with variability visualization, in the form of violin plots [57]. Each plot consists of violin-like histograms. Each histogram describes the marginal distribution of a particular component: thickness of each "violin" at certain ordinate value is proportional to the number of data points in the neighborhood of this value (Figure 12 left). We see that degrees with the high average ("strong") have high variability as well, because they are used frequently but apparently in different ways; and rare (or "weak") degrees have low variability, because they are just not used (Figure 13). Otherwise, smeared density plots for

strongest degrees for the major chord (I, III, V) reveals no pattern, because the plot gives no information about joint distribution.

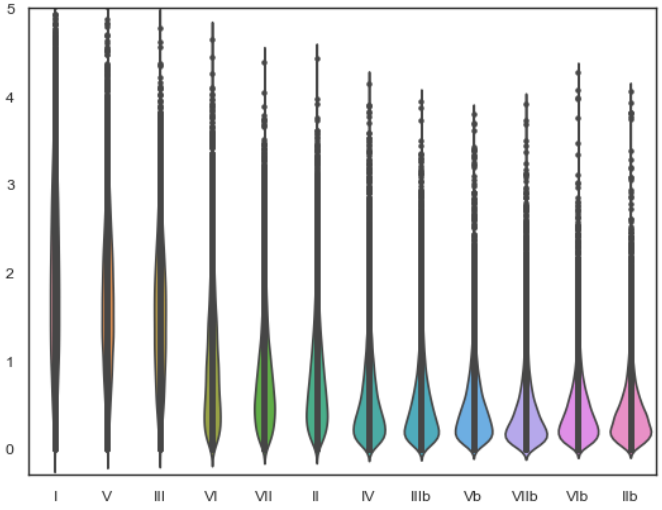

Figure 13: Beat centered chroma distribution for Major chord in JAAH dataset. Shown as violin plot, where scale degrees are ordered according to their "strengths" and simultaneously demonstrate decrease in variability

4.1 Chroma as Compositional Data

Because we are interested in the balance between pitch classes, but not in the absolute chroma components, let's consider l^1-normalized chroma vectors which are distributed on 12-part simplex and present a typical example of Compositional Data (see Background Section 2.3.1). The simplex has 12 vertexes which correspond to 12 degrees of the chromatic scale. The simplex is bounded by 220 triangle sides, where each triangle represents unique triple of scale degrees, e.g., I - III - V.

To gain some intuition about the visualization, let's consider single chroma vector extracted from audio of a major triad chord played on guitar: $X = (3.77, 0.18, 0.08, 0.01, 3.25, 0.17, 0.08, 2.42, 0.14, 0.34, 0.11, 0.58)$. After normalization it becomes: $X_{l^1} = (0.34, 0.02, 0.01, 0. \ , 0.29, 0.02, 0.01, 0.22, 0.01, 0.03, 0.01, 0.05)$. We take subcomposition for degrees I, III and V: $X_{I,III,V} = \frac{(X_I, X_{III}, X_V)}{X_I + X_{III} + X_V} = (0.4, 0.34, 0.26)$ and plot this point on a ternary diagram (Figure 14 left). Similarly, we could plot sampling distribution for a hundred of chords played on guitar. We split ternary diagram to triangle bins and color each bin according to number of vectors which

dropped in (Figure 14 right). It easily can be seen, that the distribution has mode about point (0.4, 0.3, 0.3), and in general fifths degree is presented less then first and even the third. Thus, a convenience of the ternary plots is shown: data for three components could be presented unambiguously on a two dimensional plane.

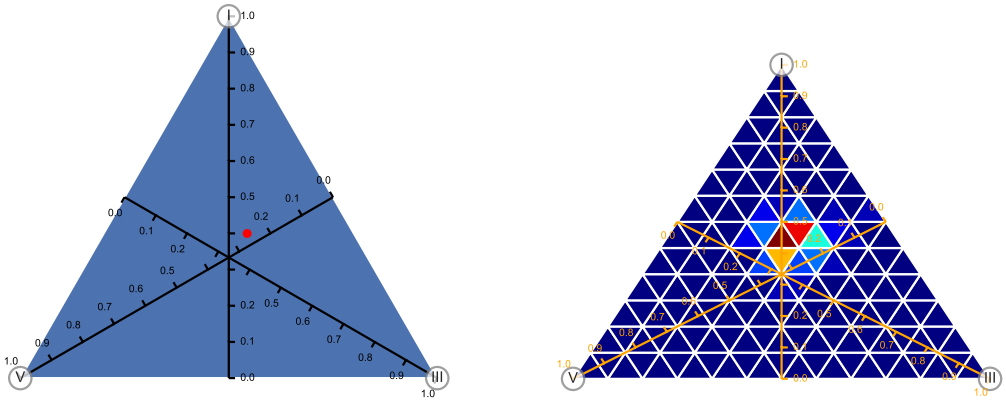

Figure 14: Projection of the chroma of triads played on guitar to ternary diagram: single chord(left), sampling distribution(right)

Projections for different triplets could be combined to obtain a picture in more dimensions. For the sake of better comprehension and compactness, it's reasonable to arrange ternary plots by uniting identical vertexes and edges. I suggest to combine projections for strongest degrees. If we select the most strong seven degrees for the particular chord type and build projection for each three adjacently ranked degrees, resulting triangles could be arranged as a hexagon, e.g.: the strongest degree in the middle, second-ranking "at seven o'clock", and the rest — counter clockwise (Figure 15 left).

Supposedly, not all 220 triangles are equally interesting. To proof the idea, let's build a similar hexagon for the weakest seven degrees (Figure 15 right). While the chroma distribution for strong degrees has distinct modes and concentrated patterns, which differ from chord to chord (see also Figure 16), the distributions for weakest degrees looks like a plateau and quite similar for all chord types (see also Figure 17). This fact conforms with the idea of tonal hierarchies [56]: for major, minor and dominant 7th chord stable pattern of nearly seven strong degrees is used, while the rest five

tones are used rarely and uniformly at random.

Half-diminished and diminished chords has the fuzziest distribution patterns (because we have the lowest content of these chord in the dataset: about 2%); major and minor patterns are similar (except for the degrees involved): triad is the most important; but dominant 7th is different from them: seventh degree is as significant as first, fifth and third, and the spread is higher.

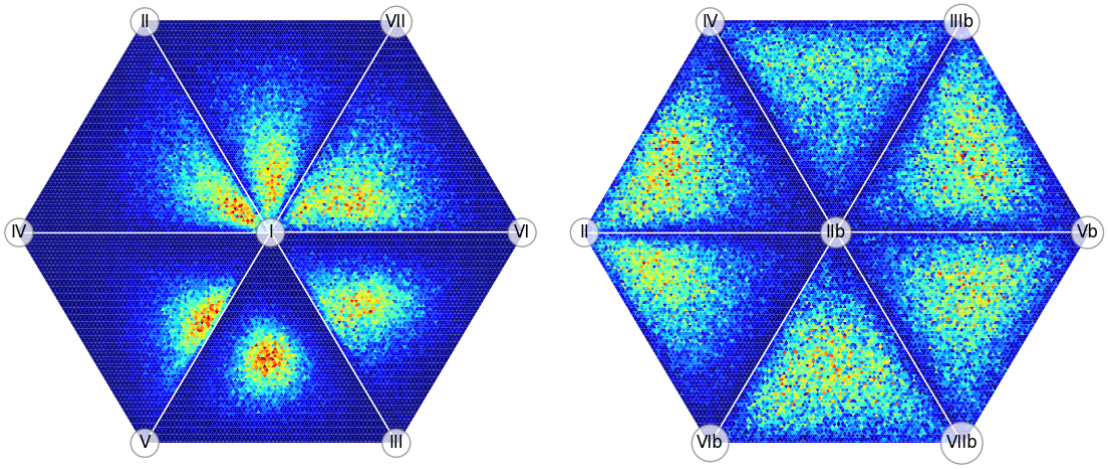

Figure 15: Combined ternary plots for strongest (left) and weakest (right) degrees for major chords in JAAH dataset.

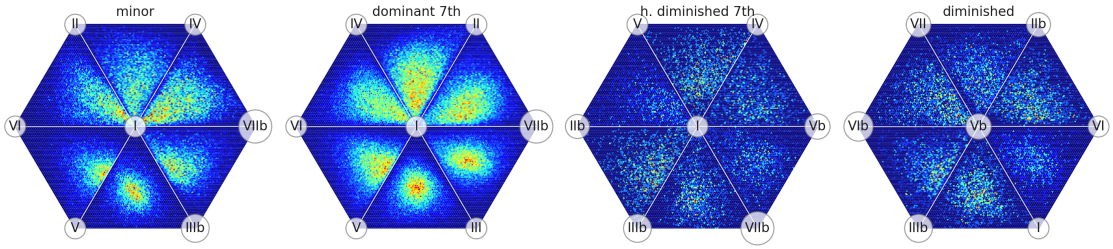

Figure 16: Combined ternary plots for strongest degrees for minor, dominant 7th, half-diminished 7th and diminished chords in JAAH dataset.

The other possible application of these plots is to compare chord profiles in different styles or performances. Let's briefly compare major chord rendering in 'Dinah" by Django Reinhardt (left) and "The Girl from Ipanema" by Stan Getz and Joao Gilberto (Figure 18). In "Dinah", the major triad is balanced and is the most prominent, while other degrees much weaker than root. In "Girl...", all points seem to run away from the root. This conforms with basic features of Django Reinhardt's

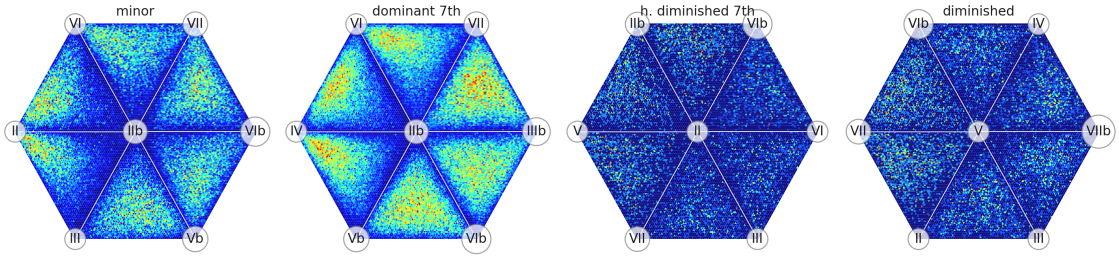

Figure 17: Combined ternary plots for weakest degrees for minor, dominant 7th, half-diminished 7th and diminished chords in JAAH dataset.

gypsy jazz simplistic approach to harmony and harmonic ambiguity and complexity of bosa-nova.

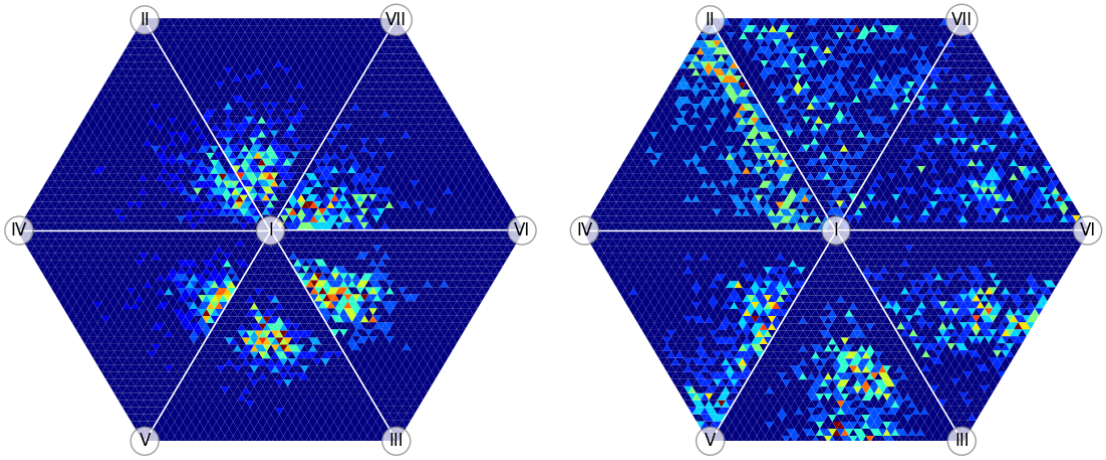

Figure 18: Combined ternary plots for strongest (on average in corpus) degrees in major chord for “Dinah” by Django Reinhardt (left) and “The Girl from Ipanema” by Stan Getz and Joao Gilberto.

Chapter 5

Automatic Chord Estimation (ACE) Algorithms with Application to Jazz

5.1 Evaluation Approach. Results for Existing Chord Transcription Algorithms

Here we apply existing ACE algorithms to our JAAH dataset introduced in Chapter 3. We adopt the MIREX[1] approach to evaluating algorithms' performance. The approach supports multiple ways to match ground truth chord labels with predicted labels, by employing the different chord vocabularies introduced by Pauwels [22]. The distinctions between the five chord classes defined in 3.3.1 are crucial for analyzing jazz performance. More detailed transcriptions (e.g., a distinction between maj6 and maj7, detecting extensions of dom7, etc.) are also important but secondary to classification into the basic five classes. To formally implement this concept of chord classification, we develop a new vocabulary, called "Jazz5," which converts chords into the five classes according to the flowchart in Figure 10.

For comparison, we also choose two existing MIREX vocabularies: "Sevenths" and "Tetrads," because they ignore inversions and can distinguish between major, minor and dom7 classes (which together occupy about 90% of our dataset). However, these

[1] `http://www.music-ir.org/mirex/wiki/2017:Audio_Chord_Estimation`

Vocabulary	Coverage %	Chordino Accuracy %	CREMA Accuracy %
"Jazz5"	99.88	32.68	40.26
MirexSevenths	86.12	24.57	37.54
Tetrads	99.90	23.10	34.30

Table 4: Comparison of coverage and accuracy evaluation for different chord dictionaries and algorithms.

vocabularies penalize differences within a single basic class (e.g., between a major triad and a major seventh chord). Moreover, the "Sevenths" vocabulary is too basic; it excludes a significant number of chords, such as diminished chords or sixths, from evaluation.

We choose Chordino[2], which has been a baseline algorithm for the MIREX challenge over several years, and CREMA[3], which was recently introduced in [58]. To date, CREMA is one of the few open-source, state-of-the-art algorithms which supports seventh chords.

Results are provided in the Table 4. "Coverage" signifies the percentage of the dataset which can be evaluated using the given vocabulary. "Accuracy" stands for the percentage of the covered dataset for which chords were properly predicted, according to the given vocabulary.

We see that the accuracy for the jazz dataset is almost half of the accuracy achieved by the most advanced algorithms on datasets currently involved in the MIREX challenge[4] (which is roughly 70-80%). Nevertheless, the more recent algorithm (CREMA) performs significantly better than the old one (Chordino) which shows that our dataset passes a sanity check: it does not contradict technological progress in Audio Chord Estimation. We see from this analysis that the "Sevenths" chords vocabulary is not appropriate for a jazz corpus because it ignores almost 14% of the data. We also note that the "Tetrads" vocabulary is too punitive: it penalizes up to 9% of predictions which are tolerable in the context of jazz harmony analysis. We

[2]http://www.isophonics.net/nnls-chroma
[3]https://github.com/bmcfee/crema
[4]http://www.music-ir.org/mirex/wiki/2017:Audio_Chord_Estimation_Results

provide code for this evaluation in the project repository.

5.2 Evaluating ACE Algorithm individual Components Performance on JAAH Dataset.

In this section I describe a gradual construction of a new ACE algorithm. I will make the simplest possible choices to fit the concept. As it was mentioned, our dataset is not big enough so training Neural Network based algorithms on it could lead to overfitting, besides chord type distribution is rather skewed which complicates training. So I consider algorithms based on "classical" machine learning. Algorithm scheme is shown at Figure 19.

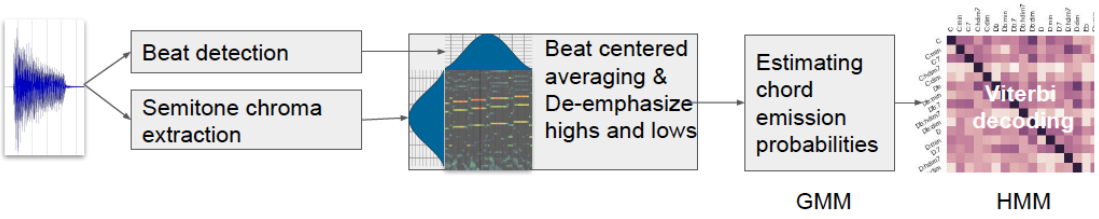

Figure 19: ACE algorithm for evaluating its' components contribution on JAAH dataset.

As we argue in 2, in musical context the harmony is considered in respect to metrical grid of beats, but not in terms of continuous time. Jazz music usually has steady pulsation, which is detected by state of the art beat detection algorithms well enough. We came to this conclusion during annotation of JAAH dataset: automatic beat detection provides a good base for manual beats annotation, which often doesn't even need to be corrected. Thus I use madmom[5] automatic beat detection as a part of the algorithm, and consider beats as only possible events for a chord change.

I decided to use NNLS[6] algorithm for chroma features extraction. It shows best results in our preliminary experiments for open source chroma features evaluation. Since inversions mostly are not important in jazz, I consider single chroma vector for the whole frequency range, but attenuate low and high frequencies as explained

[5] https://madmom.readthedocs.io/en/latest/
[6] http://www.isophonics.net/nnls-chroma

Smoothing	Accuracy
Inter-beat averaging	36.65
Beat centered chroma (0.9 sec Hanning window)	39.53

Table 5: Chroma smoothing influence on the overall algorithm performance.

in [59].

Initially I calculated beat-synchronous chromagram as described in [30]: chroma vector for the beat was estimated as average of chromas between this and the next beat. But performance is improved, if segments larger then one beat were used for averaging. I centered the segment around the beat and use "tapered" window for averaging, so chromas which are close to the beat receive more weight than chromas which are far. Results of the experiments are shown in Table 5 Thus, for averaging we use Hanning window function with the peak on the beat, which improves the performance by almost 3%.

I use GMM implementation from scikit learn package[7] to model chroma vector probability distribution for each chord type. Then the probabilities are plugged into Viterbi decoding as emission probabilities for the chord types. Two options for HMM's Transition matrix are used:

- the matrix estimated according to chord bigrams frequencies and average harmonic rhythm
- only average harmonic rhythm is taken into account, thus the matrix only promotes self-transition, but all other transitions have equally low probability.

Results are shown in Table 6. Interestingly, that this step significantly improves the performance for jazz (accuracy increased by 1.5 times), while in experiments by Cho and Bello [30] conducted with pop and rock music data, this step if applied to beat-synchronous chroma doesn't gain much, and most of the effect is explained by setting high self-transition probabilities.

[7] `http://scikit-learn.org/stable/modules/mixture.html`

Transition Matrix	Accuracy
Uniform (bypass decoding)	25.55%
Promotes self-transition	34.51%
Based on bigrams and harmonic rhythm	39.53%

Table 6: HMM Transition matrix influence on the overall algorithm performance.

We are only in the beginning of jazz-oriented ACE algorithm development, but it's already clear, that jazz music have certain specificity which could be exploited. In particular, improving window for per-beat chroma estimation and Sequence Decoding look promising.

Chapter 6

Conclusions

6.1 Conclusions and Contributions

In the scope of this thesis, I designed and participated in the creation of Jazz Audio-Aligned Harmony dataset (JAAH) which is publicly available at `https://github.com/MTG/JAAH`. A new method for estimating the performance of ACE algorithms with respect to specificity of jazz music is developed. Its implementation is available in our branch of Johan Pauwels's MusOOEvaluator at `https://github.com/MTG/MusOOEvaluator`. With this, we aim to stimulate MIR and corpus-based musicological researches targeting jazz. These results along with an analysis of the performance of existing algorithms on JAAH dataset are published at ISMIR2018 conference [60].

A new way of visualizing chroma distributions for different chord types is proposed. It allows to obtain a compact two-dimensional representation of the most important projections of the twelve-dimensional distribution density.

I also implemented a testbed for building and evaluating ACE algorithms and explore the importance of its components. In particular, it was experimentally shown that

- beat-centered chroma gives more accuracy in chord prediction than chroma averaged for inter-beat intervals. It supports the idea that chord tones are

played mainly on beats.

- Sequence decoding stage is more important for jazz then for rock and pop music, which emphasize importance of repetitive harmonic structures in jazz.

The code for the latest results is openly available at `https://github.com/seffka/jazz-harmony-analysis`.

6.2 Future Work

Further work includes growing the dataset by expanding the set of annotated tracks and adding new features. Local tonal centers are of particular interest because we could then implement chord detection accuracy evaluation based on jazz chord substitution rules and train algorithms for simultaneous local tonal center and chord detection.

Since it was shown that Sequence Decoding is essential for the ACE algorithm performance for jazz, HMM could be replaced with a more sophisticated model. Adaptive window per-beat chroma extraction could be implemented. Also, the algorithm could be extended with downbeat detection, search for hypermetric structures and cycles in a chord sequence. For the Pattern Matching, instead of Gaussian Mixture Model, which treats all the pitch classes equally, a mixture of specific distributions could be used, which assumes the hierarchy of strong and weak degrees in each chord segment.

List of Figures

List of Tables

Bibliography

[1] Serra, X. The computational study of a musical culture through its digital traces. *Acta Musicologica* **89**, 24–44 (2017).

[2] http://www.music-ir.org/mirex/wiki/2017:Audio_Chord_Estimation (2017).

[3] Strunk, S. Harmony (i). The New Grove Dictionary of Jazz, pp. 485–496 (1994).

[4] Berliner, P. *Thinking in Jazz: The Infinite Art of Improvisation* (University of Chicago Press, Chicago, 1994).

[5] Kernfeld, B. D. *The story of fake books : bootlegging songs to musicians* (Scarecrow Press, 2006). URL `https://books.google.es/books/about/The{_}Story{_}of{_}Fake{_}Books.html?id=oiMJAQAAMAAJ{&}redir{_}esc=y`.

[6] Martin, H. *Jazz harmony.* Ph.D. thesis, Princeton (1980).

[7] Martin, H. Jazz Harmony: A Syntactic Background. *Annual Review of Jazz Studies* **8**, 9–30 (1988).

[8] Levine, M. *The Jazz Piano Book* (Sher Music, 1989).

[9] Levine, M. *The Jazz Theory Book* (Sher Music, 2011). URL `https://books.google.es/books?id=iyNQpJ4oaMcC`.

[10] Aebersold, J. *Jamey Aebersold Jazz – How to Play Jazz and Improvise, Vol 1: The Most Widely Used Improvisation Method on the Market!, Book & 2 CDs.*

Jamey Aebersold Jazz (Jamey Aebersold Jazz, 2015). URL https://books.google.ru/books?id=QZHrQwAACAAJ.

[11] Pachet, F., Suzda, J. & Martinez, D. A comprehensive online database of machine-readable lead-sheets for jazz standards. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 275–280 (2013). URL https://www.semanticscholar.org/paper/A-Comprehensive-Online-Database-of-Machine-Readable-Pachet-Suzda/d4efb71516d76129b92bc11167be1a1188addea2.

[12] Schuller, G. & Collection, D. L. C. *Early Jazz: Its Roots and Musical Development*. Early Jazz: Its Roots and Musical Development (Oxford University Press, 1986). URL https://books.google.ru/books?id=PfwfMTWBGgYC.

[13] Steedman, M. J. A Generative Grammar for Jazz Chord Sequences. *Music Perception: An Interdisciplinary Journal* **2**, 52–77 (1984). URL http://mp.ucpress.edu/cgi/doi/10.2307/40285282http://mp.ucpress.edu/content/2/1/52.

[14] Granroth-wilding, M. *Harmonic analysis of music using combinatory categorial grammar*. Ph.D. thesis, University of Edinburgh (2013).

[15] Salley, K. & Shanahan, D. T. Phrase Rhythm in Standard Jazz Repertoire: A Taxonomy and Corpus Study. *Journal of Jazz Studies* **11**, 1 (2016). URL http://jjs.libraries.rutgers.edu/index.php/jjs/article/view/107.

[16] McDermott, J. H. & Oxenham, A. J. Music perception, pitch, and the auditory system (2008). URL https://www.sciencedirect.com/science/article/pii/S0959438808001050?via{%}3Dihub.

[17] Mauch, M. *Automatic Chord Transcription from Audio Using Computational Models of Musical Context*. Ph.D. thesis (2010). URL http://www.mendeley.com/research/automatic-chord-transcription-audio-using-computational-models-musical-context/

[18] Harte, C. *Towards automatic extraction of harmony information from music signals*. Ph.D. thesis, Queen Mary, University of London (2010). URL `http://qmro.qmul.ac.uk/jspui/handle/123456789/534`.

[19] Humphrey, E. J. & Bello, J. P. Four timely insights on automatic chord estimation. In *International Society for Music Information Retrieval Conference (ISMIR)* (2015). URL `https://nyuscholars.nyu.edu/en/publications/four-timely-insights-on-automatic-chord-estimation`.

[20] McVicar, M., Santos-Rodríguez, R., Ni, Y. & De Bie, T. Automatic chord estimation from audio: A review of the state of the art. *IEEE Transactions on Audio, Speech and Language Processing* **22**, 556–575 (2014). URL `http://ieeexplore.ieee.org/document/6705583/`.

[21] Harte, C., Sandler, M., Abdallah, S. & Gómez, E. Symbolic representation of musical chords: A proposed syntax for text annotations. In *Proc ISMIR*, vol. 56, 66–71 (2005). URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.9009{&}rep=rep1{&}type=pdf`.

[22] Pauwels, J. & Peeters, G. Evaluating automatically estimated chord sequences. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 749–753 (IEEE, 2013). URL `http://ieeexplore.ieee.org/document/6637748/`.

[23] de Clercq, T. & Temperley, D. A corpus analysis of rock harmony. *Popular Music* **30**, 47–70 (2011). URL `http://www.journals.cambridge.org/abstract{_}S026114301000067X`.

[24] Ni, Y., McVicar, M., Santos-Rodriguez, R. & De Bie, T. Understanding effects of subjectivity in measuring chord estimation accuracy. *IEEE Transactions on Audio, Speech and Language Processing* **21**, 2607–2615 (2013). URL `http://ieeexplore.ieee.org/document/6587770/`.

[25] Koops, H. V., Haas, B. d., Burgoyne, J. A., Bransen, J. & Volk, A. Harmonic subjectivity in popular music. Tech. Rep. UU-CS-2017-018, Department of Information and Computing Sciences, Utrecht University (2017).

[26] Fujishima, T. Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music. *ICMC Proceedings* **9**, 464–467 (1999).

[27] Ni, Y., McVicar, M., Santos-Rodriguez, R. & De Bie, T. An end-to-end machine learning system for harmonic analysis of music. *IEEE Transactions on Audio, Speech and Language Processing* **20**, 1771–1783 (2012). URL `http://ieeexplore.ieee.org/document/6155600/`. `arXiv:1107.4969v1`.

[28] Humphrey, E. J., Cho, T. & Bello, J. P. Learning a robust Tonnetz-space transform for automatic chord recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 453–456 (IEEE, 2012). URL `http://ieeexplore.ieee.org/document/6287914/`.

[29] Korzeniowski, F. & Widmer, G. A fully convolutional deep auditory model for musical chord recognition. *IEEE International Workshop on Machine Learning for Signal Processing, MLSP* **2016-November**, 13–16 (2016). URL `http://arxiv.org/abs/1612.05082http://dx.doi.org/10.1109/MLSP.2016.7738895`. `1612.05082`.

[30] Cho, T. & Bello, J. P. On the relative importance of individual components of chord recognition systems. *IEEE Transactions on Audio, Speech and Language Processing* **22**, 477–492 (2014).

[31] Mauch, M. & Dixon, S. Approximate note transcription for the improved identification of difficult chords. In *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, 1, 135–140 (2010). URL `https://www.eecs.qmul.ac.uk/{~}simond/pub/2010/Mauch-Dixon-ISMIR-2010.pdf`.

[32] Khadkevich, M. & Omologo, M. Reassigned spectrum based feature extraction for automatic chord recognition. *EURASIP Journal on Audio, Speech, and Mu-*

sic Processing (2013). URL http://asmp.eurasipjournals.com/content/2013/1/15.

[33] Gómez, E. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing* **18**, 294–304 (2006). URL http://pubsonline.informs.org/doi/10.1287/ijoc.1040.0126.

[34] Müller, M. & Ewert, S. Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Ismir, 215–220 (2011). URL http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:CHROMA+TOOLBOX+:+MATLAB+IMPLEMENTATIONS+FOR+EXTRACTING+VARIANTS+OF+CHROMA-BASED+AUDIO+FEATURES{#}0.

[35] O'Hanlon, K. & Sandler, M. B. Compositional chroma estimation using powered Euclidean distance. In *2016 24th European Signal Processing Conference (EUSIPCO)*, 1237–1241 (IEEE, 2016). URL http://ieeexplore.ieee.org/document/7760446/.

[36] Broze, Y. & Shanahan, D. Diachronic Changes in Jazz Harmony: A Cognitive Perspective. *Music Perception: An Interdisciplinary Journal* **3**, 32–45 (2013). URL http://www.jstor.org/stable/10.1525/mp.2013.31.1.32http://www.jstor.org/page/info/about/policies/terms.jsp.

[37] Gauvin, H. L. " The Times They Were A-Changin' " : A Database-Driven Approach to the Evolution of Harmonic Syntax in Popular Music from the 1960s. *Empirical Musicology Review* **10**, 215–238 (2015). URL http://emusicology.org/article/view/4467.

[38] Hedges, T., Roy, P. & Pachet, F. Predicting the Composer and Style of Jazz Chord Progressions. *Journal of New Music Research* **43**, 276–290 (2014). URL http://www.tandfonline.com/doi/abs/10.1080/09298215.2014.925477.

[39] Cannam, C., Landone, C., Sandler, M. & Bello, J. P. The Sonic Visualiser: A Visualisation Platform for Semantic Descriptors from Musical Signals. In *Proc.*

of the 7th International Conference on Music Information Retrieval, 324–327 (Victoria, Canada, 2006).

[40] Di Giorgi, B., Zanoni, M., Sarti, A. & Tubaro, S. Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony. In *Multidimensional Systems (nDS), 2013. Proc. of the 8th International Workshop on*, 1–6 (2013). URL http://ieeexplore.ieee.org/document/6623838/.

[41] Deng, J. & Kwok, Y.-k. A hybrid gaussian-hmm-deep-learning approach for automatic chord estimation with very large vocabulary. In *Proc. 17th International Society for Music Information Retrieval Conference*, 812–818 (2016). URL https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/058{_}Paper.pdf.

[42] Burgoyne, J. A., Wild, J. & Fujinaga, I. An Expert Ground-Truth Set for Audio Chord Recognition and Music Analysis. In *12th International Society for Music Information Retrieval Conference*, ISMIR, 633–638 (2011).

[43] Pfleiderer, M., Frieler, K. & Abeßer, J. *Inside the Jazzomat - New perspectives for jazz research* (Schott Campus, Mainz, Germany, 2017).

[44] Balke, S. *et al.* Bridging the Gap: Enriching YouTube Videos with Jazz Music Annotations. *Frontiers in Digital Humanities* **5** (2018). URL http://journal.frontiersin.org/article/10.3389/fdigh.2018.00001/full.

[45] Déguernel, K., Vincent, E. & Assayag, G. Using Multidimensional Sequences For Improvisation In The OMax Paradigm. In *13th Sound and Music Computing Conference* (Hamburg, Germany, 2016). URL https://hal.inria.fr/hal-01346797.

[46] Murphy, K. P. & P. Murphy, K. Machine Learning : A PROBABILISTIC PERSPECTIVE. *SpringerReference* 73–78,216–244 (2011). URL https://mitpress.mit.edu/books/machine-learning-1http://link.springer.com/chapter/10.1007/978-94-011-3532-0{_}2{%}5Cnhttp://www.springerreference.com/

index/doi/10.1007/SpringerReference{_}35834{%}0Ahttp://www.cs.ubc.ca. 0-387-31073-8.

[47] Comas-Cufi, M., Martin-Fernandez, J. A. & Mateu-Figueras, G. Log-ratio methods in mixture models for compositional data sets. *SORT* **40**, 349–374 (2016). URL http://www.raco.cat/index.php/SORT/article/view/316149.

[48] van den Boogaart, K. G. & Tolosana-Delgado, R. Fundamental Concepts of Compositional Data Analysis. In *Analyzing Compositional Data with R*, 13–50 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013). URL http://link.springer.com/10.1007/978-3-642-36809-7{_}2.

[49] Humphrey, E. J. *et al.* Jams: a Json Annotated Music Specification for Reproducible Mir Research. In *Proc. of the International Society for Music Information Retrieval (ISMIR)* (2014).

[50] Bogdanov, D. *et al.* Essentia: an audio analysis library for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR'13)*, 493–498 (Curitiba, Brazil, 2013). URL http://hdl.handle.net/10230/32252.

[51] Gioia, T. *The Jazz Standards: A Guide to the Repertoire* (Oxford University Press, 2012). URL https://books.google.es/books?id=oPuMQx9GZVcC.

[52] The Smithsonian Collection of Classic Jazz. Smithsonian Folkways Recording (1997).

[53] Jazz: The Smithsonian Anthology. Smithsonian Folkways Recording (2010).

[54] Martin, H. *Charlie Parker and Thematic Improvisation* (Institute of Jazz Studies, Rutgers–The State University of New Jersey, 1996).

[55] Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F. & Widmer, G. madmom: a new Python Audio and Music Signal Processing Library. In *Proc. of the 24th ACM International Conference on Multimedia*, 1174–1178 (Amsterdam, The Netherlands, 2016).

[56] Krumhansl, C. L. & Cuddy, L. L. A Theory of Tonal Hierarchies in Music. In *Springer Handbook of Auditory Research*, vol. 36, 4473–4482 (Springer New York, 2010). URL http://link.springer.com/10.1007/978-1-4419-6114-3{_}3http://www.springerlink.com/index/10.1007/978-1-4419-6114-3.

[57] Hintze, J. L. & Nelson, R. D. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician* **52**, 181–184 (1998). URL http://www.tandfonline.com/doi/abs/10.1080/00031305.1998.10480559.

[58] Mcfee, B. & Bello, J. P. Structured Training for Large-Vocabulary Chord Recognition. In *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, 188–194 (2017). URL https://bmcfee.github.io/papers/ismir2017{_}chord.pdf.

[59] Mauch, M. & Dixon, S. MIREX 2010: Chord detection using a dynamic bayesian network. In *International Society for Music Information Retrieval Conference (ISMIR)* (2010). URL http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=946D14A8B6BB0853D7460290C9151526?doi=10.1.1.368.2014http://www.music-ir.org/mirex/abstracts/2010/MD1.pdf.

[60] Eremenko, V., Demirel, E., Bozkurt, B. & Serra, X. Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research. *International Society for Music Information Retrieval Conference* (2018). URL http://mtg.upf.edu/node/3896.