

# Automatic Rhythm Analysis of Indian Art Music

A data-driven Bayesian approach

Ajay Srinivasamurthy

---

TESI DOCTORAL UPF / 2016

Director de la tesi

**Dr. Xavier Serra Casals**  
Music Technology Group  
Dept. of Information and Communication Technologies



By Ajay Srinivasamurthy

<http://www.ajaysrinivasamurthy.in/phd-thesis>  
<http://compmusic.upf.edu/phd-thesis-ajay>

Licensed under Creative Commons Attribution - NonCommercial  
- NoDerivs 4.0 Unported



You are free to share – to copy, distribute and transmit the work under the following conditions:

- **Attribution** – You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Non-commercial** – You may not use this work for commercial purposes.
- **No Derivative Works** – You may not alter, transform, or build upon this work.

The court's PhD was appointed by the rector of Universitat Pompeu Fabra on ..... , 2016.

Chairman

Member

Member

Member

Secretary

The doctoral defense was held on 01 Sep 2016, at the Universitat Pompeu Fabra and scored as .....

PRESIDENT

MEMBERS

SECRETARY



To the rhythms in the natural world, from which the  
rhythms in music emerge...

---



---

# **Abstract**

To be completed ...



---

# **Resum**

In catalan



---

# **Resumen**

In Spanish



---

# Preface

To be completed ...



---

# Acknowledgements

To be completed ...



---

# Contents

<b>Abstract</b>	vii
<b>Preface</b>	xiii
<b>Acknowledgements</b>	xv
<b>Contents</b>	xvii
<b>List of Symbols</b>	xxi
<b>List of Figures</b>	xxv
<b>List of Tables</b>	xxix
<b>1 Introduction</b>	1
1.1 Context and relevance . . . . .	2
1.2 Motivation . . . . .	5
1.3 Scope and objectives . . . . .	8
1.4 Organization and thesis outline . . . . .	13
<b>2 Background</b>	17
2.1 Rhythm: terminology . . . . .	17
2.2 Music Background . . . . .	19
2.2.1 Indian art music . . . . .	20
2.2.2 Rhythm and Percussion in Carnatic music .	21

2.2.3	Rhythm and Percussion in Hindustani music	27
2.2.4	Carnatic and Hindustani music: A comparison	33
2.2.5	Percussion in Beijing Opera	35
2.3	Automatic rhythm analysis: A review	40
2.3.1	Onset detection	40
2.3.2	Instrument identification	43
2.3.3	Tempo estimation	43
2.3.4	Beat tracking	44
2.3.5	Time signature estimation	46
2.3.6	Downbeat tracking	46
2.3.7	Meter tracking and inference	48
2.3.8	Evaluation measures	48
2.3.9	Rhythm similarity measures	50
2.3.10	Domain-specific approaches	51
2.3.11	Percussion pattern transcription and discovery	52
2.4	Relevant technical concepts	54
2.4.1	Bayesian Models	54
2.4.2	Inference in Bayesian models	54
2.4.3	Speech Recognition Technologies and Tools	55
<b>3</b>	<b>Automatic rhythm analysis in Indian art music</b>	<b>57</b>
3.1	Challenges and Opportunities	58
3.1.1	Challenges	58
3.1.2	Opportunities	62
3.1.3	Characteristics of Indian Art Music	63
3.2	Research problems in rhythm analysis	64
3.2.1	Building data corpora	65
3.2.2	Automatic rhythm annotation	66
3.2.3	Rhythm and percussion pattern analysis	73
3.2.4	Rhythm based audio segmentation	77
3.2.5	Ontologies for rhythm concepts	79
3.2.6	Rhythm similarity measures	80
3.2.7	Symbolic music analysis	81
3.2.8	Evaluation and Integration	82
3.2.9	Extensions to other music cultures	83
3.3	Thesis problems: A formulation	84
3.3.1	Meter inference and tracking	84

3.3.2	Percussion pattern transcription and discovery . . . . .	87
3.3.3	Datasets for research . . . . .	89
3.3.4	A note on terminology and style . . . . .	90
3.4	In search of automatic rhythm analysis methods . . . . .	91
3.4.1	Cycle length estimation . . . . .	93
3.4.2	Downbeat tracking . . . . .	102
3.4.3	Discussion . . . . .	103
<b>4</b>	<b>Data corpora for research</b>	<b>107</b>
4.1	CompMusic research corpora . . . . .	109
4.1.1	Criteria for creation of research corpora . . . . .	110
4.1.2	Carnatic music research corpus . . . . .	111
4.1.3	Hindustani music research corpus . . . . .	119
4.1.4	Creative Commons music collections . . . . .	123
4.2	Test datasets . . . . .	124
4.2.1	Carnatic music rhythm dataset . . . . .	125
4.2.2	Hindustani music rhythm dataset . . . . .	138
4.2.3	Tabla solo dataset . . . . .	158
4.2.4	Mridangam datasets . . . . .	159
4.2.5	Jingju percussion instrument dataset . . . . .	162
4.2.6	Jingju percussion pattern dataset . . . . .	164
4.2.7	Other evaluation datasets . . . . .	165
<b>5</b>	<b>Meter inference and tracking</b>	<b>169</b>
5.1	The meter analysis tasks . . . . .	170
5.2	Preliminary experiments . . . . .	173
5.2.1	Meter tracking using dynamic programming	174
5.3	Bayesian models for meter analysis . . . . .	183
5.3.1	The bar pointer model . . . . .	184
5.3.2	Model extensions . . . . .	197
5.3.3	Inference extensions . . . . .	205
5.4	Experiments and results . . . . .	210
5.4.1	Experiment parameters . . . . .	211
5.4.2	Meter inference . . . . .	215
5.4.3	Meter tracking . . . . .	217
5.4.4	Informed meter tracking . . . . .	221
5.4.5	Summary of results . . . . .	224
5.5	Conclusions . . . . .	228

<b>6 Percussion pattern transcription and discovery</b>	<b>231</b>
6.1 Approaches . . . . .	232
6.2 The case of Beijing Opera . . . . .	234
6.2.1 Percussion pattern classification . . . . .	238
6.2.2 Results and discussion . . . . .	242
6.3 The case of Indian art music . . . . .	244
6.3.1 Pattern library generation . . . . .	245
6.3.2 Automatic transcription . . . . .	249
6.3.3 Approximate pattern search . . . . .	251
6.3.4 Results and discussion . . . . .	255
<b>7 Applications, Summary and Conclusions</b>	<b>263</b>
7.1 Applications . . . . .	263
7.1.1 Dunya . . . . .	266
7.1.2 Sarāga . . . . .	268
7.2 Contributions . . . . .	270
7.3 Conclusions and Summary . . . . .	272
7.4 Future directions . . . . .	275
<b>Appendix A List of Publications</b>	<b>279</b>
<b>Appendix B Resources</b>	<b>281</b>
<b>Appendix C Glossary</b>	<b>283</b>
<b>Bibliography</b>	<b>287</b>
<b>Index</b>	<b>307</b>

---

# List of Symbols

The following is a list of different symbols used in the dissertation along with a short description of each symbol (To be reorganized, better defined and regrouped)

Symbol	Description
$N_p$	Number of particles in the particle filter
$K$	Number of audio frames in a piece
$M$	“Number” of metrical position states (update)
$N$	“Number” of tempo states (update)
$R$	Number of rhythm patterns
$V$	Number of sections
$B$	Number of beats in a <i>tāla</i> cycle/bar
$\phi$	Position in metrical cycle/bar/section
$\dot{\phi}$	Rate of change of position in metrical cycle/bar/section
$r$	Rhythmic pattern indicator
$v$	Section ( <i>vibhāg</i> ) indicator variable
$\mathbf{y}$	Feature vector at frame $k$
$\mathbf{x}$	A vector of latent variables at frame $k$
$\mathcal{B}_q$	The set of all beats in a music piece $q$
$\mathcal{S}_q$	The set of all <i>samas</i> /downbeats in a music piece $q$

---

<b>Symbol</b>	<b>Description</b>
$\mathcal{O}_q$	The set of all akşaras/tatum pulses in a music piece $q$
$\lambda$	The set of HMM parameters
$\mathcal{N}(\mu, \Sigma)$	Normal distribution with mean $\mu$ and covariance matrix $\Sigma$
$\mathcal{P}$	The set of all syllabic percussion patterns
$\mathcal{A}$	The set of all percussion syllables
$N_s$	The total number of percussion syllables
$N_a$	The total number of percussion patterns
$a$	A percussion syllable
$A$	A percussion syllable in the set
$\mathbf{A}$	A percussion pattern
$t$	Time variable (measured in second)
$L$	Length of a pattern in syllables
$\mathbf{G}$	The tempogram matrix
$f[n]$	Audio signal
$f_p[n]$	Percussion enhanced audio signal
$T_f$	Length of a piece (in minutes/seconds)
$\tau_s$	Inter-sama interval
$\tau_b$	Inter-beat interval
$\tau_o$	Inter-akşara/mātrā interval or akşara/mātrā pulse period
$\mathfrak{p}$	Precision
$\mathfrak{r}$	Recall
$\mathfrak{f}$	f-Measure
$\mathfrak{I}$	Information Gain
$\mathfrak{C}$	Correctness measure
$\mathfrak{A}$	Accuracy measure
$\mathbb{A}$	Pattern transition matrix
$\mathbb{B}$	Section transition matrix
$\mathbb{R}$	Width Across Reference matrix
$\mathbb{Q}$	Width Across Query matrix

---

<b>Symbol</b>	<b>Description</b>
$\mathbb{H}$	RLCS length of match matrix
$\sigma$	RLCS weight matrix
$CML_t$	The $CML_t$ measure
$CML_c$	The $CML_c$ measure
$AML_t$	The $AML_t$ measure
$AML_c$	The $AML_c$ measure
$\Theta$	Overlap measure
$\mathbb{1}$	Indicator function
$h$	Frame hop size used for analysis
$\Delta_\phi$	Delta tempo used in discretizing the tempo grid
$\Lambda$	HMM parameter set

---



---

# List of Figures

1.1	Automatic rhythm analysis from audio . . . . .	10
1.2	A suggested reading order for the chapters . . . . .	15
2.1	Four popular Carnatic tālas . . . . .	23
2.2	Structure of Tiśra nađe ādi tāla . . . . .	25
2.3	Four popular Hindustani tāls . . . . .	29
2.4	Ēktāl in dṛṭ lay . . . . .	29
2.5	Percussion patterns in jīngjū . . . . .	37
2.6	Functional units of a rhythm description system . . . . .	41
3.1	Relevant automatic rhythm analysis problems in Indian art music . . . . .	65
4.1	The number of artists by the number of their performances in the Carnatic music corpus . . . . .	115
4.2	Coverage of the Carnatic artists . . . . .	117
4.3	Histogram of $\tau_s$ in the CMR <sub>f</sub> dataset . . . . .	128
4.4	Histogram of $\tau_b$ in the CMR <sub>f</sub> dataset . . . . .	129
4.5	Histogram of median normalized $\tau_s$ in the CMR <sub>f</sub> dataset	130
4.6	Histogram of median normalized $\tau_b$ in the CMR <sub>f</sub> dataset	131
4.7	Computation of the spectral flux feature . . . . .	132
4.8	Rhythm patterns in ādi tāla learned from CMR <sub>f</sub> dataset	134
4.9	Rhythm patterns in ādi tāla learned from CMR dataset	134
4.10	Rhythm patterns in rūpaka tāla learned from CMR <sub>f</sub> dataset	135
4.11	Rhythm patterns in rūpaka tāla learned from CMR dataset	135

4.12 Rhythm patterns in miśra chāpu tāla learned from CMR <sub>f</sub> dataset . . . . .	135
4.13 Rhythm patterns in miśra chāpu tāla learned from CMR dataset . . . . .	136
4.14 Rhythm patterns in khaṇḍa chāpu tāla learned from CMR <sub>f</sub> dataset . . . . .	136
4.15 Rhythm patterns in khaṇḍa chāpu tāla learned from CMR dataset . . . . .	136
4.16 Histogram of $\tau_s$ in the HMR <sub>I</sub> dataset . . . . .	142
4.17 Histogram of $\tau_o$ in the HMR <sub>I</sub> dataset . . . . .	143
4.18 Histogram of $\tau_s$ in the HMR <sub>s</sub> dataset . . . . .	144
4.19 Histogram of $\tau_o$ in the HMR <sub>s</sub> dataset . . . . .	145
4.20 Histogram of median normalized $\tau_s$ in the HMR <sub>I</sub> dataset . . . . .	146
4.21 Histogram of median normalized $\tau_o$ in the HMR <sub>I</sub> dataset . . . . .	147
4.22 Histogram of median normalized $\tau_s$ in the HMR <sub>s</sub> dataset . . . . .	148
4.23 Histogram of median normalized $\tau_o$ in the HMR <sub>s</sub> dataset . . . . .	149
4.24 Rhythm patterns in tīntāl learned from HMR <sub>I</sub> dataset . . . . .	150
4.25 Rhythm patterns in tīntāl learned from HMR <sub>s</sub> dataset . . . . .	150
4.26 Rhythm patterns in ēktāl learned from HMR <sub>I</sub> dataset . . . . .	151
4.27 Rhythm patterns in ēktāl learned from HMR <sub>s</sub> dataset . . . . .	151
4.28 Rhythm patterns in jhaptāl learned from HMR <sub>I</sub> dataset . . . . .	151
4.29 Rhythm patterns in jhaptāl learned from HMR <sub>s</sub> dataset . . . . .	152
4.30 Rhythm patterns in rūpak tāl learned from HMR <sub>I</sub> dataset . . . . .	152
4.31 Rhythm patterns in rūpak tāl learned from HMR <sub>s</sub> dataset . . . . .	152
5.1 An illustration of percussion enhancement . . . . .	175
5.2 Block diagram of the tāla tracking algorithm proposed by Srinivasamurthy and Serra (2014) . . . . .	176
5.3 Estimated time varying tempo curve with tempogram . . . . .	178
5.4 The meter analysis models used in the dissertation . . . . .	185
5.5 An illustration of the bar pointer model . . . . .	187
5.6 An illustration of SP-model transition matrices . . . . .	204
5.7 Results of statistical significance testing of meter analysis results on Indian art music datasets . . . . .	225
6.1 Waveform and spectrogram of jīngjù percussion strokes	235
6.2 Waveform and spectrogram of the pattern shānchuí . . . . .	236
6.3 Block diagram: Jīngjù percussion pattern classification	240

6.4	Block Diagram: Percussion pattern discovery in Indian art music . . . . .	246
6.5	An example of tabla percussion pattern . . . . .	249
6.6	An example of mridangam percussion pattern . . . . .	249
7.1	A screenshot of the recording page of Dunya . . . . .	267
7.2	Screenshots of Sarāga . . . . .	269



---

# List of Tables

2.1	Structure of Carnatic tālas . . . . .	23
2.2	The syllables used in Carnatic music percussion . . . . .	26
2.3	Structure of Hindustani tāls . . . . .	28
2.4	The tabla bōls used in Hindustani music . . . . .	30
2.5	The ṭhēkās for popular Hindustani tāls . . . . .	31
2.6	Syllables used in Beijing opera percussion . . . . .	36
3.1	Performance of meter estimation using GUL algorithm . .	97
3.2	Performance of cycle length estimation using PIK al- gorithm . . . . .	97
3.3	Accuracy of cycle length recognition using KLA algorithm	99
3.4	Accuracy of cycle length recognition using SRI algorithm	99
3.5	Accuracy of cycle length recognition using compara- tive approaches . . . . .	101
3.6	Accuracy of downbeat tracking in Carnatic music subset	103
4.1	Coverage of the Carnatic music corpus . . . . .	114
4.2	Completeness of the Carnatic music corpus . . . . .	118
4.3	Coverage of the Hindustani music corpus . . . . .	121
4.4	Completeness of the Hindustani music corpus . . . . .	121
4.5	CMR <sub>f</sub> dataset description . . . . .	125
4.6	Tāla cycle length indicators for CMR <sub>f</sub> dataset . . . . .	125
4.7	CMR dataset description . . . . .	127
4.8	Tāla cycle length indicators for CMR dataset . . . . .	127
4.9	HMR <sub>f</sub> dataset description . . . . .	138

4.10	Tāl cycle length indicators for HMR <sub>f</sub> dataset . . . . .	139
4.11	HMR <sub>1</sub> dataset description . . . . .	139
4.12	Tāl cycle length indicators for HMR <sub>1</sub> dataset . . . . .	140
4.13	HMR <sub>s</sub> dataset description . . . . .	140
4.14	Tāl cycle length indicators for HMR <sub>s</sub> dataset . . . . .	141
4.15	The tabla solo dataset . . . . .	159
4.16	Anantapadmanabhan Mridangam Strokes dataset . . . . .	161
4.17	The mridangam solo dataset . . . . .	161
4.18	The Jingju percussion instrument dataset . . . . .	163
4.19	The Jingju percussion pattern dataset . . . . .	164
5.1	Results of akṣara period tracking on CMR <sub>f</sub> dataset . . . . .	180
5.2	Results of sama tracking on CMR <sub>f</sub> dataset . . . . .	181
5.3	Summary of the meter analysis models and inference algorithms . . . . .	210
5.4	Results of meter inference with the bar pointer model . . . . .	215
5.5	Results of meter tracking with the bar pointer model . . . . .	217
5.6	Results of meter tracking with a mixture observation model (MO-model) . . . . .	219
5.7	Results of meter tracking with the section pointer model . . . . .	219
5.8	Results of meter tracking with inference extensions to the bar pointer model . . . . .	220
5.9	Tempo informed meter tracking results on Indian music datasets . . . . .	222
5.10	Tempo-sama informed meter tracking results on Indian music datasets . . . . .	223
5.11	Summary of meter analysis performance on Indian art music datasets . . . . .	224
5.12	Summary of meter tracking performance of inference extensions . . . . .	227
5.13	Comparing the meter tracking performance of AMPF <sub>0</sub> and AMPF <sub>m</sub> algorithms . . . . .	228
6.1	Transcription and classification results on JPP dataset . . . . .	242
6.2	Confusion matrix for percussion pattern classification in JPP dataset . . . . .	244
6.3	Query tabla percussion patterns . . . . .	248
6.4	Query mridangam percussion patterns . . . . .	248

6.5	Automatic transcription results on tabla solo dataset (Flat start Hidden Markov models (HMMs)) . . . . .	256
6.6	Automatic transcription results on tabla solo dataset (Isolated start HMMs) . . . . .	256
6.7	Performance of approximate pattern search on tabla solo dataset . . . . .	258
6.8	Automatic transcription results on the mridangam solo dataset . . . . .	260
6.9	Performance of approximate pattern search on mridan- gam solo dataset . . . . .	261



# Introduction

...the most necessary, most difficult and principal thing in music, that is time...

---

W. A. Mozart from *Mozart: The Man and the Artist, as Revealed in his own Words* by Friedrich Kerst, trans. Henry Edward Krehbiel (1906)

We live in a multicultural world that is replete with rich sources of data and information that keep increasing every passing day. The present day **Information and Communication Technologies (ICT)** and tools help us to generate, organize, interact, interpret, consume, assimilate the data and information and enhance our experience with the data, information and knowledge of the world. The technology needs in a multicultural context are evolving to cater to the complex sociocultural contexts in which these technologies and tools are being built and used.

Music is an integral part of our lives and is being produced and consumed at an ever increasing rate. The consumption channels and practices of music have changed significantly over the last two decades. With music going digital, there are large collections of music available on demand to users, which provides a great opportunity to enhance our experience interacting with music. The interaction with music has grown beyond just listening into an enriching and engaging experience with the music content. In such a scenario, there are significant efforts to build automatic tools and

technologies to enhance our experience with large (and ever increasing in size) music collections. Music being a sociocultural phenomenon necessitates these automatic tools to be aware and adapt to such a context and cater to specific music cultures, music producers (musicians and artists) and the widely diverse audiences.

Music Information Research (**MIR**) aims to develop tools and applications for representation, understanding, analysis, and synthesis of music. Though a new and interdisciplinary field of research, it has a significant community working on various problems within the purview of **MIR**. **MIR** focuses on understanding and modeling what music is and how it functions. Its basic aim is to develop veridical and effective computational models of the whole music understanding chain, from sound and structure perception to the kinds of high-level concepts that humans associate with music, such as melody, rhythm, harmony, structure, mood and other possibly subjective attributes and characteristics. Automatic music analysis in **MIR** aims to ‘make sense’ of music and extract useful, musically relevant and semantically meaningful information from music pieces and music collections.

Rhythm is an essential and fundamental concept in music. Music manifests as musical events unfolding in time, and the arrangement of these events in time constitutes the rhythm of the music piece. These events can be grouped and organized in several layers to create complex rhythmic structures and patterns. Automatic rhythm analysis aims to estimate these rhythmic structures and patterns from music. The work presented in this dissertation is at the crossroads of music technology and automatic analysis of music, focusing on rhythm analysis, aiming at domain specific analysis approaches within a multicultural context.

We further describe the context and motivation for the thesis. The scope and objectives of the thesis are clearly identified. The final section describes the organization of the dissertation in detail.

## 1.1 Context and relevance

In the last two decades, **MIR** has received significant attention from the research community and has addressed several relevant research

problems advancing the field of sound and music computing. However, the current research in MIR has been largely limited to eurogenetic (popular) music<sup>1</sup> cultures and do not generalize to other music cultures of the world. The approaches have not been developed within a multicultural context and are incapable of extending to the wide variety of music cultures we encounter.

There is still a wide gap between what can accurately be recognised and extracted from music audio signals and the high level semantically meaningful concepts that human listeners associate with music. Current attempts at narrowing this semantic gap are only producing small incremental progress. One of the main reasons for this lack of major progress seems to be the bottom-up approach currently being used, in which features are extracted from audio signals and higher-level features or labels are then computed by analysing and aggregating these features. The limitation here being the lack of infusion of higher level music knowledge directly into automatic analysis.

The CompMusic project (Serra, 2011) was conceived in such a context. CompMusic<sup>2</sup> (Computational Models for the Discovery of the World’s Music) is focused on the advancement in the field of MIR by approaching a number of current research challenges from a culture specific perspective to build domain specific approaches. CompMusic aims to develop information modelling techniques of relevance to several non-Western music cultures and in the process contributing to the overall field of MIR. Five different music cultures are being studied in the project: Hindustani (North India), Carnatic (South India), Turkish-makam (Turkey), Arab-Andalusian (Maghreb), and Beijing Opera (China).

CompMusic aims to challenge the current Western centered information paradigms, advance our information technology research,

---

<sup>1</sup>The term eurogenetic music was introduced by Srinivasamurthy, Holzapfel, and Serra (2014) to avoid the misleading dichotomy of Western and non-Western music. The discussed theoretical constructs of western music are motivated by music of the European common practice period. We use the word “genetic” rather with its connotation as “pertaining to origins”, coined in 1831 by Carlyle from Gk. genetikos “genitive”, from genesis “origin”, and not in its biological sense as first applied by Darwin in 1859 (<http://www.etymonline.com>). The term was proposed by Prof. Robert Reigle (MIAM, Istanbul) in personal communication.

<sup>2</sup><http://compmusic.upf.edu>

and contribute to our rich multicultural society. The motivation behind CompMusic is that the information technologies used for music processing have typically targeted the western music traditions, and current research is emphasizing this bias even more. However, to develop technologies that can deal with the richness of our world's music, there is a need to study and exploit the unique aspects of other musical cultures.

CompMusic further identifies that 'making sense' of music is much more than decoding and parsing an incoming stream of sound waves into higher-level musical objects such as onsets, notes, beats, melodies and harmonies. Music is embedded in a rich web of cultural, historical, commercial and social contexts that influence how it is interpreted and categorised. Though all music traditions share common characteristics, each one can be recognized by particular features that need to be identified and preserved. Many qualities attributed to a piece of music by listeners and musicians cannot solely be explained by the content of the audio signal itself. It is clear that high-quality automatic music description and understanding can only be achieved by also taking into account additional information external to the music. This also constitutes a fruitful opportunity for interdisciplinary research involving engineers, musicians, musicologists and psychologists to collaborate and contribute their valuable knowledge.

Looking at the problems emerging from various musical cultures will not only help those specific cultures, but we will open up our existing computational methodologies, making them much more versatile. It will emphasize the limitations of the current methodologies and present open issues. In turn, it will also help preserve the diversity of our world's culture. The research results of CompMusic are integrated into Dunya (Porter, Sordo, & Serra, 2013), which is a web-based software application that lets users interact with an audio music collection through the use of musical concepts that are derived from a specific music culture. The users can also access all the research results and extracted features through a web API.

Within the field of MIR there are many research problems that can benefit from a culture specific perspective. CompMusic focuses on the extraction of features from audio music recordings related to melody and rhythm, and on the semantic analysis of the

contextual information of those recordings. The goal is to characterize culture specific musical facets of each repertoire and to develop musically meaningful similarity measures with them. The research in CompMusic is data driven, thus it revolves around corpora. One of the goals of CompMusic is to construct a research corpus for each music tradition(Serra, 2014). The types of data gathered are mainly audio recordings and editorial metadata, which are then complemented with descriptive information such as editorial metadata, scores and/or lyrics as available.

The work presented in this dissertation is conducted within the context of the CompMusic project but focusing on automatic rhythm analysis research problems for Indian art music from a data driven perspective using signal processing and machine learning approaches. The dissertation imbines and inherits all the goals and context of CompMusic project as applied to rhythm analysis.

A meaningful computational analysis for rhythm characterization should consider cultural aspects attached to it (Serra et al., 2013). Through a culture-aware and domain specific approach to computational rhythm modeling of Indian art music, we will also get better insights into the current MIR tools which would improve their performance. We will be able to develop better algorithms, newer methodologies and techniques for the study of world's music and reach out to a much larger part of our multicultural world. The development of these models would also allow cross cultural comparative studies between different musical systems, enriching the present knowledge of world's music and provide interesting sociocultural, cognitive, and musical perspectives. Such an approach is relevant since it aims to push ahead the boundaries of automatic rhythm analysis to address current challenges and be more inclusive to address varied needs of different music cultures of the world.

## 1.2 Motivation

Rhythm is a fundamental aspect of music. Music has repeating structures and patterns, with several musical events organized in time. It is primarily an event-based phenomenon and detecting and characterizing musical events and their transitions is an important task. The automatic analysis of these musical events can provide

us useful insights into music and help us to derive semantically meaningful higher level concepts.

Rhythmic structures are often well defined when musical events are organized in several layers of hierarchy - leading to metrical structures. The metrical structures provide a fundamental framework in time to organize events and hence play a pivotal role. Most melodic and rhythmic phrases, lyrical lines, harmonic changes are organized around the metrical structures and hence the estimation of different aspects of the meter hence is an important MIR task. Estimating the note onsets, tempo, beats and downbeats are useful and necessary for any further analysis of music. Though each of these aspects can be extracted in an isolated fashion, there is significant interplay between these entities and hence a holistic approach to describing all these aspects of meter is an approach that needs to be explored further.

There are additional structures often in music at a longer time scale than the metrical structure. These are structural components (e.g. verse, chorus, bridge, intro, outro, solo) are well defined in many music forms as different sections of a music piece. Segmenting a song at these section boundaries is also a useful task for summarizing the audio or for structural analysis of music pieces. Such a structural analysis can benefit from metrical analysis of a music piece since most of these sections are aligned with metrical boundaries in the song.

Music is also replete with rhythmic patterns at several different levels. Music is expressed through these grouping of events into rhythmic patterns and hence are very fundamental to understanding rhythm. The rhythmic patterns can also be indicative of the underlying musical structure. Understanding and analysis of these patterns would help in a comprehensive computational description of the music piece and then used in several applications.

Analysis of rhythmic structures and patterns hence is an important research task in MIR. Tools developed for rhythm analysis can be useful in a multitude of applications such as intelligent music archival, enhanced navigation of music collections, content based music retrieval, and for an enriched and informed listening of music. All these tools will further be integrated into Dunya, a culture aware music navigation. The target audience for such tools span a wide range: serious music listeners who wish for an enhanced ex-

perience with music, music students who wish to learn more about the music they are listening to, musicians who can use these tools to better promote their music, musicologists who can use tools in their work, and music collectors and record labels who can organize, archive and present their music better.

With large and ever growing music collections, the need of the hour are innovative ways for meaningful organization and navigation through these large collections. Large music collections would mean an automatic analysis is desired over manual curation that can be tedious, time consuming and highly resource intensive. In addition to the metadata associated with music recordings, using the underlying musical concepts to organize music collections is the best approach in such a case for better search and discovery within the collection. This necessitates defining similarity (or distance) measures between these recordings that can be used to group and collate recordings. In addition to the context based similarity (that uses mainly editorial metadata) that is predominantly used today, there is a need to develop content based similarity (using audio content). Further, navigating within a recording would also mean that these similarity measures are needed within the piece for different parts of the piece.

As specified earlier, as meaningful navigation and retrieval can be achieved better using the sociocultural context of the music with all its unique features and specificities - using culture specific similarity measures. Rhythmic features are a component of the overall similarity measures for such a task and rhythm similarity measures can hugely benefit from automatic rhythm analysis of rhythmic structures and patterns. The culture specificity applies at several levels - it applies to identifying unique challenges for the current day ICTs making them specific and meaningful, applies to research approaches for automatic analysis of music, and to the methodologies of combining information from several data sources to define meaningful similarity measures.

With a significantly sophisticated rhythmic framework, Indian art music poses a big challenge to the current state of the art in automatic rhythm analysis(Srinivasamurthy, Holzapfel, & Serra, 2014). There are several important automatic rhythm analysis tasks in Indian art music that have not been studied. With such complexities, developing approaches for rhythm analysis is Indian art music can

help to identify the limitations of current approaches to improve their performance to make them better and more general. As emphasized earlier, there is significant gap between the current capabilities of the music technologies used in commercial services and the needs of our culturally diverse world. This is evident in Indian art music - where the existing technologies fall short of utilizing even the basic musical characteristics and limit our music listening experience. Being well established art music traditions with a significant audience around the world, Indian art music traditions are ideal candidates to develop culturally aware automatic rhythm analysis methods.

It is important to comment that in the pursuit of culture specific methodologies, it is illusionary to believe that specialist systems can be developed for each of the musics of the world. Therefore, a more rational approach is to develop culture specific methods that are also generalizable and adaptive to other contexts and musics.

The motivation for culture specific automatic rhythm analysis in Indian art music stems from all the above described reasons. In addition to the above, to the best of our knowledge, this is the first thesis to comprehensively address automatic rhythm analysis problems in Indian art music and hence would open up the way for further research on the topic. Being an unexplored area of research, it is important and necessary to clearly identify the scope and objectives of this dissertation.

### 1.3 Scope and objectives

The work presented in thesis stands at the intersection of audio music processing, machine learning, music theory, musicology, and the application of enriched music listening aiming at automatic rhythm analysis. Automatic rhythm analysis is itself a broad area of research and hence it is quite necessary to define and delimit the scope of the research presented in the dissertation, while identifying the research questions and the objectives of the thesis clearly. The broad objectives of the presented research are listed below:

- To identify and formulate relevant automatic rhythm analysis problems in Indian art music. Convert musical definitions into engineering formulations amenable to quantitative

analysis using signal processing and machine learning approaches. To identify challenges and opportunities in automatic rhythm analysis of Indian art music.

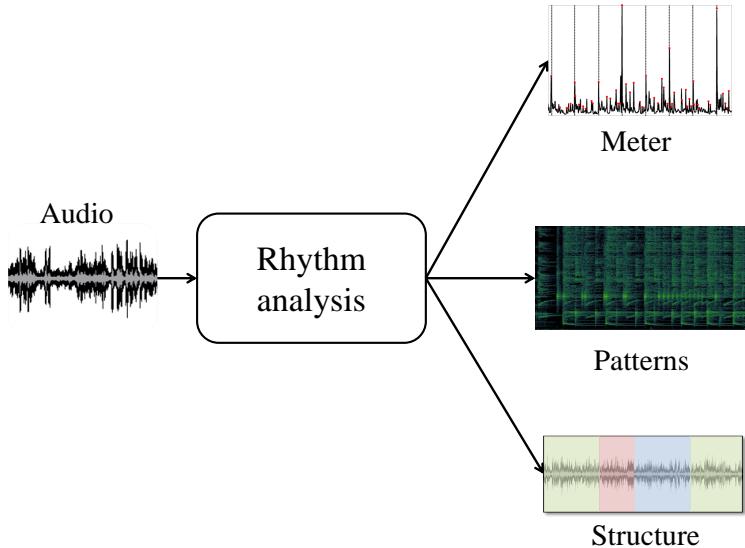
- To build useful annotated collections of Indian art music (both audio and symbolic) with a focus on rhythm, for future research in automatic rhythm analysis
- To create and construct culture-specific computational rhythm models for Hindustani and Carnatic music
- To develop novel signal processing and machine learning methods for analysis for rhythm analysis of Indian Art music
- To devise and develop specific rhythm similarity measures for a better discovery of Indian Art music
- To extend and generalize the specific models to other relevant music cultures being studied in the context of CompMusic project - Beijing Opera and Makam Music of Turkey (MMT).

To explain the scope of the thesis, a long and comprehensive title that defines the scope of the work presented can be written as:

*Culture-aware and domain-specific signal processing and machine learning approaches for automatic analysis, description and discovery of rhythmic structures and patterns in musical audio collections of Indian Art Music*

In alignment with the goals of CompMusic, the final goal of such an analysis is to define culture specific and musically meaningful rhythm similarity measures within a music repertoire. The focus of the thesis is on Indian art music. While there are five different music traditions under study in CompMusic project, the thesis aims to explore extensions to some relevant rhythm analysis problems in Beijing Opera.

The thesis explores data-driven engineering approaches for analysis of audio music recordings as the primary source of information. The audio recording is hence at the centre of analysis, with several types of rhythm related information extracted from a recording. Figure 1.1 shows an example of such a paradigm, showing



**Figure 1.1:** Example of automatic rhythm analysis from audio recordings estimating meter, rhythmic patterns and structure from audio recordings. The approaches in the dissertation follow a similar flow.

meter, patterns and structure extracted from an audio recording of a music piece. Other possible mediums of music dissemination such as scores, lyrics and contextual information is considered secondary in the scope of the thesis, though scores are occasionally used. The approaches explored in the thesis are primarily audio signal processing and Bayesian machine learning methods, exploring mostly supervised learning methods to develop novel rhythm analysis algorithms. Semantic analysis, which is the other core research area of CompMusic is not the focus of this thesis.

The thesis aims to bring in as much musical knowledge to the methods as possible, including and using all the known attributes of music. The goal is to build domain specific and informed signal processing and machine learning methods, so that the extracted information is musically relevant and useful. Bayesian methods provide an effective framework to bring in higher level music knowledge into models, in terms of model structure and priors.

The emphasis of the thesis is on data and methods. The data-driven approaches need good quality datasets, which have been

careful compiled and annotated within the context of the thesis. The algorithms work on real world representative music collection - organized curated collections of music that are accessible.

The thesis focuses only on music analysis and not on music generation, composition and synthesis. The generative models used for analysis in the thesis can however be used for such a task if needed, though it is not the focus of the thesis. In terms of our interaction and experience with music, the thesis focuses mainly on enhanced music listening. Though some of the tools and methods can be useful to both teachers and students of music, it is not the focus of the thesis.

The work presented is done on well studied art music cultures of India and borrows from the significant musicological literature already available. The thesis however aims to help musicologists with these rhythm analysis tools. The data and the methods presented in the thesis can be used by musicologists for large scale corpus level musicological analysis. There are illustrative examples of such analyses in the dissertation, but the thesis does not aim to make any significant musicological conclusions. The work in the thesis also borrows from consultation with several musician collaborators over the course of thesis. The analysis methods developed in the thesis do not aim to replace expert musician opinions, but only work within the framework provided by musicians and musicologists to enhance our experience with music. The problems formulated and addressed in the thesis are on concepts that have well grounded definitions and agreement among the musician and musicology community.

In addition to exploring novel approaches to automatic rhythm analysis, the thesis aims to answer the following questions within the context of rhythm analysis:

1. It is hypothesized that automatic analysis of rhythmic structures and patterns from audio signals needs specific methodologies that make use of knowledge about the underlying musically meaningful rhythmic structures. To what extent does incorporating higher level knowledge affect the performance of automatic analysis ? What kinds of higher level information is useful and leads to a better performance ? How can such higher level information be included in the framework of Bayesian models to develop

novel rhythm analysis algorithms ?

2. How do the existing rhythm analysis methods designed with different rhythmic structures extend to complex metrical structures in Indian art music ? What limitations can we identify of the existing state of the art ?
3. It is hypothesized that instead of a component-wise disjoint approach to estimating different components of rhythm, it might be useful to jointly estimate all the relevant components together in a single framework. It is expected to utilize the interplay between the components to lead to a better estimation. Does a holistic approach work better or is it better to estimate individual components separately. Which component of rhythm is better estimated with other components, which component can be independently estimated ?
4. It still remains an open question if we need more specialist approaches, or more general approaches that are able to react to a large variety of music. Generally, it appears desirable to have generic approaches that can be adapted to a target music using machine learning methods. What are some such methods, and how can they be useful to adapt it to different music cultures ?
5. Indian art music and several other music traditions of the world have developed a syllabic percussion systems to define and describe percussion patterns, which provide a language for percussion in those music cultures. What is the utility of these syllabic percussion systems in automatic percussion pattern analysis and discovery ?
6. Given the availability of useful annotated test datasets, one of the questions to ask is to see if the annotations and the data themselves give rise any meaningful and valid musicological conclusions.

Broadly, the thesis identifies the challenges and opportunities in automatic rhythm analysis of Indian art music, formulates several rhythm analysis tasks, addresses the issues with building datasets for rhythm analysis, and then focuses on the tasks of meter and percussion pattern analysis.

The scope of the thesis within CompMusic is to provide with the tools and methods to be part of the comprehensive set of content based analysis methods for the music cultures under study, with the final goal of utilizing these analysis methods to define musically relevant similarity measures.

The major strategy of CompMusic is open and reproducible research - to be open in sharing ideas, goals, results, data and code as widely as possible. All the data, code and results presented in the thesis will be available openly or be accessible to the research community. Whenever possible, resources will be provided to reproduce the results of the thesis. The data and code will be shared with the community through open source platforms under open licenses. The open dissemination strategy is one of primary objectives of the thesis.

## 1.4 Organization and thesis outline

The dissertation has seven chapters. Each chapter is written on a major topic of the thesis and is aimed to be self contained with a short introduction, content and a summary. This is to minimize dependence across chapters to the extent possible.

After an introduction to the thesis in Chapter 1, Chapter 2 provides an overview of the music background and review of the state of the art as needed for the thesis. Chapter 3 is focused on identifying and discussing several novel automatic rhythm analysis problems in Indian art music. Chapter 4 presents all the rhythm related datasets compiled as a part of CompMusic project that will be used for various rhythm analysis tasks. Chapter 5 and Chapter 6 are the main chapters of the dissertation discussing the topics of meter inference and percussion pattern discovery, respectively. Chapter 7 presents some of the applications and conclusion with pointers for future work. In addition, the links to resources from the thesis (data, code, examples) are listed in Appendix B. There are several new non-standard terms in the thesis including unfamiliar terms related to Indian art music which are all listed and defined in a glossary in Appendix C.

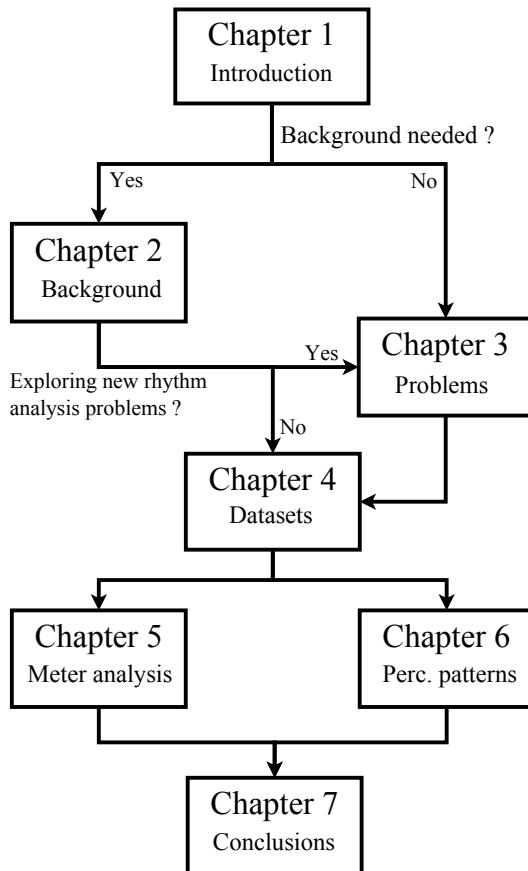
Chapter 2 provides an overview of the background material necessary for the thesis. It introduces concrete terminology of rhythm

concepts and a basic introduction to rhythm in Indian art music and Beijing Opera. It then provides an overview of the state of the art in automatic rhythm analysis in MIR. The chapter ends with a brief overview of the technical concepts useful to understand the thesis work better. The content in chapter is compiled and presented from several external sources.

Chapter 3 identifies several challenges and opportunities to automatic rhythm analysis in Indian art music. The chapter aims to present all identified relevant research problems, while only a subset of them is addressed in the thesis. For these problems, the chapter also presents an overview of the state of the art when available. It further elaborates and formulates the thesis problems that are addressed in detail in the next chapters and presents an evaluation of the state of the art for some of these tasks on Indian art music. The large part of the content of the chapter is derived from several discussions with collaborators of CompMusic, musicians, musicologists and listeners on what they consider are important rhythm analysis problems, and from the paper by Srinivasamurthy, Holzapfel, and Serra (2014).

Chapter 4 describes the efforts of the CompMusic in compiling and annotating the CompMusic research corpora and test datasets. The chapter also presents a systematic framework to elucidate a set of design principles to build data corpora for research in Indian art music. All the annotated rhythm related datasets are described in detail emphasizing on the research problems in which they are useful. Other state of the art datasets that are used in the thesis are also described. Apart from being useful as test datasets to evaluate algorithms and approaches, annotated datasets are also useful to infer musically meaningful observations. Hence a corpus level statistical analysis of relevant test datasets is also presented to draw some interesting observations. The datasets described in the chapter are a collective effort of the CompMusic team as indicated with each dataset. Some of the content is described in the paper by Srinivasamurthy, Koduri, Gulati, Ishwar, and Serra (2014).

Chapter 5 presents the primary contribution of the thesis and describes one of the main research problems addressed. The chapter focuses on the problem of meter analysis and describes several approaches to the task in the context of both Carnatic and Hindustani music of India. The chapter proposes novel Bayesian models and



**Figure 1.2:** A suggested reading order for the chapters of the dissertation

novel inference algorithms for different levels of informed meter analysis, with a comprehensive evaluation on annotated datasets. The content of the chapter is derived from the current state of the art in meter analysis, along with some of the recently published papers (Srinivasamurthy & Serra, 2014; Srinivasamurthy, Holzapfel, Cemgil, & Serra, 2015, 2016; Holzapfel, Krebs, & Srinivasamurthy, 2014) and unpublished results.

Chapter 6 presents the other important contribution of the thesis and describes the task of percussion pattern transcription and discovery in Indian art music. The work presented in the chapter is preliminary and exploratory, but demonstrates the effective use of syllabic percussion systems in representation, transcription

and discovery of percussion patterns in percussion solo recordings. An evaluation is provided on both tabla and mridangam drum solo recordings datasets. As a test case, experiments on percussion pattern classification in Beijing Opera are also presented. A part of the results presented in the chapter are derived from the papers by Gupta, Srinivasamurthy, Kumar, Murthy, and Serra (2015) and Srinivasamurthy, Caro, Sundar, and Serra (2014). Chapter 7 presents some of the applications and conclusions. The chapter summarizes the results from different chapters and presents pointers for future work.

Each chapter is self contained and can be read in isolation with sufficient background. However, the following order is recommended for reading, and further summarized in Figure 1.2. Starting with Chapter 1, if the reader has sufficient background on rhythm in Indian music and rhythm analysis, Chapter 2 can be skipped. For a researcher starting out and exploring new problems and resources, Chapter 3 and Chapter 4 might be more interesting. Chapter 5 and Chapter 6 focus on separate research problems and can be read independently. Chapter 4 might be necessary to understand the evaluations presented in Chapter 5 and Chapter 6. Chapter 7 might be useful to understand some of the applications in more detail and Appendix B and Appendix C can be used as quick guides for resources and term definitions, respectively.

To the best of our knowledge, this dissertation is the first comprehensive attempt in computational analysis of rhythm for Indian art music. By addressing the problems discussed in this dissertation within the context of CompMusic project, we aim to develop useful tools and algorithms for automatic rhythm analysis of Indian art music. Integrated into Dunya, we hope that these tools will provide an enriched experience to the listener, enhanced through a cultural context. In the process, we also hope to obtain a better understanding and provide deeper insights into the nature of rhythm in Indian art music, and contribute to improving the state of the art in MIR.

# Chapter 2

## Background

... mere metrical measurement is not *tāla*. It is a harmonious correlation of discipline and freedom.

---

Shankar (1999, p. 61)

The chapter provides the background necessary for the music and technical work presented in the dissertation. The main aims of this chapter are:

1. To describe relevant rhythm related concepts in Indian art music and Beijing Opera
2. To establish a consistent terminology for several rhythm related music concepts
3. To present an overview of the state of the art in automatic rhythm analysis problems that will be addressed in this dissertation

### 2.1 Rhythm: terminology

As observed already many decades ago, discussions about rhythm tend to suffer from inconsistencies in their terminology (Sachs, 1953). Let us therefore try to locate definitions for some basic terms, in order to establish a consistent representation in the dissertation. György Ligeti, a European composer who showed a high

interest in rhythmic structures in music of African cultures, defined rhythm as “every temporal sequence of notes, sounds and musical *Gestalten*”, while he referred to meter as “a more or less regular structuring of temporal development” (Ligeti, 2007). According to that definition, rhythm is contained in all kinds of music, while pulsating rhythm which is subordinated to a meter is not found in all music. Kolinski (1973) describes the regular structuring caused by meter as organized pulsation, which functions as a framework for rhythmic design.

Pulsations in meter are organized into hierarchical levels of differing time spans, usually demanding the presence of pulsation at least on three levels; these levels are referred to as subdivisions, beats, and measures: from short to long time-span (London, accessed 19.12.2012). The pulsation at the beat level was referred to as primary rhythmic level by Cooper and Meyer (1960), and they define it as the lowest level on which a complete rhythmic group is realized. The same authors identify this level with a subjective notion as well, by referring to it as the level at which the beats are felt and counted. As listeners tend to count the beats at varying levels (Moelants & McKinney, 2004; Parncutt, 1994), and what can be considered a complete rhythmic group can be argued as well, we are confronted with a significant amount of ambiguity in determining this level for a piece of music. Finding a clear terminology is further hampered by the fact that the pulsation at the beat level is commonly also referred to as “beat” or “pulse” as observed by (London, 2004, accessed 19.12.2012).

A shortcoming of most attempts to describe rhythm and meter is the assumption about pulsation to consist of recurring, precisely equivalent and equally-spaced stimuli (Lerdahl & Jackendoff, 1983). Such preconditions cause difficulties when analyzing music of other cultures since in many cases, mutual equidistance becomes an exception rather than the rule, taking the important role of additive meters in Indian, Greek and Turkish music as one example.

In order to obtain a consistent terminology, we consider meter as being an organization of pulsation into different levels related to the time-spans of the individual pulsations. Note that this may include an irregular pattern of unequal time-spans at some level, e.g. due to the presence of an additive meter. We will consider

pulsations on three levels. On the (lower) subdivision level, we will refer to subdivision pulsation or subdivision pulses, depending on if we refer to the time series that constitutes the pulsation or to the individual instances of the pulsation, respectively. On the beat level, we will differentiate between the beat pulsation, and beats as its instances (instead of the inconvenient term of “beat pulses”). On the bar level, the term pulsation is not appropriate due to the often larger time-span at this level. Therefore, we use the notions of bar length to describe the time-span at this level and downbeat as the beginning of a bar. The overall structure of a meter is defined by the time-span relations between these three levels. Typically, the time-span relation between bar and beat level is denoted as the bar length (in beats), and the relation between beat and subdivision level as the subdivision meter (in subdivision pulses).

It was observed by e.g. Clayton (2000) that meter in music causes a transformation of time from linear development to a repetitive and cyclic structure. This happens because its levels of pulsation constitute what we want to refer as metrical cycles. Throughout the text, we put emphasis on terms containing the notion of a cycle (such as the measure cycle for the cycle repeating on every downbeat), a notion suitable in the musical context of Indian music. In addition, in Indian music, the beats or the subdivisions of a bar can be grouped to form musically defined metrical structures, broadly called sections of the bar. The metrical structures in Indian music are discussed next, with an introduction to the Indian art music traditions of Hindustani and Carnatic music.

## 2.2 Music Background

This section describes the primary music cultures that are the focus of study in this dissertation. The focus is on the rhythm and percussion related concepts in these music cultures. This section is not a comprehensive treatment of the subject, and is just sufficient to follow the rest of the chapters of the dissertation. Additional references are provided that have an in depth discussion of the concepts presented.

### 2.2.1 Indian art music

Hindustani (Hindustāni) and Carnatic (Karnātaka) music are two of the most predominant art music traditions in India. Hindustani music is spread mainly over the northern parts of the Indian sub-continent (northern and central parts of India, Pakistan, Nepal, and Bangladesh), which is a huge geographic area with diverse cultures that influence the music. Carnatic music is predominant mainly in Southern parts of the Indian subcontinent (South India and Sri Lanka). Both these musics have a long history of performance and continue to exist and evolve in the current sociocultural contexts. Both of them have a large audience and significant musicological literature that can used to formalize MIR problems for these musics. The presence of a large dedicated audience and a significant musicological literature are a good motivation to study these music cultures from a computational perspective and build tools and methods for automatic analysis for melodic and rhythmic analysis in these music cultures.

While the two musics differ in performance practices, they share similar melodic and rhythmic concepts. The melodic framework is based on *rāg* in Hindustani music and *rāga* in Carnatic music. The rhythmic framework is based on cyclic metrical structures called the *tāl* in Hindustani music and *tāla* in Carnatic music.

A note on disambiguating terminology is in order. Both the words *tāl* and *tāla* are Sanskritic in origin and have the same literal meaning of a “hand-clap”. The difference apparent in transliteration is due to their usage in Hindustani and Carnatic music. We use the language Hindi for all the terms of Hindustani music, and the South Indian language Kannada for all the terms of Carnatic music. Hence the roman transliteration of terms change accordingly. We will use the term Indian art music to refer collectively to Carnatic and Hindustani music. For consistency and convenience, when it is clear from the context, we will use the word *tāla* to mean both *tāl* and *tāla* when we refer collectively to the two Indian musics, and use the respective terms while referring to each music culture individually.

In Carnatic music, a concert, called a *kachēri*, is the natural unit of Carnatic music and used as the main unit of music distribution. A concert has one or more lead artists - mainly vocal, *vīṇā* (commonly

spelled veena), violin, or flute, melodic accompaniment (mainly violin), and one or more percussion accompaniments - mainly *mṛdaṅgam*. Carnatic music is predominantly composition based and most commercial releases are concerts, comprising of several pieces that are improvised renderings of compositions. Vocal music is predominant in Carnatic music and most of the compositions are to be sung. Even in instrumental music, the lead artist aims to mimic vocal singing (Viswanathan & Allen, 2004). The *rāga* and *tāla* are the most important metadata associated with a composition and hence a recording of the composition. Each composition is composed in one or more *rāgas* and *tālas*.

Due to the wider geographic extent of Hindustani music, it is diverse with several different music styles falling under its gamut. Several different styles of Hindustani music exists, but in the dissertation we focus mainly on *khyāl*, the most popular style of Hindustani music. A typical *khyāl* performance has lead vocals or a lead instrument (such as sitar, sarod, flute, santoor), a melodic accompaniment (a harmonium or a *sāraṅgi*), and a percussion accompaniment tabla. In *dhrupad* style, *pakhāvaj* is the main percussion accompaniment. The artists in Hindustani music belong to what are called *gharānās*, the stylistic schools. Though all the *gharānā* use the same music concepts and basic style, each of them have their own nuances that are well distinguished and documented citeXX.

Rhythm in Indian art music revolves around the central theme of the *tāla* and hence it is the primary focus of this dissertation. Though the idea and purpose of a *tāla* in both music cultures is the same, there are significant fundamental differences in performance practices and terminology. An in depth description of rhythm concepts in Indian art music is described in the following section, highlighting these similarities and differences.

### 2.2.2 Rhythm and Percussion in Carnatic music

Sambamoorthy (1998) provides a comprehensive description of *tālas* in Carnatic music. In Carnatic music, the *tāla* provides a broad structure for repetition of music phrases, motifs and improvisations. It consists of fixed length time cycles called *āvartana* which can be referred to as the *tāla* cycle. In an *āvartana* of a *tāla*, phrase refrains, melodic and rhythmic changes occur usually at the be-

ginning of the cycle. An *āvartana* is divided into basic equidistant time units called *akṣaras*. The first *akṣaras* pulse of each *āvartana* is called the *sama*. The *sama* is often accented, with notable melodic and percussive events. Each *tāla* also has a distinct, possibly non-regular division of the cycle period into sections called the *aṅga*. The *aṅgas* serve to indicate the current position in the *āvartana* and aid the musician to keep track of the movement through the *tāla* cycle. A movement through a *tāla* cycle is explicitly shown by the musician using hand gestures, based on the *aṅgas* of the *tāla*.

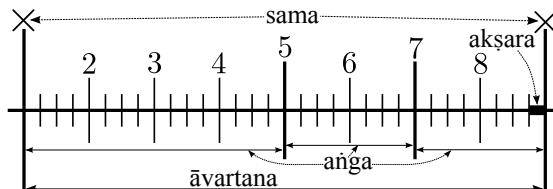
The common definition of an isochronous (equally spaced in time) beat pulsation, as the time instances where a human listener is likely to tap his/her foot to the music, is likely to cause problems in Carnatic music. Due the explicit hand gestures, listeners familiar to Carnatic music tend to tap to an non-isochronous sequence of beats in certain *tālas*. Hence we need an adapted definition of a beat for the purpose of a common ground, defined as a uniform pulsation. It is to be noted that however that an equidistant beat pulsation can later help in obtaining the musically relevant possibly irregular beat sequence that is a subset of the equidistant beat pulses. The *akṣaras* in an *āvartana* are grouped into equal length units, which we will refer to it as the beats of the *tāla*. The perceptual foot tapping time “beats” are a subset of this uniform beat pulsation.

The sub-division grouping structure of the *akṣaras* in a beat is called the *naḍe* (also spelled *naḍai*) or *gati*. The most common *naḍe* is *caturaśra*, in which a beat is divided into 4 *akṣaras*. Another important aspect of rhythm in Carnatic music is the *edupu*, the “phase” or offset of the lead melody, relative to the *sama* of the *tāla*. With a non-zero *edupu*, the composition does not start on the *sama*, but before (*atīta*) or after (*anāgata*) the beginning of the *tāla* cycle. This offset is predominantly for the convenience of the musician for a better exposition of the *tāla* in certain compositions. However, *edupu* is also used for ornamentation in many cases. Though there are significant differences in terms of scale and length, as an analogy, the concepts of *akṣara*, the beat, and the *āvartana* of Carnatic music bear analogy to the subdivision, beat and the bar metrical levels of Eurogenetic music. Further, *aṅga* are the possibly unequal length sections of the *tāla*, formed by grouping of beats.

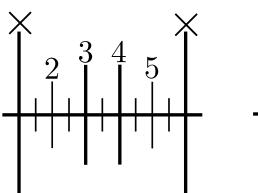
Carnatic music has a sophisticated *tāla* system which incorpo-

Tāla	# beats	nađe	# Akşara
Ādi	8	4	32
Rūpaka	3	4	12
Miśra chāpu	7	2	14
Khanḍa chāpu	5	2	10

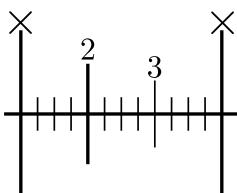
**Table 2.1:** Structure of Carnatic tālas, showing the akşaras per beat (an indicator of nađe), the number of beats and akşaras in each cycle



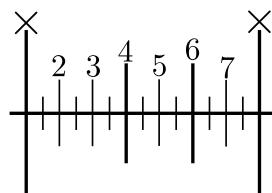
(a) Ādi tāla, illustrated



(b) Khanḍa chāpu tāla



(c) Rūpaka tāla



(d) Miśra chāpu tāla

**Figure 2.1:** An āvartana of four popular Carnatic tālas, showing the akşaras (all time ticks), beats (numbered time ticks), aṅgas (long and bold time ticks) and the sama (×). Ādi tāla is also illustrated using the terminology used in the dissertation.

rates the concepts described above. There are seven basic tālas defined with different aṅgas, each with five variants leading to the popular 35 tāla system (Sambamoorthy, 1998). Each of these 35 tālas can be set in five different nađe, leading to 175 different combinations. However, most of these tālas are extremely rare in performances with just over a ten tālas that can be regularly seen in concerts. A majority of pieces composed in four popular tālas - ādi, rūpaka, miśra chāpu, and khanḍa chāpu. The structure of those four popular tālas in Carnatic music are described in Table 2.1 and illustrated in Figure 2.1, all in caturaśra nađe (division of a beat into two

or four akṣaras). The different concepts related to the *tālas* of Carnatic music are also illustrated<sup>1</sup> in Figure 2.1a. The figure shows the akṣaras with time-ticks, beats of the cycle with numbered longer time-ticks, and the sama in the cycle using  $\times$ . The aṅga boundaries are highlighted using bold and long time-ticks e.g. ādi tāla has 8 beats in a cycle, with 4 akṣaras in each beat leading to 32 akṣaras in a cycle, while rūpaka tāla has 12 akṣaras in a cycle, with 4 akṣaras in each of its 3 beats.

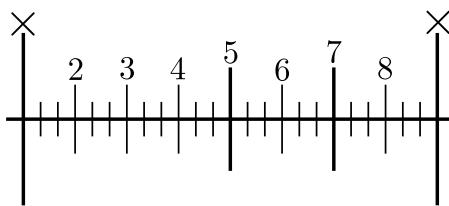
The case of non-isochronous beat *tālas*, miśra chāpu and khaṇḍa chāpu, need a special mention here. Figure 2.1d shows miśra chāpu to consist 14 akṣaras in a cycle. The 14 akṣaras have a definite unequal grouping structure of 6+4+4 (or 6+8 in some cases) and the boundaries of these groups are shown with visual gestures, and hence form the beats of this tāla (Sambamoorthy, 1998). However, in common practice, miśra chāpu can also be divided into seven equal beats. In this dissertation, we consider miśra chāpu to consist of seven uniform beats as numbered in Figure 2.1d, with beats  $\times$ , 4 and 6 being visually displayed. Similarly, khaṇḍa chāpu has 10 akṣaras in a cycle grouped into two groups as 4+6. In the scope of this dissertation, khaṇḍa chāpu can be interpreted to consist of 5 equal length beats.

In the dissertation, we focus on the most popular *tālas* for analysis, all of whom are in caturaśra naḍe. But for completion, an example of tiśra naḍe, where each beat is divided into 3 (or 6) akṣara is illustrated for ādi tāla in Figure 2.2. We can clearly see that a tiśra naḍe ādi tāla has 8 beats of 3 akṣaras each, leading to 24 akṣaras in a cycle.

Most performances of Carnatic music are accompanied by the percussion instrument mridangam (*mṛdaṅgam*), a double-sided barrel drum. There could however be other percussion accompaniments such as ghaṭam (the clay pot), khañjira (the Indian tambourine), thevil (a two sided drum) and mōrsing (the Indian jaw harp), which follow the mridangam closely. All these instruments (except the khañjira) are pitched percussion instruments and are tuned to the tonic of the lead voice. Since the progression through the *tāla* cycles is explicitly shown through hand gestures, the mridangam is

---

<sup>1</sup>Some audio examples illustrating these *tālas* at <http://compmusic.upf.edu/examples-taala-carnatic>



**Figure 2.2:** The structure of *tiśra nađe ādi tāla*, to contrast with the popularly used *caturaśra nađe*. Note the three akṣara beats, and only 24 akṣara cycle, as compared to the 32 akṣara cycle of its *caturaśra nađe* counterpart.

provided with substantial freedom of rhythmic improvisation during the performance. The *tāla* only provides a metrical construct, within which several different rhythmic patterns can be played and improvised.

The solo performed with the percussion accompaniments, called a *tani-āvartana*, demonstrates the wide variety of rhythms that can be played in the particular *tāla*. The solo performance by the percussion ensemble follows the main piece of the concert. The solo is an elaborate rhythmic improvisation within the framework of the *tāla*, but with much improvisation on the percussion patterns. The *tani* strives to present a showcase of the *tāla* with a variety of percussion and rhythmic patterns that can be played in the *tāla*. The percussion instruments duel and complement each other in a solo of each instrument, with all instruments coming together to a cadential end. The patterns played can last longer than one *āvartana*, but stay within the framework of the *tāla*. A *tani-āvartana* is a showcase of the skill and talent of the percussion artists. It is replete with a variety of percussion patterns and hence is very useful for analysis of percussion patterns. The *tani* is often performed with a subset of the percussion instruments. The mridangam is always present, while the other instruments are optional.

Percussion in Carnatic music is organized and transmitted orally with the use of onomatopoeic syllables (called *solkaṭṭus*) representative of the different strokes of the mridangam. An oral recitation of these syllables is itself an art form called *konnakōl*, and is often a part of a *tani-āvartana*. The syllables used belong to mridangam, but is widely used with other percussion instruments used in

ID	Syllable	Description
1	AC	A semi ringing stroke on the right head
2	ACT	AC with TH/TM
3	CH	A ringing stroke on the right head
4	CHT	CH with TH/TM
5	DM	A strong ringing stroke on the right head
6	DH3	A closed stroke on the right head (variant-1)
7	DH3T	DH3 with TH
8	DH3M	DH3 with TM
9	DH4	A closed stroke on the right head (variant-2)
10	DH4T	DH3 with TH/TM
11	DN	A pitched resonant stroke on the right head
12	DNT	DN with TH/TM
13	LF	Long finger stroke on the right head
14	LFT	LF with TH/TM
15	NM	A sharp pitched stroke on the right head
16	NMT	NM with TH/TM
17	TH	A closed bass stroke on the left head
18	TA	A closed sharp stroke on the right head
19	TAT	TA with TH/TM
20	TM	An open bass stroke on the left head
21	TG	Pitch modulated bass stroke on the left head

**Table 2.2:** The syllables (*solkaṭṭus*) used in mridangam grouped based on timbre and the symbol we use for the syllable group in this dissertation. The last column also provides a short description. The strokes are combinations of left+right strokes on the mridangam.

Carnatic music. These syllables vary across schools, but provide a good representation system to define, describe and discover percussion patterns. We explore the use of these syllables for MIR tasks further in the dissertation.

We consulted a senior professional Carnatic percussionist for the complete set of strokes that can be played with the mridangam. The stroke syllables of the mridangam represent the combined timbre of the left and right drum heads, and hence over 45 different strokes can be played on the mridangam. However, many of the timbrally similar strokes can be grouped together into syl-

lable groups, assuming timbral grouping to be sufficient for discovery of timbrally similar percussion patterns. This timbre based grouping enables us to work with the variability in syllables across different schools. The syllable groups, the symbol we use for them in this dissertation, and a short description is shown in Table 2.2. The different stroke names are not indicated in the table since they vary, and we will refer to the syllable groups as syllables in this work. However, we carefully note that the syllables also have a functional role, and such a timbre based grouping is an approximation done only for computational analysis approaches.

### 2.2.3 Rhythm and Percussion in Hindustani music

Clayton (2000) provides a comprehensive introduction to rhythm in Hindustani music. The definition of *tāl* in Hindustani music is similar to the *tāla* in Carnatic music. A *tāl* has fixed-length cycles, each of which is called an *āvart*. An *āvart* is divided into isochronous basic time units called *mātrā*. The *mātrās* of a *tāl* are grouped into sections, sometimes with unequal time-spans, called the *vibhāgs*. *Vibhāgs* are indicated through the hand gestures of a *thālī* (clap) and a *khālī* (wave). The first *mātrā* of an *āvart* (the downbeat) is referred to as *sam*. The first *mātrā* of the cycle (*sam*) is highly significant structurally, with many important melodic and rhythmic events happening at the *sam*. The *sam* also frequently marks the coming together of the rhythmic streams of soloist and accompanist, and the resolution point for rhythmic tension (Clayton, 2000, p. 81).

There are also tempo classes called *lay* in Hindustani music which can vary between *ati-vilaribit* (very slow), *vilaribit* (slow), *madhya* (medium), *dṛ̥t* (fast) to *ati-dhṛ̥t* (very fast). Depending on the *lay*, the *mātrā* may be further subdivided into shorter time-span pulsations. However, since these pulses are not well defined, we consider *mātrā* to be lowest level pulse in the scope of this dissertation.

As with Carnatic music, even in Hindustani music, there are significant differences to the terminology describing meter in Eurogenetic music. The definition of beat pulsation, as foot tapping

Tāl	# vibhāg	# mātrās	mātrā grouping
Tīntāl	4	16	4,4,4,4
Ēktāl	6	12	2,2,2,2,2,2
Jhaptāl	4	10	2,3,2,3
Rūpak tāl	3	7	3,2,2

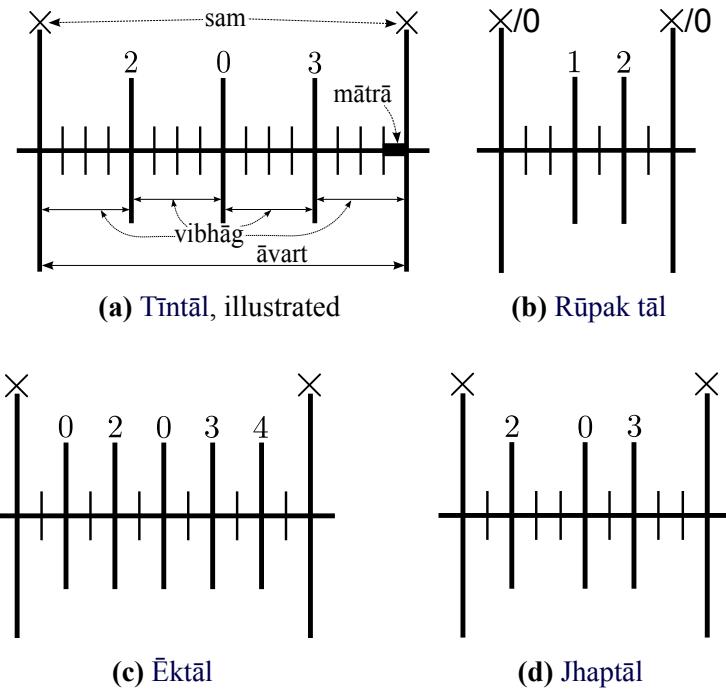
**Table 2.3:** Structure of Hindustani tāls. For each tāl, the number of vibhāgs and the number of mātrās in each āvart is shown. The last column of the table shows the grouping of the mātrās in the āvart into vibhāgs, and the length of each vibhāg, e.g. each avart of rūpak tāl has three vibhāgs consisting of three, two, two mātrās respectively.

instances in time, is also a problem with Hindustani music. Depending on the lay, the mātrā can be defined to be the subdivisions (for dṛt lay) or as beats (for vilambit and madhya lay). To maintain consistency, using accepted conventions, we note that the concepts of mātrā and the āvart of Hindustani music bear analogy to the beat and the bar metrical levels of Eurogenetic music. This implies that there is no well defined subdivision pulsation defined in Hindustani music. The possibly unequal vibhāgs are the sections of the tāl.

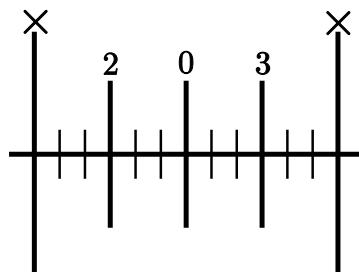
There are over 70 different Hindustani tāl defined, while about 15 tāl are performed in practice. Figure 2.3 shows four popular Hindustani tāls - tīntāl, ēktāl, jhaptāl, and rūpak tāl. The structure of these tāls are also described in Table 2.3. The figure also shows the sam (shown as  $\times$ ), and the vibhāgs, indicated with thālī/khālī pattern using numerals. A khālī is shown with a 0, while the thālī are shown with non-zero numerals. The thālī and khālī pattern of a tāl decides the accents of the tāl. The sam has the strongest accent (with certain exceptions) followed by the thālī instants. The khālī instants have the least accent.

A jhaptāl āvart has 10 mātrās with four unequal vibhāgs (Figure 2.3d), while a tīntāl āvart has 16 mātrās with four equal vibhāgs (Figure 2.3a). We can also note from Figure 2.3b that the sam is a khālī in rūpak tāl, which has 7 mātrās with three unequal vibhāgs.

The special case of ēktāl needs additional mention here. ēktāl has six equal duration vibhāgs and 12 mātrās in a cycle as shown in Figure 2.3c. However, in dṛt lay, an alternative structure emerges,



**Figure 2.3:** An āvart of four popular Hindustani tāls, showing the mātrās (all time ticks), vibhāgs (long and bold time ticks) and the sam (×). Tīntāl is also illustrated using the terminology used in this article.



**Figure 2.4:** An alternative structure of Ēktāl in dṛt lay

which is represented as four equal duration vibhāgs of three mātrās each as shown in Figure 2.4. For consistency, we use the structure as shown in Figure 2.3c in this dissertation.

Hindustani music uses the tabla as the main percussion accom-

ID	Bōls	Symbol	Description
1	D, DA, DAA	DA	
2	N, NA, TAA	NA	
3	DI, DIN, DING	DIN	
4	KA, KAT, KE, KI, KII	KI	
5	GA, GHE, GE, GHI, GI	GE	
6	KDA, KRA, KRI, KRU	KDA	
7	TA, TI, RA	TA	
8	CHAP, TIT	TIT	
9	DHA	DHA	
10	DHE	DHE	
11	DHET	DHET	
12	DHI	DHI	
13	DHIN	DHIN	
14	RE	RE	
15	TE	TE	
16	TII	TII	
17	TIN	TIN	
18	TRA	TRA	

**Table 2.4:** The bōls used in tabla are shown in first column, grouped by similarity of timbre. The symbol we use for the syllable group in the dissertation is shown in the second column and a short description of the timbre is shown in the third column. **Table to be completed.**

paniment. It consists of two drums: a left hand bass drum called the *bāyān* or *diggā* and a right hand drum called the *dāyān* that can produce a variety of pitched sounds. Similar to mridangam, the tabla repertoire is transmitted using onomatopoeic oral mnemonic syllables called the *bōl*.

Similar to leady melody in Hindustani music, tabla has different stylistic schools called *gharānās*. The repertoires of major *gharānās* of tabla differ in aspects such as the use of specific *bōls*, the dynamics of strokes, ornamentation and rhythmical phrases (Beronja, 2008, p. 60). But there are also many similarities due to the fact that the same forms and standard phrases reappear across these repertoires (Gottlieb, 1993, p. 52).

×					<b>2</b>			
×	2	3	4	DHA	5	6	7	8
DHA	DHIN	DHIN	DHA	DHA	DHIN	DHIN	DHA	DHA
<b>0</b>				<b>3</b>				
9	10	11	12	NA	13	14	15	16
DHA	TIN	TIN	NA	NA	DHIN	DHIN	DHA	DHA
<b>(a) Tintāl</b>								
×			<b>0</b>			<b>2</b>		
×	2	DHA	3	GE	4	TIRAKITA	5	6
DHIN	DHIN	DHA	GE	TIRAKITA	TUN	NA		
<b>0</b>			<b>3</b>			<b>4</b>		
7	8	9	DHA	GE	10	TIRAKITA	11	12
KAT	TA	DHA	GE	TIRAKITA	DHIN	NA		
<b>(b) Ēktāl</b>								
×		<b>2</b>		<b>0</b>		<b>3</b>		
×	2	DHI	3	DHI	4	NA	6	7
DHI	NA	DHI	DHI	NA	TI	NA	DHI	DHI
<b>(c) Jhaptāl</b>				8	9	10		
×/0				<b>1</b>		<b>2</b>		
×	2	3		4	5	6	7	
TIN	TIN	NA		DHI	NA	DHI	NA	
<b>(d) Rūpak tāl</b>								

**Table 2.5:** The *thēkās* for four popular Hindustani *tāls*, showing the *bōl* for each *mātrā*. The *sam* is shown with  $\times$  and *vibhāgs* boundaries are separated with a vertical line. Each *mātrā* of a cycle has equal duration.

The *bōls* of the tabla vary marginally within and across *gharānās*, several *bōls* can represent the same stroke on the tabla. To address this issue, we grouped the full set of 38 syllables into timbrally similar groups resulting into a reduced set of 18 syllable groups as shown in Table 2.4. Though each syllable on its own has a func-

tional role, this timbral grouping is presumed to be sufficient for discovery of percussion patterns. For the remainder of the dissertation, we limit ourselves to the reduced set of syllable groups and use them to represent patterns. For convenience, when it is clear from the context, we call the syllable groups as just syllables and denote them by the symbols in Table 2.4. A brief description of the timbre is also provided for each syllable.

Tabla acts as the timekeeper during the performance indicating the progression through the *tāl* cycles using pre-defined rhythmic patterns (called the *ṭhēkā*) for each *tāl*. The lead musician improvises over these cycles, with limited rhythmic improvisation during the main piece. The *ṭhēkās* are specific canonical tabla *bōl* patterns defined for each *tāl* as illustrated in Table 2.5. However, the musician playing tabla improvises these patterns playing many variations with filler strokes and short patterns.

To showcase the nuances of the *tāl* (the rhythmic framework of Hindustani music) as well as the skill of the percussionist with the tabla, Hindustani music performances feature tabla solos. A tabla solo is intricate and elaborate, with a variety of pre-composed forms used for developing further elaborations. There are specific principles that govern these elaborations (Gottlieb, 1993, p. 42). Musical forms of tabla such as the *ṭhēkā*, *kāyadā*, *palaṭā*, *rēlā*, *pēskār* and *gat* are a part of the solo performance and have different functional and aesthetic roles in a solo performance. A percussion solo shows a variety of improvisation possible in the framework of the *tāl*, with the role of timekeeping taken up by the lead musician during the solo. Miron (2011), Clayton (2000), A. E. Dutta (1995), Beronja (2008), and Naimpalli (2005) provide a more detailed discussion of *tāl* in Hindustani music including *ṭhēkās* for commonly used *tāls*<sup>2</sup>.

In Hindustani music, the tempo is measured in *mātrās* per minute (MPM). The music has a wide range of tempo, divided into tempo classes called *lay* as described before. The mainly performed ones are the slow (*vilāmbit*), medium (*madhya*), and fast (*dṛ̥t*) classes. The boundary between these tempo classes is not well defined with possible overlaps. In this dissertation, after consultation with a professional Hindustani musician, we use the commonly agreed tempo

---

<sup>2</sup>Some audio examples at <http://compmusic.upf.edu/examples-taal-hindustani>

ranges for these classes: *vilaribit lay* for a median tempo between 10-60 MPM, *madhya lay* for 60-150 MPM, and *dṛ̥t lay* for >150 MPM. This large range of allowed tempi means that a *tāl* cycle in Hindustani music lasts from less than 2 second to over a minute long. A *mātrā* in *vilaribit lay* can last about 6 second, and to maintain a continuous rhythmic pulse, several filler strokes are played on the tabla. Hence the surface rhythm apparent from audio recordings can be quite different from the underlying metrical structure.

#### 2.2.4 Carnatic and Hindustani music: A comparison

We compare and contrast some of the rhythm related concepts in Carnatic and Hindustani music, so that it can be used for better comparison of MIR approaches for these musics.

Both the music traditions are oral traditions, with a lot of scope for improvisation. Even a fixed composition is interpreted with significant freedom by the musicians, as long as they adhere to the framework of the *rāga* and *tāla*. The concept of cyclical metrical structures is shared by both music cultures, while the components of the *tāla* are less similar. The first pulse of the cycle is important in both cultures and has significant melodic and rhythmic events. The sections of the *tāla* cycle need not be equal in duration. The *tāla* does not change over single piece, but since Hindustani music recordings are distributed as full concerts, there is a possible change of piece in the middle of the recording, with a change of *tāl* and/or *lay*. Neither of the music cultures use a metronome during performance, which means that the responsibility of maintaining a regular pulse rests with the musicians. This leads to a flexible time varying nature of tempo, with most often the tempo increasing (and the piece getting “faster”) with time. The range of tempo in Hindustani music is enormous (from 10 MPM to over 350 MPM), while Carnatic music is performed in a smaller range of tempo. This has the implication that while *tāl* cycles can be quite long in Hindustani music, while Carnatic music *tāla* cycles are shorter (often shorter than 15 second).

In Hindustani and Carnatic music, the percussion accompaniments tabla and mridangam are tuned to the tonic of the lead musi-

cian. Both these instruments are capable of producing a rich variety of timbres. The playing style depends on the composition or the lead melody being rendered, and both are improvised during performance. Both have specific *ṭhēkās* for the exposition of the *tāla*, though *ṭhēkās* are a little more flexibly defined in Carnatic Music. Both tabla and mridangam have their own set of onomatopoeic mnemonic syllables that provide a language for percussion, which even has to the art form of reciting these syllables in a performance. Representing percussion patterns with these syllables is musically well-defined and an accurate representation of those patterns.

The surface rhythm in both the music cultures provide cues to the underlying *tāla* structures. In Hindustani music, tabla is a very important cue to the underlying *tāl* progression. All *tāls* have a definite accent and tabla stroke pattern defined by the *ṭhēkā* which is mostly followed except in improvisatory passages. The surface rhythm consists of these accents and specific strokes, but is also replete with other strokes, fillers and expressive additions. Filler strokes are employed in slow pieces with long cycles. In Carnatic music, as discussed earlier, the progression through the *tāla* is shown through visual gestures and hence there is no need for definitive cues in the surface rhythm. However, the percussion phrases played on the mridangam, the melodic phrases and the lyrics of the composition provide cues to the underlying *tāla*. Unlike tabla strokes, mridangam strokes are less indicative of the current position in the cycle of a *tāla*.

Unmetered forms of music exist in both the music cultures. The most important unmetered form in Hindustani music is the *ālāp* and in Carnatic music is the *ālāpana*, both of which are melodic improvisational forms based on a *rāga*. An understanding of the rhythmic behavior of unmetered forms is far from trivial for musicologists and even practicing musicians (Clayton, 1996). Widdess (1994) presented an interesting discussion of the notion of pulsation in *ālāps* and a disagreement about it among performers. For this reason, we believe that rhythmic analysis of unmetered forms should be reserved for a study more from a musical perspective and hence we do not consider it in this dissertation.

### 2.2.5 Percussion in Beijing Opera

The main focus of this dissertation is Indian art music, however, within the context of CompMusic, there are other music cultures that share similar music concepts and hence are suitable candidates for test and extend our approaches to those musics. Beijing Opera is one such music culture that shares the concept of a syllabic percussion system, similar to Indian art music. However, the syllabic percussion system in Beijing Opera is simpler and more well defined than for Indian art music, and hence is a test case to validate our approaches to percussion pattern transcription and discovery. A basic introduction to percussion in Beijing Opera is provided, since some of our approaches to with percussion pattern analysis are first proposed on it and then extended to Indian art music.

Beijing opera (Jīngjù, 京剧), also called Peking Opera, is one of the most representative genres of Chinese traditional performing arts, integrating theatrical acting with singing and instrumental accompaniment. It is an active art form and exists in the current social and cultural contexts, with a large audience and significant musicological literature. One of the main characteristics of Beijing opera aesthetics is the remarkable rhythmicity that governs the acting overall. From the stylized recitations to the performers' movements on stage and the sequence of scenes, every element presented is integrated into an overall rhythmic flow. The main element that keeps this rhythmicity is the percussion ensemble, and the main means to fulfill this task is a set of predefined and labeled percussion patterns.

The percussion ensemble in *jīngjù* establishes and maintains the rhythm in a performance and guides the progression of sections in an aria. Firstly, the percussion provides a base to indicate the rhythmic modes, called the *banshi*, and accompanies the singing voice. Secondly, the percussion ensemble plays different kinds of pre-defined, fixed, labeled patterns that create a context for different parts of the aria. They signal important structural points in the play. A performance starts and ends with percussion patterns, they generally introduce and conclude arias, and mark transition points within them. They accompany the actors' movements on stage and set the mood of the play, the scene, the aria or a section of the aria.

Syllables	Instruments	Symbol
bā (巴, 八), běn (本), dā (答), dà (大), dōng (冬, 咚), duō (哆), lóng (龙), yī (衣)	bǎngǔ	DA
lái (来), tái (台), lìng (另)	xiǎoluó	TAI
qī (七), pū (扑)	náobó	QI
qiē (切)	náobó+xiǎoluó	QIE
cāng (仓), kuāng (匡), kōng (空)	dàluó + <náobó> + <xiǎoluó>	CANG

**Table 2.6:** Syllables used in Beijing opera percussion and their grouping used in this dissertation. Column 2 shows the instrument combination used to produce the syllable, with the instrument shown between <> being optional. Column 3 shows the symbol used for the syllable group in this dissertation.

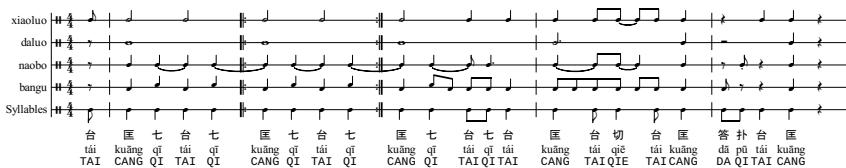
The percussion patterns in *jīngjù* music can be defined as sequences of strokes played by different combinations of the percussion instruments, and the resulting variety of timbres are transmitted using oral syllables as mnemonics. The percussion ensemble is formed mainly by five instruments played by four musicians. The *ban* (a wooden clapper) and the *danpigu* (a wooden drum struck by two wooden sticks) are played by one single performer, and are therefore known by a conjoint name, *bǎngǔ* (clapper-drum). The other three instruments are the *xiǎoluó* (small gong), the *dàluó* (big gong) and the *náobó* (cymbals) (Lee & Shen, 1999; Wichmann, 1991).

*Bangu* has a high pitched drum-like sound while the rest of three instruments are metallophones with distinct timbres<sup>3</sup>. Each of the different sounds that these instruments can produce individually, either through different playing techniques or through different dynamics, as well as the sounds that are produced by a combination of different instruments have an associated syllable that represent them (Mu(穆文义), 2007). In *jīngjù*, several syllables can be mapped to a single timbre. This many-syllable to one-timbre mapping is useful to reduce the syllable space for computational

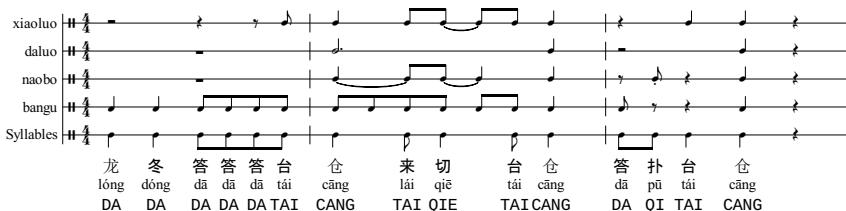
<sup>3</sup>A few annotated audio examples of these instruments can be found at <http://compmusic.upf.edu/examples-percussion-bo>



(a) dǎobǎn tóu 【导板头】



(b) mǎn chángchuí 【漫长锤】



(c) duótóu 【夺头】

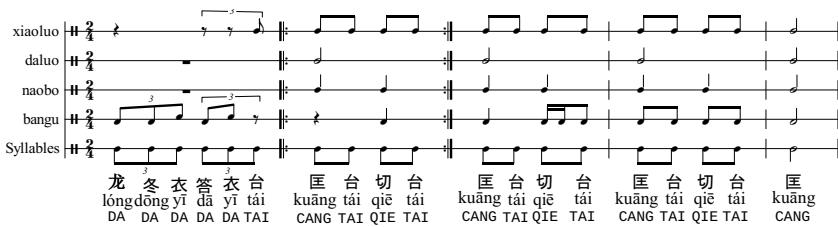
**Figure 2.5:** Scores for percussion patterns in *jīngjù*, showing the instruments and a syllabic representation of the pattern using the unmapped syllables, and the mapped syllable groups used in this dissertation.

analysis of percussion patterns.

We first mapped each syllable to one or several of the instrument categories considered for analysis, as explained by (Tian, Srivastava, & Sandler, 2014), without considering differences in playing technique or dynamics. Based on inputs from expert musicologists, we then grouped the syllables with similar timbres into five syllable groups - DA, TAI, QI, QIE, and CANG, as shown in Table 2.6. Every individual stroke of the *bǎngǔ*, both drum and clappers, have been grouped as DA. In the rest of the syl-



#### (d) xiǎoluó duótóu 【小锣夺头】



#### (e) shǎnchuí 【闪锤】

**Figure 2.5:** Scores for percussion patterns in *jīngjù*, showing the instruments and a syllabic representation of the pattern using the unmapped syllables, and the mapped syllable groups used in this dissertation (contd...)

lable groups, the *bǎngǔ* can be played simultaneously or not. The single strokes of the *xiǎoluó* and the *náobó* are called **TAI** and **QI** respectively, and the combined stroke of these two instruments together is the syllable **QIE**. Finally, any stroke of the *dàluó* or any combination that includes *dàluó* has been notated as **CANG**. This mapping to a reduced set of syllable groups is only for the purpose of computational analysis. For the remainder of the dissertation, we limit ourselves to the reduced set of syllable groups and use them to represent the patterns. For convenience, when it is clear from the context, we call the syllable groups as just syllables, and denote them by the common symbol in column 3 of Table 2.6. Hence, in the current task, there are five syllable groups.

Each percussion pattern is a sequence of syllables in their pre-established order, along with their specific rhythmic structure and dynamic features. A particular feature of the oral syllabic system for Beijing opera percussion that makes it especially interesting is

that the syllables that form a pattern refer to the ensemble as a whole, and not to particular instruments. Each particular pattern thus has a single unique syllabic representation shared by all the performers.

In practice, there is a library of limited set of named patterns (called luógǔ jīng, 锣鼓经) that are played in a performance, with each of these having a specific role in the arias. Although a definite agreed number for the total number of these patterns is lacking, some estimations, like in (Mu(穆文义), 2007), suggest the existence of around ninety of them. Figure 2.5 shows the scores for five predominantly used percussion patterns in jīngjù - dǎobǎn tóu, mǎn chángchuí, duótóu, xiǎoluó duótóu, and shǎnchuí<sup>4</sup>. The figure also shows how a possible transcription in staff notation, adapted from the scores provided by Mu(穆文义) (2007), can be simplified in a single line by the oral syllabic system. Hence, the use of these oral syllabic sequences simplify and unify the representation of these patterns played by an ensemble.

Since one of the main functions of the patterns is to accompany the movements of actors on stage, the overall length and the relative duration of each stroke can vary notably, which makes it difficult to set a stable pulse or a definite meter. The time signature and the measure bars used in Figure 2.5, as suggested by Mu(穆文义) (2007), are only indicative and fail to convey the rhythmic flexibility of the pattern. Furthermore, many patterns (such as shǎnchuí shown in Figure 2.5e) accompany scenic movements of undefined duration. In these cases, certain syllable sub-sequences in the pattern are repeated indefinitely, e.g. the sub-sequence cāng-tái-qiē-tái in the pattern shǎnchuí can be repeated indefinitely until the scene complete.

From this brief introduction, it can be seen that there are similarities between the percussion systems in jīngjù and Indian art music. Jīngjù can be used a test case for approaches to percussion pattern transcription and discovery in syllabic percussion systems.

---

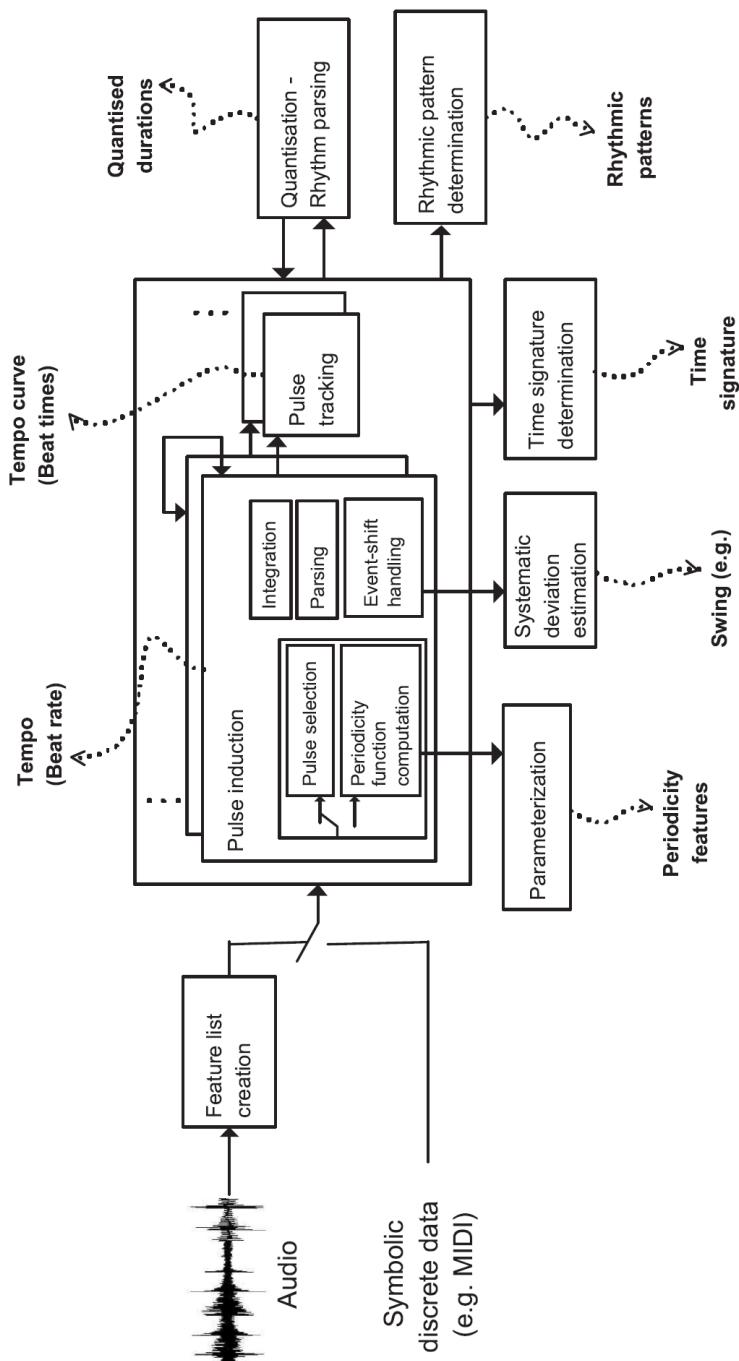
<sup>4</sup>These pattern scores are also listed at <http://compmusic.upf.edu/b0-perc-patterns>

## 2.3 Automatic rhythm analysis: A review

Automatic rhythm analysis has been an important research area within MIR, with over a decade of research on several relevant rhythm and percussion related problems. A review of the state of the art in the relevant rhythm research problems is presented to provide a basis for further work proposed in the dissertation. The review of the previous works in this section applies in general to relevant rhythm problems and not specific to Indian art music. A more detailed review of approaches specific to Indian art music, and an evaluation of the state of the art on Indian art music is discussed in Chapter 3. Several researchers have suggested a decomposition of rhythm description into complementary modules, each considering a specific task, and possibly using information from other modules (Guoyon, 2005; Gouyon & Dixon, 2005). An example of such a rhythm description system is shown in Figure 2.6. Starting from audio and/or music scores, the system shows several rhythm analysis modules that give out important rhythm analysis outputs such as tempo, beats, swing, time signature, and rhythmic patterns. Though such a rhythm description for Indian art music would involve significant changes, this system nevertheless provides a suitable basic framework to start formulating the problems.

### 2.3.1 Onset detection

Musical note/stroke onset detection is the most fundamental pre-processing task for most rhythm analysis problems. Within the task of onset detection, we can include the task of extracting features from audio that are indicative of onsets, and the approaches to obtain the onsets from those features. A musical note/stroke onset is defined as the single instant that marks a detectable start of a extended transient of the note/stroke, when the music audio signal evolves quickly in a non-trivial manner over a short time (Bello et al., 2005). In simpler words, onsets mark the start of a melodic note or a percussion stroke. Onsets mark important musical events in time and the automatic detection of onset events is an essential part of many music signal analysis algorithms and has various ap-



**Figure 2.6:** Functional units of a rhythm description system as described by Guoyon (2005, figure reproduced with permission)

plications in identification, retrieval, musicological analysis, audio editing and coding, content-based processing and many other applications.

A detailed tutorial on onset detection methods is provided by Bello et al. (2005). Onset detection needs transient detection in audio signals. The transients can be measured either in amplitude, energy, phase, frequency, and several other signal parameters. Most approaches to onset detection involve a signal pre-processing step, followed by a signal reduction (extracting features that are indicative of transients) into an onset detection function, and a peak-picking step that estimates the onset times as the peaks on the onset detection function.

The pre-processing step is optional and aims to enhance relevant parts of the signal. Signal reduction often involves framewise short time Fourier transform based analysis of signals, often in multiple frequency bands using filter banks to capture frequency information from different instruments in the audio signal. The result of such a feature extraction is an onset detection function, also sometimes called a novelty function. The peaks of the onset detection function are then the onsets.

There are several methods to compute the onset detection function, based on signal features and probabilistic models. The signal features include time domain features such as amplitude envelope, that works well for percussive onsets. More popular features are spectral features that measure some form change in spectral amplitudes and energy. The spectral flux feature is the most often used one, which measures a positive change (for onsets, negative change would indicate offsets) of spectral energy across frames of audio. The spectral flux can be computed both with the magnitude of STFT or the complex STFT, across adjacent frames or across a local set of frames. There are several ways a spectral flux can be computed, described by Bello et al. (2005), and further improved by many researchers, e.g. as LogFilt-SpecFlux by Böck, Krebs, and Schedl (2012).

The definition of an onset could become ambiguous in the case of the instruments having longer transient times without sharp bursts of energy rises. Vos and Rasch (1981) approached this issue by introducing the concept of perceptual onset as the time when the most salient metrical feature of the music signal is perceived relative to

its physical onset. Dixon (2006) examined and proposed improvements to the then state of the art spectral methods. Klapuri (1999) proposed a method utilizing band-wise processing and a psychoacoustic model of intensity coding to detect perceptual onsets.

**To be completed:** For Indian art music, we are primarily interested in onsets of percussion instruments mridangam and tabla, within the heterophonic music signal. We can do HPSS for that.

**Harmonic percussive source separation ? A short description ?**

**References:** Fitzgerald: (Fitzgerald, 2010), Balaji Thoshkhana (Thoshkahna & Ramakrishnan, 2011), Ono(Ono, Miyamoto, Le Roux, Kameoka, & Sagayama, 2008), HPR model: (Serra, 1989)

### 2.3.2 Instrument identification

**Write about instrument identification and tracking from mixtures - NMF and other methods. In brief.**

### 2.3.3 Tempo estimation

Tempo estimation refers to estimating the period of the predominant pulse in the music recording, at the correct metrical level of the beat. The definition of such a tempo is not clear and there can be disagreement on the correct metrical level. Further, in pieces where tempo can change over time, it is necessary to estimate a time varying tempo curve instead of single tempo estimate for a music piece. Despite the metrical ambiguity, tempo estimation is a useful task for further analysis in beat and meter tracking.

Tempo estimation algorithms use some form of periodicity estimation using mid-level features extracted from audio, mostly the onset detection functions. An autocorrelation of such a novelty function is a basic measure of periodicity. Following the onset detection functions, there are distinctly two different approaches that have been used for tempo induction. Some methods, such as the system proposed by Dixon (2007) are pulse selection methods that measure the inter onset intervals (IOI) and use them to estimate the tempo. An IOI histogram has peaks at the periodicity of the beat period, which can then be measured. However, there exists significant metrical ambiguity in such approaches since the IOI histograms are often multimodal. The other approaches, such as the

ones by Klapuri, Eronen, and Astola (2006); Davies and Plumley (2007); Ellis (2007) derive a periodicity function from the detection function, which provides an estimate of tempo.

Several mid-level features have been used to estimate a time varying tempo curve: a few examples of such features include novelty functions used for structural segmentation (Foote, 2000), Tempogram (Grosche & Müller, 2011b) and Predominant Local Pulse (Grosche & Müller, 2011a).

### 2.3.4 Beat tracking

In the context of MIR, beat tracking is commonly defined as determining the time instances where a human listener is likely to tap his/her foot to the music. Several approaches have been proposed for beat tracking on musical audio in a wide variety of genres. Conventional beat tracking algorithms generally use three main sub components - feature extraction, tempo induction, and beat induction. The rhythm features extracted are typically based on onsets and onset detection functions. A good overview of several beat tracking algorithms is provided by Holzapfel, Davies, Zapata, Oliveira, and Gouyon (2012).

Dixon (2007) uses a multiple agent architecture using a collection of tempo hypotheses, which are all tested for continuity to obtain the set of beat locations. Ellis (2007) developed a beat tracking algorithm based on dynamic programming, which computes a global set of optimal beat candidates, given an accent signal and a tempo estimate. The algorithm pursues a tradeoff between the temporal continuity of beats and the salience of the detection function using the dynamic programming approach. The main drawback of the algorithm is the assumption of a constant tempo, which causes problems for music with varying tempo. Further, errors in tempo estimation translate to an incorrect estimation of the beats. Wu et al. (2011) also proposed a similar dynamic programming approach to beat tracking, but with extensions to handle a time-varying tempo. Davies and Plumley (2007) proposed a context dependent beat tracking algorithm which handles varying tempo, by providing a two state model in which the first state tracks the tempo changes and provides continuity, while the second state tracks the beat pulses maintaining contextual continuity, assuming a constant tempo.

The algorithm proposed by Klapuri et al. (2006) estimates the musical meter jointly at three metrical levels of bar, beat and subdivision, which are referred to as measure, tactus and tatum, respectively. A time frequency analysis computes accent signals in four frequency bands, which are aimed at emphasizing changes due to note onsets in the signal. A bank of comb filter resonators is applied for periodicity analysis to each of the four accent signals. The periodicities thus found are processed by a probabilistic model that incorporates musical knowledge to perform a joint estimation of the tatum, tactus, and measure pulsations.

**Todo: Write about (Peeters & Papadopoulos, 2011)**

Böck and Schedl (2011) proposed a data driven approach to beat tracking using context-aware neural networks. A recurrent neural network with Long Short-Term Memory cells, called an LSTM network (Hochreiter & Schmidhuber, 1997), can learn contextual information and can classify and predict time series when there are long time lags of unknown size between important events. Mel-spectrogram based spectral features and their relative differences were used to train a bidirectional LSTM network to perform a frame by frame beat classification of a signal. The network outputs a beat activation function directly using the input signal and an autocorrelation function was then used to determine the predominant tempo to eliminate the erroneously detected beats and complement the missing beats.

Ensemble approaches have also been proposed for beat tracking, which uses mutual agreement between several beat trackers to improve beat tracking performance (Holzapfel et al., 2012). The approach is useful to identify pieces that are difficult for beat tracking, and also to create a dataset of such difficult pieces.

Beat tracking is an important MIR task and has been a part of Music Information Retrieval EXchange (MIREX) challenges. There are also several datasets that have been used for evaluating beat tracking algorithms, such as the SMC dataset citeXX, Ballroom dataset citeXX, Gainzworth dataset citeXX, and more recently, the GTZAN dataset with beat downbeat annotations citeXX: peeters2015ismir. more datasets. There is however no standard comprehensive dataset for testing beat tracking performance, and we use Ballroom dataset for baselining our approaches in this dissertation.

Despite a significant effort, beat tracking algorithms still need to be significantly improved for use in practical systems. They suffer from metrical level ambiguities and poor generalizability to other musical genres. The beats are assumed to be isochronous, which is another limitation of the beat tracking algorithms so far. However, several improvements have been suggested to improve the performance. J. Zapata and Gómez (2013) explore the use of voice suppression to improve beat tracking performance. A mutual agreement of several beat trackers can also be used to assign a confidence level to the beat tracking performance and identify samples difficult for beat tracking Holzapfel et al. (2012); J. R. Zapata, Holzapfel, Davies, Oliveira, and Gouyon (2012).

### 2.3.5 Time signature estimation

Automatic rhythm annotation problems apart from onset detection, beat and tempo tracking have been less explored by the MIR community. Gainza (2009) use beat tracking to perform musical meter detection for western music using a beat similarity matrix based approach and Foote and Uchihashi (2001) suggested a new beat spectrum for rhythm analysis. Uhle and Herre (2003) extend the tempo tracking framework for time signature and micro-time estimation on percussive music.

In the method proposed by Pikrakis, Antonopoulos, and Theodoridis (2004), a time signature is estimated from a self-distance matrix computed from Mel-frequency cepstral coefficients (MFCC) of the audio signal. To this end, minima in the distance matrix are assumed to be caused by repetitions related to the metrical structure of the piece. Hence, this algorithm does not track pulsations in a piece, but relies on existence of patterns caused by general repetitions in the MFCC features. Because MFCC features capture timbral characteristics, it can be stated that similarities in local timbre are used by the algorithm. The algorithm was tested on East-European music styles, including Greek traditional dance music.

### 2.3.6 Downbeat tracking

The methods described in this section were developed for the identification of downbeats within sequences of beats. So far mainly

music with a 4/4 time signature was focused upon in evaluations, usually in the form of collections of Eurogenetic popular and/or classical music.

The approach presented by Davies and Plumbley (2006) is based on the assumption that percussive events and harmonic changes tend to be correlated with the downbeat position. Therefore, they partition an audio signal into beat segments and compute an STFT of each segment, neglecting frequencies above 1.4 kHz. Then the magnitude differences between all neighboring blocks are computed. Subsequently, for a given bar length in beats, the sequence of bar length distant segments that is related to the maximum spectral change is chosen as downbeats.

Hockman, Davies, and Fujinaga (2012) presented an algorithm for detecting downbeats in music signals, specifically at hardcore, jungle, and drum and bass genres of music. Their approach combines information from low level onset event information, periodicity information from beat tracking, and high-level information from a regression model trained with classic breakbeats. The approach is an extension of a downbeat detection system proposed by Jehan (2005) that applies support vector regression. The features of the regression consist of Mel-frequency spectral coefficients, loudness descriptors, and chroma features, all computed for the separate beat segments. The extension proposed by Hockman et al. comprises a post-processing of the regression, a combination with a low-frequency onset detection, and a beat-time weighting. While the post-processing compensates for spurious downbeat detections, the combination of the regression with a low-frequency onset feature is motivated by the fact that strong bass drums tend to be located at the downbeat for the form of music they considered.

Downbeat estimation has been addressed as a part of beat tracking (Klapuri et al., 2006; Peeters & Papadopoulos, 2011) resulting in a joint estimation of beats and downbeats. In both cases, a probabilistic framework is used to estimate the downbeats from the beats.

**To be completed: Write about downbeat tracking with neural networks. Work of Bock, Simon and Juan Bello.**

### 2.3.7 Meter tracking and inference

Most of the approaches presented so far considered the task of beat tracking and downbeat tracking as separate tasks. The task of estimating the tempo, beats and the downbeats is what we refer to as meter tracking. Recent approaches in meter tracking have successfully applied Bayesian models that jointly estimate beat and downbeats together, using rhythmic patterns generated learned from onset detection function as the features (Krebs, Böck, & Widmer, 2013; Böck, Krebs, & Widmer, 2014; Krebs, Holzapfel, Cemgil, & Widmer, 2015). Some early approaches have explored particle filtering and approximate inference for beat tracking task (Hainsworth & Macleod, 2003). **Elaborate more here on recent Bayesian methods and those using deep learning for the task.**

### 2.3.8 Evaluation measures

There are several measures that have been proposed for measuring the accuracy of performance of beat and downbeat trackers (Davies, Degara, & Plumbley, 2009). Starting with an annotated dataset with beat marked audio, these measures consider the accuracy of beat locations estimated, continuity of beats, and the metrical level at which the beats were tracked. There have also been information theoretic measures proposed based on the entropy of beat tracking errors, which is measure of the extent of correlation between the annotations and the estimated beat locations. McKinney, Moelants, Davies, and Klapuri (2007) present a survey of the performance of several beat tracking algorithms using multiple accuracy measures<sup>5</sup>. For our evaluation, we will use the f-measure, Information Gain, CMLt, and AMLt measures, that are characterized by a set of diverse properties and are often used in beat tracking evaluations in MIREX<sup>6</sup>. The measures are now defined for beat tracking, but extend to downbeat tracking as well, with the same tolerances.

---

<sup>5</sup>An implementation of the evaluation measures is available at <http://code.soundsoftware.ac.uk/projects/beat-evaluation/>

<sup>6</sup>e.g. MIREX 2012, [http://www.musicir.org/mirex/wiki/2012:Audio\\_Beat\\_Tracking](http://www.musicir.org/mirex/wiki/2012:Audio_Beat_Tracking)

For a music piece, given the ground truth beat times and the estimated beat sequence, a beat is marked correctly detected if it lies inside a tolerance window around a ground truth annotation. The f-measure (denotes as  $f$  in this dissertation) is a number between 0 and 1 computed as the harmonic mean of the popular information retrieval performance metrics - *precision* and *recall*. Precision ( $\prec$ ) is the ratio between the number of correctly detected beats and all detected beats, while recall ( $\tau$ ) is the ratio between the number of correctly detected beats and the total annotated beats. The f-measure can take a maximum value of 1, while beats tapped on the off-beat relative to annotations will be assigned an f-measure of 0. Estimated beats with time-spans either half or double the annotated time-span are penalized with a value of 0.667.

The  $CML_t$  measure (Correct Metrical Levels, total) is a number between 0 and 1, is the ratio between the number of correctly estimated beats divided by the number of annotated beats. It takes the value of 1 only for sequences that coincide with the annotations. It does not penalize discontinuities in beat tracking as the  $CML_c$  (Correct Metrical Levels, continuity required) measure, but penalizes any beats tracked at half or double time-spans of the annotated metrical level.  $AML_t$  (Allowed Metrical Levels with no continuity required) is also a number between 0 and 1, where beat sequences are considered as correct if the beats occur on the off-beat, or are double or half of the annotated tempo, allowing for metrical ambiguities. The value of this measure is then the ratio between the number of correctly estimated beats divided by the number of annotated beats. Similar to f-measure, small mis-alignments in the estimated beats are allowed for by applying tolerance windows before computing the  $CML_t$  and  $AML_t$  measures.

Information Gain ( $I$ ) aims at determining if there exists any kind of relation between the estimated beats and the annotations, and indicates how much information the beats provide about the annotations. It uses the entropy of the beat error distribution and can be interpreted as an information theoretic measure. This measure is a numerical score that takes a value of 0 bits only for completely unrelated sequences and by using the default setting of 40 bins in the beat error histogram, a maximum value of 5.3 bits for highly related beat sequences. Timing errors are calculated between an annotation and all beat estimations within a one-beat length win-

dow around the annotation. Then, a beat error histogram is formed from the resulting timing error sequence. A numerical score is derived by measuring the K-L divergence between the observed error histogram and the uniform distribution.

### 2.3.9 Rhythm similarity measures

Defining and extracting music similarity is one of the primary areas of MIR. An important component of defining overall similarity between two music pieces is rhythmic similarity. Similarity measures to compare rhythms have been explored both with audio and symbolic scores. These rhythmic similarity measures are quite useful in computational musicology to compare rhythms.

Rhythm similarity measures have been used to classify and compare rhythms, trace ancestry of rhythms using phylogenetic analyses, to match prototypical rhythm patterns to their micro-variations. Toussaint (2004) discusses several measures and compares them based on how much insight they provide about the inter-relationships that exist among families of rhythm.

One approach to compare rhythmic content of music is by using onset patterns (OP), as initially presented by Pohle, Schnitzer, Schedl, and Knees (2009). Starting from a magnitude spectrum obtained from the Short-Time Fourier Transform (STFT) of a monophonic piece of music, a set of energy coefficients are computed in 32 logarithmically spaced frequency bands. A band-wise accent signal is then derived by applying a moving average filter and half wave rectification to each of the 32 bands. A second STFT operating on longer time scale (8 second window with 1 second hop) is applied to each band-wise accent signal. This way, a description of periodicities referred to as OP features (Holzapfel, Flexer, & Widmer, 2011) is obtained for 5 bands per octave, and 5 periodicity octaves from 30 bpm to 960 bpm. The rhythm of a whole sample is described by the mean of the OP obtained from the various segments of this sample. Pohle et al. (2009) showed that combining rhythmic descriptors with a timbral component improved the performance of the task of rhythm similarity computation on the “Ballroom Dancers” collection.

Holzapfel and Stylianou (2011) use the scale transform to compute rhythm descriptors to classify Greek traditional dances and

Turkish traditional songs. The first step is a computation of an accent signal. To this end, the sum of the 32 band-wise accent signals used for the OP features are applied to obtain a single vector describing the note onset characteristics. Then, within the moving windows of eight seconds length, autocorrelation coefficients are computed from this accent signal and then transformed into the scale domain by applying a discrete Scale Transform. For one piece, the mean of the Scale Transform Magnitudes (STM) obtained from all the analysis windows are the STM descriptors of the rhythmic content of the piece. Both the mapping onto a logarithmic axis of the magnitudes in the second STFT in the OP features, and the application of a Scale transform in the STM features provide varying degrees of robustness to tempo changes. Holzapfel and Stylianou (2011) provide more details and the exact computation of parameters of the two descriptors. The scale transform is also shown to capture relevant properties of *usuls* (metrical framework in Turkish makam music) and has been used for classifying symbolic traditional Turkish music scores to their *usuls* (Holzapfel & Stylianou, 2009). Holzapfel et al. (2011) discuss improved descriptors for rhythm similarity.

Fouloulis, Papadelis, Pastiadis, and Papanikolaou (2010) present a system containing two artificial neural networks in cascade - a self-organizing neural network (called SARDNET) and a Multi-Layer Perceptron - that receives a sequence of temporal intervals (performed rhythm pattern) as input and maps it into a given set of prototypical rhythm patterns showing strong evidence that this type of network architecture may be successful to compute similarity between a prototypical rhythm pattern and its micro-variations. Parry and Essa (2003) proposed a similarity metric based on rhythmic elaboration that matches rhythms that share the same beats regardless of tempo or identicalness. Rhythmic elaborations can help an application decide where to transition between songs.

### 2.3.10 Domain-specific approaches

Including domain specific music knowledge to build culture specific algorithms is an important focus of the dissertation. From the extracted low level audio features, we can use domain specific prior knowledge to derive mid-level representations. A few examples of

such mid-level representations include novelty functions used for structural segmentation (Foote, 2000), Tempogram and Predominant Local Pulse (Grosche & Müller, 2011b). **repeated in tempo estimation also, to be improved.** Though these functions are not generally built using domain specific parameters, we can easily extend them to incorporate priors based on the music culture, e.g. the kernel size in Novelty computation.

There are several machine learning algorithms that can include domain specific priors into their modeling parameters. Most probabilistic graphical models allow for including some form of priors and encode complex relationships. Simple examples of these include the kinds of priors and relationships that can be encoded using a hidden Markov model. Hidden semi-Markov models allow us to encode explicit timing information into the algorithm, which might be very useful for tracking rhythmic events, as explored with some promise for Chord recognition by Chen, Shen, Srinivasamurthy, and Chordia (2012). Dynamic Bayesian network (DBN) (Murphy, 2002) based models have been successfully applied for beat and downbeat tracking, and hold significant promise. Context-aware Neural Networks, as discussed in Section 2.3.4, might also be useful to bring the modeling capabilities of Neural networks to modeling structured data such as music.

### 2.3.11 Percussion pattern transcription and discovery

Music transcription addresses the analysis of an acoustic musical signal so as to write down the pitch, onset time, duration, and source of each sound that occurs within it (Klapuri & Davy, 2006). Percussion transcription focuses on percussion and aims to transcribe an audio recording, typically a percussion solo, into a sequence of symbolic drum stroke indicators. Though promising results have been achieved in percussion transcription (Gillet & Richard, 2004b; Paulus & Virtanen, 2005; Fitzgerald & Paulus, 2006), state of the art music transcription systems are still clearly inferior to skilled human annotation in their accuracy.

Most works on music transcription have focused on melodies of pitched instruments. However, recent years have witnessed a grow-

ing interest for transcribing non-pitched percussive instruments. The percussion instruments investigated in music transcription and onset detection tasks fall into two main types: membranophones, such as drums that have a stretched membrane or skin, and idiophones, such as cymbals that produce sound from their own bodies (Fletcher & Rossing, 1998).

To address the problem of percussion transcription, some event-based systems (Gillet & Richard, 2004b; Gouyon, Herrera, & Cano, 2002; Goto & Muraoka, 1994; Gillet & Richard, 2008) have been proposed which segment the input signal into events informed by the percussion and then extract and classify features from these segments to uncover its musically meaningful content, such as onsets **explain more on these papers**. An alternative to this approach is to rely on source separation-based methods to decompose the input audio signal into basis functions that capture the overall spectral characteristics of the sources. Commonly used source separation techniques and tools such as independent component analysis (ICA) and Non-negative Matrix Factorization (NMF) have proven to be useful in percussion onset detection tasks, especially when analyzing mixtures of different percussion instruments (Paulus & Virtanen, 2005; Smaragdis, 2004a, 2004b; Abdallah & Plumbley, 2003).

Nakano, Ogata, Goto, and Hiraga (2004) explored drum pattern retrieval using vocal percussion, using an HMM based approach. They used onomatopoeia as the internal representation for drum patterns, with a focus on retrieving known fixed sequences from a library of drum patterns with snare and bass drums. Kapur, Benning, and Tzanetakis (2004) explored query by BeatBoxing, aiming to map the BeatBoxing sounds into the corresponding drum sounds. A distinction to be noted here is that in vocal percussion systems such as BeatBoxing, the vocalizations form the music itself, and not a means for transmission as in the case of oral syllables in Indian art music percussion. More recently, Paulus and Klapuri (2009) proposed the use of connected HMMs for drum transcription in polyphonic music. This approach aimed to transcribe individual drums (bass, snare, hi-hat) and not overall timbres due to combinations, and no reference to syllabic percussion is made. However all these approaches have indirectly and implicitly used some form of syllabic representations for drum patterns.

Probably move this to approx string matching section Transcription is often inaccurate with many errors, and any pattern search on transcribed data needs to use approximate search algorithms. There are several attempts to deal with search in symbolic sequences (Typke, Wiering, & Veltkamp, 2005). Well explored techniques such as longest common subsequence (LCS) do not consider the local correlation while searching for a subsequence (Lin, Wu, & Wang, 2011). To overcome this limitation, Lin et al. (2011) proposed a novel Rough Longest Common Subsequence (RLCS) method for music matching.

## 2.4 Relevant technical concepts

Only as is relevant to the thesis. If necessary, to be written at the very end to just give enough background for people to understand the thesis. Mainly to point to different books to read these concepts.

### 2.4.1 Bayesian Models

Bayesian Networks, Markov networks, specific cases of Naive Bayes, HMMs, DBNs, and architectures.

### 2.4.2 Inference in Bayesian models

Bayesian Models are models that use Bayes rule to compute probabilities. Exact inference, Viterbi algorithm

### Sequential Monte Carlo and Sampling methods

Approximate Monte Carlo methods. Particle filtering, the idea and basics. MPF and AMPF described later on, within the context of Meter tracking in Chapter 5.

### 2.4.3 Speech Recognition Technologies and Tools

#### Approximate String Matching

Write about tools and methods for speech recognition, how to use them, and where they will be useful in this task. Evaluation measures in speech technologies.

A short summary of the chapter to be added here.



# Automatic rhythm analysis in Indian art music

A problem well stated is a problem half-solved

---

Charles Kettering

The formulation of the problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill

---

Albert Einstein

**Introduction to be made better** The aims of the chapter are:

1. To identify, present, and discuss main challenges to automatic rhythm analysis in Indian art music
2. To identify, present, and discuss main opportunities in automatic rhythm analysis in Indian art music
3. To identify several interesting, important and relevant research problems within the context of Indian art music and identify key challenges in addressing them, as a means to provide pointers for further future work in rhythm analysis.

4. From the relevant problems, identify a subset of research problems and formulate them in detail, to be addressed in the scope of this dissertation.
5. To present an overview of the state of the art in automatic rhythm analysis of Indian art music, and present an evaluation of the existing state of the art applied to rhythm analysis tasks in Indian art music.

## 3.1 Challenges and Opportunities

There are significant challenges to automatic rhythm analysis in Indian art music. We discuss challenges and opportunities from the standpoint of the state of the art and musical relevance. **A longer introduction ?**

### 3.1.1 Challenges

The most important challenge when addressing automatic rhythm analysis in Indian art music is the inconsistency in definition of rhythmic concepts. Though we can draw analogies between the hierarchical metrical pulsation structure of bar, beats, and subdivisions to the *āvartana/āvart*, beat/*mātrā*, and the *akṣara* of Indian art music, these analogies are mostly approximations that try to force-fit these concepts to the components of a *tāla* and not exactly equivalent. Though, for the ease of readability and clarity of presentation, we will use the commonly used terms, but it's necessary to be aware of the differences and handle them as such. The lack of objective definitions, and even approximate definitions from an engineering perspective are absent, and the main challenge is to first develop consistent engineering definitions for these concepts prior to developing algorithms for analysis.

We identified the *tāla* cycles (at the level of *āvartana* or *āvart*) as the most important and musically relevant cycles in Indian art music. But the theoretical frameworks of the *tāla* described previously also show cyclical structures at time-spans different from the *tāla* cycle. There exist sub-cycles that can be perceived at the section level, the *vibhāg* level in certain *tāls*. A *tīntāl* can be seen to have

four sub-cycles in an *āvart*, one at each *vibhāg*. Similarly, Carnatic music has sub-cycles at the level of *aṅga*, and further at the beat level defined by the *nade*, e.g. *rūpaka tāla* (See Figure 2.1c) can be seen to be comprised of three sub-cycles of four *akṣaras* each. This implies that depending on the metrical levels we focus upon, the metric structure is determined by either duple or triple relations. While this is not a distinct feature of meter in Indian music, it is encountered quite frequently here. Furthermore, in Carnatic music, metric structure might also vary within a piece while maintaining the same *tāla*. For example, though *rūpaka tāla* is generally defined with four *akṣaras* in a beat and three beats in an *āvartana*, it might change within a piece to be grouped as 4 units of 3 *akṣaras* (giving the “feel” of a ternary meter), without changing the cycle length. For the purpose of analysis in this dissertation, we consider *rūpaka* to have the structure shown in Figure 2.1c. This further indicates that ideally, the metrical structure of the piece needs to be estimated at all levels, taking into account possible changes in the metrical structure. This flexibility in interpretation of a *tāla* and the presence of additional metrical sub-cycles can be a significant challenge to MIR systems.

A specific composition can be rendered in different *tālas*. Even though the melody is the same and the total *akṣaras* add up to the same value, the listener experience varies with different *tālas*. In Carnatic music, *Avadhaana pallavi* is one such form of singing a composition set to two *tālas*. The lead musician uses hand gestures to indicate both the *tālas* at the same time, a difficult task for the musician. These compositions are rare but worth a mention in this context to emphasize the fact that the *tālas* of a musical piece is a perceived notion of periodicity, and an objective formulation of the *tālas* provides only a incomplete picture. As with most other musical concepts, the notion of a *tālas* involves a significant amount of subjectivity.

An important aspect of meter in Indian art music is the presence of pulsation at some metrical level with unequal time-spans. The *vibhāgs* in Hindustani music and *aṅgas* in Carnatic music are examples of such possibly non-isochronous pulsations. Such forms of additive meter have so far not been widely considered for computational analysis and present additional challenges.

Neither of Carnatic and Hindustani music traditions have the

notion of an absolute tempo. An expressive performance without a metronome, coupled with a lack of annotated tempo for a piece can lead to a single composition being performed in different tempi, at the convenience of the musician. This lack of a definite tempo value and the choice of a wide variety of tempo classes further complicate the choice of a relevant timescale for tracking *tāla* cycles, causing further metrical level ambiguity.

In Hindustani music, the *āvart* cycle lengths vary from 1.5 second in *ati-dhṛt* (very fast) *tīntāl* to 65 second in *ati-vilambit* (very slow) *ēktāl* (Clayton, 2000, p. 87). Long time scales such as these are far too long to be perceived as single entities (Clayton, 2000), since they are beyond the range of the phenomenon called the perceptual present, which is about 5 seconds long (Clarke, 1999). At such long time scales, the rhythm of the piece is rather characterized by the grouping structure of the piece (Lerdahl & Jackendoff, 1983). Such long cycles are replete with filler strokes to maintain a continuity in pulse, which leads a dense surface rhythm, on top of a time-sparse *mātrā* pulsation. This implies that algorithmic approaches for rhythm analysis that are solely based upon estimation of pulsation from surface rhythm might not be capable to analyze the temporal structure in presence of such long cycles. Carnatic music has a smaller range of tempi and the *tālas* are more concretely defined, and hence the choice of time scale is an easier problem. With a wide range of tempo, cycles as long as a minute, and non-isochronous subdivisions of the cycle, Indian art music is a suitable case to experiment with for extending the horizon of the state of the art in meter analysis.

**MIR** algorithms have difficulty tracking metrical structures that have expressive timing and varying tempo (Holzapfel et al., 2012). Due to the freedom of improvisation and the absence of a metronome, there are local tempo variations, and a possible increase/decrease of tempo through the piece over time. Both of these are not anomalies but accepted characteristics of Indian art music, and can be a potential source of challenge for **MIR** algorithms, with repercussions in tempo tracking, music similarity matching, and drum transcription tasks.

In Carnatic music, the *tāla* only provides a basic structural skeleton to play rhythmic patterns, with significant scope for improvisation. Several different rhythmic patterns without grouping dif-

ferent from the canonical structure of a *tāla* can be performed, as long as the basic cyclical structure and length is maintained, e.g. a musician might decide to play a rhythmic pattern that can grouped as 7,7,4,6,8 *akṣaras* (adds up to 32 *akṣaras*) in a cycle of *ādi tāla*. Several such rhythmic combinations are allowed and is a part of the music, which leads a variety of rhythms played within the basic skeletal structure of a *tāla*. Hindustani music, except a drum solo, has less rhythmic improvisation compared to Carnatic music, but is still significant.

A performance of Carnatic or Hindustani music does not use any form of music scores, while some skeletal scores are used mainly in teaching and music training. This implies that a universally agreed on system of written music does not exist, while there are several efforts to standardize melodic notation for accurate transmission in Hindustani music (Bhatkhande, 1990; Jha, 2001) and Carnatic music (Ravikiran, 2008). There are also recent efforts to create machine readable representations (Chordia, n.d.; Srinivasamurthy & Chordia, 2012b). However, the use of scores itself is limited since the scores are only indicative. This problem extends to representation of percussion patterns too. The use of the tabla and mridangam syllables is an accurate way to representing percussion patterns, but the syllables themselves vary across schools, geographic regions, and languages. This is a potential challenge in the use of syllabic system to represent percussion patterns.

In summary, there is a need for concrete engineering definitions for rhythm concepts in Indian art music. The cycle lengths in Indian music can be meaningfully tracked at multiple time levels, and distinguishing between these multiple time levels is difficult due to the wide variety of tempo classes. The absence of an absolute annotated tempo and expressive tempo are further challenging. Especially for Hindustani music, the presence of additive meters is expected to pose challenges to existing analysis approaches. The significant scope for improvisation leads a wide variety of rhythmic patterns interpreted freely, while a basic adherence to *tāla* structure is maintained. Since the scores are only indicative, there are no standardized representation systems for both melodic and rhythmic patterns, which is a necessity to be addressed. Finally, we must not forget that we attempt to track the *tāla* as a theoretical concept in music performance. However, in both music cultures, artists can

be assumed to deviate from such concepts, which results in a divergence between surface rhythm and theoretical framework that is hard to conceive in any kind of rhythm analysis using only audio.

### 3.1.2 Opportunities

There are several unique features in Indian art music which open new opportunities to pursue novel directions of research in MIR. The challenges outlined also open up new opportunities to propose novel methodologies for automatic rhythm analysis, and improve the current state of the art in MIR. The complex rhythmic framework of the *tāla* necessitates a holistic approach to rhythm description, and will be useful in rhythm analysis of various other music cultures based on similar metrical structures, such as the *usul* in Turkish makam music.

In this dissertation, we mainly consider audio for rhythmic analysis. But the associated notations, lyrics and information regarding musical form also carry rhythm information which can be used for a combined approach to rhythm analysis. The scores, though indicative, can be used to provide prior information to systems and hence are useful.

The onomatopoeic syllables of tabla (*bōls*) and that of mridangam (*solkattu*) define a language for Indian percussion and play a very important role in defining rhythmic structures and percussion patterns in Indian art music. These syllables, which can be considered as the “solfege” of percussion are standardized for Tabla, while less so for Mridangam and form an essential part of percussion training. In Hindustani music, the *thēkās* are defined using these *bōls* and hence these *bōls* can be used to track the movement through the *āvart*. The oral recitation of percussion patterns using these syllables in both Carnatic (*konnakōl*) and Hindustani music are an important part of percussion solos and used extensively in the performances of Indian art dance forms such as Kathak (using Hindustani *bōl*) and Bharatanātyam (using *solkattu*). This system of syllabic percussion is sophisticated and the rhythmic recitation of the syllables requires high skills.

These syllables take an important role in drum transcription tasks, a new way of addressing percussion patterns, where a signif-

icant analogy exists to speech and language. We can draw methodologies and approaches from the mature research area of speech technologies to address percussion pattern transcription and discovery. The percussion solo performance in both Carnatic and Hindustani music are a rich source of typical rhythm and percussion patterns for the corresponding *tāla* and an analysis of these solos can be very useful for extracting these patterns for analysis.

Another important aspect of Carnatic music is that the progression through the *āvartana* is explicitly shown through visual hand gestures. In a performance context, these visual cues play an important role in communication of rhythm among the performers, as well as between the performers and the audience. Listeners often are able to track through complex *tāla* cycles because of these gestures. In fact, in many concerts, these visual cues become a part of expressiveness of the musician and the appreciation of the audience, and hence is a part of the complete experience of the music. Since these cues consist mainly of claps, they can be quite sonorous and it is not very surprising that they can be audible in some recordings. A multi-modal approach to rhythm analysis can be done from video recordings of Carnatic music concerts, a problem that is interesting, but beyond the scope of this dissertation.

In summary, the complex metrical structures and syllabic percussion systems open up several opportunities for novel methods of automatic rhythm analysis in Indian art music. In addition, a complete description of rhythm for effective discovery and experience of music involves integrating various sources of information such as audio, scores, lyrics, visual cues, information about musical form, and other culture specific aspects. Tools for rhythm analysis need to combine these data-sources in order to arrive at a more consistent analysis than by just taking into account the audio signal.

### 3.1.3 Characteristics of Indian Art Music

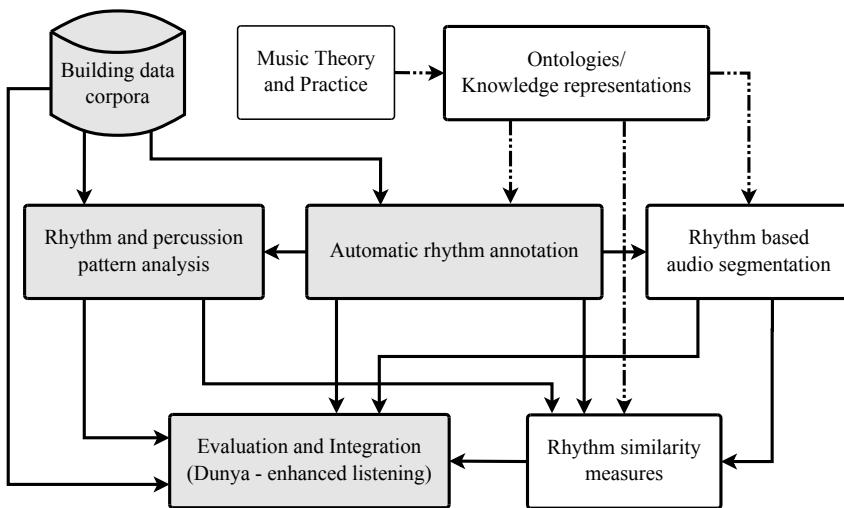
**Include a figure and explain from it. Explain the characteristics of the music signals from an signal processing perspective.** We specify some basic signal characteristics of Indian art music that will be useful to extract relevant audio features for rhythm analysis. Indian art music is predominantly melodic and heterophonic, with usually two (sometimes more) simultaneous melodic voices - a

lead melody and a melodic accompaniment. There is a drone in the background that provides the tonic for the performance. The percussion instruments tabla and mridangam have a bass drum head, and a treble drum head that is pitched. The pitched drum head is tuned to the tonic of lead musician. The pitched strokes can be sharp or sustained. **to be completed...**

## 3.2 Research problems in rhythm analysis

With the background provided so far, several relevant and interesting automatic rhythm analysis problems in Indian art music are identified and discussed. For each problem, we briefly describe the problem, explain its relevance, identify any specific challenges, discuss possible approaches, and review any prior existing work for the problem. Some allied research problems not directly in the scope of rhythm analysis, but have related applications or could benefit from rhythm analysis are also discussed for the sake of completeness. Many of the rhythm analysis problems have not been addressed before, while there have been attempts in MIR aim to solve similar problems in eurogenetic music. While some basic rhythmic feature extraction methods such as onset detection can be easily extended to Indian art music, more complex tasks have to reformulated with the context. Hence, some of the existing general rhythm descriptors such as onset detectors and tempo estimators are deemed to be useful to develop specific algorithms.

Rhythm is characterized by structures and patterns. The structures provide the basis for patterns, through which rhythms are played. The problems presented here revolve mainly around these two concepts, and are categorized into several groups, with the final goal of using all these component problems to define musically meaningful and useful rhythm similarity measures for Indian art music. There are several sub-problems that lead towards the final goal. The problems span the whole range of complexity, starting from basic tasks processing the audio signal to abstract tasks requiring extensive music knowledge. The categorization of problems presented here is only for the purpose of presentation, and the



**Figure 3.1:** Relevant automatic rhythm analysis problems in Indian art music. The solid lines indicate the flow of data and signals. The dot-dash lines indicate the flow of high level information such as parameters and priors. The subset of problems focused in this dissertation are shown in gray.

problems in each category cannot be addressed in isolation. There is significant interplay and overlap between the groups - with problems benefiting from outputs of other problems in a different group, e.g. onset detection can help in both meter analysis and pattern discovery tasks.

At the outset, from a literature review, we see that automatic rhythm analysis of Indian music has been attempted only recently, and concrete methods for Hindustani music and Carnatic music do not exist as yet. Koduri, Miron, Serra, and Serra (2011) also elucidate several unsolved problems in rhythm analysis of Carnatic music. Figure 3.1 shows an overview of the research problems. It also shows the flow of data and information across the important units. We now describe each problem in detail.

### 3.2.1 Building data corpora

A significant part of data-driven research using signal processing and machine learning approaches to rhythm analysis needs good quality data. Data corpora that are representative of the music cul-

ture under study are essential for building and testing such approaches. The data sources comprise of audio, metadata accompanying audio, music scores, lyrics, manual and automatic annotations, and linked data on the internet. One of the main problems addressed in this dissertation is building suitable data corpora for rhythm analysis research, a problem that is described further in Section 3.3.3. Building useful datasets also involves building tools for rhythm annotation and developing machine readable representations for the annotations and metadata for an effective linking and integration of data sources.

### 3.2.2 Automatic rhythm annotation

Automatic rhythm annotation encompasses a broad set of problems that aim to annotate and tag audio recordings with several rhythm related metadata and tags. The tags could be descriptor tags that are not time aligned with audio, such as the tags related to the components of the *tāla* and median tempo descriptors. There could one or more such tags associated with each recording. The rhythm annotations could also be time aligned, indicating the locations of several rhythmic events in audio recordings, such as a time varying tempo, beats, and downbeats. The common tasks of tempo and beat tracking, downbeat tracking can be classified as automatic rhythm annotation problems.

Automatic rhythm annotation, in the context of Indian art music can be defined as the estimation of the characteristics of the *tāla*. For Carnatic music, the most important rhythm related tags include estimating the median tempo of the piece (in *akṣaras* per minute or beats per minute), the length of the cycle (in number of beats or *akṣaras*) and the *tāla* label (and hence the implicit metrical structure), the *nāde* (and hence the subdivision structure), and the *eḍupu* of the piece. For Hindustani music, the most important rhythm related tags include the median tempo of the piece (in *mātrās* per minute), the *lay* class, the cycle length (in *mātrās*) and the *tāl* label. Estimating the time varying tempo curve, the *akṣara* pulse locations, the beats, the *aṅga* (section) boundaries and the *sama* instants are the important time aligned annotation problems in Carnatic music. The most important problems in Hindustani music are

the estimation of the *mātrā* pulsation, the *vibhāg* boundaries, and the *sam* instants.

Automatic rhythm annotation is an important rhythm analysis topic, and there are several applications in which these rhythm annotations are useful, such as music autotagging, rhythm based segmentation of audio, beat aligned processing of music, audio summarization, music transcription, and different rhythmic pattern analyses. Tracking the components of the *tāla* through a music piece is essential for most other rhythm description tasks such as segmentation and extraction of rhythmic patterns to define similarity. Each of these problems are now described in detail. It is to be noted again that many rhythm annotation problems can be jointly addressed, estimating several components together, e.g. the *tāla*, tempo, beats and the *sama* can be jointly estimated, in a task that we call as meter inference.

### Tāla recognition

Tāla recognition, defined as tagging an audio recording with a (and possibly more than one) *tāla* tags is a central research problem of Indian art music. It is the most basic information for listeners to follow the rhythmic structure of the music piece. As the most important rhythm related metadata associated with a recording, knowing the *tāla* is useful for archival, navigation, and enriched listening with large audio music collections of Indian art music.

Since there are only a limited set of *tālas* in Indian art music, *tāla* recognition can be formulated as a classification task based of features of the *tāla* estimated from audio. Barring a few exceptions, most compositions are composed in only one *tāla*, and hence an audio recording with the performance of the composition has only one *tāla* tag. However, audio recordings with long concerts with multiple compositions performed can have multiple *tāla* tags, with the additional problem of marking the regions of audio where these compositions are performed. Further, many recordings start with an *ālāpana*, which is an unmetered section and hence no *tāla*. Hence, Tāla recognition has to first be preceded by such segmentation of audio recording into parts that have only one or no *tāla*. Exceptions can occur despite that, when some rare compositions can be performed in two different *tālas*, a case that is uncommon

and only has artistic significance. Such marginal cases are beyond the scope of engineering approaches in this dissertation.

Tāla recognition is a subjective task that needs prior music knowledge. It can be achieved through a set of proxy tasks, all of which help in identifying the tāla. The clues to identifying a tāla are related to its structure and the rhythmic patterns played in it. The patterns (such as the thēkā) played are indicative of the tāla, but many patterns are also shared across many tālas. From an MIR perspective, it is harder to recognize these patterns, but instead, it is easier to recognize these patterns if the tāla is known. The structure attributes of the tāla that can be used to identify the tāla are the cycle length and when defined, the subdivision meter (in Carnatic music). Estimating the cycle length in beats (or mātrās) can help significantly to identify the tāla. However, there are several tāla in both Carnatic and Hindustani music that have the same length, e.g. ēktāl and cautāl both have 12 mātrās in a cycle(Clayton, 2000), and hence additional information is needed to disambiguate them. Nevertheless, cycle length estimation is an important task that has received some attention from the research community, with the analogous tasks in eurogenetic music being time signature estimation and meter estimation.

In Indian music, we can track well-defined cycles at several levels of the meter. As described earlier, though the aim of the task is to estimate the length of tāla cycle (length of a āvart/āvartana), the algorithms might track a different time scale, which need not correspond to the tāla cycle. One of such other specific levels we are interested in, apart from the tāla cycle is the sub-division within the beat pulsation, i.e. the periodic structure dominating the rhythmic content between the beat instances. In Carnatic music, this corresponds to the nađe estimation. Further, we need to point out here that there is a clear definition of the subdivision meter in the form of nađe in Carnatic music. However, such an explicit subdivision meter for Hindustani music is not clearly given by the theoretical framework. Though the tāla cycle is an important part of rhythmic organization, it is not necessary that all phrase changes occur on the sama. In ādi tāla for example, most of the phrase changes occur at the end of the 8 beat cycle, there are compositions where some phrase changes and strong accents occur at the end of half-cycle or the phrase might span over two cycles (16 beats).

Since the *tāla* cycles have a periodicity, prior approaches in Indian art music track the periodicity in pulsation. Gulati, Rao, and Rao (2011) and Gulati, Rao, and Rao (2012) proposed a method for meter detection from audio for Indian music, and classify a piece as belonging to duple (2/4/8), triple (3/6), or septuple(7) meter. A mel-scale frequency decomposition of the signal is used to drive a two stage comb filter bank. The filter bank output is used to estimate the subdivision time-spans and the meter of the song. It was tested on an Indian film music database with encouraging results. This is one of the first proposed approaches to rhythm modeling applied specifically to Indian music. However, the algorithm was tested on a different repertoire than Hindustani and Carnatic music. The algorithm only aims to classify into these broad meter classes and does not attempt to assign a *tāla* label, which is more complex than such a classification. Though proposed for Indian music, the algorithm is general in approach and does not consider any specific characteristics of rhythm in Indian music. Miron (2011) addressed the problem of *tāl* recognition in Hindustani music, based on recognizing the *ṭhēkā* played on the tabla. Using a labeled corpus of Hindustani music with tabla accompaniment, he explored segmentation and stroke recognition in a polyphonic context, and concluded that recognizing Hindustani *tāls* is a challenging task.

Srinivasamurthy, Subramanian, Tronel, and Chordia (2012) also proposed a culture-specific beat tracking based approach to *tāla* description, and applied it to Carnatic music. They proposed a system to describe meter in terms of the time-span relations between pulsations at bar, beat and *akṣara* levels. The tempo estimation in the algorithm, which is adapted from the algorithm by Davies and Plumley (2007), is modified to peak at 90 BPM allowing a wide range of tempi (from 20 bpm to 180 bpm). The algorithm applies the beat tracker proposed by Ellis (2007) with the estimated tempo as input. The algorithm then uses a beat similarity matrix and inter onset interval histogram to automatically extract the sub-beat structure and the long-term periodicity of a musical piece, from which a set of rank ordered candidates could be obtained for the *nāde* and *āvartana* length. The algorithm was tested on a manually annotated Carnatic music dataset consisting of 86 thirty second song snippets of both vocal and instrumental music with different instrumentation, set to different *tālas* and *nāde*. The algorithm was also tested

on an Indian light classical music dataset consisting of 58 semi-classical songs based on popular Hindustani *rāgas*. Though formulated using the knowledge of the *tāla*, the algorithm does not make an effort to resolve metrical level ambiguities, which can severely affect the accuracy since the performance of the algorithm depends mainly on reliable beat tracking at the correct metrical level.

Rhythmic similarity can also be used in *tāla* recognition, with the assumption that the rhythmic patterns played in a *tāla* across recordings are similar. Rhythmic similarity can be applied to the task by assigning an unknown piece to a class of rhythm it is deemed to be most similar to, based on some low level signal features. Given *tāla* annotated audio examples, rhythmic features can be extracted from audio and used to learn models of rhythm similarity that can classify *tālas*. If the classes of rhythm present in a collection have distinct cycle lengths, we can also obtain the length of the cycle for an unknown piece through this classification.

Practically, most commercially released music collections provide the *tāla* of each piece as editorial metadata. The name of the *tāla* is present in most pieces in the collection. However, it is often not present in archived recordings, or personal music collections, and open music collections. Since most of the work in this dissertation is with commercial music recordings, we already have access to *tāla* tags of these music recordings and hence the task of *tāla* recognition is redundant. We do not work explicitly on *tāla* recognition problem in this dissertation, but however, it is expected to be a byproduct of other automatic rhythm annotation tasks.

### **Lay classification in Hindustani music**

With the wide range of tempo divided into tempo classes, *lay* classification into *vilambit*, *madhya* and *dṛ̥t* (and possibly extended range of *lay* classes) is a useful problem. Since surface rhythm is not an accurate indicator of the underlying tempo, a knowledge of the *lay* class can significantly help in reducing metrical level errors in tracking the tempo and the *tāl*. Since surface rhythm can be misleading, *lay* classification needs to combine features from melody and identify specific tabla stroke timbres to determine the actual underlying tempo class.

### **Edupu estimation in Carnatic music**

Edupu estimation in Carnatic music is a unique problem. Though the edupu is a metadata of the composition, unlike the tāla label, it is never recorded as standard metadata and hence needs estimation. When not available as metadata, edupu estimation needs to be addressed based on accents and salience of the beats and their correlation with the lyrics and melodic phrases. Estimating edupu might be necessary for correct alignment of the samas since in pieces with a non-zero edupus, it is likely that the melodic changes tend to occur at the edupus point rather than at the sama of the tāla cycle. This might lead to confusion in sama tracking algorithms. We also noticed that non-zero edupus pieces tend to give poorer performance in tasks such as cycle length recognition (Srinivasamurthy, Holzapfel, & Serra, 2014). However, with a robust sama tracking algorithm which can handle different rhythmic patterns, the effect of non-zero edupu is less. To the best of our knowledge, this problem has not been addressed by the research community so far. Since the problem of edupu estimation is very specific to Carnatic music, it is of limited interest and is not addressed in this dissertation as well. But when required, we will examine the effect of non-zero edupu on other rhythm analysis tasks.

### **Tāla tracking**

Tāla tracking refers to a set of problems that aim to track the different components of the tāla (the metrical structure) over time in an audio recording, and estimate several time aligned annotations related to the meter. By tracking these tāla components, a complete description of the metrical structure of the piece at different hierarchical levels can be achieved - tracking the cycles as described by the theoretical framework over time. From such a task, all the components such as tempo, akṣaras, beats and mātrās, sections, and sama can be obtained. Tāla tracking is an important automatic rhythm annotation task and is the first step towards any further structural analysis of the music pieces.

Tāla tracking can be done without any prior knowledge of the music piece, in which case identifying the tāla is an implicit step in the process. We call such an uninformed tracking as meter in-

ference - to identify the meter type, and estimate the time varying tempo, beats and the *sama* (downbeats) all together. The *tāla* tracking algorithms can greatly benefit from knowing the *tāla*, tracking a known metrical structure. We call such a task as meter tracking (in contrast to meter inference) - given the *tāla*, estimating the time varying tempo, beats, and downbeats. We can categorize the sub-tasks in *tāla* tracking as tempo tracking, beat tracking, and *sama* tracking.

**Tempo tracking:** Tempo tracking aims to track the time-varying tempo over the piece of song, measured in *akṣaras/beats/mātrās* minute. The knowledge of time varying tempo is useful to track the beats. Even a good estimate of median tempo helps in tracking the beats at the correct metrical level. Median tempo is a good indicator of tempo and can be used as a rhythm tag on the music piece. As described earlier, the tempo changes both locally and over time, and the algorithms for tempo estimation need to be robust to these changes.

**Beat tracking:** In the context of MIR, as noted earlier, beat tracking is commonly defined as determining the time instances where a human listeners are likely to tap their foot to the music. This definition is likely to cause problems in our context, as for example in Carnatic *khaṇḍa chāpu tāla*, listeners familiar to the music tend to tap an irregular sequence of pulses, at the section level, instead of the faster regular pulsation. Also, depending on the *lay*, listeners of Hindustani music tap on either the *mātrā* level for *vilambit* and *madhya lay*, or at the *vibhāg* level for *dṛt lay* (Clayton, 2000, p. 91). In such a case, the more appropriate task of tracking the possibly irregularly spaced beats is more relevant for Indian art music.

Despite these ambiguities, we pursue the task in the present context using a more adapted definition of a beat for the purpose of consistency, defined as a uniform pulsation defined at the “beat” (as defined in Section 2.2.2) level for Carnatic music, and at the *mātrā* level for Hindustani music. This definition of an equidistant beat pulsation can later help in deriving the musically relevant possibly non-isochronous beat sequence that is a subset of the equidistant pulses. This approximation is further inconsequential if the whole cycle along with the *sama* are tracked, using which any pulsation within the beat - uniform or non-uniform can be derived out of that

information. The possibly irregular pulse sequence is a subset of the uniform pulsation estimated from the algorithms. The *vibhāg* or *aṅga* boundaries also coincide with a subset of the beats and hence can be derived from the estimated beat locations. A task that is specific to Carnatic music is *akṣara* pulse tracking, estimating the subdivisions of the beat, an algorithm for which is elaborated in Section 5.2.

**Sama (downbeat) tracking:** The information about where a *tāla* cycle begins provides us with the ability to comprehend most of melodic, rhythmic and structural development of a piece, which is typically synchronized with the phase of the *tāla* cycle. This corresponds to detecting the *sama* (or *sam*) instants of the *tāla* cycle. In Hindustani music, the *sama* is highly significant structurally, as it frequently marks the coming together of the rhythmic streams of soloist and accompanist, and the resolution point for rhythmic tension (Clayton, 2000, p. 81). In Carnatic music, most of the phrasing and improvisations, both melodic and rhythmic, are tied with the *sama* and hence its relevant and meaningful to explore tracking the *sama* as a primary problem in automatic rhythm annotation. Note that while the term downbeat has been mostly applied to eurogenetic music, we apply it here as well because it generally denotes the pulse at the beginning of a bar. The downbeat does not necessarily correspond to the strongest accent in the cycle. In this sense, downbeat in Indian art music and Eurogenetic music are likely to be concepts with different meaning.

It is to be noted that all the above components can be tracked together jointly, instead of individually. Such meter tracking algorithms are an important focus of the dissertation. Automatic rhythm annotation one the main topics of this dissertation, and a subset of the problems described above is the subject matter of Chapter 5. To the best of our knowledge, the problem of meter tracking for Indian art music is addressed for the first time in the work presented in this dissertation. The subset of problems that will be explored deeper in this dissertation are better formulated in Section 3.3.1.

### 3.2.3 Rhythm and percussion pattern analysis

While *tāla* provides a framework and structure, the rhythm and percussion patterns form the content through which the metrical struc-

tures and rhythms are illustrated, and hence form the other main component of rhythm analysis. Rhythm patterns mainly refer to the temporal arrangement of different events with different accents, while percussion patterns include a temporal arrangement of different percussion timbres. To distinguish, percussion patterns are rhythmic patterns, but rhythmic patterns need not contain only percussion, and can be from melody or percussion.

A pattern is defined as a temporal sequence of events and hence necessary to estimate onsets of various instruments in music, since that creates of time-aligned sequence of note/stroke events which can be further used to obtain both rhythmic and percussion patterns. Some important sub-problems in pattern analysis are instrument-wise onset detection, pattern transcription, and pattern discovery, each of which is described further. Transcription aims to map an audio recording into a time aligned sequence of symbols (strokes, accents, e.g.). The problem of discovery is more open ended and aims to automatically retrieve interesting patterns and insights about those patterns, in a data-driven way.

### **Instrument-wise onset detection**

**To be completed...**

### **Rhythm pattern analysis**

Rhythm patterns extracted from audio recordings are representative patterns of the *tāla*, and hence useful for both automatic *tāla* recognition and meter tracking. The most relevant rhythmic patterns are cycle length rhythmic patterns - patterns that are played in once cycle of the *tāla*. Shorter patterns, played within a cycle mostly act as rhythmic atoms to make up the whole cycle and are played more often. There are however, long rhythmic patterns played on mridangam/tabla and accentuated through melody which can last many *āvartana/āvart*. Automatic discovery of rhythm patterns can be used to define content based rhythmic similarity between pieces of music, which is expected to be more relevant than metadata based similarity. Automatic extraction of rhythm patterns can also be a tool for musicologists to study various rhythm patterns in a larger corpora.

Rhythmic patterns are closely tied to the *tāla*, and further within a *tāla* into position within the cycle. Hence, a pattern discovery system can significantly benefit from all forms of *tāla* related metadata. Consequently, rhythm pattern discovery needs *tāla* annotated (with time-aligned sama and beats) datasets for a better performance. A systematic automatic study of rhythm patterns in Indian art music is lacking, and an effort towards that is a part of the dissertation.

### Percussion pattern transcription and discovery

Percussion pattern transcription are mainly useful on audio recordings with percussion solo, and aims to transcribe the audio recording into a time-aligned sequence of drum stroke labels, and in the case of Indian art music, into the percussion syllables. Percussion transcription is a sub-problem of the more general music transcription, that is based on rhythm and percussion. Transcription of a solo into symbolic syllables is an example of audio segmentation at a fine grain level. Transcription of solo performances can be very useful for percussion training. Since Indian art music is mostly improvised, the need for such a fine grained transcription system is limited, except for music education and performance analysis applications. However, such a fine grained transcription can be used to automatically further discover percussion patterns and develop rhythm similarity measures using such discovered patterns.

Many music traditions around the world have developed particular systems of oral mnemonics for transmission of the repertoire and the technique. D. Hughes (2000) coined the term *acoustic-iconic mnemonic* systems for these phenomena, and described their use in different genres of traditional Japanese music. As he points out, the core aspect of these systems is that the syllables are chosen for the similarity of their phonetic features with the acoustic properties of the sounds they are representing, establishing an iconic relationship with them. Therefore, these systems are essentially different from those of solmization (A. Hughes & Gerson-Kiwi, accessed April 28, 2014), like for instance the syllables of solfège, of the Indian svaras or the Chinese gongche notation, which are nonsensical in relation to the acoustic phenomena they represent.

The use of the aforementioned systems for the transmission of percussion is wide extended among traditional musics. D. Hughes

(2000) metions the *shōga* used for the set of drums of *Noh* theatre. In Korea, the young genre of *samul nori*, a percussion quartet of drums and gongs, draws on traditional syllabic mnemonics for the transmission of the repertoire. Furthermore, these systems are also known to be used in Turkish traditional music and Javanese music.

The benefits of using oral syllabic systems from an MIR perspective are both the cultural specificity of the approach and the accuracy of the representation of timbre, articulation and dynamics. The characterization of these percussion traditions need to consider elements that are essential to them such as the richness of their palettes of timbres, subtleties of articulation, and the different degrees and transitions of dynamics, all of which is accurately transmitted by the oral syllables.

As discussed earlier, the onomatopoeic percussion system in Indian art music provides a language for percussion and hence is the most musically meaningful way to represent percussion patterns of tabla and mridangam. However, there are some challenges to percussion transcription. The syllables of tabla and mridangam are not unique across all the schools. Since these syllables closely mimic the timbre and dynamics of the drum stroke, the mapping between the strokes to syllables is not unique and one to one, with several different syllables mapping to one stroke timbre. Further, within a percussion solo, a syllabic pattern can also be loosely interpreted, leading to further complexity.

Percussion pattern transcription can be formulated a supervised learning task, using labeled training data to build syllable stroke models, which can then be used to transcribe a test recording. Percussion pattern discovery is an unsupervised task, aiming to extract percussion patterns from audio and/or scores in an unsupervised way, though some priors can be used. Music scores of percussion solos (represented by syllables) is used for percussion training. Such scores can be used for symbolic analysis of percussion patterns, and used to discover percussion patterns from score corpora, a task that much less complex than extracting them from audio. We can then use these patterns and search for them in a long percussion solo recordings. Such an approach with pattern discovery from scores followed by pattern search in audio is explored further in this dissertation.

A scientific study of Indian percussion instruments can be traced

back to the study of acoustics of Indian drums by Sir C. V. Raman (Raman & Kumar, 1920; Raman, 1934). In the last decade, most of the MIR work with Hindustani music percussion has focused on drum stroke transcription, creative modeling for automatic improvisation of tabla and predictive modeling of tabla sequences. The first attempt at tabla stroke transcription was done by Gillet and Richard (2004a). Parag Chordia (Chordia, 2005a, 2005b) focused on automatic transcription of strokes from solo tabla music. He developed a new encoding scheme for transcription of tabla bols called the `**bol` format based on the humdrum syntax (Huron, 2002). (Rae & Chordia, 2010) developed an automatic tabla improviser. Recent approaches have been mainly on sequence modeling of rhythm sequences (Gillet & Richard, 2007). Extending further, most of work using tabla sequences has been in a predictive modeling setup using the multiple viewpoint modeling framework (Chordia, Albin, & Sastry, 2010; Chordia, Sastry, & Şentürk, 2011; Chordia, Sastry, Mallikarjuna, & Albin, 2010; Sastry, 2012). Miron (2011) explored segmentation and transcription of tabla strokes within the context of `tāl` recognition in Hindustani music.

The work with Carnatic percussion has been limited so far. Motivated by the work of Raman (1934), Anantapadmanabhan, Bellur, and Murthy (2013) used a Non-negative matrix factorization based approach to a decomposition of mridangam strokes into its modes, and used them for transcription. The work was further extended using cent-filterbank based features to make transcription independent of tonic (Anantapadmanabhan, Bello, Krishnan, & Murthy, 2014). More recently, Kuriakose, Kumar, Sarala, Murthy, and Sivaraman (2015) proposed an algorithm for mridangam stroke transcription and evaluated it on an annotated dataset of mridangam solos (see Section 4.2.4 for the dataset). Percussion pattern transcription and discovery in both mridangam and tabla solos is one of the problems addressed in the dissertation and is formulated more comprehensively in Section 3.3.2.

### 3.2.4 Rhythm based audio segmentation

Segmentation problems refer to a broad category of problems which involve the labeling segments of audio with a label/tag. Segmen-

tation can be done at several levels, based on different music concepts. Segmentation problems are useful since they provide additional metadata to navigate through music collections and within a single piece, and to further develop similarity measures. Audio segmentation is not the main focus of the dissertation, but several rhythm related segmentation problems are described briefly.

A music recording can be segmented based on the instruments that are playing, a problem that can also be described as instrument tracking in audio. In Indian art music, this is useful to segment a piece into structural segments that are known to have specific instruments, e.g. an *ālāpana* only has melodic instruments playing, while a percussion solo only has percussion instruments. In the context of Indian art music, Ranjani and Sreenivas (2015) recently proposed an approach to track different instruments from a mixture.

Segmentation can also be useful for segmenting a concert into the pieces that were performed in it, which is useful for archival. Segmentation at a structural level within a piece aim to segment a piece into the different sections of the piece, and useful for navigation and similarity. An applause based segmentation of Carnatic music concerts was proposed by Sarala and Murthy (2013), which was also extended to intra-piece segmentation into sections of the piece. Hindustani *khyāl* music concert recordings are often presented as a single recording with multiple pieces, performed in possibly different *lay* and *tāl*. Segmenting Hindustani concert recordings based mainly on rhythm features using a modified tempogram was proposed by Vinutha and Rao (2014), and structural segmentation using additional audio features was proposed by Verma, Vinutha, Pandit, and Rao (2015). A recent approach to estimate reliable tempo that aids in rhythmic segmentation was applied to Sarod (a fretless stringed instrument in India) concerts by Vinutha, Sankagiri, and Rao (2016). A rhythm based segmentation of such an audio recording is also useful for *tāla* tracking on the recording. Tempo or *lay* class based segmentation can use a novelty function for onset detection and detect changes in tempo to segment audio.

Segmentation of recordings at the time scale where we can define meaningful rhythmic phrases is relevant. The span of these phrases are closely tied to the metrical positions of the *tāla* cycle. These phrases can characterize the rhythm of the piece and would be instrumental to measure rhythmic similarity between two pieces

of music. Given some form of automatic rhythm annotations, such as the sama and the beats, extracting rhythmic patterns and rhythm phrase boundaries in the music piece, e.g. *thēkā* segmentation aims to segment the piece at *thēkā* changes. More generally, it encompasses the task of segmentation at rhythm phrase changes. Further, though the *tāla* of a song is fixed, the *nade* could change through the song, and *nade* based segmentation of audio would be further useful for structure segmentation of the song. Rhythm in both Carnatic and Hindustani music is highly improvised with a possibility of wide variety of rhythms. However, there are regions in the music piece with well defined structures that contain rhythmic phrases characteristic of the *tāla*. Identifying these “regions of stable rhythm” would be helpful in rhythm annotation tasks. Further, these stable regions can be used to extract representative rhythm templates for measuring rhythmic similarity.

### 3.2.5 Ontologies for rhythm concepts

Though not a topic addressed in the dissertation, ontology engineering (also called as knowledge engineering) aims to integrate human knowledge into computer systems to solve complex problems that require human expertise (Brachman & Levesque, 2004; Gómez-Pérez, Fernández-López, & Corcho, 2004; Berners-Lee, Hendler, & Lassila, 2001). An ontology specifies concepts, attributes, relations, constraints, and instances in a domain. Since music is a complex and varied phenomenon with many perspectives, a cultural domain specific ontology is needed to define the relationships that pertain to a specific type of music. *Tāla* ontologies are knowledge representations of rhythm. They encode the relationships that exist among the rhythmic concepts of Indian art music. Built using the knowledge of music theory and practice, the ontologies would be useful for querying complex rhythmic relationships between the pieces. The ontologies complement the features derived from the data with music knowledge based relationships that can be used for defining rhythmic similarity. e.g. using a *tāla* ontology and the knowledge of cycle length, it might be easier to identify the *tāla* from audio. Further, the ontologies will also be useful to create specific models with priors obtained using from the ontology. In summary, ontologies can be built both for a direct use for nav-

igation and inference, and for building domain specific machine learning algorithms.

Previous work on ontologies have been mainly in organizing music and metadata (Swartz, 2002; Raimond, 2008). The Comp-Music project aims to develop ontologies for all the music cultures under study, some examples include the work by Koduri and Serra (2013); Koduri (2014). Building comprehensive ontologies needs expertise in music theory and ontology languages, an effort that is beyond the scope of this dissertation. In this dissertation however, we use basic knowledge representations to incorporate prior information in several rhythm analysis tasks.

### 3.2.6 Rhythm similarity measures

Rhythm similarity measures aim to use the rhythm descriptors, metadata, and segmentation information to provide an objective similarity value between two phrases, two music pieces, or two parts of the same piece. Developing culture-specific similarity measures is one of the final goals of the CompMusic project, and rhythm similarity is a major component of it. Since rhythmic similarity is not a very concrete notion, we need definitive and objective measures of similarity, especially in a multi-cultural setting. This would necessitate the use of knowledge based approaches for similarity modeling. Developing an in depth study of rhythm similarity measures is not addressed in this dissertation. However, some possible directions of research towards the goal are discussed.

The onus of developing new similarity measures clearly lie on the choice of metrics which correspond to rhythm similarity as perceived through musically relevant concepts - based on *tāla* and the rhythmic patterns. The *tāla* ontologies provide the empirical *a priori* music theory based models for similarity. As an evidence for the prior from metadata and audio, we need novel mid level features obtained from both the automatic rhythm annotations and the rhythmic phrases extracted using audio segmentation. These mid level features provide a semantic abstraction that is in between the well defined but less definitive signal level features and the abstract high level music theory based features. These features can be then used to define objective measures of rhythm similarity. These features are a combination of the parameters computed on the whole

piece as well as those computed on each rhythmic phrase that has been extracted from the piece. This way, we will be able to define measures and compute similarity between rhythmic phrases, between music pieces and between parts of the same music piece.

With the automatic rhythm annotations, rhythm based segmentation tasks can be used to extract characteristic patterns of the piece. With the *tāla* information, we can then make a library of rhythm patterns that can be used for measuring rhythmic similarity. Since melodic and rhythmic phrases are closely tied to *tāla* cycles, we can use the sama and beat markers to segment the audio into relevant phrases. Each phrase can then be characterized using the notes/strokes, duration and their salience. Further, using intra piece similarity between these phrases, we can aim to perform structural segmentation of the piece.

With the rhythmic phrases extracted from each piece, we can cluster these pieces based on empirical distance measures to form families of phrases with phylogenetic relationships with some basic characteristic phrases of a *tāla*. This would be the initial approach to defining measures of similarity from data. We also define empirical distance measures based on music theoretic concepts such as *tāla*, *nade*, *edupu*, *lay* classes, and the measures obtained from the data can also be used to refine these empirical measures. We can also cross test the data derived measures and empirical measures on the data to evaluate and improve their performance. The empirical measures and the data derived measures can be combined using the inference obtained from ontologies and then used to define culture specific measures of rhythm similarity. These culture-specific measures will finally have to be evaluated using listening tests with trained musicians, and both experienced and non-experienced listeners.

### 3.2.7 Symbolic music analysis

Symbolic music scores in Indian art music are not comprehensive and are only indicative. They are almost never used in performance, but mainly in music training and archival. There are no standard notation systems for melody or percussion, in both Hindustani and Carnatic music, that are widely accepted and used.

Rhythm related information encoded in scores of compositions is limited to the *tāla* and *akṣara* or *mātrā* durations. In Carnatic music, with a knowledge of the composition, the percussionist closely follows the composition. Though the percussion accompaniment is largely improvised, the score implicitly encodes the note durations and the set of possible *ṭhēkās* played during the composition. Thus a rhythmic analysis on symbolic scores using note durations and *sama* boundaries provide a good starting point for tasks such as audio to score alignment, and structure similarity problems. Due to the large deviation of performed music from the indicative scores, score analysis can at best be good starting points towards rhythm analysis. Further, there is no comprehensive collection of machine readable music scores in Indian art music. We do not therefore explicitly work on symbolic score analysis, but make use of the available scores when they provide useful information.

The syllabic scores of tabla and mridangam are useful to discover percussion patterns from symbolic data, a problem that is further addressed. Automatic score analysis research in the context of melody, rhythm, and percussion are very few in Indian art music. Symbolic scores are used for different melodic analyses by Koduri, Ishwar, Serrá, and Serra (2014) and Ranjani and Sreenivas (2013) for Carnatic music, and by Srinivasamurthy and Chordia (2012a) for Hindustani music vocal compositions creating a machine readable Hindustani music score dataset. As described earlier, tabla *bōls* sequences have been used for predictive modeling of solo tabla performances using the multiple viewpoint modeling framework (Chordia, Sastry, et al., 2010; Chordia, Albin, & Sastry, 2010).

### 3.2.8 Evaluation and Integration

The algorithms and measures developed as a part of the dissertation need comprehensive evaluation. Most of the automatic rhythm annotation tasks are well defined have groundtruth, and hence are suitable for automatic evaluation using information retrieval measures. However, they require substantial amount of good quality annotated datasets, which need to be built. Percussion pattern transcription also can be evaluated using measures borrowed for speech recognition research. Audio segmentation for rhythmic phrases is

not very well defined and objective performance measures need to be defined, based on their usefulness in defining rhythm similarity measures.

Rhythm similarity is the hardest to formulate and evaluate, since a significant amount of human subjectivity is involved. The best evaluation for Rhythm similarity is through listening tests, with the defined measures and the target audience. Since listening tests are both time consuming and need a lot of responses before reaching concrete conclusions. Since these measures are not concrete, the most effective strategy would be to iteratively improve these measures with feedback from listening tests, or use proxy tasks as a measure of rhythm similarity.

**Integration:** Dunya (Porter et al., 2013) is a web-based software application that lets users interact with an audio music collection through the use of musical concepts that are derived from a specific music culture. Dunya is the best showcase of research resulting from this dissertation. Dunya can be used to visualize all the automatically generated rhythm annotations and segmentation of a music piece. This leads to an enriched experience in listening with a better understanding of the underlying rhythmic processes in the piece. Further, Dunya provides an interface to integrate the ontologies and data derived measures of similarity. It also provides an interface to integrate rhythm similarity measures developed in the thesis to other similarity measures (such as melodic and timbral) to be developed in the CompMusic project to provide a complete system for similarity based navigation of music collections. The rhythm similarity measures are a part of the suite of similarity measures being developed as a part of CompMusic project. These measures need to be combined to provide an overall similarity measure, which will be the basis for navigation through the music collections of Dunya.

### 3.2.9 Extensions to other music cultures

The algorithms in the dissertation are developed with the possibility of extensions to rhythm analysis of other music cultures within the context of CompMusic project. Turkish makam music is based on rhythmic cycles called *usul*. An usul is a rhythmic pattern of a certain length that defines a sequence of strokes with varying ac-

cent. An usul is analogous to *tāla*, but is less complex than the *tāla* system. Hence, most of the algorithms developed for Indian music extend to makam music. In Beijing opera, *banshi* represent the metrical patterns to set lyrical couplets into music. A rhythmic analysis of Beijing Opera, such as tracking the banshi through an aria is an analogous task to *tāla* tracking.

Beijing Opera percussion shares the concept of a syllabic percussion system, which is more simpler and more well defined than Indian art music. It hence is an ideal pilot case for percussion transcription work with syllabic representation of percussion patterns, a topic of study in Chapter 6. Despite the rich musical heritage and the size of audience, little work has been done for computational analysis of Beijing opera from an MIR perspective. It has been studied as a target in some genre classification works (Zhang & Zhou, 2003) and the acoustical properties of Beijing opera singing has been studied (Sundberg, Gu, Huang, & Huang, 2012).

### 3.3 Thesis problems: A formulation

With an overview of the relevant research problems, some challenges in them, possible approaches and the state of the art for those problems, a subset of those problems that are addressed in this dissertation are formally defined and discussed. For these problems, we formulate the research question, discuss any assumptions with justification, discuss the terminology used, and give a basic idea about the approach. The problems across Hindustani and Carnatic music are very analogous, but all the experiments are done separately for each music culture - implicitly assuming that the music culture of the piece is known *a priori*. A detailed discussion of the approaches, experiments and results is presented in subsequent chapters.

#### 3.3.1 Meter inference and tracking

The main problem addressed in the work presented in this dissertation is meter analysis of audio recordings. Meter analysis is an umbrella term used for the problems of meter inference and meter inference. To the best of our knowledge, a comprehensive auto-

matic meter analysis has not been researched in Indian art music and hence the primary goal of the dissertation is to propose and present meter analysis approaches for Indian art music. In addition, we also ask the research question if building culture specific models of *tāla* and informed meter analysis, providing additional information about the *tāla* *a priori* into meter analysis approaches, can improve their performance, leading to more accurate tracking of the components of the *tāla*. The additional information provided is explored at different levels - with meter inference is most uninformed to tempo-sama-informed meter tracking, the most informed meter analysis problem. We also aim to effectiveness and applicability of Bayesian models for meter analysis in the case of Indian art music. We mainly explore supervised approaches for meter analysis, in specific focusing on meter inference and meter tracking problems. We explore using rhythm features extracted from audio indicative of rhythmic events, using models of the *tāla* that can effectively model its structure.

In the scope of the work presented in this dissertation, the music culture to which the audio recording belongs to - Carnatic or Hindustani music, is known *a priori*. The audio recordings are assumed to have a percussion instrument playing, mainly the mridangam in Carnatic music and tabla in Hindustani music. This implies that only metered forms of music are analyzed, leaving out the unmetered melodic improvisations (e.g. *ālāpana*). We restrict our scope in Hindustani music to *khyāl* performances. The music recording is assumed to have been already segmented into pieces that are in a single *tāla*, e.g. long recordings with multiple piece are segmented into pieces with one *tāla* and presented to the algorithms. We don't make an assumption that the audio file presented has the starting of the piece - any excerpt of audio of any length can be presented for analysis, as long as it is in a single *lay* (Hindustani music) and *tāla*. This assumption mainly stems from the limitation of our approaches in handling changing *tālas* through a piece. However, targeting automatic meter analysis of large music archives, segmentation into excerpts with the a single *tāla* is less complex than meter analysis. Most commercial releases already are segmented into pieces and hence it is a fair assumption. We do not assume any restrictions on tempo range and its variability in time over the piece.

We restrict our work to four popular *tālas* that span a majority of recordings in Indian art music. For Carnatic music, we restrict our work to *ādi*, *rūpaka*, *miśra chāpu*, and *khaṇḍa chāpu tāla*s, and for Hindustani music to *tīntāl*, *ēktāl*, *jhaptāl*, and *rūpak tāl*. Since our approaches are supervised, this restriction is mainly due to the lack of availability of annotated training data in less popular *tāla*s - the rare *tāla*s have very few examples available even in large music archives. The performance of the approaches is likely to extend to other *tāla*s as well, provided we have sufficient training data. From a practical standpoint, these four *tāla*s will cover over ***XX% citation needed*** of the compositions in both Carnatic and Hindustani music, and hence such a restriction is justified.

Let a music recording  $q$  be represented as an audio signal  $f[n]$  and can be reduced by frame-wise analysis to a feature vector sequence  $\mathbf{y}_k$ , for  $k = \{1, 2, 3, \dots, K\}$ , where  $K$  is the total number of audio frames. Let the set of time instants of beats/*mātrās* labeled with their position in the cycle be  $\mathcal{B}_q$ , and the set of *sama/sam* time instants be denoted as  $\mathcal{S}_q$ . In addition, the set of *akṣara* pulses in a Carnatic music recording be denoted as  $\mathcal{O}_q$ . The time varying sequence of tempo value estimates, called a tempo curve can be measured in beat/*mātrā* pulse period (or inter-beat/*mātrā* interval)  $\tau_{b,k}$ , or as cycle period (or inter-*sama/sam* interval)  $\tau_{s,k}$ .  $60/\tau_{b,k}$  would represent the same value in beats per minute (BPM) or *mātrās* per minute (MPM). For Carnatic music, tempo can additionally be measured in *akṣara* pulse period (or inter-*akṣara* interval)  $\tau_{o,k}$ , where  $60/\tau_{o,k}$  would have a unit of *akṣaras* per minute (APM). The beats are labeled with their position in the cycle. Given that the section (*aṅga* or *vibhāg*) boundaries are a subset of the set of beats, the beat number and beat time can be used to obtain the section boundaries in a straightforward way selecting only those beats with labels corresponding to section boundaries. The approaches, experiments and results for meter inference and tracking problem are presented in Chapter 5.

### 3.3.2 Percussion pattern transcription and discovery

The problem of discovery of percussion patterns in percussion solo recordings is the second problem that is addressed in this thesis. Not being the primary problem, it is explored to a lesser extent and most experiments presented contain preliminary results, needing further work. The approach we explore in this dissertation is to use syllables to define, transcribe, and search for percussion patterns. The goal in the dissertation is to test the effectiveness and relevance of percussion syllables in representation and modeling of percussion patterns for automatic transcription and discovery. Since these syllables have a clear analogy to speech and language, we present a speech recognition based approach to transcribe a percussion pattern into a sequence of syllables.

We assume that the percussion solos have been segmented out of the concert/performance, since structural segmentation is not a problem that is addressed in this dissertation and some prior methods can be used for the task (Sarala & Murthy, 2013). We focus only on tabla and mridangam solos in Hindustani and Carnatic music, respectively, since they form a majority of the recordings. Percussion solos with other instruments (e.g. khanjira, ghatam, and morsing) in Carnatic music is left for future work.

The syllabic percussion system in both Carnatic and Hindustani music provides a musically relevant representation system for percussion patterns. However, there are considerable differences in names of syllables that represent a specific stroke timbre - that vary across regions and schools. Hence, while using syllables for representation, we aim to base percussion pattern definitions on stroke timbres and not on specific syllable names. To that effect, we group syllables that represent similar timbre, and use these syllable groups to represent percussion patterns. Though each syllable on its own has a functional role, this timbral grouping is presumed to be sufficient for discovery of percussion patterns. Though this leads some form of reduced representation and not a rich representation using the whole set of percussion syllables, it leads to a smaller subset of syllables - that can be trained with lesser training data. Further, it makes the definition of patterns more concrete from timbral perspective, removing ambiguities - similar sounding

patterns will have the same representation. The syllable grouping for mridangam and tabla, along with the datasets that were created for percussion transcription research are presented in Section 4.2.4 and Section 4.2.3, respectively. It is to be noted that this syllable grouping is only for the ease of representation for the task of automatic transcription and discovery.

Let the set of syllables be denoted as  $\mathcal{A} = \{A_1, A_2, \dots, A_{N_s}\}$ , where  $N_s$  is the total number of syllables. A percussion pattern is not well defined and varied definitions can exist. In this work, we use a simple definition of a percussion pattern - as a sequence of syllables and their time-stamps. A pattern  $\mathbf{A}$  indexed by  $j$  is defined as  $\mathbf{A}_j = [a_1, a_2, \dots, a_{L_j}]$ , where  $a_i \in \mathcal{A}$  and  $L_j$  is the length of  $\mathbf{A}_j$ . Though, for defining patterns, it is important to consider the relative and absolute durations of the constituent syllables, as well as the metrical position of the pattern in the *tāla*, we use a simple definition and leave a more comprehensive definition for future work.

The pattern transcription and discovery problem is addressed using both audio and syllabic scores. From an analysis of symbolic scores, we build a set of syllabic query patterns of different lengths,  $\mathcal{P} = \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_{N_a}\}$ , where  $N_a$  is the number of query patterns. Different strategies can be used such libraries of percussion patterns, but we extract the most frequent patterns and consider them as representative.

Given an audio recording  $f[n]$ , it is first transcribed into a sequence of time-aligned syllables  $\mathbf{A}^*$  using syllable level timbral models. Hence,  $\mathbf{A}^* = [(t_1, a_1), (t_2, a_2), \dots, (t_{*}, a_{L^*})]$ , where  $t_i$  is the onset time of  $a_i$  and  $L^*$  is the length of the transcribed sequence. The task of syllabic transcription has a significant analogy to connected word speech recognition using word models. Syllables are analogous to words and a percussion pattern to a sentence - a sequence of words.

Given a query pattern  $\mathbf{A}_q$  of length  $L_q$  from the set  $\mathcal{P}$ , we perform an approximate search for the pattern in the output syllabic transcription  $\mathbf{A}^*$  to obtain the locations  $\{t_q^{(k)}\}$  of the patterns in the audio recording. Syllabic transcription is often not exact and it can have common transcription errors such as insertions, substitutions and deletions, to handle which we need an approximate search al-

gorithm. An analogous task for this search in speech research is keyword spotting in speech, where a known word (or a phrase) is searched in a longer piece of speech recording citeXX.

The whole approach can be formulated as a discovery problem with percussion solo recordings and percussion scores - to discover characteristic audio percussion patterns from these recordings. The characteristic patterns are first discovered automatically from scores, and the audio training data is used to build timbre models for the syllables. Given a new test recording, the timbre models are used to transcribe the recording, and symbolic percussion patterns are then searched in the transcribed score. The approaches, experiments and results for both tabla and mridangam solos are presented in Chapter 6.

### 3.3.3 Datasets for research

Building such data corpora scientifically for MIR itself is a research problem (Serra, 2014; Peeters & Fort, 2012). Setting up criteria for selection and curation of music, and designing datasets for research are to be done with objective parameters that can measure the goodness of a corpus for a particular research task. One of the primary aims of the CompMusic project is to build such data corpora and make it available for research. Collection of good quality data and easy access to both audio and metadata is essential for reproducibility of research and to further the work presented in the dissertation.

For developing algorithms, we focus on commercial quality audio from CDs, with manually edited metadata. The CompMusic audio collection is comprehensive for both Carnatic and Hindustani music and includes rhythm related metadata such as the *tāla*, rhythmic form and the *lay*. For tasks such as automatic rhythm annotation and rhythm segmentation, we need rhythm annotated audio data. Starting from the vast CompMusic collection, we build a comprehensive rhythm annotated sub-collection with beat and *sama* level annotations. For a rhythm based segmentation task, time aligned segment boundaries are needed as appropriate.

For both Carnatic and Hindustani music, we aim to build an annotated audio sub-collection that representative of the real world performance practices. The pieces chosen need to span all the *tālas*, *lay*, and forms as needed by experiments. The datasets built in the

context of this dissertation are further elaborated in detail in Chapter 4. An interesting corpus wide cycle length rhythmic pattern analysis using the rhythm annotated music corpora is also presented in Section 4.2.1 for Carnatic music and Section 4.2.2 for Hindustani music. Interesting and valid musicological inferences can be drawn from such an analysis, showing the potential of such a methodology.

### 3.3.4 A note on terminology and style

The dissertation presents work with several acronyms, unfamiliar and uncommon terms. Hence, before presenting the main chapters of the dissertation describing approaches and experiments, a note on terminology is in order.

As mentioned earlier, the terms related to music related concepts use Kannada for Carnatic music terms and Hindi for Hindustani music terms. The latin transliteration of these terms are according to the ISO 15919 (Transliteration of Devanagari and related Indic scripts into Latin characters) standard (ISO/TC, 2001). For consistency, when it is clear from context, we use Carnatic music terminology to refer collectively to the two Indian art musics, e.g. we use the word *tāla* to mean both *tāl* and *tāla* when we refer collectively to the two Indian art musics. We use the respective terms while referring to each music culture individually. Further we will use the term Indian art music to refer collectively to Carnatic and Hindustani music.

When clear from context, we use the commonly used eurogenetic music terminology and the specific Indian music terminology interchangeably, primarily to enhance readability to an unfamiliar reader. E.g., the term meter tracking and *tāla* tracking are equivalent, the term syllable and *bōl* are interchangeable in Hindustani music, a bar or a cycle is used to mean an *āvartana* or *āvart* of a *tāla*. Such interchangeable use, however, assumes only the limited equivalence between these terms as defined in Section 2.2, and hence the distinction still is to be clearly maintained.

The algorithms and datasets presented in the dissertation are all identified using acronyms, but a consistency is maintained throughout the dissertation. Meter analysis is an umbrella term encompassing both meter inference and meter tracking.

The writing style followed in the thesis is a mixture of both active and passive usage. Since most of the dissertation derives content from published research papers describing the work done by collaborative teams, the word “we” refers to the author and in cases the co-authors and collaborators in research papers. When presenting results and making observations, the word “we” further includes the reader. However, the main original contributions of the thesis are emphasized appropriately, wherever needed.

## 3.4 In search of automatic rhythm analysis methods

To conclude the chapter, we present an evaluation of the performance of some existing approaches in MIR applied to automatic rhythm annotation tasks in Indian art music (Srinivasamurthy, Holzapfel, & Serra, 2014). The evaluation presented here is an early evaluation of the algorithms, and the goal of such an evaluation is not to compare performance of these algorithms with the proposed approaches. The goal is to obtain insights into the nature of rhythm in these cultures and the challenges to rhythm analysis, and to learn about the capabilities and limitations of the existing approaches when applied to Indian art music, and use these insights in proposing novel approaches.

Many of these approaches were not proposed to handle the rhythmic structures encountered in Indian art music, and hence their performance is at best sub-optimal. The algorithms and the evaluation had to be adapted to a common ground in which an evaluation could be done, and hence the evaluations are not strict and comprehensive, but still provide insights into the approaches.

We further focus on the problems that are not explicitly addressed in later chapters. In specific, meter estimation (cycle length estimation) and downbeat tracking are evaluated here. These two tasks however are implicitly addressed within the task of meter inference in Chapter 5. Cycle length estimation task is used as a proxy for *tāla* recognition. Downbeat tracking is an important focus of this dissertation, but we approach it together as a part of meter analysis, while the approaches evaluated here attempt downbeat

tracking as an independent task.

If existing methods from MIR are capable of handling the following tasks in a satisfying way for Indian art music, we will be able to automatically analyze the content of these music signals in a well-structured way. However, as recent research results show (Holzapfel et al., 2012), these tasks are far from being solved even for the Eurogenetic forms of music, for which most methods have been presented. We evaluate several approaches for each of the three tasks and analyze which of those are promising in their results and can provide directions for future work. It should be pointed out here that there are algorithmic approaches which tackle more than one of the tasks in a combined way (Klapuri et al., 2006, e.g.). We will report the accuracy of individual tasks for such systems in our experiments as well. Further, it is also to be noted that we use only audio and its associated metadata in these tasks, because none of the available methods is capable of combining audio processing with the several other cues that were specified in Section 3.1.2.

## Datasets for evaluation

The recordings used for evaluation in this section are a subset of the bigger CompMusic collection that is described in detail in Chapter 4. The CompMusic collection is a comprehensive collection representative of Indian art music, and in the context of this section, we use only a subset of the audio collection and the associated rhythm metadata from commercially available releases. The audio recordings are short clips extracted from full length pieces.

In order to evaluate the algorithms, we need collections of audio recordings that are annotated in various aspects. For cycle length recognition we only need high-level information about the *tāla*, which decides the length of the *tāla*. For the tasks of downbeat tracking however, we need low-level annotations that specify the alignment between organized pulsation and music sample. Because no such annotated music collection is available, a collection of samples had to be manually annotated. As the process of manual annotation is very time consuming, we decided to compile a bigger set of recordings with high-level annotation and selected a smaller set of recordings for the evaluation and downbeat tracking.

For both Hindustani and Carnatic music, recordings from four popular *tālas* were selected for evaluation. The Carnatic dataset has 61, 63, 60, and 33 pieces in *ādi*, *rūpaka*, *mīśra chāpu*, and *khaṇḍa chāpu tāla*s, respectively. The Hindustani dataset has 62, 61, 19, and 15 pieces in *tīntāl*, *ēktāl*, *jhaptāl*, and *rūpak tāl*, respectively. The Hindustani dataset has compositions in three *lay* classes - *vilāribit*, *madhya* and *dṛ̥t*. In the datasets, the pieces are 2 minute long excerpts sampled at 44100 Hz. Though the audio recordings are stereo, they are down-mixed to mono since none of the algorithms evaluated in this study make use of additional information from stereo audio and primarily work on mono. They include instrumental as well as vocal recordings. The *tāla/tāl* annotation of these pieces were directly obtained from the accompanying editorial metadata contained in the CompMusic collection.

The downbeat recognition task is evaluated only on Carnatic dataset, and *ādi* and *rūpaka tāla*. Thirty two examples in *ādi tāla* and thirty four examples in *rūpaka tāla* of Carnatic music have beat and sama instants manually annotated, and we refer to the Carnatic low-level-annotated dataset. Similar to the Carnatic dataset, Carnatic low-level-annotated dataset also consists of two minute long excerpts. All annotations were manually done using Sonic visualizer (Cannam, Landone, & Sandler, 2010) by tapping along to a piece and then manually correcting the annotations.

### 3.4.1 Cycle length estimation

The algorithms which will be evaluated for cycle length estimation can be divided into two substantially different categories. On one hand, we have approaches that examine characteristics in the surface rhythm of a piece of music, and try to derive an estimate of cycle length solely based on the pulsations found in that specific piece - called self-contained approaches in this section. The self-contained approaches evaluated are:

**GUL algorithm:** The meter estimation algorithm proposed by Gulati et al. (2012) that focused on music with a regular divisive meter. Further, the algorithm only considers a classification into double, triple, or a septuple meter. Therefore, we had to restrict the evaluation to those classes that are based on such a meter. For Carnatic

music, *ādi*, *rūpaka*, and *miśra chāpu tālas* have a double, triple and septuple meter respectively. In Hindustani music, *tīntāl*, *ēktāl*, and *rūpak tāl* were annotated to belong to double, triple and septuplet meter classes.

**PIK algorithm:** The time signature estimation algorithm proposed by Pikrakis et al. (2004). The approach presents two different diagonal processing techniques and we report the performance for both methods (Method-A and Method-B). As suggested by Pikrakis et al., we also report the performance using a combination of the two methods.

**KLA algorithm:** The meter analysis algorithm by Klapuri et al. (2006) can be used for cycle length recognition task by using the bar, beat, and subdivision interval durations. Ideally, dividing the inter-downbeat interval by the inter-beat interval should present us with the bar length in beats. However, we explore the use of the bar-beat, beat-subdivision, and bar-subdivision interval relations to estimate cycle length and evaluate how well they coincide with the known cycle lengths of a piece.

**SRI algorithm:** Similar to KLA algorithm, the long term periodicity and the sub-beat structure estimated by the algorithm proposed by Srinivasamurthy et al. (2012) can be used for cycle length recognition, and we explore the use of the bar-beat, beat-subdivision, and bar-subdivision interval relations to estimate cycle length. For the present evaluation, the tempo estimation in the algorithm, which is adapted from Davies and Plumbley (2007), is modified to peak at 90 BPM. Further, the tempo analysis was modified to include a wide range of tempi (from 20 BPM to 180 BPM).

On the other hand, there are rhythmic similarity approaches that can give an insight into the rhythmic properties of a piece by comparing with other pieces of known rhythmic content. To this end, we will use the music collections that contain pieces with known cycle lengths. There, we can determine the rhythmic similarity of an unknown piece to all the pieces in our collection. We can then assign a cycle length to the piece according to the observation of the cycle lengths of other similar pieces. The approaches based on

rhythm similarity measures (called Comparative approaches in this section) evaluated are:

**OP:** The approach proposed by Pohle et al. (2009) that uses Onset Patterns (OP) as the rhythm similarity measure.

**STM:** The approached proposed by Holzapfel and Stylianou (2011) that uses Scale Transform Magnitudes (STM) as the rhythm similarity measure.

### Evaluation criteria

For comparative approaches, we apply a 1-nearest-neighbor classification in a *leave-one-out* scheme, and report the accuracy for a dataset. For self-contained approaches, we examine the accuracy of the outputs obtained from various algorithms.

We note that the algorithms **PIK** and **GUL** consider short time scales for cycle lengths and may track cycles of shorter length than the measure cycle. Hence, as explained in Section 3.1.1, the algorithms may track meter at a subdivision level. As the algorithms are not specifically designed to perform the task of cycle length recognition as defined in Section 3.2.2, the evaluation has to be adapted to the algorithms. For example, **GUL** classifies the audio piece into three classes - duple, triple, and septuple meter. For this reason, samples in the the dataset are labeled as being duple, triple or septuple based on the **tāla** for evaluating **GUL**. Rhythm classes in the datasets that do not belong to any of these categories are excluded from evaluation.

We are primarily interested in estimating the cycle length at the **āvart/āvartana** level for Indian music and at the **usul** level in Turkish music, a problem related to estimating the measure length in Eurogenetic music. However, as explained in Section 3.1.1, cycles may exist at several metrical levels, with especially Carnatic **tālas** having equal subdivisions at lower metrical levels in many cases. In connection with the fact that the measure cycles might extend over a long period of time, these shorter cycles contribute an important aspect to forming what can be perceived as beats. For the evaluations on Carnatic music in this section, we will refer to the subdivision meter and the cycle length as given in Table 2.1.

Since there is no well-defined subdivision meter in Hindustani music, we will refer to only the cycle length in number of *mātrās* from Table 2.3.

For *KLA* and *SRI* algorithms we report the accuracy of estimating the annotated cycle length at the Correct Metrical Level (CML). We also report the Allowed Metrical Levels (AML) accuracy considering cycle length estimates by the algorithms to be correct that are related to the annotated cycle length by a factor of 2 or 0.5, which is referred to as doubling or halving, respectively. For cycle lengths which are odd we only consider doubling of cycle length estimates in AML. Halving and doubling of cycle lengths can be interpreted as estimating sub-cycles and supra-cycles related to the annotated cycle length by a multiple, and can provide insights on tempo estimation errors committed by the algorithms. Though the *tāla* cycle is an important part of rhythmic organization, it is not necessary that all phrase changes occur on the *sama*. In *ādi tāla* for example, most of the phrase changes occur at the end of the 8 beat cycle, there are compositions where some phrase changes and strong accents occur at the end of half-cycle or the phrase might span over two cycles (16 beats). Hence, in this case a cycle length of 4, 8, or 16 would be acceptable, depending on the composition. This needs to be considered when we evaluate the performance of algorithms.

### **Self-contained approaches**

We differentiated between self-contained and comparative approaches, and the self-contained approaches are divided into two types of methods. The first type attempts to estimate the meter or the time signature based on repetitions observed in the signals, while the second type aims at tracking the pulsations related to those repetitions. We will start our evaluations with methods that belong to the first type (*GUL*, *PIK*), and evaluate then the tracking methods (*KLA*, *SRI*).

Table 3.1 shows the accuracies for the three datasets, using the types of rhythms which can be processed by the algorithm. The performance on Carnatic music is better than the performance on Hindustani music. A detailed analysis revealed that the performance on *rūpaka* (ternary) is only 65.08%, which leads to considerable de-

Dataset	Accuracy (%)
Carnatic (without <i>khaṇḍa chāpu</i> )	75.27
Hindustani (without <i>jhaptāl</i> )	49.30

**Table 3.1:** Performance of meter estimation using *GUL* algorithm.

Dataset	Method-A	Method-B	Combined
Carnatic	52.53	49.30	64.06
Hindustani	35.67	53.50	57.96

**Table 3.2:** Performance of cycle length estimation using *PIK* algorithm. The Method-A and Method-B refer to the two methods suggested by Pikrakis et al. (2004). All values are in percentage.

crease in the performance on Carnatic music. This poorer performance can be attributed to the ambiguity between duple and triple meter that is an intrinsic property of this *tāla* (see Section 3.1.1). Furthermore, the performance on Hindustani music was found to be poor on *rūpak tāl* and *ēktāl* while the performance on just *tīntāl* is 80.64%. This can be attributed to the fact that there are very long cycles in Hindustani music in *vilambit lay*, where the long subdivision time-spans restrains the algorithm from a correct estimation. In most of such cases in *ēktāl* and *rūpak tāl*, the estimated meter is a duple meter, which might be related to the further division of the *mātrās* using filler strokes.

Pikrakis algorithm (*PIK*) looks for measure lengths between 2 and 12. We report the accuracy accepting an answer if it is correct at one of the metrical levels. For example, for *ādi tāla* and *tīntāl*, 4/4, 8/4, 4/8, 8/8 are all evaluated to be correct estimates, because 4 is the subdivision meter, and 8 is the length of the *āvartana* (cycle length). Further, the algorithm outputs an estimation for every 5 second frame of audio, and therefore time signature of a song is obtained by using a majority vote for a whole song. The performance is reported as the accuracy of estimation (% correctly estimated) for both the diagonal processing methods (Method-A and Method-B) in Table 3.2. As suggested by Pikrakis et al., we also use both methods to combine the decision and it improves the performance, as can be seen from the table. The performance on Carnatic music

is better than that on Hindustani music. Though the performance on Hindustani dataset is poor, further analysis shows that for *tīntāl*, the accuracy is 74.19%. PIK algorithm performs better in the cases where the meter is a simple duple or triple, while the performance is worse with other meters. For example, *miśra chāpu* (length 7) has an additive meter and the cycle can be visualized to be a combination of 3/4 and 4/4. On that class the PIK algorithm estimates most of *miśra chāpu* pieces to have either a 3/4 meter or a 4/4 meter.

To evaluate the tracking methods, we can compare the pulsations estimated by the algorithms with the ground truth annotations at all three metrical levels to determine if the large possible tempo ranges cause the beat to be tracked at different levels of the meter. From the estimates obtained from KLA for downbeats, beats and subdivision pulses on a specific piece, we define the following time-spans: let  $T_c$  denote the median cycle duration (inter-downbeat interval),  $T_b$  the median beat duration, and  $T_a$  the median subdivision duration. We use a different terminology compared for these as compared to  $\tau_s$ ,  $\tau_b$ , and  $\tau_o$  defined in Section 3.3.1 to highlight the difference that these approaches evaluated here were not specifically designed for Indian art music. We then compute the cycle length estimates as,

$$L_{cb} = \left\lfloor \frac{T_c}{T_b} \right\rfloor \quad L_{ca} = \left\lfloor \frac{T_c}{T_a} \right\rfloor \quad L_{ba} = \left\lfloor \frac{T_b}{T_a} \right\rfloor$$

where  $\lfloor . \rfloor$  indicates rounding to the nearest integer. We examine which of the three estimates more closely represents the cycle length. We report both the CML and AML accuracy of cycle length recognition. Table 3.3 shows the recognition accuracy (in percentage) of KLA algorithm separately for  $L_{cb}$ ,  $L_{ca}$ , or  $L_{ba}$  as the cycle length estimates.

We see in Table 3.3 that there is a large difference between CML and AML performance, which indicates that in many cases tracked level is related to the annotated level by a factor 2 or 1/2. We also see that for Hindustani music, the cycle length is best estimated using  $L_{ca}$ , with the CML accuracy being very low or zero when we use the other cycle length estimates instead. As discussed earlier, in Hindustani music, the cycle length is defined as the number of *mātrās* in the cycle. However, in the case of *vilārbīt* pieces, the *mātrās* are longer than the range of the tatum pulse time-span

Dataset	CML (%)			AML (%)		
	$L_{cb}$	$L_{ca}$	$L_{ba}$	$L_{cb}$	$L_{ca}$	$L_{ba}$
Carnatic	11.06	8.76	4.15	34.10	45.16	25.81
Hindustani	0.00	25.4	-	45.22	46.50	-

**Table 3.3:** Accuracy of cycle length recognition using [KLA](#) algorithm. Subdivision meter ( $L_{ba}$ ) in Hindustani music is not well-defined and hence omitted.

Dataset	CML (%)			AML (%)		
	$L_{cb}$	$L_{ca}$	$L_{ba}$	$L_{cb}$	$L_{ca}$	$L_{ba}$
Carnatic	3.69	0.46	6.45	40.55	50.69	14.28
Hindustani	14.64	9.55	-	43.95	55.41	-

**Table 3.4:** Accuracy of cycle length recognition using [SRI](#) algorithm. Subdivision meter ( $L_{ba}$ ) in Hindustani music is not well-defined and hence omitted.

estimated by the algorithm and hence the performance is poor. Interestingly, we see a good performance when evaluated with  $L_{cb}$  only with [tīntāl](#), which resembles the Eurogenetic 4/4 meter, with an AML accuracy of 88.71% in spite of the CML accuracy being zero. In fact, it is seen that  $L_{cb}$  is always four in the case of a correct estimation (AML), which is the estimate of the number of [vibhāgs](#) in the [tāl](#). Further, it follows from Klapuri et al. (2006, Figure 8) that relation between neighboring levels in [KLA](#) cannot be larger than 9, which implies longer cycle length estimates (as needed by e.g. [ēktāl](#) or [tīntāl](#)) could possibly appear only in the  $L_{ca}$  length.

The CML accuracy in Carnatic dataset with  $L_{cb}$  is hence better than the other cycle length estimates, showing that [KLA](#) tracked correct tempo in a majority of cases in Carnatic music. However, the performance is poor because the algorithm often under-estimates the cycle length. Further, in [tālas](#) of Carnatic music that have two akṣaras in a beat ([khaṇḍa chāpu](#) and [miśra chāpu](#)),  $L_{ca}$  is a better indicator of the cycle length than  $L_{cb}$ , since akṣaras are closer to the estimated subdivision duration. In general,  $L_{ba}$  performs poorly compared to  $L_{ca}$  or  $L_{cb}$ , which is not astonishing since the cycle

lengths we are looking for are longer than the estimated subdivision meter. Summing up, none of the estimated meter relations can serve as a robust estimate for the *āvartana* cycle length.

[SRI](#) algorithm estimates the cycle length at two metrical levels using the beats tracked by Ellis (Ellis, 2007) beat tracker, one being at the cycle level (bar length in beats), and the second at the beat level (subdivision meter, or *nađe*). The algorithm computes a list of possible candidates for the subdivision meter and bar length, ordered by a score. We consider the top candidate in the list and compute the cycle length estimates  $L_{cb}$ ,  $L_{ba}$ , and the  $L_{ca}$ , assuming that the beats tracked by Ellis beat tracker correspond the beat duration  $T_b$ . Similar to [KLA](#) algorithm, we present the CML and AML accuracy of performance in Table 3.4.

We see that there is large disparity between the CML and AML accuracy, which indicates that the beat tracker and the correct beat are related by a factor of 2 or 1/2. In general, the algorithm performs poorly, which can be mainly attributed to errors in tempo and beat tracking. The tempo estimation uses a weighting curve that peaks at 90 beats per minute, which is suitable for Carnatic music, but leads to an incorrect estimation of cycle length for Hindustani music. A beat tracking based approach as the [SRI](#) algorithm might in general not be well suited for Hindustani music which often includes long cycles that might be more reflected in the structure of melodic phrases than in pulsation and rhythmic aspects.

The poor performance on Carnatic music can in part be also attributed to variation in percussion accompaniment, which is completely free to improvise within the framework of the *tāla*. Further, the algorithm is based on the implicit assumption that beats at the same position in a measure cycle are similar between various recurrences of the cycle. For certain music pieces where there are no inherent rhythmic patterns or the patterns vary unpredictably, the algorithm gives a poorer performance. For Carnatic music, the algorithm specifically estimates the subdivision-meter (*nađe*), as the number of *akṣaras* per beat. Using  $L_{ba}$  as an estimate of the *nađe*, we obtain a reasonably good performance comparable to [GUL](#) with an accuracy of 39.63% and 79.72% at CML and AML (of the subdivision meter), respectively. We see that a reasonable performance when demanding an exact numerical result for the meter (CML) is only reached for the *nađe* estimation in Carnatic music.

Dataset	OP (%)	STM (%)
Carnatic	41.0	42.2
Hindustani	47.8	51.6

**Table 3.5:** Accuracy of cycle length recognition using comparative approaches

We observe that the duration of cycles in seconds is often estimated correctly, but the presence or absence of extra beats causes the estimated length in beats to be wrong. Ellis beat tracker is sensitive to tempo value and cannot handle small tempo changes effectively. This leads to addition of beats into the cycle and the cycle length in many cases were estimated to be one-off from the actual value, though the actual duration of the cycle (in seconds) was estimated correctly.

## Comparative approaches

The comparative approaches are based on a description of periodicities that can be derived from the signal without the need to perform meter tracking. Performances of the two evaluated methods, **OP** and **STM**, is the average accuracy in a 1-nearest neighbor classification. It tells us how often a piece found to be most similar to a test piece belongs actually to the same class of rhythm as the test piece. The results of this classification experiment are depicted in Table 3.5. It is apparent that the comparative approaches lead to a performance significantly better than random, which would be 25% for our compiled four-class datasets. In fact, accuracies are in the same range as the results of the **PIK** algorithm, with **PIK** performing better on Carnatic music (64.1% instead of 42.2%). This might indicate the potential of combining self-contained and comparative approaches, because none of the approaches evaluated for cycle length recognition provides us with a sufficient performance for a practical application.

### 3.4.2 Downbeat tracking

So far mainly music with a 4/4 time signature was focused upon in evaluations, usually in the form of collections of Eurogenetic popular and/or classical music. Hence, we will address the questions if such approaches can cope with the lengths of cycles present in our data and if Indian art music poses challenges of unequal difficulty. The approaches evaluated are:

**DAV algorithm:** The algorithm proposed by Davies and Plumbley (2006) that assumes that percussive events and harmonic changes tend to be correlated with the downbeat.

**HOC algorithm:** The algorithm proposed by Hockman et al. (2012) for downbeat tracking in hardcore, jungle, and drum and bass genres of music.

It is apparent that both systems are conceptualized for styles of music with notable differences to Indian art music. The system by Davies and Plumbley (2006) is mainly sensitive to harmonic changes, whereas Indian art music does not incorporate a notion of harmony similar to the Eurogenetic concept of functional harmony. On the other hand, the system by Hockman et al. (2012) is customized to detect the bass kick on a downbeat, which will not occur in the music we investigate here. As the latter system contains this low-frequency feature as a separate module, we will examine the influence of the low-frequency onsets and the regression separately our experiments.

## Evaluation results

The evaluation metrics we use are the same as the continuity-based approach applied by Hockman et al. (2012). This measure applies a tolerance window of 6.25% of the inter-annotation-interval to the annotations. Then it accepts a detected downbeat as correct, if

1. The detection falls into a tolerance window.
2. The precedent detection falls into the tolerance window of the precedent annotation.
3. The inter-beat-interval is equal to the inter-annotation-interval (accepting a deviation of the size of the tolerance window).

Method	ādi (8)	rūpaka (3)
DAV	21.7	41.2
HOC-SVM	22.9	42.1
HOC	49.9	64.4

**Table 3.6:** Accuracy of downbeat tracking on Carnatic music. The cycle lengths are indicated in parentheses next to the *tāla*.

In Table 3.6, we depict the downbeat recognition accuracies (in %) for the two systems. The results are given separately for each of the two *tālas* in the Carnatic low-level-annotated dataset. The HOC algorithm was applied with and without emphasizing the low-frequency onsets, denoted as HOC and HOC-SVM, respectively. The DAV algorithm has the lowest accuracies for all presented *tālas*. This is caused by the focus of the method on changes in harmony that is related to chord changes - concepts not present in Indian art music. However, the results obtained from HOC are more accurate and allows for an interesting conclusions that taking onsets in the low-frequency region into account improves recognition for all contained rhythms. However, Carnatic music with its wide rhythmic variations and its flexible rhythmical style seems to represent a more difficult challenge for downbeat recognition, with the range of accuracy smaller than that reported for electronic dance music (Hockman et al., 2012). Pieces without such phenomenal cues are very likely to present both automatic systems and human listeners with a more difficult challenge when looking for the downbeat. Furthermore, the accuracies depicted in Table 3.6 can only be achieved with known cycle length, and correctly annotated beats, which is a big limitation.

### 3.4.3 Discussion

We summarize and discuss the key results of the evaluation. The results provide us with useful insights to indicate promising directions for further work. At the outset, the results indicate that the performance of evaluated approaches is not adequate for the presented tasks, and that methods that are suitable to tackle the culture specific challenges in computational analysis of rhythm need to be

developed.

Cycle length estimation is challenging in Indian art music since cycles of different lengths exist at different time-scales. Although we defined the most important cycle to be at the *āvart* and *usul* level, the other cycles, mainly at the beat and subdivision level, also provide useful rhythm related information. The evaluated approaches *PIK* and *GUL* estimate the subdivision meter and time signature. This is possible to an acceptable level of accuracy, when restricting to a subset of rhythm classes with relatively simple subdivision meters. Though they do not provide a complete picture of the meter, they estimate the underlying metrical structures at short time scales and can be used as pre-processing steps for estimating longer and more complex cycles.

Both *SRI* and *KLA* aimed to estimate the longer cycle lengths but show a performance that is inadequate for any practical application involving cycle length estimation. Tempo estimation and beat tracking have a significant effect on cycle length estimation, especially in the self-contained approaches and also needs to be explored further. The comparative approaches show that the applied signal features capture important aspects of rhythm but are not sufficient to be used standalone for cycle estimation. A combination of self-contained and comparative approaches might provide useful insights in rhythm description Indian art music through mutual reinforcement. Developing methods that take the culture-specific properties of rhythm into account is therefore necessary to proceed towards a reliable computational rhythm analysis.

Downbeat tracking, *i.e.* the estimation of the instant of beginning of the bar was explored using *HOC* and *DAV* algorithms. The downbeat detectors evaluated here needed an estimation of beats and the cycle length of the piece, which are difficult to estimate presently. Since downbeat information can help in estimating the cycle length and also beat tracking, an independent approach to downbeat tracking will be very useful. A joint estimation of the beat, cycle length, and downbeat might be a potential solution since each of these parameters are mutually useful for estimating the others. A combination of bottom up and top down knowledge based approach which performs a joint estimation of these parameters is to be explored further, using models that better represent the underlying metrical structures.

Long āvart cycles is a significant challenge in Hindustani music. For tāls with a very long cycle duration, estimating the correct metrical level is essential and methods that aim at tracking short time-span pulsation will be not adequate due to the grouping structure of the glstaal. With a wide variety of rhythms, coupled with the perceptual edupu, Carnatic music poses a difficult challenge in sama tracking. Since there is no time adherence to a metronome, tempo drifts are common and lead to small shifts in the sama instants.

For estimating the components of meter from audio, we need signal descriptors that can be used to reliably infer the underlying meter from the surface rhythm in audio. The availability of such descriptors will greatly enhance the performance of automatic annotation algorithms. At present, we have suitable audio descriptors for low level rhythmic events such as note onsets and percussion strokes, but better descriptors for higher level rhythmic events is necessary.

The inadequate performance of the presented approaches leads us to explore the specific problems more comprehensively. It also motivates us to explore varied and unconventional approaches to rhythm analysis. Though we considered beat tracking, cycle length estimation and downbeat tracking as separate independent tasks, it might be better to consider a holistic approach and build a framework of methods in which the performance of each element can be influenced by estimations in another method. Ironically, we see from the HOC algorithm, a downbeat detector for electronic dance music, that sometimes the most rigid specialization leads to good performance on apparently completely different music. Thus, it still remains an open question if we need more specialist approaches, or more general approaches that are able to react to a large variety of music. Generally, it appears desirable to have generic approaches that can be adapted to a target music using machine learning methods, that can adapt flexibly to the underlying rhythmic structures.

**A chapter summary to be added here**



# Data corpora for research

Data is a precious thing and will last longer than the systems themselves.

---

Tim Berners-Lee, inventor of the World Wide Web

**Todo: Fix the glossary issues for this chapter**

**Todo: Move pattern figures closer to the text describing them**

Computational approaches in MIR need data for developing algorithms and for testing approaches. A carefully designed data collection is critical for the success of these approaches. To develop such MIR approaches and advance knowledge, there is a need for research corpora that can be considered authentic and representative of the real world.

A research corpus is an evolving collection of data that is representative of the domain under study and can be used for relevant research problems. A good data corpus includes data from multiple sources and can even be community driven. In the context of MIR, since it's practically infeasible to work with the whole universe of music, a research corpus acts as a representative subset for research. Hence, algorithms and approaches developed and technologies demonstrated on the research corpus can be assumed to generalize to real world scenarios.

A test corpus or a test dataset is often a subset of the research corpus, possibly with additional metadata for use in a specific research task. In experiments, test corpora are used to develop tools, and to evaluate and improve their performance. Computational approaches are developed using these datasets and then extended to the research corpus. Hence test corpora can even consist of synthetic data that can be used for testing. Unlike the research corpus, a test corpus is fixed for use in a specific experiment. A test corpus can evolve, but each version of the dataset used in a specific experiment is retained for better reproducibility of research results.

Building a research corpus itself is a research problem and has been studied in many fields such as linguistics, speech and biomedical language processing (Wynne, 2005; Pan & Weng, 2002; Cohen, Ogren, Fox, & Hunter, 2005). There are also many central repositories of corpora such as the Linguistic Data Consortium (LDC) by Liberman and Cieri (1998)<sup>1</sup> for language resources and PhysioBank<sup>2</sup> for physiological signals. Other open repositories of data such as MusicBrainz<sup>3</sup> or Wikipedia themselves can be used as research corpora for different MIR related tasks.

There have been efforts to compile large collections of music related data, e.g. the Million Song Dataset (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011), which is a good research corpus for several MIR tasks on contemporary popular music. However, despite the importance of a good research corpus in MIR, the problem of building it has received little attention by the research community. There have been no studies on a systematic way to compile and curate a research corpus. Recently, Peeters and Fort (2012) presented a unified way to describe annotated MIR test datasets. Serra (2014) elucidated a set of design principles to build and compile a research corpus, based on a set of primary considerations such as Purpose, Coverage, Completeness, Quality and Reusability. We use these primary considerations to develop a corpus for MIR in Indian Art Music.

In this chapter, we address some of these concerns and focus on a systematic compilation and analysis of data for research. The cri-

---

<sup>1</sup><https://www.ldc.upenn.edu/>

<sup>2</sup><http://www.physionet.org/physiobank/>

<sup>3</sup><http://musicbrainz.org/>

teria and the evaluation methodology discussed here can be used to systematically build a representative and comprehensive research corpora and test datasets. Our primary focus in the chapter would be on Indian Art Music, while other test datasets that are relevant to the thesis are also presented and discussed. The main aims of the chapter are:

1. To describe and discuss the research corpora and the test datasets that have been built as a part of CompMusic, relevant for this thesis - emphasizing on the research problems and tasks in which these datasets can be used. In addition, other state of the art datasets that are used in the thesis are also presented in brief for completeness.
2. To present a systematic framework and elucidate a set of design principles to curate and compile a research corpus, and then use those principles to illustrate a methodology to measure the goodness of the Carnatic and Hindustani research corpora.
3. To present corpus level statistical analyses of relevant test datasets, to see if we can draw musically meaningful inferences from those analyses.

As we described earlier, the research corpora are growing entities through continued efforts. Hence, the numbers presented are only indicative and are of secondary importance. We primarily emphasize on presenting a scientific approach to develop a corpus and evaluate its suitability for a particular set of research tasks. We emphasize on methodologies that can be used to evaluate a corpus on the aspects of coverage and completeness. Apart from the description of the corpora, a methodology for evaluation of the corpus is an important contribution of this chapter. We further note that in addition to the sources described in this article, there are several other sources that can be used for computational research in Indian Art Music, and eventually could be a part of the corpus.

## 4.1 CompMusic research corpora

Musics of the world might share some basic concepts such as melody and rhythm, but some salient aspects can be described completely

only by considering the specificities of that music culture. For such studies, in the context of the CompMusic project, Serra (2011) emphasized the need for culture specific research corpora to develop approaches that utilize the important aspects of the music culture.

Working with five music traditions of the world, the data driven methodologies in CompMusic primarily involve signal processing, machine learning and semantic web technologies. Hence, there has been a significant effort towards the design and compilation of research corpora for relevant problems in the music cultures being studied. This effort complements the primary aim of CompMusic, which is to build culture specific computational methodologies for better exploration of music collections through meaningful music concepts and automatically extracted melody, rhythm and semantic descriptors.

In this chapter, we focus mainly on Indian art music. The Turkish music research corpus has been presented in detail by Atlı, Uyar, Şentürk, Bozkurt, and Serra (2014) and Caro and Serra (2014) have described the Beijing Opera *jīngjù* research corpus comprehensively. We first discuss the criteria for creating research corpora, and then describe the Carnatic and Hindustani music research corpora.

### 4.1.1 Criteria for creation of research corpora

Serra (2011) listed the primary criteria for creating research corpora, which are described in brief here.

**Purpose** A research corpus is built for a specific purpose and it is necessary to define the research problem(s) and the approaches that will be used. In CompMusic, we wish to develop methodologies to extract musically meaningful music features from audio recordings, mainly related to melody and rhythm. The research corpus has to be aligned to this purpose.

**Coverage** The coverage of a corpus is a measure of representativeness of the corpus with respect to several relevant concepts that we wish to study. For our quantitative approach, we need sufficient samples of each instance for the data to be statistically representative and significant. For rhythm analysis, we

need to have audio recordings, plus appropriate accompanying metadata covering different rhythms and metrical structures present in the music culture.

**Completeness** Completeness refers to the completeness of the accompanying metadata for each audio recording. Since the research corpus contains data from many different sources, ensuring completeness of audio and metadata is important for its use for different research tasks.

**Quality** The data in the corpus needs to be good quality, the audio needs to be well recorded and the accompanying metadata must be accurate, obtained from reliable sources and validated by experts. The manual and automatic annotations on audio files must be carefully done and verified independently.

**Reusability** The reusability of research corpora and datasets and reproducibility of research results is necessary for continued and sustainable research using these datasets, with the effect of improving the research corpora and research results. Reusability can be addressed by emphasizing the use of open sources of information, and providing a platform for easy access of data for research. For editorial metadata, we use MusicBrainz.

All the music cultures under study can be described in terms of musical concepts, music content and the music community. The elements of the corpora can be associated with one or more of these categories and hence useful for computational tasks in these three aspects. Central to each of the corpus is an audio music recording with its metadata. We first present the Carnatic Music research corpus followed by the Hindustani music corpus. All audio in both the corpora are stereo recordings sampled at 44.1 kHz and stored as 160 kbps mp3 files for ease of transmission and storage.

### 4.1.2 Carnatic music research corpus

The Carnatic music research corpus mainly comprises of audio recordings, its associated editorial metadata, lyrics, scores, contextual information on music concepts, and community (social) information

from online music forums and other sources. Audio recordings, editorial metadata, scores, and lyrics are the content used by signal processing and machine learning approaches. Contextual information and the forum discussions form the music concepts and community information used for semantic analysis.

There are several considerations in collecting a corpus of Carnatic music. Given that a *kachēri* (concert) is the natural unit of Carnatic music and the main unit of music distribution, most commercial releases are concerts, comprising of several pieces that are improvised renderings of compositions. Vocal music is predominant and even in instrumental music, the lead artist aims to mimic vocal singing. The *rāga* and *tāla* are the most important metadata associated with a composition and hence a recording of the composition.

Based on these considerations, we consulted expert musicians and musicologists, such as T M Krishna<sup>4</sup> to arrive at a representative collection of Carnatic music audio. The main institutional reference for Carnatic music is the **Madras Music Academy (MMA)**<sup>5</sup>, which is a premier institution dedicated to Carnatic music and organizes the annual music conference in Chennai, India. The annual Carnatic music festival is one of the largest music festivals in the world, with a significant part of the Carnatic music community taking part in it. The **MMA** has been driving scholarly research and opinion in Carnatic music. The **MMA** has a panel of experts that formulates the procedure and standards for the selection of artists for the music festival. The **MMA** has been recording concerts and its archive can be considered a standard repository of Carnatic music. However, the archive is not openly available online. We thus followed the musical criteria followed by the **MMA** and procured the audio from commercially available releases. Though Carnatic music is spread across South India, the choice of **MMA** as an institutional reference has an influence on the research corpus introducing a bias towards the music scene in Chennai.

We wished to compile concerts over several generations of musicians. We started with the artists that have been performing at the **MMA** in the last five years, and then expanded the collections

---

<sup>4</sup><http://www.tmkrishna.com/>

<sup>5</sup><http://musicacademymadras.in/>

to include their teachers, and popular musicians of their era. The record label *Charsur*<sup>6</sup> specializes in Carnatic music and the core of our audio collection is from their catalog of music concerts. Hence, the corpus consists of audio from commercially available releases from Charsur and other music labels. The corpus presently consists of 248 releases(concerts) with 1650 audio recordings (346 hours) spanning 1068 compositions. The number of other relevant music entities in the corpus is described in Table 4.1 (column 2). Though we focus on concerts with vocalist leads, we also have instrumental music releases (mainly with veena, violin, flute, saxophone, and mridangam in lead). The whole audio collection is commercial and easily accessible, but is not open and distributable.

The editorial metadata associated with each release has been stored and organized in MusicBrainz. The primary metadata associated with each concert is the name of the release, the lead and the accompanying artists, and the musical instruments in the concert. For each audio recording contained in the release, the relevant metadata are the artists performed on the track, the name of the composition/s and the composer, *rāga*/s, *tāla*/s, musical form/s. MusicBrainz assigns a unique identifier (MBID) for each entity in MusicBrainz, such as the artist, composer, instrument, recording, work, and a release. This helps to organize the metadata in an effective way. All the editorial metadata was entered using Roman alphabet and a roman transliteration was used when the language of the release was not English. The *rāga* and *tāla* information have been added as work attributes.

Since Carnatic music is predominantly a vocal music tradition, lyrics play an important role. A significant part of the rendition of a composition is improvised and hence the scores associated with a composition are of limited use, nonetheless important. The lyrics and scores, even though not time aligned to audio recordings, are useful for computational analysis and hence we compiled them. The primary languages in which Carnatic music is composed are Telugu, Tamil, Kannada, Sanskrit, and Malayalam. There are several published compilations of lyrics and scores for most of the currently performed compositions, such as the ones of the three most popular composers in Carnatic music: *Tyāgarāja* (e.g. T.

---

<sup>6</sup><http://www.charsur.com/>

	<b>Corpus</b>	<b>Raaga.com</b>	<b>Kutcheris</b>	<b>Charsur</b>
Rāgas	246	489 (42%)	N/A	301 (68%)
Tālas	18	16 (100%)	N/A	21 (85%)
Composers	131	598 (17%)	N/A	256 (42%)
Artists	233	501	2978	264 (48%)

**Table 4.1:** Coverage of the Carnatic music corpus. The number in parentheses is the *overlap* measure in percentage. N/A indicates data not available.

K. Govinda Rao (2009)), Šyāmā śāstri(e.g. T. K. Govinda Rao (2003b)) and Muttusvāmi dīkṣitar (e.g. T. K. Govinda Rao (2003a)). However, these compilations are not machine readable and hence not accessible for computational analysis.

There are several good online open repositories for lyrics, such as sahityam.net<sup>7</sup>, which is a wiki of lyrics of Carnatic compositions. Sahityam.net is our primary source for machine readable lyrics. It uses a uniform scheme for transliteration to Roman script and hence has minimal ambiguity. In some cases, it provides additional commentary, references, and example renditions. Sahityam.net currently hosts lyrics for about 1820 compositions of Carnatic music. Machine readable scores are more difficult to access, with no comprehensive machine readable score compilations available. A set of machine readable (HTML, Word) scores compiled by Dr. Shiv-kumar Kalyanaraman<sup>8</sup> is the main source of scores.

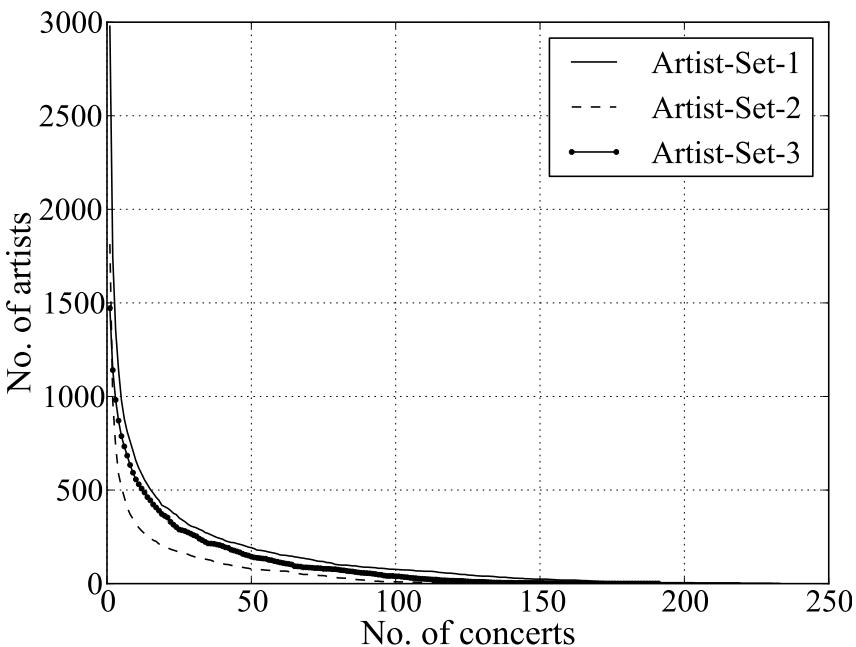
The music community and music concepts related information in the corpus form the primary source of information for semantic analysis, and come from various reliable sources on the Internet. Kutcheris.com<sup>9</sup> is an up-to-date directory of artist biographies, music venues, concerts and events. The category of Carnatic music on Wikipedia<sup>10</sup> is a source of contextual information including music concepts. We have added a lot of information and contributed to Wikipedia with the help of experts. While Wikipedia acts as an encyclopedia of music concepts providing linked information, on-

<sup>7</sup><http://www.sahityam.net>

<sup>8</sup><http://www.shivkumar.org>

<sup>9</sup><http://www.kutcheris.com>

<sup>10</sup>[http://en.wikipedia.org/wiki/Category:Carnatic\\_music](http://en.wikipedia.org/wiki/Category:Carnatic_music)



**Figure 4.1:** The number of artists by the number of their performances in the Carnatic music corpus

line music forums with discussions provide opinions from which some of these links can be inferred. The rasikas.org<sup>11</sup> Carnatic music forum is an active forum of Carnatic music listener community with useful discussions about Carnatic music concepts, concerts, and performances. It is an important source of data useful for community profiling.

## Coverage

A research corpus needs to be representative of the real world in the concepts that are primary to the music culture. The aim of a coverage analysis is to estimate the comprehensiveness of the corpus with respect to another representative reference source. For Carnatic music, a coverage analysis is presented for artists, *rāgas*, *tālas*, and composers. For artist coverage, we chose to use Kutcheris.com as the primary reference since it is up-to-date with current artists and their performances. We use the last five years of their concert

---

<sup>11</sup><http://www.rasikas.org/>

listings. Many of the artists and the concerts listed on Kutcheris.com are from Chennai. Charsur's release catalog provides information about *rāgas*, *tālas*, composers and artists. Raaga.com<sup>12</sup> is an Indian music streaming service and its Carnatic channel is another reference for *rāgas*, *tālas*, composers and artists. However, Raaga.com has many light music forms included in its Carnatic channel, some of which we have consciously excluded from our corpus. Hence it is to be noted that numbers and the analysis with Raaga.com will have an adverse influence from these other included music forms. The data from each of these reference sources was crawled from their online catalogues. The data from raaga.com was crawled in March, 2012 and from the others in March, 2014. We observed that nearly every source had duplicate entities mostly arising due to spelling variations (e.g. Tyagaraja, Tyaagaraaja). We merged the duplicates by matching the longest common subsequence in the strings and by using Damerau-Levenshtein distance.

Table 4.1 shows the coverage of the Carnatic corpus in comparison to the references. For each music entity  $i$ , we define a coverage measure called the *overlap* ( $\Theta$ ) as,

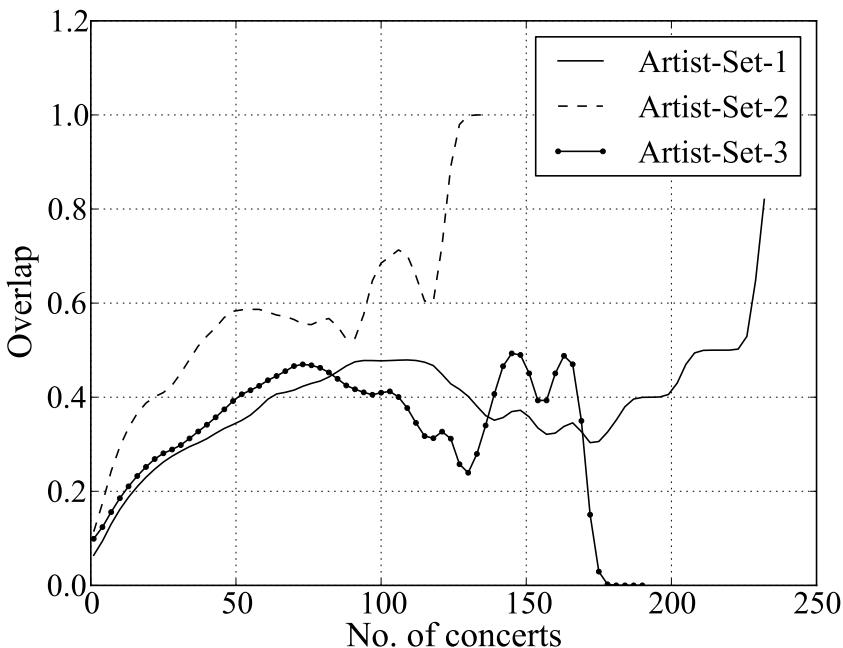
$$\Theta_i^j = \frac{|\varsigma_i^c \cap \varsigma_i^j|}{|\varsigma_i^j|} \quad (4.1)$$

where  $\Theta_i^j$  is the *overlap* measure of the entity  $i$  with reference  $j$ ,  $\varsigma_i^c$  is the set of entities in the corpus,  $\varsigma_i^j$  is the set of entities in the reference, and  $|\varsigma|$  denotes the cardinality of a set  $\varsigma$ . An *overlap* of 100% is achieved if all the elements in the reference set are present in the corpus. Table 4.1 shows the *overlap* measure for *rāgas*, *tālas* and composers for both Raaga.com and Charsur. We can see that there is a good coverage of *tālas* and a satisfactory coverage of *rāgas* in the corpus. A good coverage of *tālas* is necessary for rhythm analysis. The composer coverage with respect to Raaga.com is poor since it includes the light music composers in its set of composers.

Among the 233 artists who have at least one recording in the corpus, 74 are lead artists (lead vocal or lead instrumental). Further, we have 28 violin accompanying artists and 48 unique percussion artists in the corpus. The concerts listed by Kutcheris.com span the

---

<sup>12</sup><http://www.raaga.com>



**Figure 4.2:** Coverage of Carnatic artists. The ordinate is the overlap value of the set of artists in corpus, compared against a set of artists in Kutcheris.com who have performed in at least as many concerts as the abscissa.

whole year and all through the day. However, the evening concerts are more recognized, and we took it to be a measure of popularity of the artists. Moreover, the evening concerts during the music season lasting from November to January are ticketed. For a coverage analysis, we thus consider three categories of artists: Artists-Set-1 (all the artists), Artists-Set-2 (artists who have performed in the evening concerts, through the year) and Artists-Set-3 (artists who have performed in evening concerts between November and January). Of the 2978 total artists present in Set-1 on Kutcheris.com concert listings, there are 1814 artists in Set-2 and 1472 artists in Set-3.

The number of concerts performed by each artist is also an indicator of popularity. Though there are a large number of artists in Kutcheris, we see that the distribution of the number of concerts they have performed is exponential (Figure 4.1), e.g. there are only about 200 artists who have over 50 concerts. Hence to

Accompanying metadata	#Recordings	% of total
Lead artist	1650	100.0
Accompanying artists	1221	74.0
Rāga	959	58.1
Tāla	917	55.6
Work (Composition)	989	59.9

**Table 4.2:** Completeness of the Carnatic music corpus, showing the number of recordings in which the corresponding metadata is available.

capture this fact, we used the set of artists in the corpus and computed the *overlap* as defined in Eq. 4.1 through different subsets of artists in Kutcheris.com, sweeping over the number of concerts (at least) they have performed.

Figure 4.2 shows the *overlap*, using a set of artists that have performed at least as many concerts as the number shown on the abscissa. The *overlap* is also shown for the three categories of artists we discussed before. We can see that the *overlap* increases as we consider more frequently performing artists and becomes almost constant. The artists who have performed the most concerts are often the accompanying artists, and are few in number, which explains why the *overlap* becomes a constant, when we discount the *overlap* for more than 150 concerts. When we consider a large number of concerts, the *overlap* values are unreliable since the number of artists is less. In general, we can see that the *overlap* is better for Artists-Set-2 than Artists-Set-1 and Artists-Set-3, showing that the corpus has more representation of artists from evening concerts round the year.

## Completeness

In the context of this thesis, completeness of the corpus refers mainly to the completeness of the associated metadata for each recording, primarily from MusicBrainz. Even though carefully built, the editorial metadata associated with a release and its recordings can be incomplete. There are three possible reasons for incomplete metadata. Many releases do not provide all the required metadata on

the CD. In many releases, only the lead artist is listed, without the accompanying artists. It is seen very often that the composition information is also absent on the CD cover. The second reason is that the editorial metadata was not completely entered into MusicBrainz. This is sometimes seen with release and recording relationships that were left incomplete by the person who added the metadata. Further, since all the metadata, including the *rāga/tāla* tags, are imported and linked automatically, there can be import errors due to variations in transliterations and spelling. Multiplicity of languages used in Carnatic music further adds to these inconsistencies. These import errors are the third reason for incomplete metadata.

Missing metadata in MusicBrainz can only be completed by manually adding the missing fields to MusicBrainz. However, we are also exploring automatic metadata completion based on other relations on the release or the recording, using semantic web approaches. The missing data due to transliteration errors have been addressed to an extent by making curated lists of entities such as *rāgas* and *tālas*, and using robust algorithms for matching and linking metadata. Despite significant efforts, there are many recordings and releases that have incomplete metadata.

Table 4.2 shows the completeness of the recordings in the corpus, including all the three factors that result in incomplete metadata. All the recordings have a lead artist, but about a quarter of the recordings (429/1650) do not have accompanying artist information. *Rāga*, *tāla* and work (composition) are listed for about half the recordings. It is to be noted that these numbers reflect only the recordings for which we were completely sure of the editorial metadata. There are several recordings that have the required metadata but deemed incomplete since we could not accurately match it to a related entity in the curated lists.

### 4.1.3 Hindustani music research corpus

Similar to Carnatic music, *Rāg* and *tāl* are the fundamental music concepts in Hindustani music and hence the main theme around which the corpus has been built. Hindustani music tradition is much more diverse and heterogeneous and thus presents a significant challenge to compile a good research corpus. Though vocal

music is predominant, instrumental music in Hindustani music is also popular. The main focus in Hindustani music is on improvisation and compositions are short. For Hindustani music corpus we focus on two important vocal music styles - *Dhrupad* and *Khyāl*.

There are many institutions that have compiled huge audio archives of Hindustani music. The primary of them are the ITC Sangeet Research Academy (ITC-SRA), Sangeet Natak Academy, and the All India Radio (AIR). Each of these institutions own thousands of hours of expert curated music recordings that represent the real world performance practice. ITC-SRA is a premier music academy of Hindustani music and has taken up major efforts in the archival of music. Sangeet Natak Academy is India's national academy for music, drama and dance. AIR is the largest public broadcaster in India and has a huge archive of Hindustani music curated over many decades. AIR awards grades to musicians and its archives can be considered as a reference. None of these archives are publicly available and we compiled the audio in our corpus using these collections as a reference. We consulted expert musicians and musicologists, such as Dr. Suvarnalata Rao at the National Centre for the Performing Arts (NCPA), Mumbai, India to curate the audio collection in the corpus.

The audio collection in the corpus comprises of commercially available music releases from several music labels. It mainly consists of *khyāl* and *dhrupad* vocal music releases, though a significant number of instrumental music releases are present. The corpus presently has 233 releases with a total of 1096 recordings (300 hours). As with Carnatic music, the editorial metadata associated with each release is stored in MusicBrainz. The metadata associated with each release is the name of the release, the lead and the accompanying artists, and the musical instruments in the concert. For each audio recording in the release, the relevant metadata are the artists performed on the track, the name of the composition/s (*bandish*) and the composer/s (if composed), *rāg*/s, *tāl*/s, *lay*/s (tempo class), form/s, and section/s. All the editorial metadata was entered using Roman alphabet, following a uniform transliteration scheme for a better consistency.

Hindustani music is mainly improvised and hence lyrics and scores are not very relevant for computational analysis. Bhatkhande (1990) and Jha (2001) compiled lyrics and scores of bandishes us-

	<b>Corpus</b>	<b>ITC-SRA</b>	<b>Swarganga</b>
Artists	360	240 (19%)	629 (14%)
Rāgs	176	185 (48%)	534 (13%)
Tāls	32	N/A	59 (37%)
Works	685	N/A	1957

**Table 4.3:** Coverage of the Hindustani music corpus. The number in parentheses is the *overlap* measure in percentage. N/A indicates data not available.

<b>Accompanying metadata</b>	<b># Recordings</b>	<b>% of total</b>
Lead Artist	1096	100.0
Accompanying artist	658	39.9
Rāg	960	58.2
Tāl	627	38.0
Work (Bandiś)	576	34.9

**Table 4.4:** Completeness of the Hindustani music corpus showing the number of recordings in which the corresponding metadata is available.

ing a standardized notation for Hindustani music. However, they are not available in a machine readable form. Swarganga Music Foundation <sup>13</sup> has a good archive of rāgs, tāls and bandishes. The category of Hindustani music on Wikipedia<sup>14</sup> is a source of contextual information including music concepts of Hindustani music.

## Coverage

The methodology followed for the coverage analysis of Hindustani music is the same as followed for Carnatic music. We present the coverage analysis for artists, rāgs, tāls and compositions. The coverage analysis for Hindustani music is more complex than Carnatic music. This can be attributed to the heterogenous nature of the music repertoire, and to the lack of dedicated recording labels like

<sup>13</sup><http://www.swarganga.org/>

<sup>14</sup>[http://en.wikipedia.org/wiki/Category:Hindustani\\_music](http://en.wikipedia.org/wiki/Category:Hindustani_music)

Charsur in the case of Carnatic music. For each of these entities we choose two main references, ITC-SRA and Swarganga.

Unlike Carnatic music, the unit of music distribution in Hindustani music is not often a concert. Further, it is geographically spread over the Indian sub-continent and hence there is no single repository of Hindustani music performances, such as Kutcheris.com for Carnatic music. Therefore, it is challenging to do a comprehensive artist coverage analysis like the one presented for Carnatic music.

Table 4.3 shows the coverage of the Hindustani corpus. We see that the corpus and the chosen references have comparable number of entities, but the *overlap* is less. This is primarily because we mainly focused on recordings made in last 20-30 years to ensure good recording quality and to reflect current performance practices. On the other hand both the references focus primarily on archiving Hindustani music and hence consist of several generations of artists, infrequent *rāgs* and *tāls*, and a more comprehensive list of compositions. Further, the Hindustani corpus is mainly composed of vocal music recordings with a focus on only two styles, *khyāl* and *dhrupad*. The reference archives additionally include instrumental music and several other styles of Hindustani music.

## Completeness

The completeness of the editorial metadata for Hindustani music is shown in Table 4.4. We see that the editorial metadata for all the recordings at least includes the lead artist, and for more than half of the collection, the accompanying artists (658/1096). Roughly 90% of the corpora is annotated with the *rāg* label and more than half with the *tāl* label. Work (bandish) labels are present for nearly half of the collection (576/1096). *Ālāp* performances in Hindustani music are not compositional works, and hence should be discounted while assessing the completeness of work metadata. But due to the unavailability of such an information (*ālāp* labels), *ālāp* performances are also included in assessment and hence work completeness is an underestimate.

An important concern in research is the reproducibility of the experiments, which necessitates a corpus accessible to the research community. When possible, we emphasize the use of open repos-

itories of information such as MusicBrainz<sup>15</sup> and Wikipedia. The releases in the Carnatic<sup>16</sup> and Hindustani<sup>17</sup> corpora have been organized into collections in MusicBrainz. For audio, we use easily accessible commercial recordings. Further, the test datasets and the derived information such as annotations and extracted features are openly available<sup>18</sup>. In CompMusic, we are developing a tool for navigating through music collections called *Dunya* (Porter et al., 2013), which also acts as the central permanent online repository to store the metadata, audio, annotations and research results. *Dunya* is open source and provides an API for accessing these data.

#### 4.1.4 Creative Commons music collections

The audio in the Carnatic and Hindustani research corpora are commercial releases. Though easily accessible, they cannot be distributed openly. Since there are no open repositories of quality audio, one effort of CompMusic is to create and open audio collection released under creative commons licenses (CC BY-NC 4.0). In addition to the audio, the collection has carefully curated editorial metadata, and semi-automatically extracted melody and rhythm related annotations. Due permissions from artists have been secured for redistribution. The audio will be hosted on Internet Archive<sup>19</sup>, with both the audio and associated metadata and annotations available through the *Dunya* API.

The Creative Commons Carnatic corpus ( $\text{CMD}_o$ )<sup>20</sup> is collection of 19 vocal concerts (with more releases being added) with over 190 tracks and XX hours of music by professional Carnatic musicians. The audio in the  $\text{CMD}_o$  collection were professionally recorded at 44.1kHz sampling rate in multitrack at Arkay Convention Center, Chennai, India, and mastered professionally. The pieces from the

---

<sup>15</sup><http://musicbrainz.org/>

<sup>16</sup><http://musicbrainz.org/collection/f96e7215-b2bd-4962-b8c9-2b40c17a1ec6>

<sup>17</sup><http://musicbrainz.org/collection/213347a9-e786-4297-8551-d61788c85c80>

<sup>18</sup><http://compmusic.upf.edu/corpora>

<sup>19</sup>[www.archive.org](http://www.archive.org)

<sup>20</sup><https://musicbrainz.org/collection/a163c8f2-b75f-4655-86be-1504ea2944c2>

concerts were split into individual recordings and released as an album. Each recording has the following accompanying metadata: *rāga*, *tāla*, artists, composer, composition, and form. It has manually annotated time aligned characteristic melodic phrases and sections. In addition, it has semi-automatically extracted tonic, vocal pitch track, tempo, and time aligned *sama* annotations. The collection has **XX** *sama* annotations that can be used for rhythm analysis.

The Creative Commons Hindustani corpus (*HMD<sub>o</sub>*)<sup>21</sup> is collection of **40** vocal Hindustani music albums (with more releases being added) with over **XX** tracks and **XX** hours of music by professional Hindustani musicians, sourced from personal collections of musicians. The audio in the *HMD<sub>o</sub>* collection are stereo mp3 tracks sampled at 44.1 kHz. The tracks procured from personal collections have been grouped into musically meaningful short collections and then released as albums. Each recording in the collection has the following accompanying metadata: *rāg*, *tāl*, *lay/s*, artists, form, and if applicable, the *bandiś* and the composer. It has manually annotated time aligned characteristic melodic phrases and *lay* based sections. In addition, it has semi-automatically extracted tonic, vocal pitch track, tempo, and time aligned *sam* annotations. The collection has **XX** *sam* annotations that can be used for rhythm analysis.

The Creative Commons collections are useful for several MIR tasks. From a rhythm analysis perspective, the collection is useful for meter inference and tracking, rhythmic and percussion pattern analysis, and rhythm based structural segmentation. To the best of our knowledge, this collection is the largest *tāla* and *sama* annotated music collection of Indian Art Music.

## 4.2 Test datasets

The test datasets are designed for specific tasks and contain additional information such as annotations and derived data. They are useful for various melody and rhythm analysis tasks. We describe only those test datasets that are useful in rhythm analysis tasks. We describe each dataset briefly emphasizing the primary research task they can be used for.

<sup>21</sup><https://musicbrainz.org/collection/6adc54c6-6605-4e57-8230-b85f1de5be2b>

Tāla	# Pieces	Total Duration hours (min)	$\bar{T}_f$ min	#Ann.	#Sama
Ādi	50	4.21 (252.78)	4m51s	22793	2882
Rūpaka	50	4.45 (267.45)	4m37s	22668	7582
Miśra chāpu	48	5.70 (342.13)	6m35s	54309	7795
Khaṇḍa chāpu	28	2.24 (134.62)	4m25s	21382	4387
Total	176	16.61 (996.98)	5m4s	121602	22646

**Table 4.5:** CMR<sub>f</sub> dataset showing the total duration and number of annotations. #Sama shows the number of sama annotations and #Ann. shows the number of beat annotations (including samas).  $\bar{T}_f$  indicates the median piece length in the dataset.

Tāla	$\bar{\tau}_s \pm \sigma_s$	$\bar{\tau}_o \pm \sigma_o$	$[\tau_{s,\min}, \tau_{s,\max}]$
Ādi	$5.34 \pm 0.723$	$0.167 \pm 0.023$	[2.88, 7.07]
Rūpaka	$2.13 \pm 0.239$	$0.178 \pm 0.020$	[1.21, 3.10]
Miśra chāpu	$2.67 \pm 0.358$	$0.191 \pm 0.026$	[1.63, 3.65]
Khaṇḍa chāpu	$1.85 \pm 0.284$	$0.185 \pm 0.028$	[0.91, 2.87]

**Table 4.6:** Tāla cycle length indicators for CMR<sub>f</sub> dataset.  $\bar{\tau}_s$  and  $\sigma_s$  indicate the mean and standard deviation of the median inter-sama interval of the pieces, respectively.  $\bar{\tau}_o$  and  $\sigma_o$  indicate the mean and standard deviation of the median inter-akṣara interval of the pieces, respectively.  $[\tau_{s,\min}, \tau_{s,\max}]$  indicate the minimum and maximum value of  $\tau_s$  and hence the range of  $\tau_s$  in the dataset. All values in the table are in seconds.

### 4.2.1 Carnatic music rhythm dataset

The Carnatic music rhythm dataset (CMR<sub>f</sub> dataset)<sup>22</sup> is a rhythm annotated test corpus for many automatic rhythm analysis tasks in Carnatic Music. The collection consists of audio excerpts from the Carnatic research corpus, manually annotated time aligned markers indicating the progression through the tāla cycle, and the associated tāla related metadata. The dataset has pieces in four popular tālas

<sup>22</sup><http://compmusic.upf.edu/carnatic-rhythm-dataset>

(Table 4.5) that encompass a majority of current day Carnatic music performance. The pieces include a mix of vocal and instrumental recordings, recent and old recordings, and span a wide variety of forms. All pieces have a percussion accompaniment, predominantly Mridangam. There are also several different pieces by the same artist (or release group), and multiple instances of the same composition rendered by different artists. Each piece is uniquely identified using the MBID of the recording. The pieces are mono WAV files downmixed from stereo recordings, and sampled at 44.1 kHz. The audio is also available as downmixed mono WAV files for experiments. The audio files are full length pieces or clips extracted from full length pieces. Of the 176 audio files, 120 contain full length pieces.

There are several annotations that accompany each excerpt in the dataset. The primary annotations are audio synchronized time-stamps indicating the different metrical positions in the *tāla* cycle - the *sama* (downbeat) and other beats shown with numerals in Figure 2.1. The annotations were created using Sonic Visualizer (Cannam et al., 2010) by tapping to music and manually correcting the taps. The annotations have been verified by a professional Carnatic musician. Each annotation has a time-stamp and an associated numeric label that indicates the position of the beat marker in the *tāla* cycle. In addition, for each excerpt, the *tāla* of the piece and *edupu* (offset of the start of the piece, relative to the *sama*) are recorded. The possibly time varying tempo of a piece can be obtained using the beat and *sama* annotations.

Carnatic music rhythm dataset (*CMR<sub>f</sub>*) dataset is described in Table 4.5, showing the four *tālas* and the number of pieces for each *tāla*. The total duration of audio in the dataset is over 16.6 hours, with 121062 time-aligned beat annotations. The median length of a piece is about 5 minutes in the dataset. Table 4.6 shows a basic statistical analysis of the *tāla* cycle length indicators in the dataset, which is useful to understand the tempo characteristics and the range of the metrical cycle lengths in the dataset. *Ādi tāla* is the longest *tāla* in the dataset and hence has the highest  $\bar{\tau}_s$  among all the *tālas*. Despite no notated tempo, we can see from the values of the median inter-akṣara interval,  $\bar{\tau}_o$  and its standard deviation that the tempo in Carnatic music does not vary much across the *tālas*. The range of  $\bar{\tau}_s$  values show that a wide range of tempi are present

Tāla	# Pieces	Total Duration hours (min)	#Ann.	#Sama
Ādi	30	0.98 (58.87)	5452	696
Rūpaka	30	1.00 (60.00)	5148	1725
Miśra chāpu	30	1.00 (60.00)	8992	1299
Khaṇḍa chāpu	28	0.93 (55.93)	9133	1840
Total	118	3.91 (234.80)	28725	5560

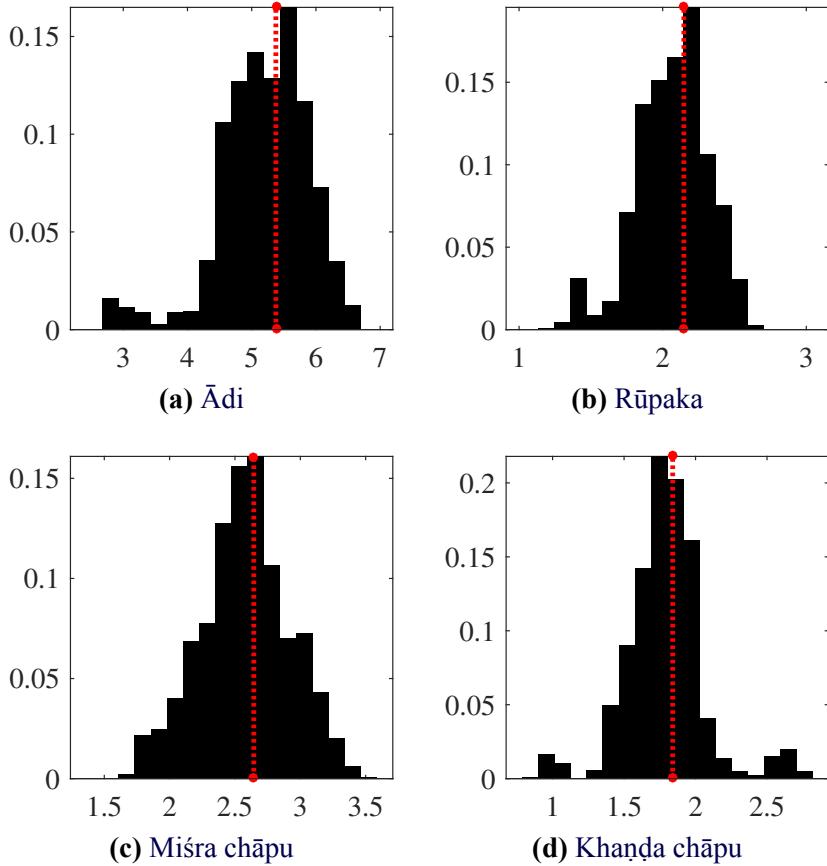
**Table 4.7:** CMR dataset showing the total duration and number of annotations. #Sama shows the number of sama annotations and #Ann. shows the number of beat annotations (including samas).

Tāla	$\bar{\tau}_s \pm \sigma_s$	$\bar{\tau}_o \pm \sigma_o$	$[\tau_{s,\min}, \tau_{s,\max}]$
Ādi	$5.32 \pm 0.868$	$0.17 \pm 0.027$	[2.88, 7.07]
Rūpaka	$2.12 \pm 0.225$	$0.18 \pm 0.019$	[1.40, 3.10]
Miśra chāpu	$2.81 \pm 0.272$	$0.20 \pm 0.019$	[2.03, 3.65]
Khaṇḍa chāpu	$1.87 \pm 0.290$	$0.19 \pm 0.029$	[1.00, 2.84]

**Table 4.8:** Tāla cycle length indicators for CMR dataset.  $\bar{\tau}_s$  and  $\sigma_s$  indicate the mean and standard deviation of the median inter-sama interval of the pieces, respectively.  $\bar{\tau}_o$  and  $\sigma_o$  indicate the mean and standard deviation of the median inter-akṣara interval of the pieces, respectively.  $[\tau_{s,\min}, \tau_{s,\max}]$  indicate the minimum and maximum value of  $\tau_s$  and hence the range of  $\tau_s$  in the dataset. All values in the table are in seconds.

in Carnatic music pieces, often over two tempo octaves. The shortest cycle in the dataset is less than second long, while the longest cycle is over 7 seconds long.

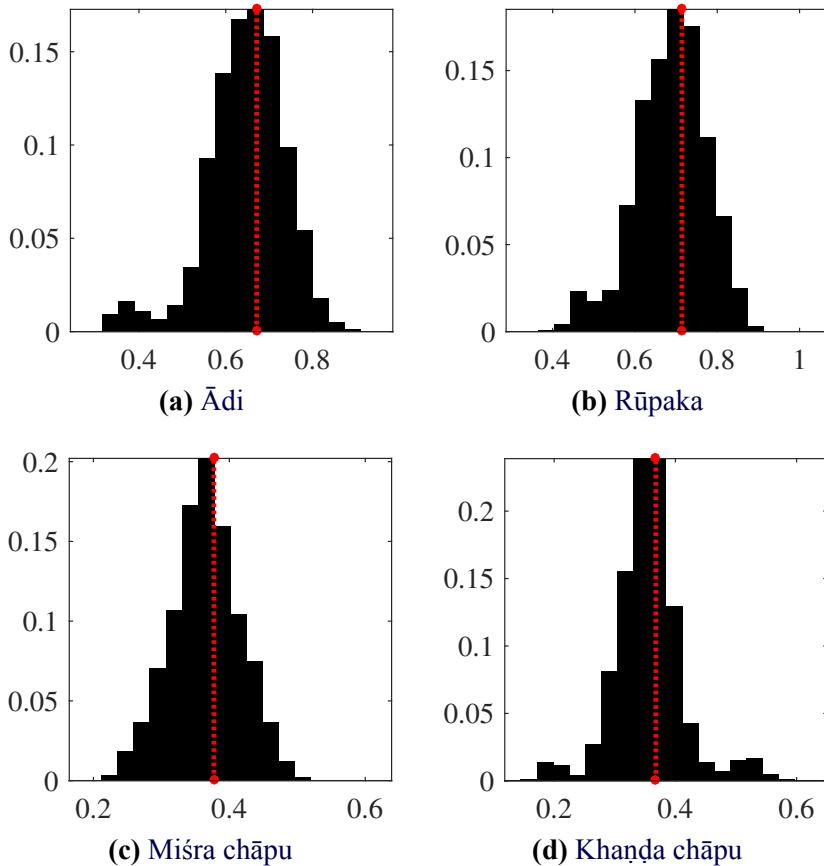
A representative subset of the CMR<sub>f</sub> dataset is also compiled as Carnatic music rhythm dataset (subset) (CMR), with two minute excerpts of pieces in CMR<sub>f</sub> (or the full piece if the piece is shorter than 2 minutes). These short excerpts additionally contain all the annotations of the full dataset, including time aligned sama and beat annotations. The smaller Carnatic music rhythm dataset (subset) (CMR) dataset will be useful for faster testing of approaches and algorithms.



**Figure 4.3:** A histogram of the inter-sama interval  $\tau_s$  in the CMR<sub>f</sub> dataset for each tāla. The ordinate is the fraction of the total count corresponding to the  $\tau_s$  value shown in abscissa. The median  $\tau_s$  for each tāla is shown as a red dotted line.

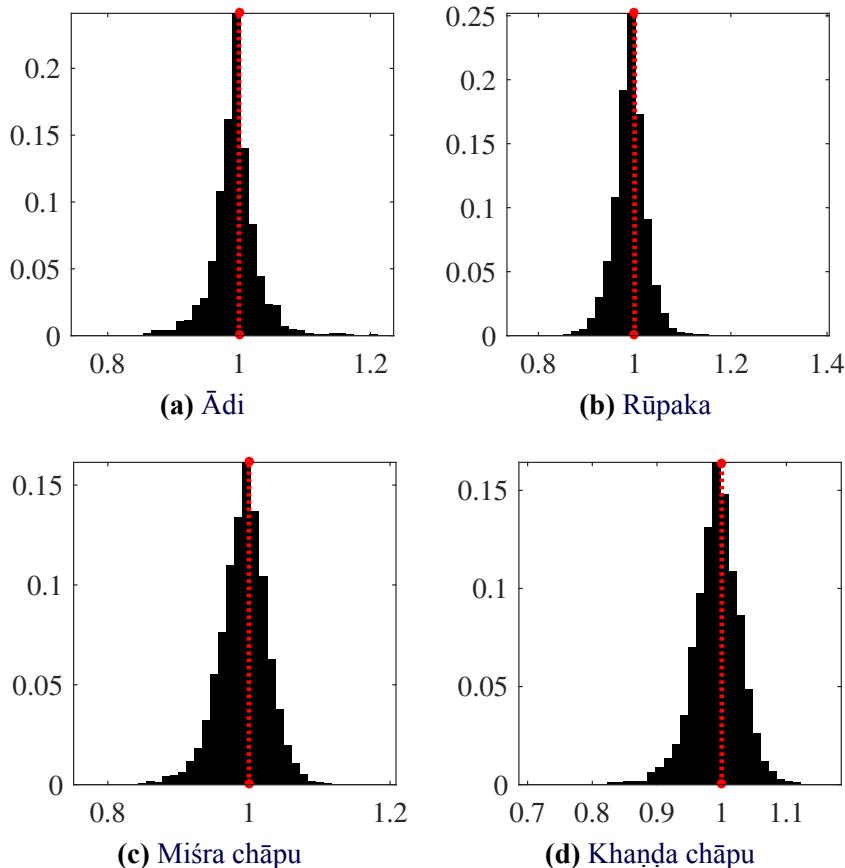
The smaller subset CMR dataset is described in Table 4.7, showing the four tālas and the number of pieces for each tāla. The total duration of audio in the dataset is about 4 hours, with 28725 time-aligned beat annotations. Table 4.8 shows a basic statistical analysis of the tāla cycle length indicators in the CMR dataset, which are similar to the indicators of CMR<sub>f</sub> dataset shown in Table 4.6, showing that CMR dataset is a representative subset of CMR<sub>f</sub> dataset.

The tempo values are not notated in Carnatic music, and the pieces are not played to a metronome. Hence the tempo varies over a piece in time. Hence, in addition to the median values tabulated in



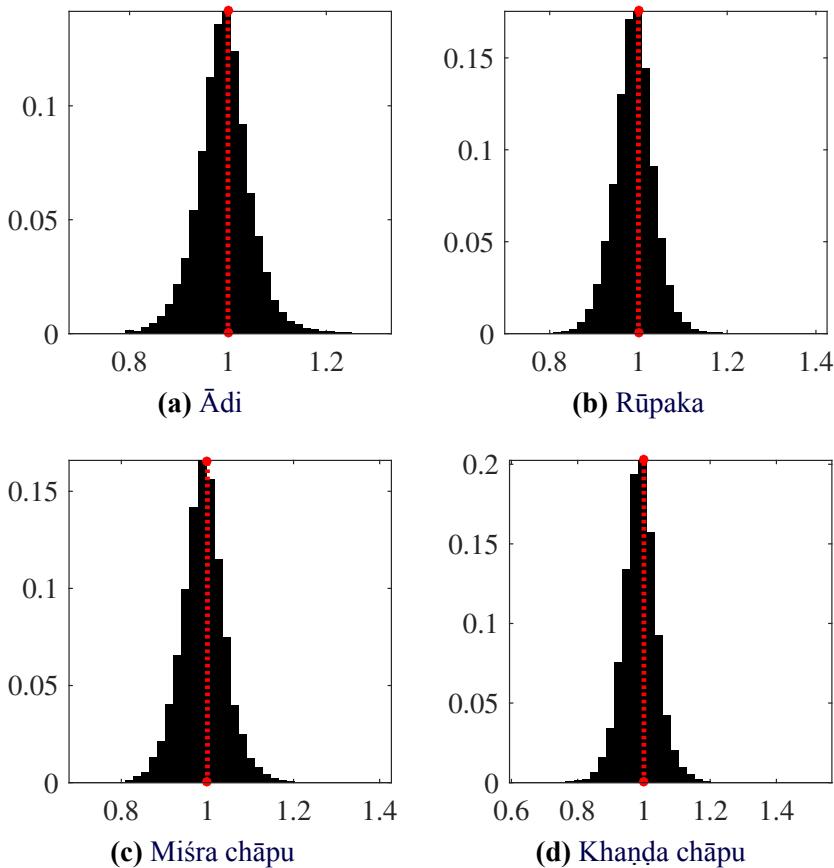
**Figure 4.4:** A histogram of the inter-beat interval  $\tau_b$  in the  $\text{CMR}_f$  dataset for each  $\text{tāla}$ . The ordinate is the fraction of the total count corresponding to the  $\tau_b$  value shown in abscissa. The median  $\tau_b$  for each  $\text{tāla}$  is shown as a red dotted line.

Table 4.6 we present further analysis of the inter-sama interval ( $\tau_s$ ) and inter-beat interval ( $\tau_b$ ) for each  $\text{tāla}$  over the whole  $\text{CMR}_f$  dataset. A histogram of  $\tau_s$  and  $\tau_b$  for each  $\text{tāla}$  is shown in Figure 4.3 and Figure 4.4 respectively. This shows the distribution of cycle lengths in the dataset over the whole range of  $\tau_s$  for each  $\text{tāla}$ , around the median value. Despite the large range of  $\tau_s$  values, the distribution in Figure 4.3 and Figure 4.4 show that the tempo often is limited to a small range of values. Though the musicians are free to choose any tempo, we empirically observe that they tend to choose a narrow range of tempo.



**Figure 4.5:** A histogram of the median normalized inter-sama interval  $\tau_s$  in the CMR<sub>f</sub> dataset for each tāla. The ordinate is the fraction of the total count corresponding to the normalized  $\tau_s$  value shown in abscissa.

To illustrate and measure the time varying tempo of music pieces in Carnatic music, we normalize all the  $\tau_s$  and  $\tau_b$  values in a piece by the median in the piece to obtain median normalized  $\tau_s$  and  $\tau_b$  values, a histogram of which is shown in Figure 4.5 and Figure 4.6, respectively. These histograms are centered around 1, since they are normalized by the median, and the spread of these histograms around the value of 1 is a measure of deviation of tempo from the median value. From the figures, it is clear that the tempo is time varying but with less than about 20% maximum deviation from the

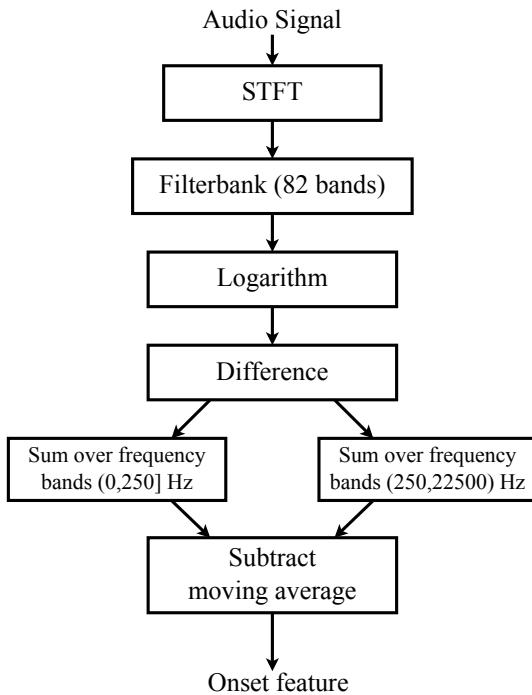


**Figure 4.6:** A histogram of the median normalized inter-beat interval  $\tau_b$  in the  $\text{CMR}_f$  dataset for each tāla. The ordinate is the fraction of the total count corresponding to the normalized  $\tau_b$  value shown in abscissa.

median tempo of the piece for all tālas.

### Rhythm patterns in $\text{CMR}_f$ and $\text{CMR}$ datasets

With a sizeable annotated corpus of Carnatic music, we can do corpora level analysis of patterns in rhythm and percussion. The idea is to showcase these patterns as the potential of dataset analysis, while showing their utility for meter tracking, musicology, performance analysis, comparative analysis.



**Figure 4.7:** Computation of the spectral flux onset feature in two frequency bands, modified from the figure by Holzapfel et al. (2014).

The aim here is not to seek all musicological insights from data, but to illustrate the possibilities of a corpus level analysis data, and how such analysis tools can help aid and advance musicology. The MIR applications of such datasets is the primary goal of the thesis and discussed in subsequent chapters. Hence, an example of corpus level musicological analysis is presented in this chapter, which amounts to a performance analysis of music in current practice from audio recordings. These analyses can corroborate several musicological inferences, and can provide additional insights into the differences between musicology, music theory and music practice. At the outset, it is necessary to note that the insights we discuss and conclusions we draw are limited by the available annotated dataset, and hence needs further validation. It is however useful to focus on the methodology, which can aid musicologists and engineers to build systems that use these patterns for different analyses.

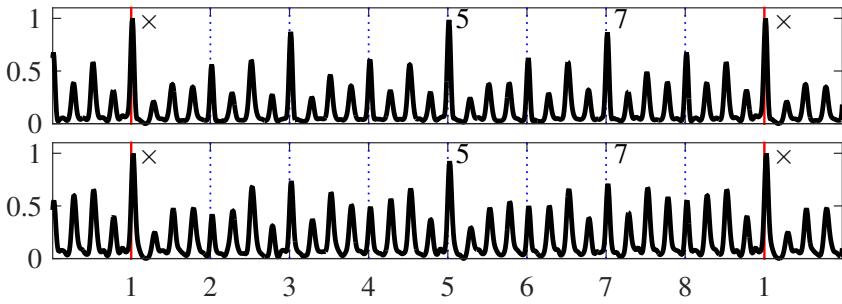
The rhythm patterns are computed using a spectral flux feature

(called LogFilt- SpecFlux as proposed by Böck et al. (2012) and used further by Krebs et al. (2013)) that is used for detecting musical onsets in audio recordings. The short time Fourier transform (STFT) of the audio signal with a window size of 46.4 ms (2047 samples of audio at a sampling rate of 44.1 kHz), FFT size of 2048 and hop size of 20 ms is computed from audio. The successive difference between frames of the logarithm of the filter bank energies in 82 different bands are then computed. Since the bass onsets have significant information about the rhythmic patterns, the features are computed in two frequency bands (Low:  $\leq 250$  Hz, High:  $> 250$  Hz) to additionally consider the bass onsets. The process of computing the spectral flux feature is outlined in Figure 4.7.

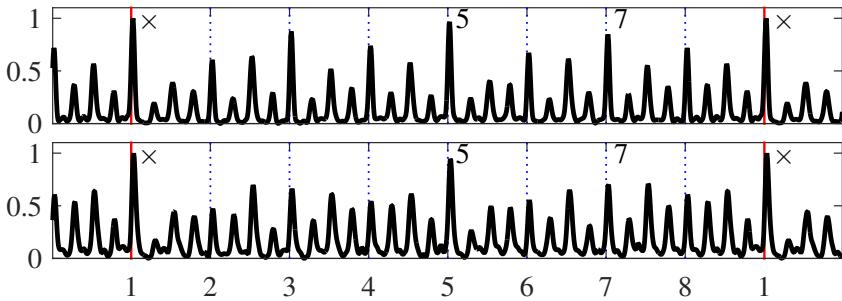
Using beat and downbeat annotated training data, the audio features from all music pieces in a specific *tāla* are then grouped into cycle length sequences, and interpolated to equal lengths using a fine grid. A mean of all the pattern instances for a specific *tāla* is computed in both the frequency bands and used as a representative rhythmic pattern illustrated here.

The Figures 4.8-4.15 show the ensemble average of cycle length patterns over all the pieces in the dataset for each *tāla*, computed using the spectral flux feature in two different frequency bands as outlined above. In each figure, the bottom pane corresponds to the low frequency band ( $y_l$ ) and the top pane corresponds to the high frequency band ( $y_h$ ). The abscissa is the beat number within the cycle (dotted lines), with 1 indicating the *sama* (marked with a red line). The start of each *aṅga* is indicated with beat numbers at the top of each pane (*sama* shown as  $\times$ ). The patterns in each figure pane is normalized so that maximum value is 1, to comment on relative onset strengths at different metrical positions of the cycle.

The rhythm patterns are indicative of mridangam strokes played in the cycle. In the figures, the bottom pane that shows the low frequency band has content from the left bass drum while the top pane has content predominantly from the right pitched drum, but additionally from the lead melody. Hence, for the purpose of this discussion, we use the terms left and right accents to refer to the accents in rhythm patterns from the bottom and top pane, respectively. The left and right accents provide interesting insights into the patterns played within a *tāla* cycle. In addition, these rhythm patterns help in meter tracking.



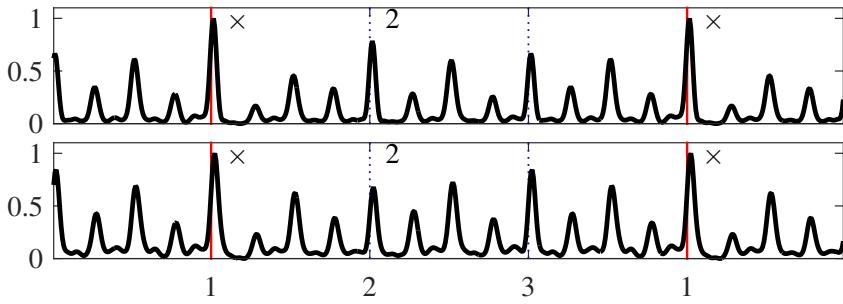
**Figure 4.8:** Cycle length rhythmic patterns learned from  $\text{CMR}_f$  dataset for  $\bar{\text{a}}\text{di t}\bar{\text{a}}\text{l}\bar{\text{a}}$ . In each of the following Figures 4.8-4.15, the patterns are computed from spectral flux feature and averaged over all the pieces in the dataset. The bottom/top pane corresponds to the low/high frequency bands, respectively. The abscissa is the beat number within the cycle (dotted lines), with 1 indicating the *sama* (marked with a red line). The start of each *āṅga* is indicated with beat numbers at the top of each pane (*sama* shown as  $\times$ ).



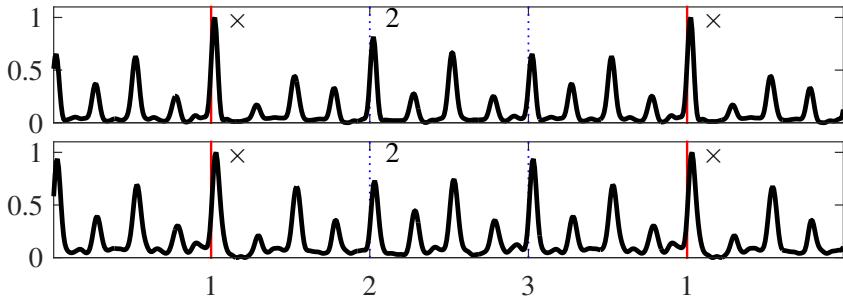
**Figure 4.9:** Cycle length rhythmic patterns learned from  $\text{CMR}$  dataset for  $\bar{\text{a}}\text{di t}\bar{\text{a}}\text{l}\bar{\text{a}}$ .

We list down and discuss some salient qualitative observations from figures for each  $\text{t}\bar{\text{a}}\text{l}\bar{\text{a}}$ , for both  $\text{CMR}_f$  dataset and its subset  $\text{CMR}$ . The Figures 4.8-4.15 show the cycle length rhythm patterns for all  $\text{t}\bar{\text{a}}\text{l}\bar{\text{a}}$ s for both  $\text{CMR}_f$  and  $\text{CMR}$  datasets. For each  $\text{t}\bar{\text{a}}\text{l}\bar{\text{a}}$ , we plot the rhythm patterns together to compare patterns across the short excerpts in  $\text{CMR}$  dataset and full length pieces in  $\text{CMR}_f$  dataset.

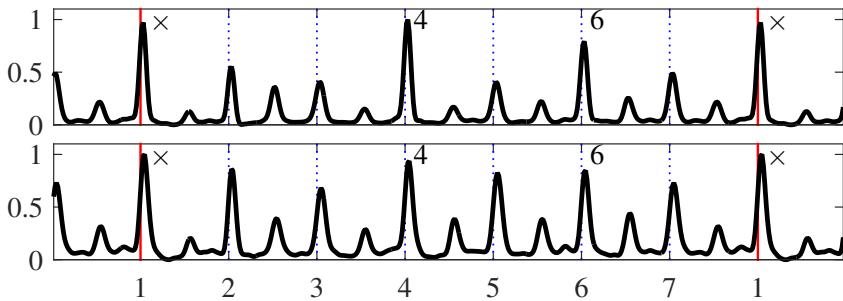
Overall, we see stronger accents on the *akṣaras*, with *sama* having the strongest accent in most cases. We can clearly see the ac-



**Figure 4.10:** Cycle length rhythmic patterns learned from  $\text{CMR}_f$  dataset for *rūpaka tāla*.

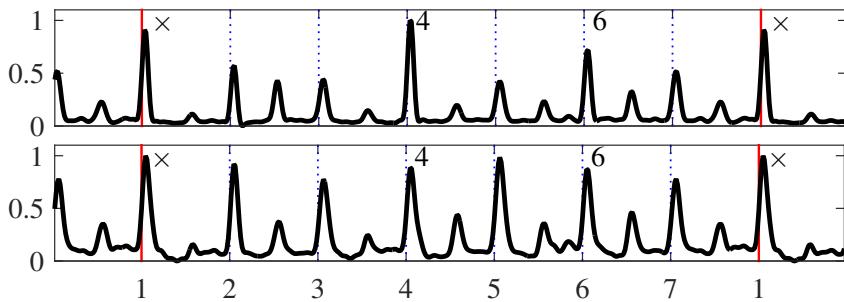


**Figure 4.11:** Cycle length rhythmic patterns learned from  $\text{CMR}$  dataset for *rūpaka tāla*.

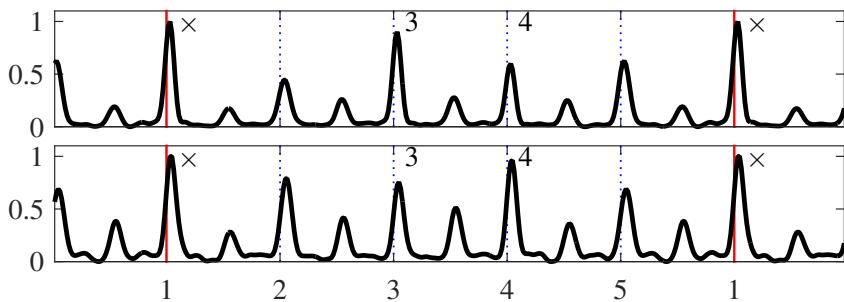


**Figure 4.12:** Cycle length rhythmic patterns learned from  $\text{CMR}_f$  dataset for *miśra chāpu tāla*.

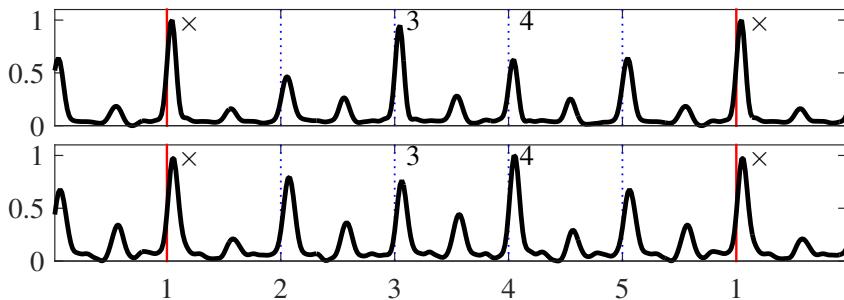
cents organized in three different strengths, reflecting the metrical levels of the *aṅga*, the beat and the *akṣara*. The two *akṣara* long



**Figure 4.13:** Cycle length rhythmic patterns learned from CMR dataset for miśra chāpu tāla.



**Figure 4.14:** Cycle length rhythmic patterns learned from CMR<sub>f</sub> dataset for khanḍa chāpu tāla.



**Figure 4.15:** Cycle length rhythmic patterns learned from CMR dataset for khanḍa chāpu tāla.

beats in the tālas miśra chāpu and khanḍa chāpu, and the four akṣara long beats in tāla ādi and rūpaka can be additionally seen. The pat-

terns and *ṭhēkās* played in Carnatic music are quite diverse, and no obvious “*tāla* pattern” can be inferred, apart from the three levels of accents.

The patterns illustrated here are kind of the average patterns that occur and do not tell us too much about the various individual patterns that might occur in specific points in particular recordings. The *tālas* are metrical structures that allow many different patterns to be played, and not a specific rhythm. It is further seen that the first *akṣara* after *sama* has softer accents. Fewer strokes are played after the *sama*, to emphasize that the *sama* has just passed and a new cycle has begun. It might also perhaps indicate some form of recovery time after the intense stroke-playing towards the end of the cycle. Further, the rhythm patterns computed using *CMR* dataset are very similar to those computed using *CMR<sub>f</sub>* dataset, showing that *CMR* is a good representative subset of the larger *CMR<sub>f</sub>*. Additionally, all the observations we make with patterns from *CMR<sub>f</sub>* extend to *CMR*. We now discuss several *tāla* specific observations.

The Figures 4.8-4.9 show the rhythm patterns for *ādi tāla*. We see that a three level hierarchy of *aṅga*, beats and *akṣaras* is well demarcated. The *akṣara* at half cycle (beat 5) has an accent as strong as the *sama*. The odd beats (marked 1, 3, 5, 7) have stronger right accents. The left accents are distributed through the cycle, with strong accents at half cycle.

The Figures 4.10-4.11 show the rhythm patterns for *rūpaka tāla*. Apart from the three level hierarchy of accents that is quite apparent, the half beat accent between the beats 2 and 3 are strong - indicating the often played 6+6 *akṣara* grouping structure of *rūpaka*, with a ternary meter.

The Figures 4.12-4.13 show the rhythm patterns for *miśra chāpu tāla*. We see that the *aṅga* boundaries have strong left and right accents showing their use as anchor points to indicate the progression through the cycle. Though defined with a 3+2+2 *akṣara* grouping structure, a 1+2+2+2 structure is often seen in *miśra chāpu tāla*, which can be observed here, based on the strong left accent on beat 2. A additional strong left accent on beat 5 shows that it is also used as an anchor.

From the rhythm patterns of *khaṇḍa chāpu tāla* shown in Figures 4.14-4.15 show a strong left accent on beat 4, which is used an anchor. A stronger right accent on beat 3 shows the progression

Tāl	# Pieces	Total Duration hours (min)	# Ann.	# Sam
Tīntāl	54	1.80 (108)	17142	1081
Ēktāl	58	1.93 (116)	12999	1087
Jhaptāl	19	0.63 (38)	3029	302
Rūpak tāl	20	0.67 (40)	2841	406
Total	151	5.03 (302)	36011	2876

**Table 4.9:** HMR<sub>f</sub> dataset showing the total duration and number of annotations. #Sam shows the number of sam annotations and #Ann. shows the number of mātrā annotations (including sams).

through the unequal aṅgas. The 2+1+2 akṣara grouping structure of khaṇḍa chāpu is often played out as 3+2 or 2+3, showing strong accents on beats 3 and 4.

### Applications of the dataset

The CMR<sub>f</sub> dataset and its subset CMR dataset are intended to be test corpora for several computational rhythm analysis tasks in Carnatic music. Possible tasks include sama and beat tracking, tempo estimation and tracking, tāla recognition, rhythm based segmentation of musical audio, structural segmentation, audio to score/lyrics alignment, and rhythmic pattern analysis. In this thesis, these two datasets are primarily used for rhythmic pattern analysis and meter inference/tracking. Most of the research results are presented for CMR and then extended to CMR<sub>f</sub> to verify their applicability to larger datasets.

#### 4.2.2 Hindustani music rhythm dataset

CompMusic Hindustani music rhythm dataset (HMR<sub>f</sub>)<sup>23</sup> is a rhythm annotated test corpus for automatic rhythm analysis tasks in Hindustani Music. The collection consists of audio excerpts from the CompMusic Hindustani research corpus, manually annotated time aligned markers indicating the progression through the tāl cycle,

---

<sup>23</sup><http://compmusic.upf.edu/hindustani-rhythm-dataset>

Tāl	$\bar{\tau}_s \pm \sigma_s$	$\bar{\tau}_o \pm \sigma_o$	$[\tau_{s,\min}, \tau_{s,\max}]$
Tīntāl	$10.36 \pm 9.875$	$0.65 \pm 0.617$	[2.32, 44.14]
Ēktāl	$30.20 \pm 26.258$	$2.52 \pm 2.188$	[2.23, 69.73]
Jhaptāl	$8.51 \pm 3.149$	$0.85 \pm 0.315$	[4.06, 16.23]
Rūpak tāl	$7.11 \pm 3.360$	$1.02 \pm 0.480$	[2.82, 16.09]

**Table 4.10:** Tāl cycle length indicators for HMR<sub>f</sub> dataset.  $\bar{\tau}_s$  and  $\sigma_s$  indicate the mean and standard deviation of the median inter-sam interval of the pieces, respectively.  $\bar{\tau}_o$  and  $\sigma_o$  indicate the mean and standard deviation of the median inter-mātrā interval of the pieces, respectively.  $[\tau_{s,\min}, \tau_{s,\max}]$  indicate the minimum and maximum value of  $\tau_s$  and hence the range of  $\tau_s$  in the dataset. All values in the table are in seconds.

Tāl	# Pieces	Total Duration hours (min)	# Ann.	# Sam
Tīntāl	13	0.43 (26)	1020	65
Ēktāl	32	1.07 (64)	967	79
Jhaptāl	6	0.2 (12)	592	59
Rūpak tāl	8	0.27 (16)	701	101
Total	59	1.97 (118)	3280	304

**Table 4.11:** HMR<sub>I</sub> dataset showing the total duration and number of annotations. #Sam shows the number of sam annotations and #Ann. shows the number of mātrā annotations (including sams).

and the associated tāl related metadata. The dataset has pieces from four popular tāls of Hindustani music (Table 4.9), which encompasses a majority of Hindustani khyāl music.

The audio recordings are chosen from the CompMusic Hindustani music collection. The pieces include a mix of vocal and instrumental recordings, new and old recordings, and to span three layas. For each taal, there are pieces in dṛṭ (fast), madhya (medium) and vilāmbit layas. All pieces have Tabla as the percussion accompaniment. All the audio recordings in the dataset are 2 min excerpts of full length pieces. Each piece is uniquely identified using the

Tāl	$\bar{\tau}_s \pm \sigma_s$	$\bar{\tau}_o \pm \sigma_o$	$[\tau_{s,\min}, \tau_{s,\max}]$
Tīntāl	$26.16 \pm 7.963$	$1.63 \pm 0.498$	[18.57, 44.14]
Ēktāl	$52.16 \pm 12.531$	$4.35 \pm 1.044$	[14.43, 69.73]
Jhaptāl	$12.30 \pm 1.935$	$1.23 \pm 0.194$	[10.20, 16.23]
Rūpak tāl	$10.28 \pm 3.050$	$1.47 \pm 0.436$	[6.95, 16.09]

**Table 4.12:** Tāl cycle length indicators for HMR<sub>1</sub> dataset.  $\bar{\tau}_s$  and  $\sigma_s$  indicate the mean and standard deviation of the median inter-sam interval of the pieces, respectively.  $\bar{\tau}_o$  and  $\sigma_o$  indicate the mean and standard deviation of the median inter-mātrā interval of the pieces, respectively.  $[\tau_{s,\min}, \tau_{s,\max}]$  indicate the minimum and maximum value of  $\tau_s$  and hence the range of  $\tau_s$  in the dataset. All values in the table are in seconds.

Tāl	# Pieces	Total Duration hours (min)	# Ann.	# Sam
Tīntāl	41	1.37 (82)	16122	1016
Ēktāl	26	0.87 (52)	12032	1008
Jhaptāl	13	0.43 (26)	2437	243
Rūpak tāl	12	0.40 (24)	2140	305
Total	92	3.07 (184)	32731	2572

**Table 4.13:** HMR<sub>s</sub> dataset showing the total duration and number of annotations. #Sam shows the number of sam annotations and #Ann. shows the number of mātrā annotations (including sams).

MBID of the recording. The pieces are stereo, 160 kbps, mp3 files sampled at 44.1 kHz. The audio is also available as downmixed mono WAV files for experiments.

There are several annotations that accompany each audio file in the dataset. The primary annotations are audio synchronized time-stamps indicating the different metrical positions in the tāl cycle. The sam and mātrās of the cycle are annotated. The annotations were created using Sonic Visualizer by tapping to music and manually correcting the taps. Each annotation has a time-stamp and an associated numeric label that indicates the mātrā position in the tāl

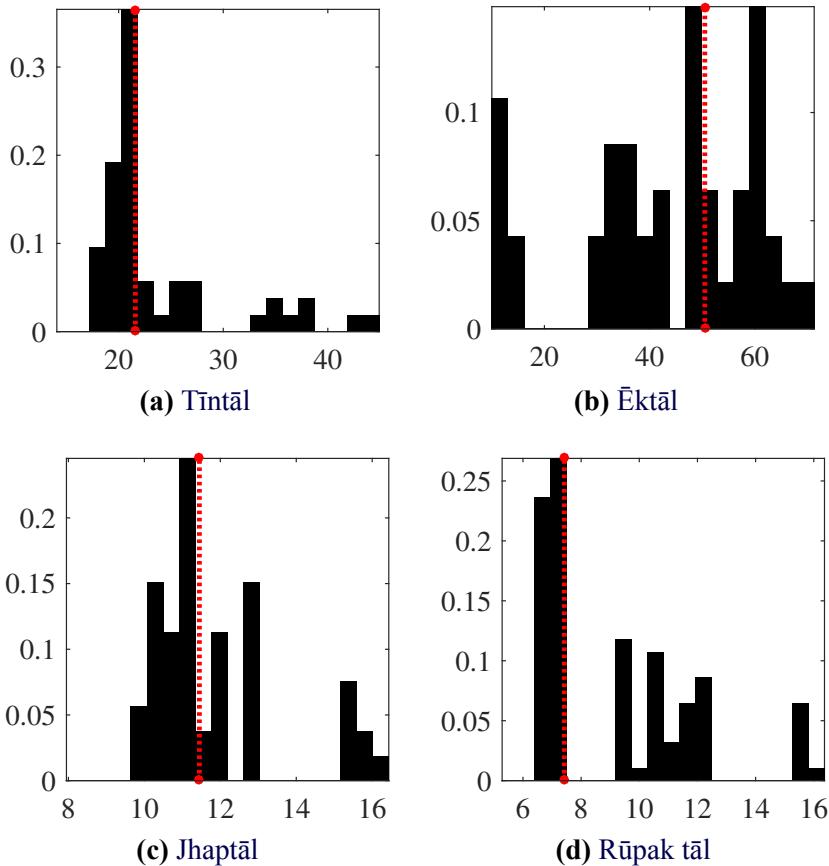
Tāl	$\bar{\tau}_s \pm \sigma_s$	$\bar{\tau}_o \pm \sigma_o$	$[\tau_{s,\min}, \tau_{s,\max}]$
Tīntāl	$5.35 \pm 1.823$	$0.33 \pm 0.114$	[2.32, 9.89]
Ēktāl	$3.17 \pm 0.471$	$0.26 \pm 0.039$	[2.23, 4.11]
Jhaptāl	$6.77 \pm 1.688$	$0.68 \pm 0.169$	[4.06, 9.97]
Rūpak tāl	$5.00 \pm 1.191$	$0.71 \pm 0.170$	[2.82, 6.68]

**Table 4.14:** Tāl cycle length indicators for HMR<sub>s</sub> dataset.  $\bar{\tau}_s$  and  $\sigma_s$  indicate the mean and standard deviation of the median inter-sam interval of the pieces, respectively.  $\bar{\tau}_o$  and  $\sigma_o$  indicate the mean and standard deviation of the median inter-mātrā interval of the pieces, respectively.  $[\tau_{s,\min}, \tau_{s,\max}]$  indicate the minimum and maximum value of  $\tau_s$  and hence the range of  $\tau_s$  in the dataset. All values in the table are in seconds.

cycle, as shown in Figure 2.3. The sams are indicated using the numeral 1. The time varying tempo of the piece can be obtained from the mātrā and sam annotations.

For each excerpt, the tāl and the lay of the piece are recorded. Each excerpt can be uniquely identified and located with the MBID of the recording, and the relative start and end times of the excerpt within the whole recording. The artist, release, the lead instrument, and the rāg of the piece are additional editorial metadata obtained from the release. There are optional comments on audio quality and annotation specifics. The annotations and the associated metadata have been verified for correctness and completeness by a professional Hindustani musician and musicologist.

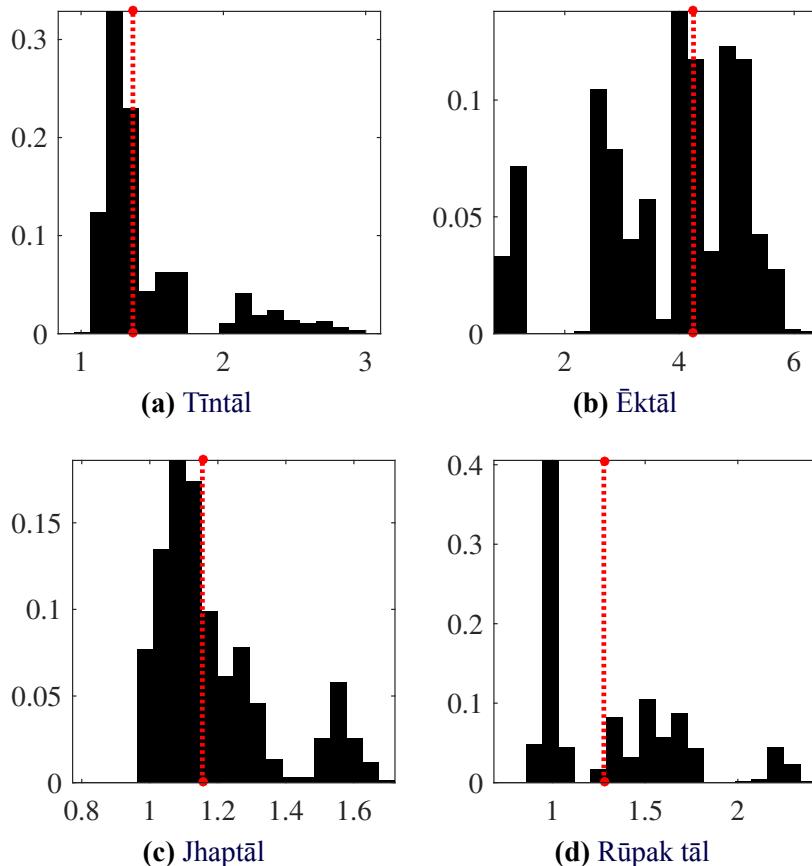
The HMR<sub>f</sub> dataset is described in Table 4.9, showing the four tāls and the number of pieces for each tāl, totaling to 151 pieces. The total duration of audio in the dataset is about 5 hours, with 36011 time-aligned mātrā annotations of which 2876 are sam annotations. Table 4.10 shows a basic statistical analysis of the tāl cycle length indicators in the dataset to understand the tempo characteristics and the range of the metrical cycle lengths in the dataset. The large range of tempi seen in Hindustani music is reflected in the dataset, with the values of median inter-sam interval  $\bar{\tau}_s$ , ēktāl cycle lengths ranging from 2.2 seconds to 69.7 seconds, which is about 5 tempo octaves. This also shows that the mātrā period can vary from less



**Figure 4.16:** A histogram of the inter-sam interval  $\tau_s$  in the HMR<sub>1</sub> dataset for each tāl. The ordinate is the fraction of the total count corresponding to the  $\tau_o$  value shown in abscissa. The median  $\tau_s$  for each tāl is shown as a red dotted line.

than 150 ms to over 6 seconds. This huge range of cycle lengths and mātrā periods is a significant challenge in Hindustani music automatic meter inference. Across different tāls, we see that tīntāl and ēktāl have the largest range of  $\overline{\tau_s}$ , since they are performed in all the lay classes, vilāmbit to dṛ̥t. Jhaptāl and rūpak tāl have smaller  $\overline{\tau_s}$  ranges.

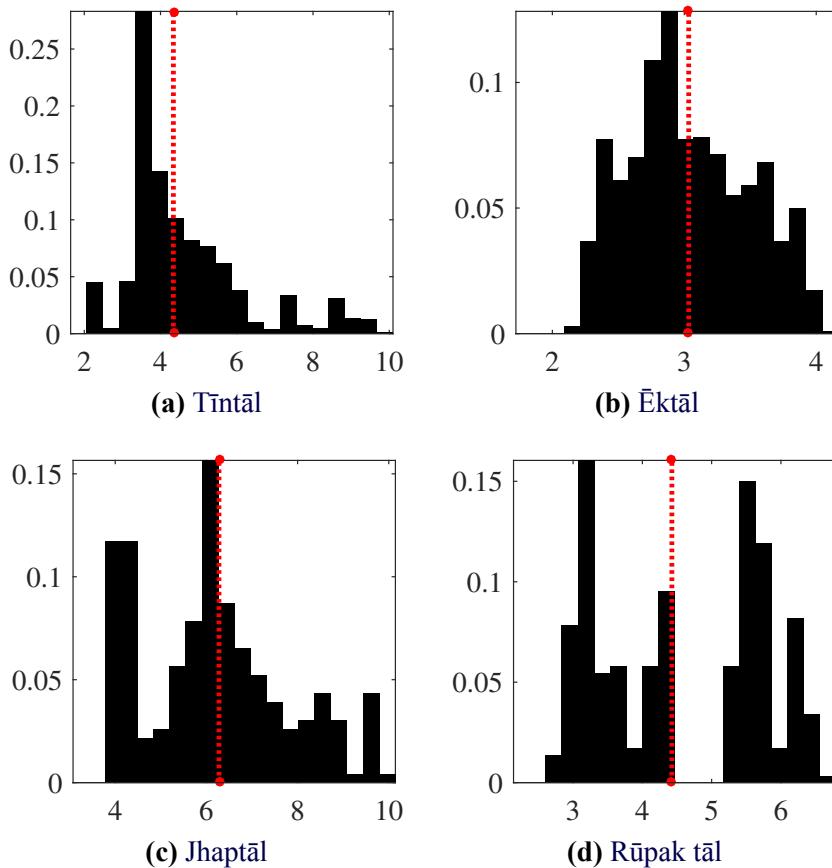
The dataset consists of excerpts with a wide tempo range from 10 MPM (matras per minute) to 370 MPM. As discussed in Chapter 2, Hindustani music divides tempo into three main tempo classes (lay). Since no exact tempo ranges are defined for these classes, we



**Figure 4.17:** A histogram of the inter-mātrā interval  $\tau_o$  in the  $HMR_1$  dataset for each tāl. The ordinate is the fraction of the total count corresponding to the  $\tau_o$  value shown in abscissa. The median  $\tau_o$  for each tāl is shown as a red dotted line.

determined suitable values, measured in mātrās per minute (MPM), in correspondence with a professional Hindustani musician as 10-60 MPM, 60-150 MPM, and >150 MPM for the slow (*vilāṁbit*), medium (*madhyā*), and fast (*dr̥t*) tempi, respectively.

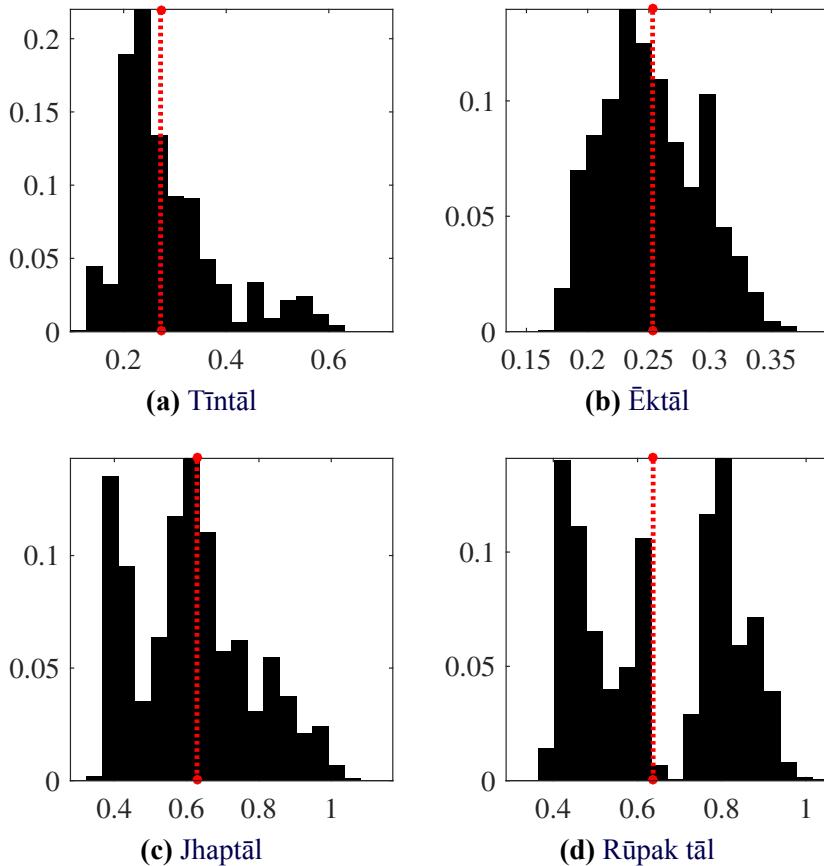
The lay of a piece has a significant effect on meter tracking and rhythm analysis due to this wide range of possible tempo. To study any effects of the tempo class, the full  $HMR_f$  dataset is divided into two other subsets - the long cycle subset called the  $HMR_1$  dataset (shown in Table 4.11) consisting of *vilāṁbit* pieces with a median tempo between 10-60 MPM, and the short cycle subset  $HMR_s$  dataset



**Figure 4.18:** A histogram of the inter-sam interval  $\tau_s$  in the HMR<sub>s</sub> dataset for each tāl. The ordinate is the fraction of the total count corresponding to the  $\tau_o$  value shown in abscissa. The median  $\tau_s$  for each tāl is shown as a red dotted line.

(shown in Table 4.13) with madhya lay (60-150 MPM) and the dṛ̥t̥ lay (150+ MPM) pieces.

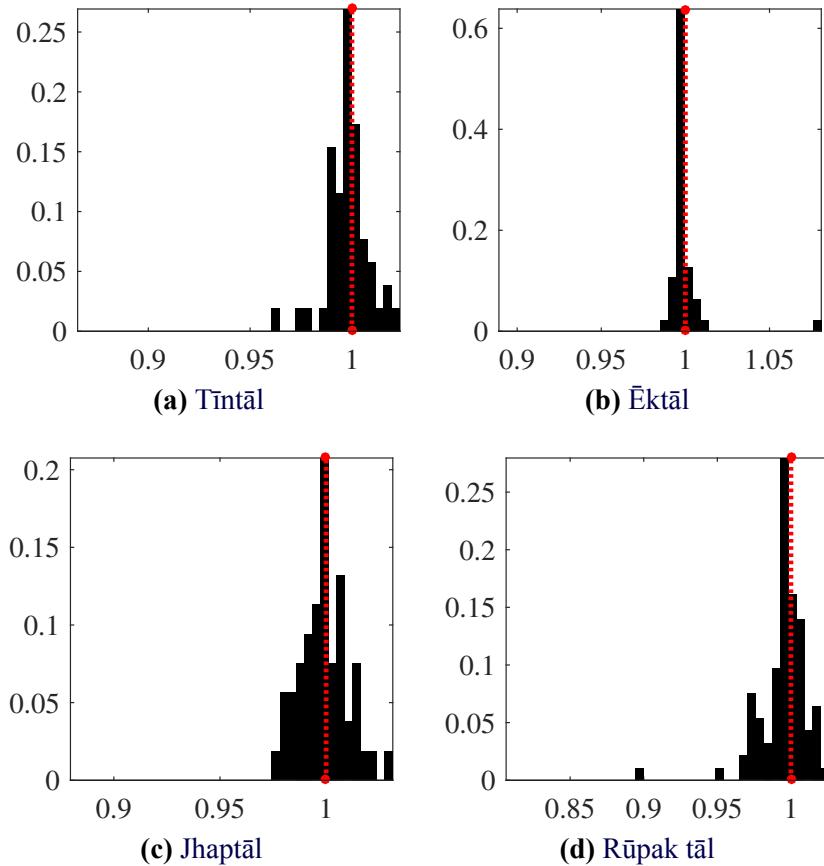
Hindustani music rhythm dataset (subset with vilāmbit and madhya lay pieces) (HMR<sub>1</sub>) dataset shown in Table 4.11 consists of 59 pieces in vilāmbit lay, with over 3200 mātrā and sam annotations. A majority of pieces are in ēktāl and tīntāl. Since it's very uncommon for a piece to be performed in vilāmbit lay jhaptāl and rūpak tāl, there are only 6 and 8 pieces for those tāls, respectively. As described with HMR<sub>f</sub>, a basic statistical analysis of the tāl cycle length indicators in Table 4.12 shows that the median inter-sam interval



**Figure 4.19:** A histogram of the inter-mātrā interval  $\tau_o$  in the HMR<sub>s</sub> dataset for each tāl. The ordinate is the fraction of the total count corresponding to the  $\tau_o$  value shown in abscissa. The median  $\tau_o$  for each tāl is shown as a red dotted line.

and its range for jhaptāl and rūpak tāl are less than that for tīntāl and ēktāl.

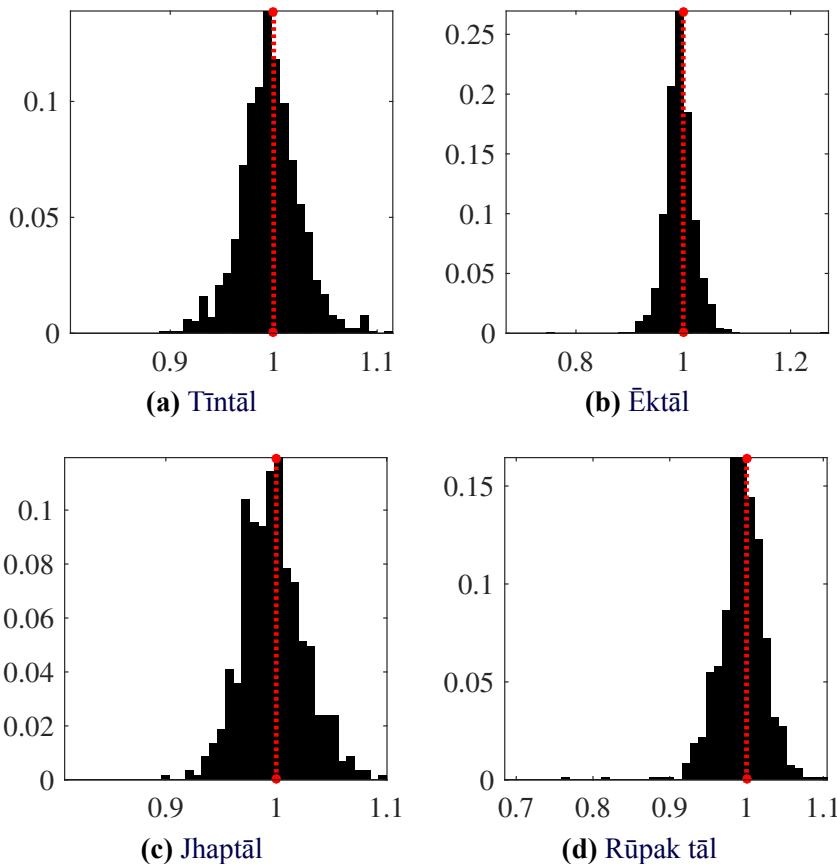
Hindustani music rhythm dataset (subset with dṛt lay pieces) (HMR<sub>s</sub>) dataset consists on 92 pieces in madhya and dṛt lay, with over 3 hours of audio and over 32000 mātrā and sam annotations. A basic statistical analysis of the tāl cycle length indicators in Table 4.14 shows that the pieces of tīntāl and ēktāl have higher tempi in the dataset. Comparing the median mātrā period for ēktāl between Table 4.12 (4.35 second) and Table 4.14 (0.26 second) shows that ēktāl is performed either in vilambit or dṛt and its rare for a



**Figure 4.20:** A histogram of the median normalized inter-sam interval  $\tau_s$  in the  $HMR_1$  dataset for each tāl. The ordinate is the fraction of the total count corresponding to the normalized  $\tau_s$  value shown in abscissa.

piece to be performed in madhya lay ēktāl.

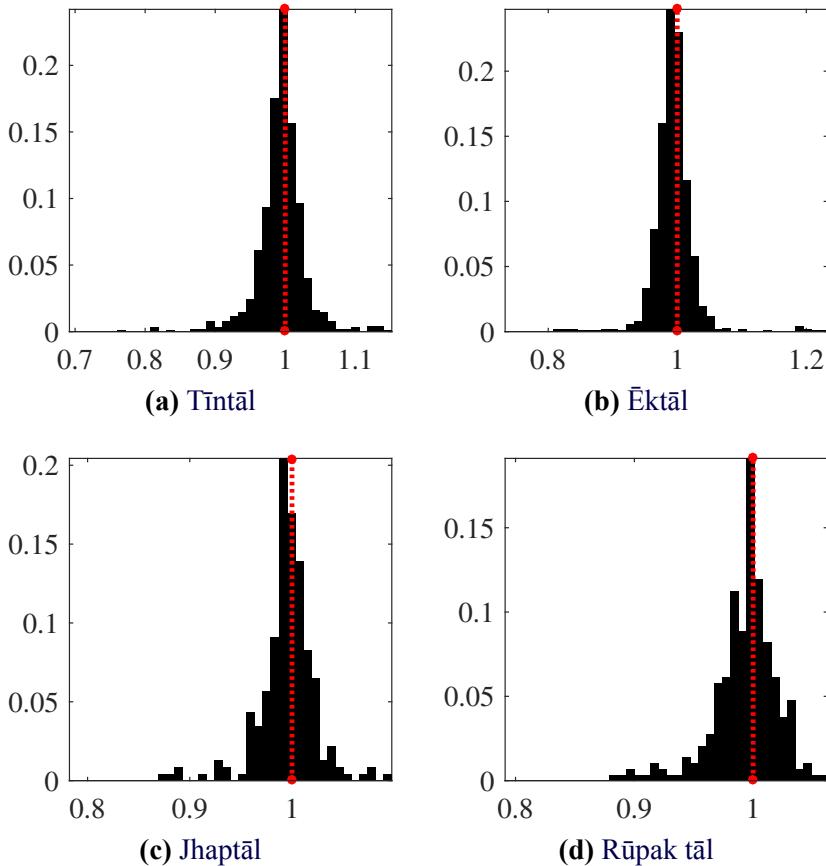
The pieces in Hindustani music have a tempo class indicated but not a specific tempo value, nor are they performed to a metronome. Hence the tempo varies over a piece in time - often the tempo increases with time. Hence, in addition to the median values tabulated in Table 4.6 we present further analysis of the inter-sam interval ( $\tau_s$ ) and inter-mātrā interval ( $\tau_o$ ) for each tāl. For better comparison, we present this analysis for each data subset  $HMR_1$  and  $HMR_s$  separately. A histogram of  $\tau_s$  and  $\tau_o$  for each tāl for  $HMR_1$  dataset is shown in Figure 4.16 and Figure 4.19, respectively, and those



**Figure 4.21:** A histogram of the median normalized inter-mātrā interval  $\tau_o$  in the HMR<sub>I</sub> dataset for each tāl. The ordinate is the fraction of the total count corresponding to the normalized  $\tau_o$  value shown in abscissa.

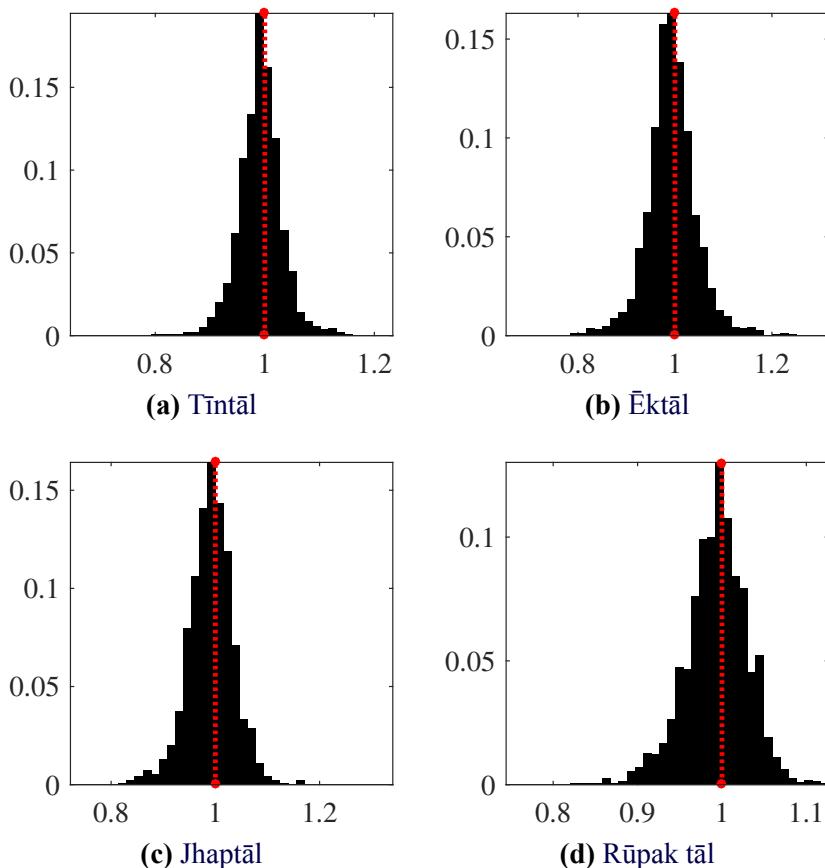
for HMR<sub>s</sub> dataset is shown in Figure 4.18 and Figure 4.19, respectively. These figures show the distribution of cycle lengths in the dataset over the whole range of  $\tau_s$  for each tāl, around the median value. The large range of  $\tau_s$  and  $\tau_o$  values and an irregular distribution spanning the whole range is seen with both datasets, unlike the Carnatic music CMR<sub>f</sub> dataset with a short tightly defined range of tempo.

In addition, similar to what was presented for Carnatic music, to illustrate and measure the time varying tempo of music pieces in Hindustani music, we normalize all the  $\tau_s$  and  $\tau_o$  values in a



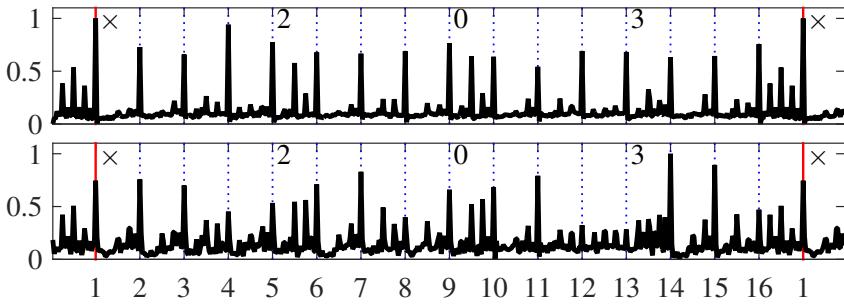
**Figure 4.22:** A histogram of the median normalized inter-sam interval  $\tau_s$  in the HMR<sub>s</sub> dataset for each tāl. The ordinate is the fraction of the total count corresponding to the normalized  $\tau_s$  value shown in abscissa.

piece by the median in the piece to obtain median normalized  $\tau_s$  and  $\tau_o$  values, a histogram of which is shown in Figure 4.20 and Figure 4.21, respectively for HMR<sub>1</sub> dataset and Figure 4.22 and Figure 4.23, respectively for HMR<sub>s</sub> dataset. These histograms are centered around 1 and normalized by the median. From the figures, it is clear that the tempo is time varying but with less than about 10% maximum deviation from the median tempo of the piece for all tāls. This is in contrast to Carnatic music where the median normalized tempo had a higher deviation ( $\sim 20\%$ ). However, this could also possibly due to the fact that the Hindustani pieces in the dataset are

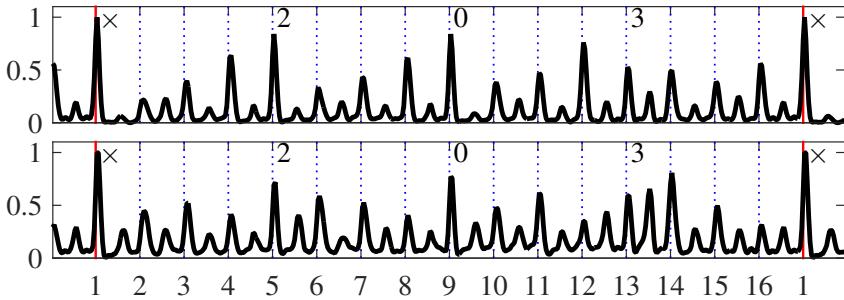


**Figure 4.23:** A histogram of the median normalized inter-mātrā interval  $\tau_o$  in the  $\text{HMR}_s$  dataset for each tāl. The ordinate is the fraction of the total count corresponding to the normalized  $\tau_o$  value shown in abscissa.

two minute short excerpts, compared to full length Carnatic pieces in the  $\text{CMR}_f$  dataset.



**Figure 4.24:** Cycle length rhythmic patterns learned from  $\text{HMR}_1$  dataset for  $\text{tīntāl}$ , computed from spectral flux feature and averaged over all the pieces in the dataset. The bottom/top pane corresponds to the low/high frequency bands, respectively. The abscissa is the  $\text{mātrā}$  number within the cycle (dotted lines), with 1 indicating the *sam* (marked with a red line). The start of each *vibhāg* is indicated at the top of each pane (*sam* shown as  $\times$ ).

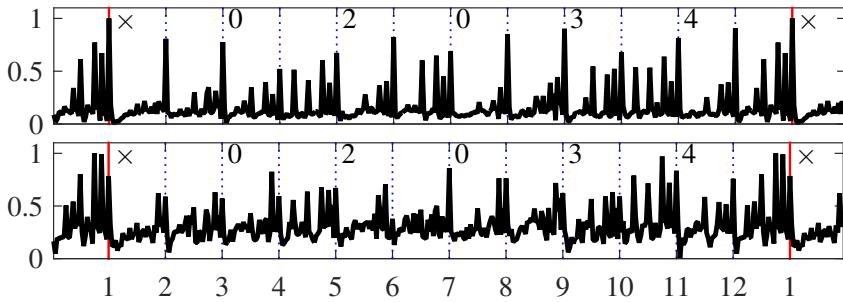


**Figure 4.25:** Cycle length rhythmic patterns learned from  $\text{HMR}_s$  dataset for  $\text{tīntāl}$ .

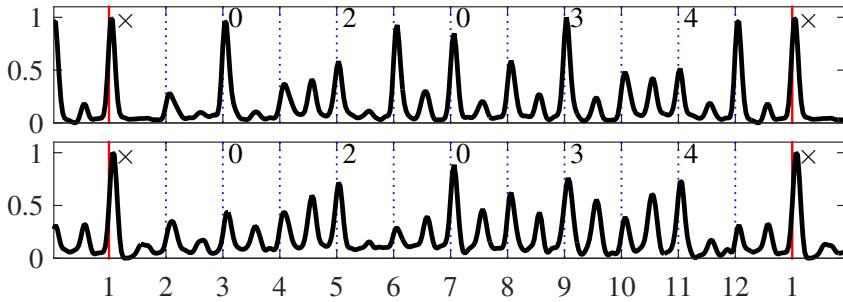
### Rhythm patterns in Hindustani rhythm datasets

Similar to Carnatic music, we do corpora level analysis of rhythm patterns in Hindustani music and draw several musicological inferences and insights, contrasting the differences between music theory and practice **explain further: what are these differences ?**. The rhythm patterns described in this section were obtained using spectral flux, identical to the process described for Carnatic music.

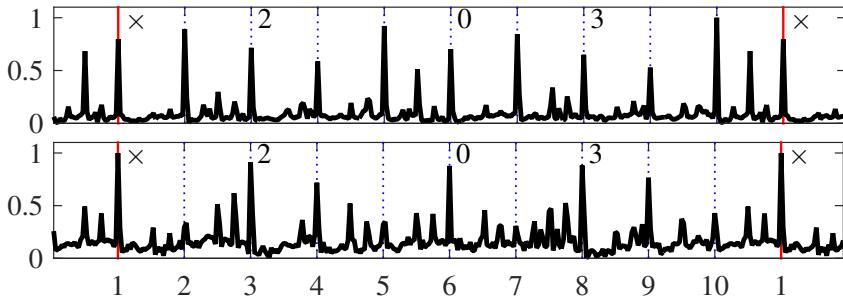
The Figures 4.24-4.31 show the cycle length rhythm patterns



**Figure 4.26:** Cycle length rhythmic patterns learned from  $\text{HMR}_1$  dataset for  $\text{\texttt{ekta\k{a}l}}$ .

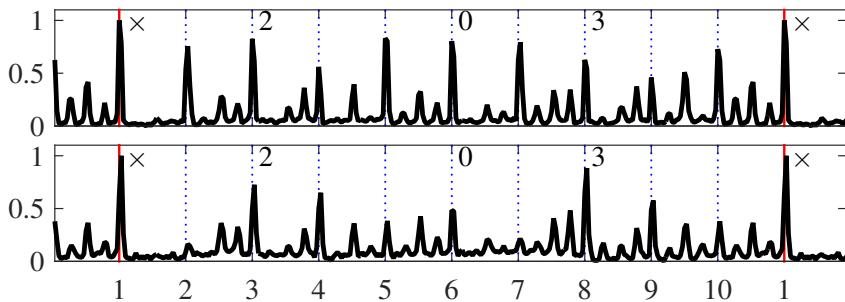


**Figure 4.27:** Cycle length rhythmic patterns learned from  $\text{HMR}_s$  dataset for  $\text{\texttt{ekta\k{a}l}}$ .

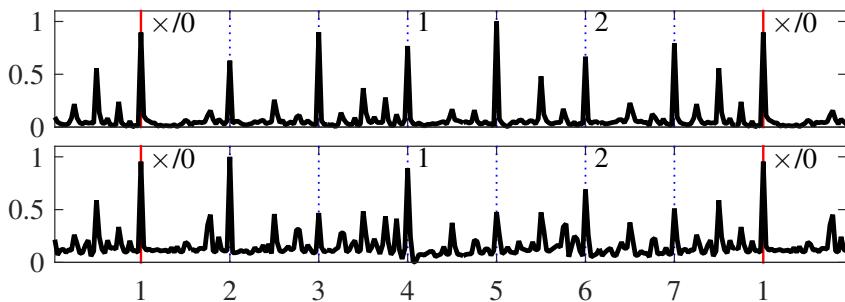


**Figure 4.28:** Cycle length rhythmic patterns learned from  $\text{HMR}_1$  dataset for  $\text{jhap\k{a}t\k{a}l}$ .

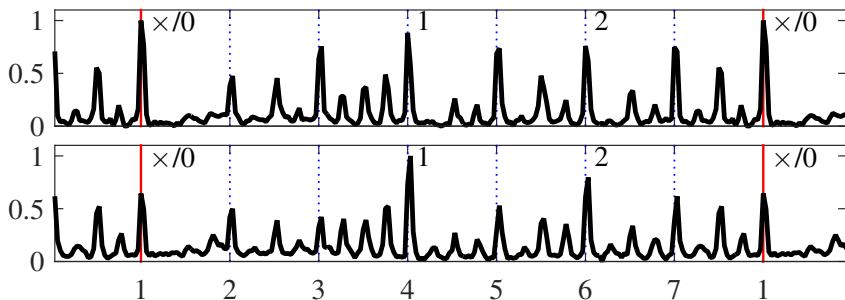
for all  $\text{t\k{a}ls}$  for both  $\text{HMR}_1$  and  $\text{HMR}_s$  datasets, using the spectral flux feature computed identically to the way it was computed for Car-



**Figure 4.29:** Cycle length rhythmic patterns learned from  $\text{HMR}_s$  dataset for *jhaptāl*.



**Figure 4.30:** Cycle length rhythmic patterns learned from  $\text{HMR}_1$  dataset for *rūpak tāl*.



**Figure 4.31:** Cycle length rhythmic patterns learned from  $\text{HMR}_s$  dataset for *rūpak tāl*.

natic music rhythm patterns, as an average over the entire dataset indicated. In each figure, the bottom pane corresponds to the low

frequency band ( $y_l$ ) and the top pane corresponds to the high frequency band ( $y_h$ ). The abscissa is the *mātrā* number within the cycle (dotted lines), with 1 indicating the *sam* (marked with a red line). The start of each *vibhāg* is indicated at the top of each pane (*sam* shown as  $\times$ ).

The rhythm patterns in Hindustani are indicative of tabla strokes played in the cycle. In the figures, the bottom pane that shows the low frequency band has content from the *bāyān* (the left bass drum) of the tabla while the top pane has content predominantly from the *dāyān* (the right pitched drum) of the tabla, but additionally from the lead melody. Hence, for the purpose of this discussion, we use the terms left and right accents to refer to the accents in rhythm patterns from the bottom and top pane, respectively. The left and right accents provide interesting insights into the patterns played within a *tāl* cycle. We additionally compare rhythm patterns across the *layas* by plotting the patterns for *HMR<sub>1</sub>* dataset (with *vilāmbit* lay pieces) and *HMR<sub>s</sub>* dataset (*madhya* and *dṛ̥t* lay pieces) - for each *tāl*, the patterns for these two data subsets are plotted in two figures one below the other.

The patterns played in a *tāl* cycle have both energy/amplitude accents due to varying strength of the tabla stroke and also timbral characteristics, due to the specific stroke played. The rhythm patterns have been generated using the spectral flux feature, which models mostly only energy, and hence can only explain energy accents with these figures. We list down and discuss some salient qualitative observations from the figures for each *tāla*, for both *vilāmbit* lay and *dṛ̥t* lay. The patterns are indicative of the surface rhythm present in these audio recordings.

Overall, from Figures 4.24-4.31, we observe across all *tāls* and *layas* that accents are stronger on the *mātrās*, with accents present even at half and fourth divisions of the matra in many cases. The *sam* most often has the strongest accent. Unlike Carnatic *tālas*, *ṭhēkās* in Hindustani music are less flexible, and hence we can infer several concrete conclusions from the rhythm patterns of Hindustani music.

Across all *tāls* in *vilāmbit* and *madhya* lay, we see additional filler strokes present between *mātrās*, showing that percussionists add further metrical subdivisions lower than the *mātrā*, though not defined in theory. These fillers are also mostly mostly towards to

second half of the *mātrā*. The 1<sup>st</sup> *mātrā* (and often the 2<sup>nd</sup> *mātrā*) is quite empty with few accents, while the last few *mātrās* of the cycle have dense accents. This places a special emphasis on the *sam*, indicating the approaching of *sam* with fillers and dense stroke playing and a short recovery period with fewer strokes after the *sam* has passed. In addition, a dense matra with many fillers is often followed by a sparsely accented *mātrā* to better contrast the progression through the *tāl* cycle, e.g. *mātrā* 9 after a quieter *mātrā* 8 in Figure 4.24. Due to the large *mātrā* period ( $\tau_o$ ) in *vilambit* and *madhya lay*, each *mātrā* acts as an anchor for timekeeping, and can be played without any effect from the previous strokes (in fast tabla playing in *dṛt*, the previous stroke can possibly affect the sound, intonation, and playing technique of the following strokes). In addition, due to a large time interval available to play the *ṭhēkā*, the tabla playing musician focuses on modulation of left bass strokes that can sustain longer. Finally, left and right hand can operate independently, which means modulation of accents through the cycle can be different for left and right accents. The left and right strokes also complement each other. Each of these effects can be observed in the patterns of *vilambit* and *madhya lay*.

In contrast, across all *tāls* in *dṛt lay*, given the short cycles, we see that *vibhāgs* are anchors. The fillers are largely restricted only to half *mātrā*, with lower accents. *Dṛt* pieces also has a relatively more relaxed timing, and the focus is on right strokes, with the left hand playing the theory defined “textbook” strokes for timekeeping. In addition, the left and right hands are in sync, which can be seen in the modulation of accents through the cycle being well correlated for both left and right accents - the left and right strokes work together here, in contrast to complementing each other as in *vilambit lay*. Furthermore, the patterns differ widely between the *lay* classes, especially for *ēktāl* and *tīntāl*.

We now present some *tāl* specific observations from the rhythm patterns for each *tāl*. Some of these observations corroborate theory while some of them show the contrast between theory and practice. These inferences mainly address tabla stroke playing during the cycles, while the effects of melody has not been considered into account. This is a valid assumption to make since these patterns are averaged over several cycles, averaging out and reducing the effect of melody on these rhythm patterns.

**Vilambit and madhya lay tīntāl:** From Figure 4.24, we see that the 14th matra has the strongest left accent, and the last mātrā (matra 16) has many fillers, both to indicate the arrival of sam - a phenomenon known in music theory as āmad (literal meaning - the approach). A strong left accent on the 9<sup>th</sup> matra is not defined in theory (the stroke in the thēkā is a right stroke NA), but often a DHA is played instead. This is a difference between theory and practice that is known to musicians and can be observed in the pattern here too. As described earlier, the right stroke fillers are fewer in mātrās 1 and 2, and the left supports the timekeeping task when the right accents are weaker there. 4<sup>th</sup> mātrā has a strong right accent, to indicate the end of the 1<sup>st</sup> vibhāg, after a filler-less mātrās 2 and 3. The beginning of the 2<sup>nd</sup> and 3<sup>rd</sup> vibhāgs, labeled 2 and 0 have larger number of fillers. The left accents between the 11<sup>th</sup> and the 14<sup>th</sup> matra are weak - with the 11<sup>th</sup> and 14<sup>th</sup> mātrā accents acting as anchors for the “quiet” created in between them. It is interesting to note the varying modulation of accent levels through the vibhāgs of the cycle. Specifically, we can see that the left and right accent envelopes through the cycle are complementary, indicating that left and right drums are complementary in vilambit lay.

**Dṛt tīntāl:** From Figure 4.25, we see that the filler strokes in dṛt tīntāl are restricted to a single filler at half mātrā positions in contrast to three of more fillers in vilambit. The accents are more regular due to higher tempi associated. Similar to vilambit, the 9<sup>th</sup> matra has a strong left accent, which again is a well known difference between theory and practice. The 11<sup>th</sup> and 14<sup>th</sup> mātrās have high left accents to support the build up of accents through mātrās 12-14 to indicate the arrival of sam (āmad). It is interesting to note the the vibhāg boundary mātrā 13 has a weaker right accent than the previous mātrā 12 right accent. The stroke on mātrā 13 is skipped and a strong left stroke on mātrā 14 is often played to indicate the approaching sam.

**Vilambit and madhya lay ēktāl:** From Figure 4.26, we see that the last matra of the cycle before the sam (mātrā 12) has dense accents, with the final filler strokes having stronger left accents than the sam. This is another example of āmad, where the approach of a sam is distinctly indicated. The mātrās 4 and 10 (both with the

ṭhēkā bōl TI RA KI TA) have equal accents in theory. However, mātrā 10 has stronger accents than 4 in practice since it is closer to the sam. TI RA KI TA is often played with more than four strokes towards the end of the matra 4 and 10. Since TI RA KI TA is dense, the mātrā following them (mātrās 5 and 11) have less fillers. In addition, only mātrās 4 and 10 have fillers through the mātrā, while the rest have fillers only towards the end. Vibhāgs 2 and 3 (spanning mātrās 3-6) and vibhāgs 5 and 6 (spanning mātrā 9-×) are identical in theory, but we can see several deviations in performance, with vibhāgs 5 and 6 having stronger left accents since they are closer to sam. Further, the strokes DHIN at mātrā 1 and mātrā 2 are identical in theory, but in practice the DHIN at mātrā 2 is played softer to differentiate it from the DHIN at the sam. The modulation of right accent levels through the cycle is interesting, with stronger accents occurring when the mātrā is less dense with lesser number of accents. This has a functional role in timekeeping - aided by stronger accents and denser mātrās, complementing each other.

**Dṛt ēktāl:** Though defined with six vibhāgs in theory, dṛt ēktāl is described better as having four vibhāgs of 3 mātrās each, as shown in Figure 2.4, with the vibhāgs starting at mātrās 1, 4, 7, and 10. As can be seen from Figure 4.27, the strong right accents due to NA stroke at mātrās 3, 6, 9 and 12 are distinctly seen. This suggests that for dṛt lay, timekeeping is done more with the sharp right strokes (e.g. ‘NA’ here) and accentuation can even be at non-vibhāg marker mātrās such as 6 and 12. Even though the last vibhāg starts on matra 10, there is strong right accent on matra 9, an indication of the approaching sam (āmad). The four strokes in TI RA KI TA is often not played in dṛt, replacing it with just two strokes TRE KE - we see only two accents in mātrās 4 and 10. In addition, due to the dense stroke playing on mātrā 4 and 10, the left accents in mātrā 6 and 12 are quiet with relatively weaker accents. Similar to vilāmbit ēktāl, though the first and second matra have equal accented DHIN stroke in theory, DHIN on the second mātrā is played considerably softer with weak accent. As with all tāls in dṛt lay, the accents on left and right through the cycle are correlated.

**Vilāmbit and madhya lay jhaptāl:** From Figure 4.28, we see that all the NA strokes (mātrās 2, 5, 7, 10) have a strong right accent and

weak left accents, as described in theory. There are filler strokes to end the *vibhāgs* at mātrās 2 and 7. This can be explained with the often played variant of the *ṭhēkā* (DHI NA TRE - KE DHI DHI NA | TI NA TRE - KE DHI DHI NA). There are further strong accented fillers on mātrās 5 and 10 that act as anchors to indicate the end of half and full cycle.

**Dṛt jhaptāl:** Figure 4.29 shows the left accents are as defined in theory (a “textbook” bāyān playing) with basic *ṭhēkā* playing. The envelope of accents through the cycle is more regular than in *vilāmbit jhaptāl*. In theory, the *vibhāg* 2 (mātrās 3-5) and *vibhāg* 4 (mātrās 8-10) are identical, but some deviations can be observed in practice.

**Vilāmbit and madhya lay rūpak tāl:** Rūpak tāl is defined in theory with no left accents on mātrās 1 and 2, but in practice left strokes are often played (with closed strokes than modulated sustained left strokes). This also implies that *rūpak tāl* having a *khālī* (0) on the *sam* does not mean it is less accented. Rūpak tāl is defined to have a 3+2+2 structure, but we see from Figure 4.30 that mātrā 2 has a strong left accent, which acts as an anchor, giving the *vilāmbit rūpak tāl* a 1+2+2+2 structure, which is close to the tapping of *miśra chāpu tāla* of Carnatic music in practice. This could also be because musicians might play with the same accent on both TIN (mātrās 1 and 2) with a KAT stroke to contrast with the NA stroke which is less left-accented. The *vibhāg* 2 (mātrās 4-5) and *vibhāg* 3 (mātrā 6-7) are identical in theory, but in practice the accents differ. Mātrā 5 has the strongest right accent (NA stroke), perhaps indicating āmad. Fillers are more on mātrā 3, to end *vibhāg* 1. In general, we also see that the fillers get more dense towards the end of *vibhāgs*.

**Dṛt rūpak tāl:** From Figure 4.31, the left strokes and accents closely follow the description in theory. The strongest left accent is on mātrā 4, as defined in theory. The *vibhāg* 2 and 3 are identical with similar accents. Interestingly, the fillers grow through the cycle, becoming more dense towards the end of the cycle.

### Applications of the HMR<sub>f</sub> dataset

The HMR<sub>f</sub> dataset and its subsets HMR<sub>l</sub> and HMR<sub>f</sub> datasets are intended to be test corpora for several computational rhythm analysis tasks in Hindustani music. Possible tasks include Possible tasks where the dataset can be used include sam and mātrā tracking, tempo estimation and tracking, tāl recognition, rhythm based segmentation of musical audio, audio to score/lyrics alignment, and rhythmic pattern discovery. In this thesis, these datasets are primarily used for rhythmic pattern analysis and meter inference/tracking. Most of the research results are presented for the three subsets separately, to contrast performance of algorithms across different lay.

#### 4.2.3 Tabla solo dataset

The Mulgaonkar Tabla Solo dataset (MTS dataset) is a parallel corpus of tabla solo compositions with time-aligned scores and audio recordings. We built a dataset comprising audio recordings, scores and time aligned syllabic transcriptions of 38 tabla solo compositions of different forms in tīntāl. The compositions were obtained from the instructional video DVD *Shades Of Tabla* by Pandit Arvind Mulgaonkar<sup>24</sup>. Out of the 120 compositions in the DVD, we chose 38 representative compositions spanning all the gharānās of tabla (Ajrada, Benaras, Dilli, Lucknow, Punjab, Farukhabad).

The booklet accompanying the DVD provides a syllabic transcription for each composition. We used Tesseract(Smith, 2007), an open source Optical Character Recognizer (OCR) engine to convert printed scores to a machine readable format. The scores obtained from OCR were manually verified and corrected for errors, adding the vibhāgs (sections) of the tāl to the syllabic transcription. The score for each composition has additional metadata describing the gharānā, composer and its musical form.

We extracted audio from the DVD video and segmented the audio for each composition from the full audio recording. The audio recordings are stereo, sampled at 44.1 kHz and have a soft harmonium accompaniment. A time aligned syllabic transcription for each score and audio file pair was obtained using a spectral flux

---

<sup>24</sup><http://musicbrainz.org/release/220c5efc-2350-43dd-95c6-4870dc6851f5>

ID	Syllable	#Inst.	ID	Syllable	#Inst.
1	DA	132	10	KI	1482
2	DHA	582	11	NA	1308
3	DHE	277	12	RE	294
4	DHET	67	13	TA	2375
5	DHI	156	14	TE	18
6	DHIN	149	15	TII	64
7	DIN	117	16	TIN	61
8	GE	961	17	TIT	43
9	KDA	95	18	TRA	64

**Table 4.15:** The tabla dataset with 8245 syllables, showing the number of instances of each syllable in the dataset. The syllable group names correspond to that presented in Table 2.4.

based onset detector (Bello et al., 2005) followed by manual correction. The dataset contains about 17 minutes of audio with over 8200 syllables. The syllables in the dataset are grouped based on timbre as described in Table 2.4, and Table 4.15 lists the number of instances in the dataset for each group syllable. The dataset is freely available for research purposes through a central online repository<sup>25</sup>. The dataset was created in collaboration with Swapnil Gupta and more details are described in the masters thesis by Gupta (2015). The dataset is useful both for building isolated stroke timbre models and for a comprehensive evaluation of tabla solo pattern transcription and discovery, as used by Gupta et al. (2015). The scores in the dataset can be used to do symbolic analysis of percussion patterns.

#### 4.2.4 Mridangam datasets

There are two percussion datasets for Carnatic music: a collection of audio examples of mridangam strokes compiled by Akshay Anantapadmanabhan, and a parallel corpus of scores and audio recordings of mridangam solos played by Padmavibhushan Dr. Umayalpuram K. Sivaraman and compiled by IIT Madras, Chennai, India.

---

<sup>25</sup><http://compmusic.upf.edu/tabla-solo-dataset>

### Mridangam stroke dataset

The Anantapadmanabhan Mridangam Strokes dataset (AMS dataset)<sup>26</sup> is a collection of 7162 audio examples of individual strokes of the mridangam in various tonics. The dataset can be used for training models for each mridangam stroke (Anantapadmanabhan et al., 2013). The dataset comprises of ten different strokes played on mridangams with six different tonic values. The audio examples were recorded from a professional Carnatic percussionist in semi-anechoic studio conditions using SM-58 microphones and an H4n ZOOM recorder. The audio was sampled at 44.1 kHz and stored as 16 bit wav files.

The dataset is described in Table 4.16, with stroke labels along rows and tonic values along columns. As can be seen from the table, the dataset uses different stroke labels names compared to the notation used in the dissertation, and hence the analogous syllabic symbol corresponding to each stroke label is also shown in the table.

### Mridangam solo dataset

The UKS Mridangam Solo dataset (UMS dataset) is a transcribed collection of two *tani-āvartanas* (solo performance by the percussion ensemble) played by the renowned mridangam maestro Padmavibhushan Umayalpuram K. Sivaraman. The audio was recorded at IIT Madras, India and annotated by professional Carnatic percussionists (Kuriakose et al., 2015).

Since percussion in Carnatic music is organized and transmitted orally with the use of onomatopoeic syllables representative of the different strokes of the mridangam, a syllabic representation of the *tani* and the patterns provides a musically meaningful representation for analysis. The dataset uses such a representation. The dataset consists of two *tani-āvartanas* played on a mridangam tuned to tonic C#, one played in *vilambita ādi tāla* (a cycle of 16 beats) and the other played in *rūpaka tāla*. Each *tani* is about 12 minutes long. Each *tani* has been segmented into short phrases and each phrase has been transcribed into its constituent strokes, represented as syllables. The transcriptions also include pauses (denoted by , )

---

<sup>26</sup><http://compmusic.upf.edu/mridangam-stroke-dataset>

Stroke	B	C	C#	D	D#	E	Total	Syl.
<b>Bheem</b>	5	3	1	0	15	25	<b>49</b>	??
<b>Cha</b>	57	50	54	67	49	53	<b>330</b>	CH
<b>Dheem</b>	127	86	78	12	111	54	<b>468</b>	DM
<b>Dhin</b>	48	48	63	12	198	113	<b>482</b>	DN
<b>Num</b>	81	98	97	18	143	60	<b>497</b>	NM
<b>Ta</b>	145	165	217	180	119	105	<b>931</b>	TA
<b>Tha</b>	200	185	211	224	196	160	<b>1176</b>	TH
<b>Tham</b>	88	80	35	29	92	50	<b>374</b>	??
<b>Thi</b>	438	334	369	283	444	345	<b>2213</b>	DH3
<b>Thom</b>	136	80	72	91	128	135	<b>642</b>	TM
<b>Total</b>	<b>1325</b>	<b>1129</b>	<b>1197</b>	<b>916</b>	<b>1495</b>	<b>1100</b>	<b>7162</b>	

**Table 4.16:** The Anantapadmanabhan Mridangam Strokes dataset. The row and column headers are the stroke labels and the tonic values, respectively. The last column shows the analogous syllable used in the dissertation. **Two syllables to be mapped, check!**

ID	Syllable	#Inst.	ID	Syllable	#Inst.
1	AC	119	12	DNT	922
2	ACT	50	13	LF	467
3	CH	114	14	LFT	12
4	CHT	112	15	NM	850
5	DM	14	16	NMT	632
6	DH3	1266	17	TH	776
7	DH3T	23	18	TA	754
8	DH3M	602	19	TAT	13
9	DH4	367	20	TM	913
10	DH4T	12	21	TG	30
11	DN	829	-	-	-

**Table 4.17:** The UMS dataset with 8877 syllables, showing the number of instances of each syllable in the dataset. The syllable group names correspond to that presented in Table 2.2.

and change in speed (denoted by { and } ). The combined duration

of both the tanis is about 24 minutes and consists of 8863 strokes. The stroke syllables are grouped based on timbre as described in Table 2.2 into syllable groups, and the dataset is described in Table 4.17, showing the number of instances for each syllable (group) in the audio recordings.

Both tanis were recorded in studio-like conditions using a Zoom H4n recorder with an SM 57 for the treble head (right) and SM 58 for the base head (left) of the mridangam. The audio files are mono, sampled at 44.1KHz, and stored in 16 bit .wav format. The audio file has been segmented into short musically relevant phrases by professional musicians. The syllabic transcription of each phrase was done by professional Carnatic percussionists. The transcription is not time aligned, but only a sequence of the strokes played in the phrase.

The dataset can be used for several MIR tasks such as onset detection, percussion transcription, rhythm and percussion pattern analysis, and mridangam stroke modeling. The dataset (audio + annotations) is freely available for research purposes <sup>27</sup> and has been recently used by Kuriakose et al. (2015) in their work.

#### 4.2.5 Jingju percussion instrument dataset

The Jingju percussion instrument dataset (JPI dataset) is an annotated collection of Beijing opera percussion instruments, with audio and time aligned onset annotations. The dataset is split into training set with audio files containing single strokes of individual percussion instruments and a test dataset that has the whole percussion ensemble playing together.

The dataset was built by Mi Tian at the Centre for Digital Music (C4DM), Queen Mary University of London. The dataset was built by recording sound samples with professional musicians in studio conditions at C4DM. The audio was recorded in mono using an AKG C414 microphone at a sampling rate of 44.1 KHz.

The dataset, shown in Table 4.18, consists of recordings of the four percussion instrument classes: bangu, daluo, naobo and xi-aoluo. Unlike pitched instruments, most idiophones cannot be tuned. These percussion instruments are made from metal casting or wood

---

<sup>27</sup><http://compmusic.upf.edu/mridangam-tani-dataset>

Dataset	Bangu	Daluo	Naobo	Xiaolu	Total
Training	59	50	62	65	<b>236</b>
Test	1645	338	747	291	<b>3021</b>

**Table 4.18:** The Jingju percussion instrument dataset (JPI dataset) showing the number of examples for each instrument in the training and test dataset.

carving hence subtle differences might exist between the physical properties of individual instruments even of the same kind. For each kind of the above instruments, sound samples of 2-4 individual instruments were recorded, played with different playing styles commonly used in Beijing Opera performances with a hope to achieve a better coverage of timbre and variations of playing techniques.

The training set consists of short audio samples with single strokes of each individual instrument that capture most of the possible timbres of the instrument that exist in Beijing Opera. For the test dataset, the individually recorded instrument examples were manually mixed together using Audacity<sup>28</sup> into 30-second long tracks, with possibly simultaneous onsets to closely reproduce the real world conditions. The examples in training and test dataset are mutually exclusive.

For the onset annotations, manual labeling of onset locations is tedious and time consuming, especially for complex ensemble music consisting of instruments with diverse properties. The onset ground truth was constructed by taking the average onset locations marked by three participants without any Beijing Opera background. Participants were asked to mark the onset locations in each recording using the audio analysis tool Sonic Visualiser (Cannam et al., 2010) displaying the waveform and corresponding spectrogram.

The set of training examples are freely available for research and reuse<sup>29</sup>. The dataset can be used for training models for each percussion instrument class, and MIR tasks such as percussion in-

<sup>28</sup><http://audacity.sourceforge.net>

<sup>29</sup><http://compmusic.upf.edu/bo-perc-dataset>

ID	Pattern Class	# Instances	$\bar{T}_f (\sigma)$
1	dǎobǎn tóu 【导板头】	66	8.70 (1.73)
2	màn chángchuí 【漫长锤】	33	13.99 (4.47)
3	duótóu 【夺头】	19	7.18 (1.49)
4	xiǎoluó duótóu 【小锣夺头】	11	8.16 (2.15)
5	shǎnchuí 【闪锤】	8	10.31 (3.26)
<b>Total</b>		<b>133</b>	<b>9.85 (3.69)</b>

**Table 4.19:** The Jingju percussion pattern dataset (JPP dataset). The last column is the mean pattern length ( $\bar{T}_f$ ) and standard deviation ( $\sigma$ ) in seconds.

strument identification, source separation, and instrument-wise enhanced onset detection, as used by Tian et al. (2014).

#### 4.2.6 Jingju percussion pattern dataset

The Jingju percussion pattern dataset (JPP dataset) is a collection of audio examples and scores of percussion patterns played by the percussion ensemble in Jingju. The dataset was built from commercial jingju aria recordings with the help of Rafael Caro, a musicologist working on jingju.

The dataset is a collection of 133 audio percussion patterns spanning five different pattern classes described in Section 2.2.5, and comprises about 22 minutes of audio with over 2200 syllables in total. The audio files are short segments containing one of the above mentioned patterns. The audio is stereo, sampled at 44.1 kHz, and stored as wav files. The segments were chosen from the introductory parts of arias, which are characteristic and important. The recordings of arias are from commercially available releases spanning various artists. The audio and segments were chosen carefully by a musicologist to be representative of the percussion patterns that occur in Jingju. The audio segments contain diverse instrument timbres of percussion instruments (though the same set of instruments are played, there can be slight variations in the individual instruments across different ensembles), recording quality and period of the recording. Though these recordings were chosen from introductions of arias where only percussion ensemble is

playing, there are some examples in the dataset where the melodic accompaniment starts before the percussion pattern ends.

The syllabic transcription of each audio pattern is obtained directly from the score of the pattern class it belonged to.

Each of the audio patterns has an associated syllable level transcription of the audio pattern. The syllabic transcription of each audio pattern is directly obtained from the score of the pattern and hence is not time aligned to the audio. In case of patterns where a sub-sequence of the pattern can be repeated (e.g. *man changchui* and *shanchui*), the additional syllables that occur due to repetitions were manually added by listening to the pattern. Though most of the dataset consists of isolated percussion patterns, there are many audio examples that contain a melodic background apart from the percussion pattern. The transcription is done using the reduced set of five syllables described in Table 2.6 and is sufficient to computationally model the timbres of all the syllables. The annotations are stored as **Hidden Markov model Toolkit (HTK)**<sup>30</sup> label files. There is also a single master label file provided for batch processing using HTK.

The annotations are publicly shared and available to all<sup>31</sup>. The audio is from commercially available releases and can be easily accessed using the associated MusicBrainz IDs. The dataset can be used for instrument-wise onset detection and percussion pattern transcription and classification(Srinivasamurthy, Caro, et al., 2014).

#### 4.2.7 Other evaluation datasets

**Are descriptions of Turkish and Cretan datasets needed, since not many results are included ?** There are other datasets on which we present some evaluation results.

##### Turkish rhythm dataset

The Turkish rhythm dataset was compiled and annotated by Andre Holzapfel (**citation needed**) and is an extended version of the annotated data used by Srinivasamurthy, Holzapfel, and Serra (2014).

<sup>30</sup><http://htk.eng.cam.ac.uk/>

<sup>31</sup><http://compmusic.upf.edu/bopp-dataset>

It includes 82 excerpts of one minute length each, and each piece belongs to one of three rhythm classes that are referred to as *usul* in Turkish Art music. 32 pieces are in the 9/8-usul *Aksak*, 20 pieces in the 10/8-usul *Circuna*, and 30 samples in the 8/8-usul *Düyük*.

### Cretan music dataset

The corpus of Cretan music consists of 42 full length pieces of Cretan leaping dances compiled and annotated by Andre Holzapfel ([citation needed](#)). While there are several dances that differ in terms of their steps, the differences in the sound are most noticeable in the melodic content, and we consider all pieces to belong to one rhythmic style. All these dances are usually notated using a 2/4 time signature, and the accompanying rhythmical patterns are usually played on a Cretan lute. While a variety of rhythmic patterns exist, they do not relate to a specific dance and can be assumed to occur in all of the 42 songs in this corpus.

### Ballroom dataset

The ballroom dataset includes beat and bar annotations audio recordings of several dance styles sourced from [BallroomDancers.com](#) and was first introduced by Gouyon et al. (2006). The beat and bar annotations were then added by Krebs et al. (2013). The ballroom dataset contains eight different dance styles (Cha cha, Jive, Quick-step, Rumba, Samba, Tango, Viennese Waltz, and (slow) Waltz) and has widely used for several MIR tasks such as genre classification, tempo tracking, beat and downbeat tracking [citeXX](#).

It consists of 697 30 seconds-long audio excerpts (sampled at 11.025 kHz) and has tempo and dance style annotations. The dataset contains two different meters (3/4 and 4/4) and all pieces have constant meter. The tempo restrictions given the dance style label from <http://www.ballroomdancers.com/Dances/> were used to annotate the beats and downbeats at the correct metrical level.

The ballroom dataset is used as a dataset to present several evaluations of the algorithms and approaches presented in thesis - to compare performance with the state of the art, and to test if the proposed approaches scale and extend to different music genres and

cultures. **Also mention other rhythm datasets: Geoffroy Peeters dataset, Gainsworth dataset, SMC dataset**

A summary of the chapter to be written here.



---

# Meter inference and tracking

...the first beat (sam) is highly significant structurally, as it frequently marks the coming together of the rhythmic streams of soloist and accompanist, and the resolution point for rhythmic tension.

---

Clayton (2000, p. 81)

Meter analysis of audio music recordings is an important MIR task. It provides useful musically relevant metadata not only for enriched listening, but also for pre-processing of music for several higher level tasks such as section segmentation, structural analysis, and defining rhythm similarity measures.

To recapitulate, meter analysis aims to time-align a piece of audio music recording with several defined metrical levels such as tatum, tactus, measure (bar). In addition, it also tags the recording with additional meter and rhythm related metadata such as time signature, median tempo and salient rhythms in the recording. Within the context of Indian music, meter analysis aims to time-align and tag a music recording with **tāla** related events and metadata.

This chapter aims to address some of these important tasks related to meter analysis within the context of Indian art music, presenting several approaches and a comprehensive evaluation of those

approaches. The main aims of the chapter are:

1. To address meter analysis tasks for the music cultures under study - Carnatic and Hindustani music. The tasks of meter inference, meter tracking, and informed meter tracking are addressed in detail - formulation of these tasks, and propose several approaches to address the tasks.
2. To present a detailed description of the state of the art and the proposed Bayesian models and inference schemes for meter analysis.
3. To present an evaluation of the state of the art meter tracking approaches based on Bayesian models and explore extensions to those approaches, for the rhythm annotated datasets of Carnatic and Hindustani music. A comprehensive performance analysis is presented for these approaches, identifying their strengths and limitations in the tasks under study.

## 5.1 The meter analysis tasks

We describe the meter analysis tasks addressed in this dissertation, from the least informed to the most informed. This order of tasks also emphasizes different practical scenarios for such tasks, and hence the results can indicate the type of task and the additional information to be provided to achieve the level of performance required for an application. We will also describe how the set of tools and approaches described in the chapter can be adapted and used in each of these tasks, making the task of meter analysis flexible to the audio data and the related additional metadata that we can obtain. We will also describe how the set of tools and approaches described in the chapter can be adapted and used in each of these tasks, making the task of meter analysis flexible to the audio data and the related additional metadata that we can obtain.

### Meter inference

Given an audio music recording, meter inference aims to estimate the rhythm class (or meter type), possibly time-varying tempo, beats

and downbeats. In the context of Carnatic music, the task of meter inference aims to recognize the *tāla*, and estimate the time varying tempo ( $\tau_o$  or  $\tau_b$ ), the beat locations, and the *sama* (downbeat) locations. Since some of the beats correspond to the *aṅga* boundaries, with the *sama* and numbered beat locations (beat number in the cycle), the *aṅga* (section) boundaries can be indirectly inferred, e.g. the beats 1 (*sama*), 5, 7 mark the start of the three sections of the *ādi tāla*. Similarly, for Hindustani music, meter inference task aims to recognize the *tāl*, and estimate the time varying tempo ( $\tau_o$ ), the *mātrā* and the *sam* locations. With the numbered *mātrā* and *sam* locations, the *vibhāg* boundaries can be indirectly inferred, e.g. the *mātrās* 1 (*sam*), 3, 6, 8 indicate the start of the four sections in *jhaptāl*. For Carnatic music, in addition to the beats, we can also estimate the sub-division *akṣaras*, which can be grouped into beats.

Without any prior information on metrical structure, meter inference is a difficult task owing to the large range of tempi and different *tālas*. The problem is further made harder due to several *tāla* having similar structure. In Carnatic music, it is quite often possible that the same composition is performed in two different *tālas*, which further can lead to confusion (**provide an example here**). From a practical application point of view, most of commercially released music in both Carnatic and Hindustani music has the name of the *tāla* as a part of the editorial metadata, and hence *tāla* recognition is a redundant task. Even within a live concert, the musician announces the *tāla* of the piece, or shows it with hand gestures in Carnatic music. Meter inference is used a baseline task to understand the complexity of uninformed meter analysis.

### Meter tracking

Given that the *tāla* of an audio music piece is often available as editorial metadata, the most relevant meter analysis task for Indian art music is meter tracking. Given an audio music recording and the rhythm class (or meter type) of the music piece, meter tracking aims to estimate the time varying tempo, the beat and the downbeat locations. In the context of Carnatic music, meter tracking aims to track the time varying tempo, beats and the *sama* from an audio music recording, given the *tāla*. For Hindustani music, given the *tāl*, the task aims to track the time varying  $\tau_o$ , the *mātrā* and *sam*. The

section boundaries of the *tāla* can be indirectly inferred as explained earlier. Assuming that the *tāla*, and hence the metrical structure is known in advance is a fair and practical assumption to make, and we explore if providing this information helps to track the metrical structure better. Meter tracking is the main problem and most comprehensively addressed task in this thesis. We explore different approaches and evaluate them on the rhythm annotated datasets for Carnatic and Hindustani music. The proposed extensions and enhancements are also evaluated for the task of meter tracking.

### Informed meter tracking

Informed meter tracking is a sub-task of meter tracking in which some additional information apart from the meter type is provided along with the audio recording. The additional information could be in the form of a tempo range, a few instances of beats and downbeats annotated, or even partially tracked metrical cycles. These additional metadata could come from manual annotation or as an output of other automatic algorithms, e.g. the median tempo of a piece can be obtained from a standalone tempo estimation algorithm, or some melodic analysis algorithms might output (with a high probability) some beats/downbeats as a byproduct. Even from a practical standpoint, it is useful to explore informed meter tracking. While it is prohibitively resource intensive to manually annotate all the beats and downbeats of a large music collection, it might be possible to seed the meter tracking algorithms with the first few beats and downbeats, which could improve meter tracking performance. For a musician or even an expert listener, it would be very easy to tap some instances of the beat and *sama*, which could then be used automatically track meter, which is a useful application. We aim to explore these questions, to see whether providing additional information can improve meter tracking performance. In specific, we explore two variants of informed meter tracking:

1. Tempo-informed meter tracking in which the median tempo of the piece is provided as an additional input to the meter tracking algorithm. Providing the median tempo intends to help reduce tempo octave errors, tracking the metrical cycles at the correct metrical level instead of tracking half and double cycles. The

median tempo can be obtained through simpler state of the art tempo estimation algorithms outlined in Section 2.3.3 (one such algorithm for Carnatic music is also described later in the chapter in Section 5.2.1). Since the tempo of a piece can vary over time, a narrow range of tempo in the piece can also be provided in addition or in lieu of the median tempo.

2. Tempo-sama-informed meter tracking in which the median tempo and the first downbeat location in the excerpt are provided as additional inputs to the meter tracking algorithm. The practical scenario for such a case is a semi-supervised meter tracking system, where a human listener can tap along to one or some of the downbeats of the piece and an automatic meter tracker can track the rest of the piece. In this thesis, we only explore the use of first downbeat of the piece in informed meter tracking.

There are other meter analysis tasks that have been addressed in MIR, such as beat tracking, and downbeat tracking from the set of known beats. The task of beat tracking as defined in the state of art is ill defined in Indian art music, due to possibly non-isochronous pulsation. We can adapt the task and track a uniform pulsation as the beat. However, since the tasks of meter inference and meter tracking aim to track all the relevant events of the metrical cycle, the task of beat tracking is subsumed in those tasks. We do not address specifically the task of beat tracking in Indian music directly, but as a sub-task of the meter tracking/inference tasks. Estimating the downbeats and the start of measure from a set of beats, as done by Davies and Plumbley (2006); Hockman et al. (2012) is also handled as a sub-task within the joint estimation of tempo, beats and the downbeats.

We now describe the approaches to these tasks, starting with some preliminary approaches followed by Bayesian models. With Bayesian models, several different extensions are proposed over the state of the art models.

## 5.2 Preliminary experiments

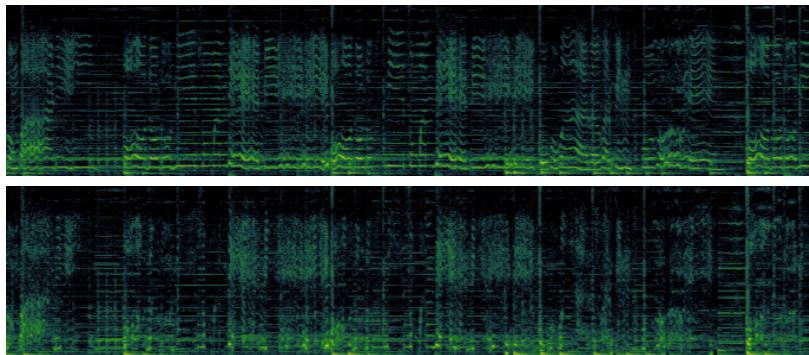
The preliminary experiments around the task of meter analysis are exploratory experiments with existing features, rhythm descriptors,

methods and algorithms to gain insights into the problem and test their relevance and utility in these tasks. The aim of including them in thesis is to gain useful insights and understand the limitations of those algorithms in meter analysis tasks in Indian art music. Only a selection of them are described here, primarily for Carnatic music, as a base for improved Bayesian models for meter analysis. We proposed a novel meter tracking algorithm in Carnatic music (Srinivasamurthy & Serra, 2014) using pre-existing tools and rhythm descriptors, which is described in detail. The features and tools are explained as a part of the proposed meter tracking algorithm, emphasizing on their utility.

### 5.2.1 Meter tracking using dynamic programming

The primary philosophy of meter tracking is to incorporate specific knowledge of the rhythmic structures we aim to estimate, which is also used in this approach. However, it aims to estimate the components of meter separately using a descriptor for each music concept. Using Carnatic music as an illustration, the algorithm estimates the *akṣara* period  $\tau_o$ , the *akṣara* pulse locations, and the *sama*. For estimating these components, a set of rhythm descriptors is first computed from the audio that indicates the possible candidates for each musical concept. The periodicity and the relationships between these structures are then utilized to estimate the components. This framework can be generalized to estimating other rhythmic structures by suitably modifying the audio descriptor for the specific music culture and the rhythmic structure under consideration.

The algorithm for Carnatic music is explained in detail in this section. A hypothesis is that the *akṣara* pulses can be estimated from the onsets of mridangam, and hence a percussion onset based rhythm descriptor (Bello et al., 2005) is useful for tracking the *akṣara* pulses. Tempogram, a mid-level tempo representation for music signals was proposed by (Grosche & Müller, 2011b), is used to track the time-varying *akṣara* period. A novelty function is computed using a self similarity matrix constructed using frame level onset and timbral features. These are then used to estimate possible *akṣara* and *sama* candidates, followed by a candidate selection

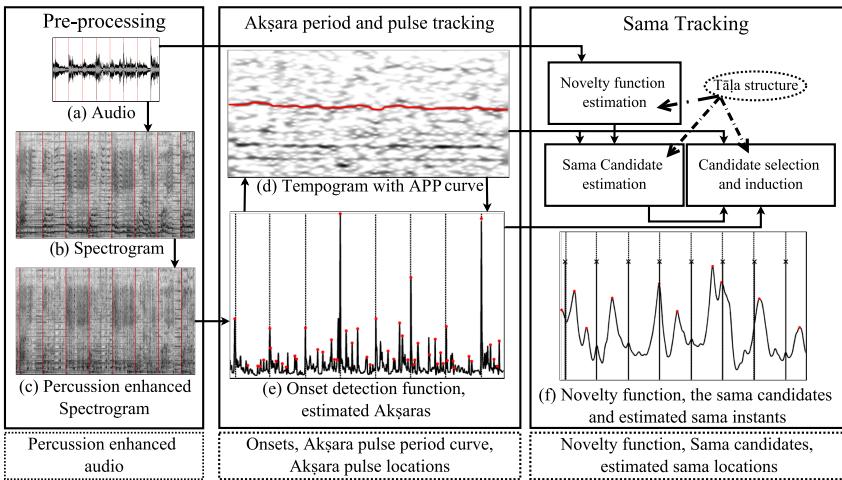


**Figure 5.1:** An illustration of percussion enhancement on a short audio excerpt of Carnatic music. The figure shows the spectrogram of the audio excerpt, before percussion enhancement (top panel) and after percussion enhancement by suppressing the lead melody (bottom panel). The lead melody is suppressed, while the tambūra (drone) is still present.

based on periodicity constraints, which leads to the final estimates. The features and the approach are explained further in detail.

### Pre-processing: Percussion enhancement

The *akṣara* pulse most often coincides with the onsets of mridangam strokes. To enhance the mridangam onsets, percussion enhancement is performed on the downmixed mono audio signal  $f[n]$ , as it has been shown to improve beat tracking performance in pieces with predominant vocals by J. Zapata and Gómez (2013). The predominant melody is estimated using the algorithm proposed by Salamon and Gómez (2012) using which the harmonic component of the signal is extracted using a sinusoidal+residual model proposed by (Serra, 1997). The percussion enhanced signal  $f_p[n]$ , with the harmonic component suppressed, is used for further processing (Figure 5.2(c)). An illustration of percussion enhancement for a short audio excerpt of Carnatic music is shown in Figure 5.1.



**Figure 5.2:** Block Diagram of the algorithm showing the signal flow and representative illustrations of different stages of the algorithm. The important outputs at each stage are also shown at the bottom. In each panel, the vertical lines that run through the panel indicate the **sama** ground truth instants. The estimated **sama/akṣara** candidates are shown with red dots and the estimated **sama** are shown with  $\times$ .

### Akṣara period and pulse tracking

The spectrogram of  $f_p[n]$  is used to compute two frame level spectral-flux based onset detection functions (Bello et al., 2005) computed every 11.6 ms. The first function ( $d_f[k]$ ) uses the whole frequency range of the spectrogram and the other function computes the spectral flux only in the range of 0-120 Hz ( $d_l[k]$ ) and captures the low frequency onsets of the left (bass)-side of the mridangam.

The function  $d_f[k]$  is used to compute a Fourier-based Tempogram proposed by (Grosche & Müller, 2011b), computed every 0.25 second using a 8 second long window (Figure 5.2(d)). If the time indexes at which the tempogram is computed is denoted with  $k$ , ( $1 \leq k \leq K$ ), the most predominant  $\tau_o$  curve can be tracked by estimating the best path  $\Gamma = \{\gamma_k : k = 1, 2, \dots, K\}$  through the tempogram matrix  $\mathbf{G}$  that provides a balance between tempogram amplitude at time index  $k$ ,  $\mathbf{G}_{\gamma_k, k}$ , and the local continuity of  $\tau_o$ . An objective function, that is an extended version of the one used by

Wu et al. (2011), is defined as shown in Eq. 5.1.

$$J_1(\Gamma, \theta_1, \theta_2) = \sum_{k=1}^K \mathbf{G}_{\gamma_k, k} - \sum_{k=1}^{K-1} \left( \theta_1 |\gamma_k - \gamma_{k+1}| + \theta_2 \mathfrak{O}\left(\frac{\gamma_k}{\gamma_{k+1}}\right) \right) \quad (5.1)$$

The function  $\mathfrak{O}(\gamma_k / \gamma_{k+1})$  is an extra penalty term to penalize tempo doubling and halving between adjacent frames, and the weights  $\theta_1 (=0.01)$  and  $\theta_2 (=10^6)$  provide different weights to the three terms. Based on observations from the CMR<sub>f</sub> dataset, the search for the best path through the tempogram is restricted between the range of 120 to 600 APM (akşaras per minute). The above objective function is solved using a dynamic programming (DP) based approach to obtain a  $\tau_o$  curve. Assuming the longest tracked  $\tau_o$  curve to be at the correct metrical level, any possible tempo doubling/halving errors that are present are corrected to obtain the final curve  $\Gamma^*$  (Figure 5.2(d),  $\Gamma^*$  is shown as a thick red line). Using the  $\tau_o$  and the tāla information, we can obtain the time varying  $\tau_s$  curve for the piece by multiplying the  $\tau_o$  by the number of akşaras in a cycle of the tāla. A further example of a tempogram and the estimated time varying  $\tau_o$  curve for a piece of Carnatic music<sup>1</sup> from CMR<sub>f</sub> dataset is shown in Figure 5.3. The figure shows the variations in tempo through a Carnatic music piece.

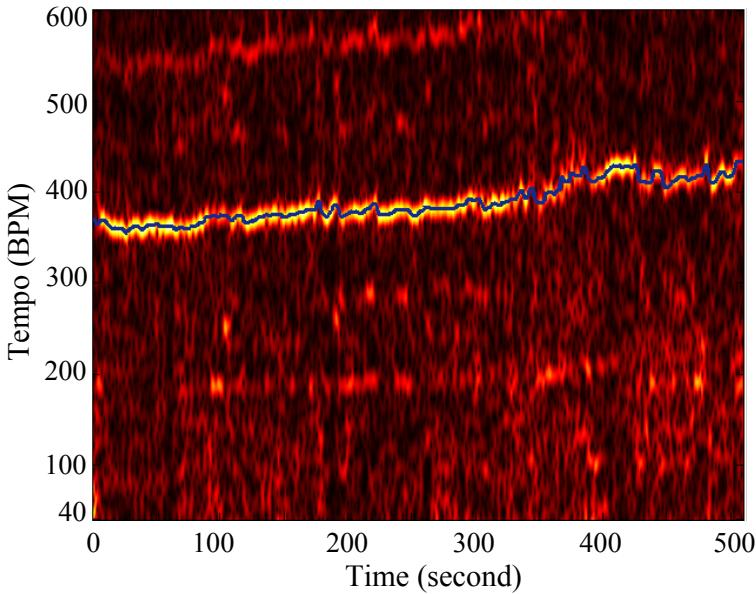
The akşara pulse locations predominantly lie on strong Mridangam onsets. The akşara pulse candidates are estimated as the peaks of the function  $d_f[k]$ . Using these  $\kappa$  candidate peaks  $\{o_i\}$ ,  $i = 1, 2, 3, \dots, \kappa$ , with locations  $t_i$  and peak amplitude  $\xi_i$ , a cost function is setup as shown in Eq. 5.2 to select the best candidates that provide a balance between the amplitude of these candidates and a periodicity provided by the estimated akşara period. The best set of candidates  $\{o_i^*\} \subset \{o_i\}$  are estimated using a DP approach (Figure 5.2(e)).

$$J_2(\{o_i\}, \delta) = \sum_{i \in \{1, 2, \dots, \kappa\}} (\xi_i + \delta \Upsilon(t_i, t_{i+1}, \Gamma)) \quad (5.2)$$

The function  $\Upsilon(t_i, t_{i+1}, \Gamma)$  is a function that returns a exponentially decaying weight based on the time difference between  $t_i$  and  $t_{i+1}$

---

<sup>1</sup>Kamalamba, a kṛti in ānandabhairavi rāga and miśra chāpu tāla, from the album Madrasil Margazhi 2005 by Aruna Sairam: <http://musicbrainz.org/recording/3baa722d-480e-4ae7-8559-a88dce41e1d4>



**Figure 5.3:** Estimated time varying tempo curve (shown in blue) plotted on top of a tempogram, for a Carnatic music piece (see footnote). In the piece, apart the local tempo variations, we can see that the tempo increases with time. The tempogram shows high values in tempo octave related bands, with the highest value (in yellow) at the estimated  $\tau_o$ .

in relation to the local *aksara* period,  $\gamma_{t_k}$ . The parameter  $\delta (=3)$  provides a tradeoff between the two terms.

### Sama Tracking

The use of Mel-Frequency cepstral co-efficients (MFCC) as features for timbral characteristics is explored. As a detection function for *sama* ( $d_s[k]$ ), a novelty function is computed through the diagonal processing of a self similarity matrix (Foote, 2000) constructed using frame level z-score MFCC features from audio (using audio processing library *Essentia* (Bogdanov et al., 2013a)) as shown in Figure 5.2(f). Based on the  $\bar{\tau}_s$  shown in Table 4.6, a checkerboard kernel with size of 7, 3, 4, and 3 seconds is used for the *tālas ādi*, *rūpaka*, *miśra chāpu* and *khanḍa chāpu* respectively so that the novelty function is computed over about an *āvartana*.

The peaks of the novelty function  $d_s[k]$  indicate a significant

change of timbre at that time. Starting with the premise that timbral change is an important indicator of **sama** location, the peaks of the novelty function are used to estimate **sama** candidates. Two methods are explored to estimate the candidates. In Method-A, to uniformly choose **sama** candidates throughout a piece, the piece is cut into segments of length 120, 40, 40 and 30 seconds for **ādi**, **rūpaka**, **miśra chāpu** and **khaṇḍa chāpu** respectively ( $\sim 10$  **āvartanas**), and the top five most prominent peaks in each segment of the piece are estimated as **sama** candidates ( $\{s_i^A\}$ ).

Another approach, Method-B, is also proposed for candidate estimation that enforces a periodicity constraint while estimating **sama** candidates. Starting from the peaks of  $d_s[k]$  and estimated  $\tau_s$  curve, for a specific peak, the **tāla** cycle is induced starting from it. The number of other peaks that would support such an induced **tāla** is assigned as the weight of the specific peak. The peaks are then rank ordered using this weight and the top ten ranked peaks are chosen as the **sama** candidates ( $\{s_i^B\}$ ). In addition, two random baseline methods RB-1 and RB-2 are created to compare the performance. In RB-1, a randomly chosen constant  $\tau_s$  between 1-8 seconds is used, and a random starting time between 0-2 seconds to induce periodic **samas**. In RB-2, the estimated  $\tau_s$  is used with 10 randomly chosen **akṣara** locations from  $\{o_i\}$  as **sama** candidates. RB-1 neither uses the  $\tau_s$ , nor the candidate estimation using  $d_s[k]$ , while RB-2 uses the estimated  $\tau_s$  but not the candidate estimation using  $d_s[k]$ .

Starting with the **sama** candidates obtained either from Method-A or Method-B, for each candidate, the **tāla** cycles are induced based on local  $\tau_s$  period obtained from the  $\tau_s$  curve. For each seed, the next and previous three estimated cycle periods are searched for onset peaks in  $d_f[k]$  that support a **sama**. If a supporting onset is found, it is marked as a **sama** and the algorithm proceeds further with the new estimated onset as the new anchor. The induction is stopped from a candidate when it does not lead to such a supporting onset. Hence for each candidate, an estimated **sama** sequence is obtained. Since all candidates are not necessarily **sama** locations, though the estimated  $\tau_s$  is right, the sequences can have different offsets.

The final step of the algorithm is to shift, align and merge these sequences obtained from each candidate. Starting with the longest

Measure	CML	AML
$\bar{\tau}_o$ estimation	81.2	98.9
$\tau_o$ tracking	80.4	96.3

**Table 5.1:** Accuracy (%) of *akṣara* period tracking on the  $\text{CMR}_f$  dataset. The values are measured using a 5% tolerance, at both correct metrical level (CML) and allowed metrical levels (AML).

sama sequence that has been estimated, other sequences are merged into this based on maximum correlation between the sequences. The merging of these sequences often leads to many sama estimates concentrated around the true location of sama due to small offsets. Since the left bass onsets on the Mridangam are often strong at the samas, all groups of sama estimates that are closer than 1/3rd of  $\tau_s$  are merged into a single sama estimate aligned with the closest left stroke onset obtained from  $d_l[k]$ . This forms the final set of sama locations  $\{s_{t_i}\}$  estimated from the candidates and the onset detection function, as shown in Figure 5.2(f) with  $\times$ .

## Results

The annotated  $\text{CMR}_f$  dataset has annotations only for beats and samas of the piece. From the sama locations, we can obtain the ground truth for  $\tau_s$  curve, and hence the ground truth for  $\tau_o$  curve. Since we do not have the ground truth for akṣara locations, we present the results only for  $\tau_o$  and sama tracking.

The performance of akṣara period tracking is measured by comparing the ground truth akṣara period curve with the estimated curve with an error tolerance of 5%. The results of median akṣara period estimation computed from the whole akṣara period curve is also reported. Further, since there can be tempo doubling and halving errors, the accuracies are reported at the annotated correct metrical level (CML) and then using a weaker AML measure that allows tempo halving and doubling (AML - allowed metrical levels). The results are presented in Table 5.1. We see that an acceptable level accuracy is achieved at CML for both median akṣara period estimation and akṣara period tracking and further, there is not a significant difference between their performances, indicating that the

Variant	$\text{p}$	$\text{r}$	$\text{f}$	$\mathfrak{I}(\text{bits})$	Cand.	Accu. (%)
Method-A	0.290	0.190	0.216	1.17		20.46
Method-B	0.246	0.202	0.215	1.25		27.85
RB-1	0.155	0.175	0.137	0.40		-
RB-2	0.228	0.200	0.206	1.11		15.3

**Table 5.2:** Accuracy of `sama` tracking. The measures  $\text{p}$ : Precision,  $\text{r}$ : Recall,  $\text{f}$ : f-measure,  $\mathfrak{I}$ : Information Gain, are shown. The values are mean performance over the whole  $\text{CMR}_f$  dataset. The last column shows the fraction (as a percentage) of the estimated sama candidates that are true samas.

algorithm can track changes in tempo effectively. Even when the `akṣara` period tracking fails at CML, the algorithm tracks a metrically related `akṣara` period, as indicated by a high AML accuracy.

For sama tracking, the accuracy of estimation is reported with a margin of 7% the annotated  $\tau_s$  of the piece. Given the ground truth and the estimated sama time sequence, we use the common evaluation measures used in beat tracking - precision, recall, f-measure and Information Gain (McKinney et al., 2007) to measure the performance. The results are shown in Table 5.2, which also shows the accuracy of sama candidate estimation. The results for RB-1 and RB-2 show mean performance over 100 and 10 experiments for each piece, respectively.

We see that the performance of sama candidate estimation and `sama` tracking is poor in general, with `samas` correctly tracked only in about a fifth of cases. The precision is higher than recall in all cases, and Information Gain is lower than a perceptually acceptable threshold (J. R. Zapata et al., 2012). Both methods perform better than RB-1, but have comparable results with RB-2, with a slightly better f-measure performance (statistically significant in a Mann–Whitney U test at  $p < 0.05$ ). This shows that the estimated inter-sama interval ( $\tau_s$ ) is useful for sama estimation, whereas candidate estimation using novelty function is only marginally useful. The poor performance can be mainly attributed to poor `sama` candidate estimation with either of Method-A or Method-B. This is further substantiated by the fact that Method-B achieves an F-measure

of 0.436 and an information gain of 1.70 bits when at least half the estimated candidates are true *samas*. This clearly shows that the performance of *sama* tracking crucially depends on *sama* candidate estimation. There are only four pieces (among all pieces with accurate  $\tau_s$  estimation) in which all the estimated candidates are true *samas*, for which an F-measure of 0.894 and a information gain of 3.51 bits is achieved. This clearly indicates that the novelty function from which the *sama* candidates were estimated is not a very good indicator of *sama*, and better descriptors need to be explored.

## Conclusions

The presented approach to meter tracking with relevant rhythm descriptors for tempo, *akṣara*, and *sama* and a hierarchical framework is promising, but has several limitations. The onset detection functions have information about surface rhythms and hence can be utilized for tempo tracking and *akṣara* pulse tracking, but the novelty function used presently is not a good indicator for *sama*. Further, it is observed that *akṣara* pulse period tracking performs to an acceptable accuracy for practical applications, while *sama* tracking is challenging and performs poorly primarily due to poor *sama* candidate estimation. Though the tempo, *akṣara* and *sama* are related, they were tracked separately. Even though information from tempo estimation was used in estimating the *sama*, a joint estimation of the meter components is desired, since it can tightly couple these related components together.

The approach uses the musical characteristics in isolation, without considering the interdependence between them. Further, many heuristic measures are used to track the components of the *tāla*. The learning from such heuristic approaches can be used to build a model that can more effectively model the underlying metrical structure, one that would consider the problem of meter inference and tracking more holistically. Such a model would also be adaptable to different metrical structures and handle variations in real world scenarios. The tracking algorithm based on dynamic programming is also ad hoc and loosely uses the tightly coupled information between the tempo, *akṣaras* and the *sama*.

Considering these insights and limitations, we explore Bayesian models for meter inference, which provide an effective probabilis-

tic framework for the task, with several useful inference algorithms and well studied formulations that can be utilized to our benefit. The framework learns from training examples and hence the large number of heuristics used in these initial experiments become unnecessary.

### 5.3 Bayesian models for meter analysis

Recently, Bayesian models have been applied successfully to meter analysis tasks(Krebs et al., 2013; Böck et al., 2014; Krebs, Holzapfel, et al., 2015). The effectiveness of such models stem from their ability to accurately model metrical structures and their adaptability to different metrical structures, music styles and variations. These advantages are supplemented by the huge literature on Bayesian models and efficient exact and approximate inference algorithms. Since metrical structures are mostly mental constructs, the use of such generative graphical probabilistic models can even perhaps be hypothesized that they closely (better than other approaches to meter analysis) emulate the mechanisms of progression through metrical cycles.

With a fundamental dependence on time, any model that aims to accurately represent metrical structures should work on sequential data from audio features, and must be able to incorporate several different variables within one probabilistic framework. A **DBN** (Murphy, 2002) is well suited in such cases, since it relates variables over time through conditional (in)dependence relations. A **DBN** is a generalization of the traditional linear state-space models such as Kalman filters and stochastic models such as the **HMM** and provide a general probabilistic representation and inference schemes for arbitrary non-linear and non-gaussian time-dependent processes.

The bar pointer model is one such **DBN** model that has been successfully applied to meter analysis. Proposed by Whiteley, Cemgil, and Godsill (2006), it has been improved since then and applied to various meter analysis tasks over different music styles (Whiteley, Cemgil, & Godsill, 2007; Krebs et al., 2013; Krebs, Holzapfel, et al., 2015; Böck et al., 2014; Holzapfel et al., 2014; Krebs, Böck, & Widmer, 2015; Srinivasamurthy et al., 2015, 2016). In the thesis,

we start with the bar pointer model and present several extensions and explore different inference schemes for those extensions, all in the context of Indian art music. The performance of such models and inference schemes are evaluated on the Carnatic and Hindustani music test datasets presented in Chapter 4, with additional evaluations to test for generalization and to baseline performance on the Ballroom dataset. The primary focus of the thesis is on the most relevant task of meter tracking, while meter inference, informed meter tracking tasks being addressed to a limited extent.

The remainder of the chapter is organized as follows. The bar pointer model is first described, explaining its model structure and inference schemes (Section 5.3.1). Extensions and enhancements to the model structure are then proposed and described in Section 5.3.2:

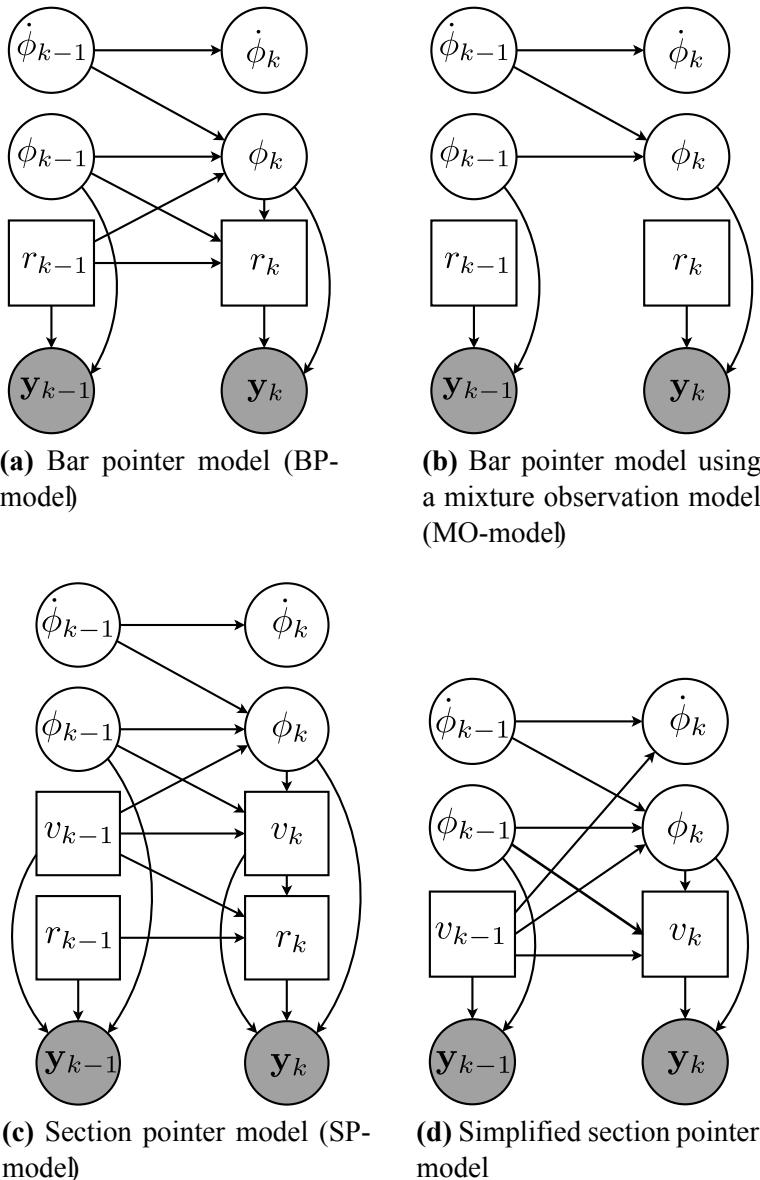
1. A simplified bar pointer model with a mixture observation model, that aims to complement observation likelihood from many rhythmic patterns.
2. The section pointer model that aims to use patterns that are shorter than bar for meter tracking, and hence might be useful to track long metrical structures.

Extensions and enhancements to inference schemes on the bar pointer model extensions are then proposed and described in Section 5.3.3:

1. End of bar rhythm pattern sampling, which proposes to defer pattern sampling to the end of the bar.
2. Hop inference for fast meter tracking, which aims to do faster inference by performing inference only when there is a significant rhythmic event in audio (such as an onset).

### 5.3.1 The bar pointer model

The bar pointer model (referred to in short as BP-model) is a generative model that has been successfully applied for meter analysis tasks. The model assumes a hypothetical time pointer within a bar, progressing at the speed of the tempo traversing through the bar and reinitializing at the end of the bar to track the next bar. The model also assumes that specific bar length rhythm patterns are played in a bar depending on the rhythmic style, and uses these patterns to



**Figure 5.4:** The meter analysis models used in the dissertation. In each of these DBNs, circles and squares denote continuous and discrete variables, respectively. Grey nodes and white nodes represent observed and latent variables, respectively.

track the progression through the bar. These rhythmic patterns can

be fixed *a priori* or learned from data to build an observation model for each position in the bar. When learned from data, the rhythmic patterns are built using a signal representation derived from audio, most often from frame level audio features to preserve the temporal information in features. Progressing through the bar, the model can hence be used to sample the observation model and generate a rhythmic pattern that is possible with the rhythm style. It allows for different metrical structures, tempi ranges, and rhythm styles, providing a flexible framework for meter analysis. Though applied only for meter analysis from audio recordings in this dissertation, the BP-model can be applied even to symbolic music. BP-model can be represented as a DBN with specific conditional dependence relations between the variables that lead to several variants and extensions of the model. The structure of the BP-model is shown in Figure 5.4a.

In a DBN, an observed sequence of features derived from an audio signal  $\mathbf{y}_{1:K} = \{\mathbf{y}_1, \dots, \mathbf{y}_K\}$  is generated by a sequence of hidden (latent) variables  $\mathbf{x}_{1:K} = \{\mathbf{x}_1, \dots, \mathbf{x}_K\}$ , where  $K$  is the length of the feature sequence (number of audio frames in an audio excerpt). The joint probability distribution of hidden and observed variables factorizes as,

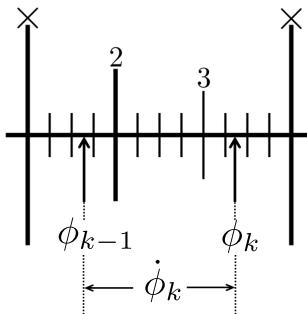
$$P(\mathbf{y}_{1:K}, \mathbf{x}_{0:K}) = P(\mathbf{x}_0) \cdot \prod_{k=1}^K P(\mathbf{x}_k | \mathbf{x}_{k-1}) P(\mathbf{y}_k | \mathbf{x}_k) \quad (5.3)$$

where,  $P(\mathbf{x}_0)$  is the initial state distribution,  $P(\mathbf{x}_k | \mathbf{x}_{k-1})$  is the transition model, and  $P(\mathbf{y}_k | \mathbf{x}_k)$  is the observation model.

## Hidden variables

In the bar pointer model, at each audio frame  $k$ , the hidden variable vector  $\mathbf{x}_k$  describes the state of a hypothetical bar pointer  $\mathbf{x}_k = [\phi_k \dot{\phi}_k r_k]$ , representing the bar position, instantaneous tempo and a rhythmic pattern indicator, respectively (see Figure 5.5 for an illustration).

- *Rhythmic pattern indicator:* The rhythmic pattern variable  $r \in \{1, \dots, R\}$  is an indicator variable to select one of the  $R$  observation models corresponding to each bar (cycle) length rhythmic pattern of a rhythm class. Each pattern  $r$  has an associated length



**Figure 5.5:** An illustration of the progression of bar position and instantaneous tempo variables over two consecutive audio frames in a cycle of *rūpaka tāla*. The effect of instantaneous tempo is greatly exaggerated for clarity in the illustration.

of cycle  $M_r$  and number of beat (or *mātrā*) pulses  $B_r$ . In the scope of this dissertation, all rhythmic patterns are learned from training data and not fixed a priori. We can infer the rhythm class or meter type (*tāla*) by allowing rhythmic patterns of different lengths from different rhythm classes to be present in the model, as used by Krebs, Holzapfel, et al. (2015). However, it is to be noted that for the problem of meter tracking, we assume that the cycle length is known and that all the  $R$  rhythmic patterns belong to the same rhythm class (*tāla*),  $M_r = M$  and  $B_r = B \forall r$ .

- *Bar position*: The bar position  $\phi \in [0, M_r]$  variable indicates a position in the bar at any audio frame and tracks the progression through the bar. Here,  $M_r$  is the length of the bar (cycle), which is also the length of the bar length rhythmic pattern being tracked. The bar position variable traverses the whole bar and wraps around to zero at the end of the bar to track the next bar. The maximum value of bar (cycle) length,  $M$ , depends on the longest bar (cycle) that is tracked. We set the length of the longest bar being tracked to a fixed value, and scale other bar (cycle) lengths accordingly.
- *Instantaneous tempo*: Instantaneous tempo  $\dot{\phi}$  (measured in positions per time frame) is the rate at which the bar position variable progresses through the cycle at each time frame, measured in bar positions per time frame. The range of the variable  $\dot{\phi}_k \in [\dot{\phi}_{\min}, \dot{\phi}_{\max}]$  depends on the length of the cycle  $M$  and the hop

size ( $h = 0.02$  second used in this thesis), and can be preset or learned from data. A tempo value of  $\dot{\phi}_k$  corresponds to a bar (cycle) length of  $(h \cdot M_r / \dot{\phi}_k)$  seconds and  $(60 \cdot B \cdot \dot{\phi}_k / (M \cdot h))$  beats (mātrās) per minute. The range of the variable can be used to restrict the range of tempi that is allowed within each rhythm class.

### Initial state distribution

The initial state distribution  $P(\mathbf{x}_0)$  can be used to incorporate prior information about the metrical structure of the music into the model. Different initializations are explored depending on the meter analysis task under consideration.

### Transition model

Given the the conditional dependence relations between the variables of the BP-model in Figure 5.4a, the transition model factorizes as,

$$P(\mathbf{x}_k | \mathbf{x}_{k-1}) = P(\phi_k | \dot{\phi}_{k-1}, r_{k-1}) P(r_k | \dot{\phi}_{k-1}) \\ P(r_k | r_{k-1}, \phi_k, \dot{\phi}_{k-1}) \quad (5.4)$$

The individual terms of the equation can be expanded as,

$$P(\phi_k | \dot{\phi}_{k-1}, r_{k-1}) = \mathbb{1}_\phi \quad (5.5)$$

where  $\mathbb{1}_\phi$  is an indicator function that takes a value of one if  $\phi_k = (\phi_{k-1} + \dot{\phi}_{k-1}) \bmod(M_{r_k})$  and zero otherwise. The tempo transition is given by,

$$P(\dot{\phi}_k | \dot{\phi}_{k-1}) \propto \mathcal{N}(\dot{\phi}_{k-1}, \sigma_{\dot{\phi}}^2) \times \mathbb{1}_{\dot{\phi}} \quad (5.6)$$

where  $\mathbb{1}_{\dot{\phi}}$  is an indicator function that equals one if  $\dot{\phi}_k \in [\dot{\phi}_{\min}, \dot{\phi}_{\max}]$  and zero otherwise, restricting the tempo to be between a predefined range.  $\mathcal{N}(\mu, \sigma^2)$  denotes a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . The value of  $\sigma_{\dot{\phi}}$  depends on the value of tempo and the length of the pattern. We set  $\sigma_{\dot{\phi}} = \sigma_n \cdot \dot{\phi}_{k-1} \cdot (M_{r_{k-1}}/M)$ , where  $\sigma_n$  is a user parameter that controls the amount of local tempo variations we allow in the music piece.

$$P(r_k | r_{k-1}, \phi_k, \dot{\phi}_{k-1}) = \begin{cases} \mathbb{A}(r_{k-1}, r_k) & \text{if } \phi_k < \phi_{k-1} \\ \mathbb{1}_r & \text{else} \end{cases} \quad (5.7)$$

where,  $\mathbb{A}$  is the  $R \times R$  time-homogeneous transition matrix with  $\mathbb{A}(i, j)$  being the transition probability from  $r_i$  to  $r_j$ , and  $\mathbb{1}_r$  is an indicator function that equals one when  $r_k = r_{k-1}$  and zero otherwise. Since the rhythmic patterns are one bar (cycle) in length, pattern transitions are allowed only at the end of the bar (cycle). When there are multiple patterns, these transition probabilities indicate the most probable movement through these patterns from bar to bar, as the piece progresses. To reflect the performance practice, the pattern transition probabilities are learned from data.

## Observation Model

The observation model aims to model the underlying rhythmic patterns present in the metrical structure being inferred/tracked, explaining the possible rhythmic events at each position in the bar. Some of the positions in a bar have a higher probability of an onset occurring than other parts (the positions corresponding to downbeats, beats, e.g.). Further, the strength of these onsets also vary depending on accent patterns of a rhythm class (which can be modeled from labeled data). The observation model used in this dissertation aims to address both these aspects (the locations and strengths of the rhythmic events), and closely follows the observation model proposed by Krebs et al. (2013).

The utility of spectral flux based rhythmic audio features was outlined in preliminary experiments Section 5.2. A similar audio derived spectral flux feature is used in this dissertation as well, identical to features used by Krebs et al. (2013), as explained in Section 4.2.1 (see Figure 4.7). Since the bass onsets have significant information about the rhythmic patterns, the features are computed in two frequency bands (Low:  $\leq 250$  Hz, High:  $> 250$  Hz).

It is assumed that the audio features depend only on the bar position and rhythmic pattern variables, without any influence from tempo. While this assumption is not completely true, it simplifies the observation model and helps to train better models with limited training data. Further, it is assumed that the audio features do not vary too much over short changes in position in cycle (e.g. the spectral flux variations within a small fraction of an *akṣara* might be negligible), which additionally helps to tie several positions to

have the same observation probability and helps train models with limited training data.

Using beat and downbeat annotated training data, the audio features are then grouped into bar length patterns. The bar is then discretized into 64<sup>th</sup> note cells (four cells per *akṣara* for Carnatic music, and four cells per *mātrā* for Hindustani music, corresponds to 25 bar positions with  $M = 1600$ ). A k-means clustering algorithm clusters and assigns each bar of the dataset to one of the  $R$  rhythmic patterns. All the features within the cell are then collected for each pattern, and maximum likelihood estimates of the parameters of a two component Gaussian mixture model (GMM) are obtained. The observation probability within a 64<sup>th</sup> note cell is assumed to be constant, and computed as,

$$P(\mathbf{y} \mid \mathbf{x}) = P(\mathbf{y} \mid \phi, r) = \sum_{i=1}^2 \pi_{\phi,r,i} \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{\phi,r,i}, \boldsymbol{\Sigma}_{\phi,r,i}) \quad (5.8)$$

where,  $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  denotes a normal distribution of the two dimensional feature  $\mathbf{y}$ . For the mixture component  $i$ ,  $\pi_{\phi,r,i}$ ,  $\boldsymbol{\mu}_{\phi,r,i}$  and  $\boldsymbol{\Sigma}_{\phi,r,i}$  are the component weight, mean (2-dimensional) and the covariance matrix ( $2 \times 2$ ), respectively.

### Inference in bar pointer model

The goal of inference in meter analysis tasks is to find a hidden variable sequence that maximizes the posterior probability of the hidden states given an observed sequence of features: a maximum *a posteriori* (MAP) sequence  $\mathbf{x}_{1:K}^*$  that maximizes  $P(\mathbf{x}_{1:K}^* \mid \mathbf{y}_{1:K})$ . The inferred hidden variable sequence  $\mathbf{x}_{1:K}^*$  can then be translated into a sequence downbeat (sama) instants ( $\phi_k^* = 0$ ), beat instants ( $\phi_k^* = i \cdot M_r / B_r$ ,  $i = 1, \dots, B_r$ ), the local instantaneous tempo ( $\dot{\phi}_k^*$ ), and the sequence of estimated rhythmic patterns  $r^*$ .

Two different inference schemes are now described, an exact inference using the Viterbi algorithm in a discretized state space, and an approximate inference using particle filters in the continuous space of  $\phi$  and  $\dot{\phi}$ , with the discrete variable  $r$ .

### Viterbi algorithm

The continuous variables of bar position and tempo can be discretized, which transforms the DBN into an HMM over the cartesian product space of the discretized variables. In the HMM, an exact inference can be performed using the Viterbi algorithm to compute the most likely sequence of hidden states given the observed data.

We follow the discretization identical to the method proposed by Krebs, Holzapfel, et al. (2015), by replacing the continuous variables  $\phi$  and  $\dot{\phi}$  by their discretized counterparts  $m$  and  $n$ , respectively, as

$$m \in \{1, 2, \dots, \lceil M_r \rceil\} \quad (5.9)$$

$$n \in \{n_{\min}, n_{\min} + 1, n_{\min} + 2, \dots, N - 1, N\} \quad (5.10)$$

Here,  $n_{\min} = \lfloor \dot{\phi}_{\min} \rfloor$  and  $N = n_{\max} = \lceil \dot{\phi}_{\max} \rceil$  is the discrete minimum and maximum tempo values allowed, where  $\lfloor \cdot \rfloor$  and  $\lceil \cdot \rceil$  denote floor and ceil operations, respectively.

With such a discretization in place, the transition model equations Eq. 5.4, Eq. 5.5 and Eq. 5.7 remain as defined. However, the tempo transition probability is redefined within the allowed tempo range as,

$$P(n_k | n_{k-1}) = \begin{cases} 1 - p_n & \text{if } n_k = n_{k-1} \\ \frac{p_n}{2} & \text{if } n_k = n_{k-1} \pm 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.11)$$

where  $p_n$  is the probability of tempo change. It is to be noted that the discretization of  $\phi$  and  $\dot{\phi}$  need not be done on an integer or on a uniform grid. It is possible that the tempo range can be non-uniformly sampled, as was proposed by Krebs, Böck, and Widmer (2015). In this dissertation, however, only a uniform discretization is explored in the context of the HMM. Viterbi algorithm (Rabiner, 1989) is then used to obtain a MAP sequence of states with the HMM. The HMM based exact inference in bar pointer model as described in the section will be denoted as  $\text{HMM}_0$  in the dissertation.

The drawback of this approach is that the discretization has to be on a very fine grid in order to guarantee good performance,

which leads to a prohibitively large state space and, as a consequence, to a computationally demanding inference. The size of the state space is  $\mathfrak{S} = M \cdot N \cdot R$  and needs an  $\mathfrak{S} \times \mathfrak{S}$  sized transition matrix. As an example, dividing a bar into  $M = 1600$  position states, with  $N = 15$  tempo states and  $R = 4$  patterns, the size of the state space is  $\mathfrak{S} = 96000$  states. The computational complexity of the Viterbi algorithm is  $O(K \cdot |\mathfrak{S}|^2)$ . Even though the state transition matrix is sparse due to lesser number of allowed transitions leading to a complexity of  $O(K \cdot M \cdot R)$ , the inference with HMM can become computationally prohibitive and does not scale well with increasing number of states. This problem can be overcome, for instance, by using approximate inference methods such as particle filters.

## Particle Filter (PF)

Particle filters (or Sequential Monte Carlo methods) are a class of approximate inference algorithms to estimate the posterior density in a state space. They overcome two main problems of the HMM: discretization of the state space and the quadratic scaling up of the size of state space with additional hidden variables. In addition, they can incorporate long term relationships between hidden variables.

In the continuous state space of  $\mathbf{x}_{1:K}$ , the exact computation of the posterior  $P(\mathbf{x}_{1:K} | \mathbf{y}_{1:K})$  is often intractable, but it can be evaluated pointwise. In particle filters, the posterior is approximated using a weighted set of points (known as particles) in the state space as,

$$P(\mathbf{x}_{1:K} | \mathbf{y}_{1:K}) \approx \sum_{i=1}^{N_p} w_K^{(i)} \delta(\mathbf{x}_{1:K} - \mathbf{x}_{1:K}^{(i)}) \quad (5.12)$$

Here,  $\{\mathbf{x}_{1:K}^{(i)}\}$  is a set of points (particles) with associated weights  $\{w_K^{(i)}\}$ ,  $i = 1, \dots, N_p$ , and  $\mathbf{x}_{1:K}$  is the set of all state trajectories until frame  $K$ , while  $\delta(x)$  is the Dirac delta function.  $N_p$  is the number of particles.

Starting with  $P(\mathbf{x}_0)$ , to approximate the posterior pointwise, we need a suitable method to draw samples  $\mathbf{x}_k^{(i)}$  and compute appropriate weights  $w_k^{(i)}$  recursively at each time step. It is further non-

trivial to sample from an arbitrary posterior distribution. A simple approach is Sequential Importance Sampling (SIS) (Doucet & Johansen, 2009), where we sample from a *proposal* distribution  $Q(\mathbf{x}_{1:K} | \mathbf{y}_{1:K})$  that has the same support and is as similar to the true (target) distribution  $P(\mathbf{x}_{1:K} | \mathbf{y}_{1:K})$  as possible. To account for the fact that we sampled from a proposal and not the target, we attach an importance weight  $w_K^{(i)}$  to each particle, computed as,

$$w_K^{(i)} = \frac{P(\mathbf{x}_{1:K} | \mathbf{y}_{1:K})}{Q(\mathbf{x}_{1:K} | \mathbf{y}_{1:K})} \quad (5.13)$$

With a suitable proposal density, these weights can be computed recursively as,

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{P(\mathbf{y}_k | \mathbf{x}_k^{(i)}) P(\mathbf{x}_k^{(i)} | \mathbf{x}_{k-1}^{(i)})}{Q(\mathbf{x}_k^{(i)} | \mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k)} \quad (5.14)$$

Following Krebs, Holzapfel, et al. (2015), we choose to sample from the transition probability  $Q(\mathbf{x}_k^{(i)} | \mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k) = P(\mathbf{x}_k^{(i)} | \mathbf{x}_{k-1}^{(i)})$ , which reduces Eq. 5.14 to

$$w_k^{(i)} \propto w_{k-1}^{(i)} P(\mathbf{y}_k | \mathbf{x}_k^{(i)}) \quad (5.15)$$

The SIS algorithm derives samples by first sampling from proposal, in this case the transition probability and then computes weights according to Eq. 5.15. Once we determine the particle trajectories  $\{\mathbf{x}_{1:K}^{(i)}\}$ , we then select the trajectory  $\mathbf{x}_{1:K}^{(i*)}$  with the highest weight  $w_K^{(i*)}$  as the MAP state sequence.

Many extensions have been proposed to the basic SIS filter (Doucet and Johansen (2009) provide a comprehensive overview) to address several problems with it. Some of the relevant extensions are briefly mentioned, emphasizing their key aspects. A more detailed description of the algorithms has been presented by Krebs, Holzapfel, et al. (2015). The most challenging problem in particle filtering is the degeneracy problem, where within a short time, most of the particles have a weight close to zero, representing unlikely regions of state space. This is contrary to the ideal case when we want the proposal to match well with the target distribution leading to a uniform weight distribution with low variance. To reduce the variance of the particle weights, resampling steps are necessary,

which replace low weight particles with higher weight particles by selecting particles with a probability proportional to their weights. Several resampling methods have been proposed, but we use systematic resampling in this dissertation as recommended by Doucet and Johansen (2009). With resampling as the essential difference, the SIS filter with resampling is called as Sequential Importance Sampling/Resampling (SISR) filter.

In meter analysis problems, due to metrical ambiguities, the posterior distribution  $P(\mathbf{x}_k \mid \mathbf{y}_{1:k})$  is highly multimodal. Resampling tends to lead to a concentration of particles in one mode of the posterior, while the remaining modes are not covered. One way to alleviate this problem is to compress the weights  $\mathbf{w}_k = w_k^{(i)}$ ,  $i = 1, \dots, N_p$  by a monotonically increasing function to increase the weights of particles in low probability regions so that they can survive resampling. After resampling, the weights have to be uncompressed to give a valid probability distribution. This can be formulated as an Auxiliary Particle Filter (APF) (Johansen & Doucet, 2008).

A particle system that is capable of handling metrical ambiguities must maintain the multimodality of posterior distribution and be able to track several hypotheses together, which SISR and APF cannot do explicitly. A system called the Mixture Particle Filter (MPF) was proposed by Vermaak, Doucet, and Pérez (2003) to track multiple hypotheses, and was adapted to meter inference by Krebs, Holzapfel, et al. (2015).

In a MPF, each particle is assigned to a cluster that (ideally) represents a mode of the posterior. During resampling, the particles of a cluster interact only with particles of the same cluster. Resampling is done independently in each cluster, while maintaining the probability distribution intact. This way, all the modes of the posterior can be tracked through the whole audio piece, and the best hypothesis can be chosen at the end. In this work, we use an identical clustering scheme using a cyclic distance measure as described by Krebs, Holzapfel, et al. (2015) to track several different possible metrical positions at a given time. We use a cyclic distance measure that can take into account the cyclic nature of the bar position  $\phi$ . By representing the bar position as a complex phasor on the unit circle, we can compute the corresponding angle  $\varphi(\phi_k) = 2\pi\phi_k/M$ .

A distance between two particles indexed by  $i$  and  $j$  can then be computed as,

$$\begin{aligned} d(i, j) = \lambda_\phi & \left[ (\cos(\varphi^{(i)}) - \cos(\varphi^{(j)}))^2 + (\sin(\varphi^{(i)}) - \sin(\varphi^{(j)}))^2 \right] \\ & + \lambda_{\dot{\phi}} \left( \dot{\phi}^{(i)} - \dot{\phi}^{(j)} \right)^2 + \lambda_r (r^{(i)} - r^{(j)})^2 \end{aligned} \quad (5.16)$$

where, the parameters  $[\lambda_\phi, \lambda_{\dot{\phi}}, \lambda_r]$  control the relative weights in the distance.

In the MPF, after an initial cluster assignment, we perform a re-clustering before every resampling step, merging or splitting clusters based on the average distance between cluster centroids. The clustering, merging and splitting of clusters is necessary to control the number of clusters, which ideally represents the number of modes in the posterior. The mixture particle filter can be combined with the Auxiliary resampling to give the Auxiliary Mixture Particle Filter (AMPF). As recommended by Krebs, Holzapfel, et al. (2015), we resample at a fixed interval  $T_s$ .

It has been clearly shown by Krebs, Holzapfel, et al. that AMPF can be effectively used for the task of meter inference and tracking. In this dissertation, the AMPF algorithm, as outlined in Algorithm 1 is used for all meter analysis tasks that need approximate inference. The AMPF algorithm with the bar pointer model as described in this section will be denoted as  $\text{AMPF}_0$  in the dissertation.

The complexity of the PF schemes scale linearly with the number of particles  $N_p$  irrespective of the size of state space, leading to an efficient inference in large state spaces. Further, compared to the HMM using Viterbi decoding that has a space complexity of  $O(K \cdot |\mathcal{S}|)$ , the PF needs to store just  $N_p$  state trajectories and weights, significantly reducing the memory requirements. An additional advantage is that the number of particles can be chosen based on the computational power we can afford, and we can make the state space larger with no or only a marginal increase in the computational requirements.

To conclude, the bar pointer model is a state of the art model useful in all the meter analysis that are addressed in the thesis. The performance of meter analysis with bar pointer model will be a baseline for all the datasets and music cultures under study. Though a state of the art model explored before, the dissertation presents a

---

**Algorithm 1** An outline of the  $\text{AMPF}_0$  algorithm (Inference in BP-model using AMPF)

---

```

1: for  $i = 1$  to  $N_p$  do
2:   Sample  $\mathbf{x}_0^{(i)} \sim P(\mathbf{x}_0)$  ▷  $\mathbf{x}_k = [\phi_k, \dot{\phi}_k, r_k]$ 
3:   Set  $w_0^{(i)} = 1/N_p$ 
4: Cluster  $\{\mathbf{x}_0^{(i)} | i = 1, 2, \dots, N_p\}$ , get cluster assignments  $\{c_0^{(i)}\}$ 
5: for  $k = 1$  to  $K$  do
6:   for  $i = 1$  to  $N_p$  do ▷  $\phi, r$ : Proposal and weights
7:     Sample  $\phi_k^{(i)} \sim P(\phi_k^{(i)} | \mathbf{x}_{k-1}^{(i)})$ , Set  $c_k^{(i)} = c_{k-1}^{(i)}$ 
8:     if  $\phi_k^{(i)} < \phi_{k-1}^{(i)}$  then ▷ Bar crossed
9:        $r_k^{(i)} \sim P(r_k^{(i)} | r_{k-1}^{(i)})$  ▷ Sample patterns
10:    else
11:       $r_k^{(i)} = r_{k-1}^{(i)}$ 
12:       $\tilde{w}_k^{(i)} = w_k^{(i)} \cdot P(\mathbf{y}_k | \phi_k^{(i)}, r_k^{(i)})$ 
13:    for  $i = 1$  to  $N_p$  do ▷ Normalize weights
14:       $w_k^{(i)} = \frac{\tilde{w}_k^{(i)}}{\sum_{i=1}^{N_p} \tilde{w}_k^{(i)}}$ 
15:    if  $\text{mod}(k, T_s) = 0$  then ▷ Cluster, resample, reassign
16:      Cluster and resample  $\{\mathbf{x}_k^{(i)}, w_k^{(i)}, c_k^{(i)} | i = 1, 2, \dots, N_p\}$ 
      to obtain  $\{\hat{\mathbf{x}}_k^{(i)}, \hat{w}_k^{(i)} = 1/N_p, \hat{c}_k^{(i)}\}$ 
17:      for  $i = 1$  to  $N_p$  do
18:         $\mathbf{x}_k^{(i)} = \hat{\mathbf{x}}_k^{(i)}, w_k^{(i)} = \hat{w}_k^{(i)}, c_k^{(i)} = \hat{c}_k^{(i)}$ 
19:        Sample  $\dot{\phi}_k^{(i)} \sim P(\dot{\phi}_k^{(i)} | \dot{\phi}_{k-1}^{(i)})$  ▷ Sample tempo
20: Compute  $\mathbf{x}_{1:K}^* = \mathbf{x}_{1:K}^{(i*)} | i^* = \text{argmax}_i w_K^{(i)}$  ▷ MAP sequence

```

---

further exploration of the model with the following novelties compared to the state of the art:

1. The bar pointer model has been extended and evaluated on Indian art music, showing its utility and discussing its limitations with the kinds of metrical structures that occur in Indian music. These learnings and insights will help improve the components of the model, pushing the state of the art ahead.
2. Even though the bar pointer model can handle multiple rhythmic patterns per rhythm class (or meter type), only one previous

study has applied it to include more than one rhythmic pattern per rhythm class(Holzapfel et al., 2014). The dissertation for the first time applies the bar pointer model to multiple rhythm patterns per rhythm class and presents a comprehensive evaluation.

3. Several novel extensions to the bar pointer model are explored and presented in the dissertation to address several shortcomings of the model, and to extend the functionality of the model.

Several extensions and enhancements to the bar pointer model can be proposed. For better organization, these extensions are grouped into two categories: model extensions that explore changes to the model structure of the bar pointer model, either by adding additional hidden variables or using different conditional independence relationships, and inference extensions that explore different inference schemes in the bar pointer model, for better and faster inference.

### 5.3.2 Model extensions

The model extensions proposed to the bar pointer model improve upon the model structure. Two different model extensions are proposed in the dissertation: a mixture observation model, and the section pointer model.

#### Bar pointer model with a mixture observation model (MO-model)

We propose a simplification to the bar pointer model that uses a diverse mixture observation model incorporating observations from multiple rhythmic patterns. The bar pointer model uses multiple rhythmic patterns for meter analysis. When the task is only to track the beats and downbeats in meter tracking (assuming the meter type is known *a priori*), tracking pattern transitions is superfluous. However, to capture the diversity of patterns, a diverse mixture observation model can be used to incorporate observations from multiple rhythmic patterns. Since all the rhythmic patterns belong to the same type of meter, we can simplify BP-model to track only the  $\phi$  and  $\phi$  variables while using an observation model that computes the likelihood of an observation by marginalizing over all the

patterns. The motivation for this simplification is two-fold: the inference is simplified with only two hidden variables, and we can increase the influence of diverse patterns that occur throughout a metrical cycle in the inference. This simplification of the BP-model that uses a mixture observation model is referred to as MO-model and is shown in Figure 5.4b.

With this simplification in the model structure in Figure 5.4b, the transition model in Eq. 5.4 now changes to,

$$P(\mathbf{x}_k \mid \mathbf{x}_{k-1}) = P(\boldsymbol{\beta}_k \mid \boldsymbol{\beta}_{k-1}) = P(\phi_k \mid \phi_{k-1}, \dot{\phi}_{k-1}) P(\dot{\phi}_k \mid \dot{\phi}_{k-1}) \quad (5.17)$$

Here,  $\boldsymbol{\beta} = [\phi, \dot{\phi}]$  is defined as the subset of the hidden variables tracked using the MO-model. The tempo transition term of the above equation remains identical to the BP-model, as in Eq. 5.6. The term for  $\phi$  also remains similar to Eq. 5.5 in the BP-model, apart from the removal of the dependence on  $r_{k-1}$  as,

$$P(\phi_k \mid \phi_{k-1}, \dot{\phi}_{k-1}) = \mathbb{1}_\phi \quad (5.18)$$

where  $\mathbb{1}_\phi$  is an indicator function that takes a value of one if  $\phi_k = (\phi_{k-1} + \dot{\phi}_{k-1}) \bmod(M)$  and zero otherwise, noting that the length of all rhythmic patterns are equal,  $M_r = M$ , for all values of  $r$ .

The observation model aims to utilize information from multiple rhythmic patterns. The MO-model uses a mixture observation model computed from Eq. 5.8 by marginalizing over the patterns, assuming equal priors.

$$P(\mathbf{y} \mid \boldsymbol{\beta}) \propto \sum_{j=1}^R P(\mathbf{y} \mid \phi, r = j) \quad (5.19)$$

This observation model makes the MO-model simpler, while giving a computational advantage. Since the observation likelihood can be precomputed, inference with MO-model requires much lower computational resources, with only a marginal increase in cost during inference with increase in number of patterns.

### Inference in MO-model

The inference in MO-model is similar to that using BP-model, by discretizing the state space to lead to an HMM and applying Viterbi

algorithm, or using particle filters. The inference in HMM can be performed with pre-computed likelihood from different rhythmic patterns from the mixture observation model, denoted to as  $\text{HMM}_m$  in this dissertation. Similarly, the AMPF with the mixture observation model extension is outlined in Algorithm 2 and is denoted as  $\text{AMPF}_m$  in the rest of the chapter.

---

**Algorithm 2** Outline of the  $\text{AMPF}_m$  algorithm (AMPF for inference in the simplified bar pointer model with a mixture observation model: MO-model)

---

```

1: for  $i = 1$  to  $N_p$  do
2:   Sample  $\beta_0^{(i)} \sim P(\phi_0)P(\dot{\phi}_0)$ ,  $w_0^{(i)} = 1/N_p$   $\triangleright \beta_k = [\phi_k, \dot{\phi}_k]$ 
3:   Cluster  $\{\beta_0^{(i)} | i = 1, 2, \dots, N_p\}$ , get cluster assignments  $\{c_0^{(i)}\}$ 
4: for  $k = 1$  to  $K$  do
5:   for  $i = 1$  to  $N_p$  do  $\triangleright \phi$ : Proposal and weights
6:     Sample  $\phi_k^{(i)} \sim P(\phi_k^{(i)} | \beta_{k-1}^{(i)})$ , Set  $c_k^{(i)} = c_{k-1}^{(i)}$ 
7:      $\tilde{w}_k^{(i)} = w_k^{(i)} \times \sum_{j=1}^R P(\mathbf{y}_k | \phi_k^{(i)}, r = j)$ 
8:   for  $i = 1$  to  $N_p$  do  $\triangleright$  Normalize weights
9:      $w_k^{(i)} = \frac{\tilde{w}_k^{(i)}}{\sum_{i=1}^{N_p} \tilde{w}_k^{(i)}}$ 
10:  if  $\text{mod}(k, T_s) = 0$  then  $\triangleright$  Cluster, resample, reassign
11:    Cluster and resample  $\{\beta_k^{(i)}, w_k^{(i)}, c_k^{(i)} | i = 1, 2, \dots, N_p\}$ 
        to obtain  $\{\hat{\beta}_k^{(i)}, \hat{w}_k^{(i)} = 1/N_p, \hat{c}_k^{(i)}\}$ 
12:    for  $i = 1$  to  $N_p$  do
13:       $\beta_k^{(i)} = \hat{\beta}_k^{(i)}$ ,  $w_k^{(i)} = \hat{w}_k^{(i)}$ ,  $c_k^{(i)} = \hat{c}_k^{(i)}$ 
14:      Sample  $\dot{\phi}_k^{(i)} \sim P(\dot{\phi}_k^{(i)} | \dot{\phi}_{k-1}^{(i)})$ 
15:  Compute  $\beta_{1:K}^* = \beta_{1:K}^{(i*)} | i^* = \text{argmax}_i w_K^{(i)}$   $\triangleright$  MAP sequence

```

---

## Section pointer model

To the best of our knowledge, the methods for meter tracking and inference so far, including the bar pointer model, have been applied and evaluated on metrical cycles of short durations. E.g., the typical duration of a 4/4 measure in popular Eurogenetic music would

last from a bit less than 2s to little more than 4s. Longer metrical cycles were reported to cause problems in existing approaches (Holzapfel et al., 2014). Interestingly, this upper duration coincides with the limit of a perceptual phenomenon referred to as *perceptual present* (Clarke, 1999), and it has been argued that longer metrical cycles might not be perceived as a single rhythmic entity (Clayton, 2000). In tracking such long metrical cycles, listeners often track shorter, but musically meaningful sections of the cycle. This motivates the use of sub-bar or sub-cycle length rhythmic patterns in meter analysis tasks. Compared to longer cycle length patterns, shorter pattern have lower variability and hence might provide better cues for meter tracking.

A similar idea was applied by Böck et al. (2014), where rhythmic patterns of beat length are learned in order to perform beat tracking. However, Böck et al. assume beats to form an isochronous sequence - an assumption that does not hold for many musics of the world, such as Indian, Turkish, Balkan, or Korean musics. Furthermore, the authors do not attempt to infer higher metrical levels, i.e. downbeat positions. By proposing a generalization to the bar pointer model, we address for the first time, the two basic limitations of the existing meter tracking approaches including the bar pointer model: the restrictions to short cycles and isochronous beat sequences (Srinivasamurthy et al., 2016). The generalization of the bar pointer model, called the section pointer model (SP-model), uses musically meaningful and possibly unequal section length rhythmic patterns in the task of meter tracking. With the new model, it is further possible to evaluate if using shorter section length rhythmic patterns can improve meter tracking compared to bar (cycle) length rhythmic patterns, in the presence of long metrical cycles.

The idea behind the section pointer model is to track sections instead of the whole bar (cycle). The rhythmic patterns are now one section in length, and hence possibly unequal in length. A pointer tracks the progression through each section, and a over-arching section identifier handles the progression through the sections of a cycle. The structure of the SP-model is shown in Figure 5.4c, and is a generalization to the bar pointer model, with the bar pointer model being a special sub-case. Hence the SP-model can be applied to arbitrary music styles in a straight forward way, just like the bar pointer model.

Both Carnatic and Hindustani music have sections within the **tāla** (*aṅga* and *vibhāg*, respectively), which are musically well defined and hence the use of section length rhythmic patterns in the task of meter analysis can be explored with meaningful cycle divisions. Hindustani music further has **tāl** cycles that last over a minute (Clayton, 2000) and hence is a good test case for the section pointer model. **improve!** The large tempo range and the filler strokes can provide a dense surface rhythm than what is expected from the underlying metrical structure. This surface rhythm can confuse the meter trackers and bias it towards the higher values of tempo, something that can be mitigated by tracking shorter section length patterns. Further, tracking large **mātrā** periods in *vilāmbit* pieces causes an unstable local tempo estimate that leads to a drifting of the tracking algorithms, which also is expected to be mitigated by tracking shorter length patterns.

In the section pointer model, a hypothetical pointer traverses each section of a metrical cycle. Hence, in addition to the variables  $\phi, \dot{\phi}, r$  of the bar pointer model, we now additionally introduce a section indicator variable. In reference to the SP-model, at each audio frame, we redefine and denote the hidden (latent) variable vector  $\mathbf{x}_k = [\phi_k, \dot{\phi}_k, r_k, v_k]$ , where:

- *Section indicator*: The section indicator variable  $v \in \{1, \dots, V\}$  is an indicator variable that identifies the section (*vibhāg* in Hindustani music or *aṅga* in Carnatic music) of a bar (**tāl/a**), and selects one of the  $V$  observation models corresponding to each section length rhythmic pattern learned from data. A rhythm class (**tāl/a**) might have many sections of different lengths. We denote the number of **mātrās/beats** in a section  $v$  by  $B_v$ .
- *Rhythmic pattern indicator*: For each section  $v$ , there are one or more associated rhythm patterns denoted by  $r$ . The rhythm pattern indicator  $r$ , along with the section indicator  $v$  select the appropriate observation model to be used. For convenience, we assume each section to be modeled by an equal number of patterns, with a total of  $R$  distributed across all the sections equally. Hence, the number of rhythmic patterns per section is given as,  $R/V$  patterns, with the assumption that  $R$  is an integer multiple of  $V$ .

- *Position in section*: The position variable  $\phi$  in the SP-model tracks the position within a section as  $\phi \in [0, M_v]$ , where  $M_v$  is the length of section  $v$ .  $\phi$  increases from 0 to  $M_v$  and then resets to 0 to start tracking the next section. We set the length of the longest section as  $M$ , and then scale the lengths of other sections accordingly.
- *Instantaneous tempo*: Instantaneous tempo variable  $\dot{\phi}$  (measured in positions per time frame) is similar to the instantaneous tempo variable of the BP-model and denotes the rate at which the position variable  $\phi$  progresses through a section at each time frame. The allowed range of the variable  $\dot{\phi}_k \in [\dot{\phi}_{\min}, \dot{\phi}_{\max}]$  depends on the frame hop size ( $h = 0.02$  second used here as before), and can be preset or learned from data. In a given section  $v$ , a value of  $\dot{\phi}_k$  corresponds to a section duration of  $(h \cdot M_v / \dot{\phi}_k)$  seconds and  $(60 \cdot B_v \cdot \dot{\phi}_k / (M_v \cdot h))$  mātrās/beats per minute.

Given the conditional dependence relations in Figure 5.4c, the transition probability in SP-model factorizes as,

$$P(\mathbf{x}_k | \mathbf{x}_{k-1}) = P(\phi_k | \phi_{k-1}, \dot{\phi}_{k-1}, v_{k-1}) P(\dot{\phi}_k | \dot{\phi}_{k-1}) \\ P(v_k | v_{k-1}, \phi_k, \dot{\phi}_{k-1}) P(r_k | r_{k-1}, v_k, v_{k-1}) \quad (5.20)$$

Each of the terms in Eq. 5.20 can be expanded as,

$$P(\phi_k | \phi_{k-1}, \dot{\phi}_{k-1}, v_{k-1}) = \mathbb{1}_\phi \quad (5.21)$$

where  $\mathbb{1}_\phi$  is an indicator function that takes a value of one if  $\phi_k = (\phi_{k-1} + \dot{\phi}_{k-1}) \bmod(M_{v_{k-1}})$  and zero otherwise. The tempo transition is given by,

$$P(\dot{\phi}_k | \dot{\phi}_{k-1}) \propto \mathcal{N}(\dot{\phi}_{k-1}, \sigma_{\dot{\phi}}^2) \times \mathbb{1}_{\dot{\phi}} \quad (5.22)$$

where  $\mathbb{1}_{\dot{\phi}}$  is an indicator function that equals one if  $\dot{\phi}_k \in [\dot{\phi}_{\min}, \dot{\phi}_{\max}]$  and zero otherwise, restricting the tempo to be between a predefined range.  $\mathcal{N}(\mu, \sigma^2)$  denotes a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . The value of  $\sigma_{\dot{\phi}}$  depends on the value of tempo and the length of the section. As before with the BP-model, we set  $\sigma_{\dot{\phi}} = \sigma_n \cdot \dot{\phi}_{k-1} \cdot (M_{v_{k-1}} / M)$ , where  $\sigma_n$  is a user parameter that controls the amount of local tempo variations we allow in the music

piece. The section transition probability is given by,

$$P(v_k \mid v_{k-1}, \phi_k, \phi_{k-1}) = \begin{cases} \mathbb{B}(v_{k-1}, v_k) & \text{if } \phi_k < \phi_{k-1} \\ \mathbb{1}_v & \text{else} \end{cases} \quad (5.23)$$

where,  $\mathbb{B}$  is the  $V \times V$  time-homogeneous transition matrix with  $\mathbb{B}(i, j)$  being the transition probability from  $v_i$  to  $v_j$ , and  $\mathbb{1}_r$  is an indicator function that equals one when  $v_k = v_{k-1}$  and zero otherwise. The pattern transitions are governed by,

$$P(r_k \mid r_{k-1}, \phi_k, \phi_{k-1}) = \begin{cases} \mathbb{A}(r_{k-1}, r_k) & \text{if } \phi_k < \phi_{k-1} \\ \mathbb{1}_r & \text{else} \end{cases} \quad (5.24)$$

where,  $\mathbb{A}$  is the  $R \times R$  time-homogeneous transition matrix with  $\mathbb{A}(i, j)$  being the transition probability from  $r_i$  to  $r_j$ , and  $\mathbb{1}_r$  is an indicator function that equals one when  $r_k = r_{k-1}$  and zero otherwise.

Section changes are permitted only at the end of the section. Since the rhythmic patterns are also one section in length, pattern transitions are allowed only at the end of a section. The matrix  $\mathbb{B}$  is used to determine the order of the sections as defined in the **tāl/a** by allowing only those defined transitions. Further,  $\mathbb{B}$  can be set to do meter tracking by including only the section transitions of a specific **tāl/a**. A larger  $\mathbb{B}$  including all the sections from all the rhythm classes can be used for meter inference as well.  $\mathbb{A}$  closely follows  $\mathbb{B}$  and has non-zero probabilities only for allowed pattern transitions. As an illustration, consider tracking **rūpak tāl** (which has three **vibhāg**  $V = 3$ ) with the SP-model and two rhythmic patterns per section (hence,  $R = 6$ . The canonical forms of the section transition matrix  $\mathbb{B}$  and  $\mathbb{A}$  can be illustrated as in Figure 5.6.

The observation model with the SP-model is similar to that of the BP-model, with the assumption that the audio features depend on the position in section, the rhythmic pattern, and the section indicator variables. The annotated data has **mātrās/beats** numbered with their position in the bar (cycle) and hence they can be used to extract section length rhythmic patterns from audio recordings. Section length patterns from each section are then clustered into  $R/V$  pattern clusters using a k-means algorithm. Each section is further discretized into 64<sup>th</sup> note cells, all features within the cell are

$$\mathbb{B} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \mathbb{A} = \begin{bmatrix} 0 & 0 & p_1 & 1-p_1 & 0 & 0 \\ 0 & 0 & p_2 & 1-p_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_3 & 1-p_3 \\ 0 & 0 & 0 & 0 & p_4 & 1-p_4 \\ p_5 & 1-p_5 & 0 & 0 & 0 & 0 \\ p_6 & 1-p_6 & 0 & 0 & 0 & 0 \end{bmatrix}$$

**Figure 5.6:** An illustration of the form of section and rhythmic pattern transition matrices for tracking *rūpak tāl* with the SP-model. The patterns with index  $\{1, 2\}$ ,  $\{3, 4\}$ ,  $\{5, 6\}$  correspond to sections 1, 2, and 3, respectively. The values  $p_1$  to  $p_6$  are learnt from training data.

accumulated and a two component GMM is fit to each cell. The observation likelihood with the SP-model can hence be computed as,

$$P(\mathbf{y} | \mathbf{x}) = P(\mathbf{y} | \phi, r, v) = \sum_{i=1}^2 \pi_{\phi, r, v, i} \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{\phi, r, v, i}, \boldsymbol{\Sigma}_{\phi, r, v, i}) \quad (5.25)$$

where,  $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  denotes a normal distribution and for the mixture component  $i$ ,  $\pi_{\phi, r, v, i}$ ,  $\boldsymbol{\mu}_{\phi, r, v, i}$  and  $\boldsymbol{\Sigma}_{\phi, r, v, i}$  are the component weight, mean (2-dimensional) and the covariance matrix ( $2 \times 2$ ), respectively. Hence, there is an observation GMM for each section, rhythmic pattern, and tied section position states.

A special case of the SP-model is when the number of sections equals the number of rhythmic patterns,  $V = R$ , with each section being modeled with just one rhythmic pattern. In such a case, the matrices  $\mathbb{A} = \mathbb{B}$  rendering the additional  $r$  variable superfluous. In such a case, the SP-model can be simplified, as proposed and applied by (Srinivasamurthy et al., 2016), to the form shown in Figure 5.4d.

### Inference in SP-model

Both exact and approximate inference schemes can be used for inference in SP-model, similar to those for BP-model. The Viterbi algorithm inference on a discretized SP-model state space is denoted as  $\text{HMM}_s$ . The AMPF inference in SP-model will be referred to in

the rest of the chapter as  $\text{AMPF}_s$  and the algorithm is outlined in Algorithm 3.

---

**Algorithm 3** Outline of the  $\text{AMPF}_s$  algorithm (AMPF inference in SP-model)

---

```

1: for  $i = 1$  to  $N_p$  do
2:   Sample  $\mathbf{x}_0^{(i)} \sim P(\mathbf{x}_0)$                                  $\triangleright \mathbf{x}_k = [\phi_k, \dot{\phi}_k, r_k, v_k]$ 
3:   Set  $w_0^{(i)} = 1/N_p$                                           $\triangleright \boldsymbol{\alpha}_k = [\phi_k, \dot{\phi}_k, v_k]$ 
4: Cluster  $\{\mathbf{x}_0^{(i)} \mid i = 1, 2, \dots, N_p\}$ , get cluster assignments
    $\{c_0^{(i)}\}$ 
5: for  $k = 1$  to  $K$  do
6:   for  $i = 1$  to  $N_p$  do                                 $\triangleright \phi, r, v$ : Proposal and weights
7:     Sample  $\phi_k^{(i)} \sim P(\phi_k^{(i)} \mid \boldsymbol{\alpha}_{k-1}^{(i)})$ , Set  $c_k^{(i)} = c_{k-1}^{(i)}$ 
8:     if  $\phi_k^{(i)} < \phi_{k-1}^{(i)}$  then                                 $\triangleright$  Section crossed
9:        $r_k^{(i)} \sim P(r_k^{(i)} \mid r_{k-1}^{(i)})$                        $\triangleright$  Sample from  $\mathbb{A}$ 
10:       $v_k^{(i)} \sim P(v_k^{(i)} \mid v_{k-1}^{(i)})$                        $\triangleright$  Sample from  $\mathbb{B}$ 
11:     else
12:        $r_k^{(i)} = r_{k-1}^{(i)}, v_k^{(i)} = v_{k-1}^{(i)}$ 
13:        $\tilde{w}_k^{(i)} = w_k^{(i)} \cdot P(\mathbf{y}_k \mid \phi_k^{(i)}, v_k^{(i)}, r_k^{(i)})$ 
14:   for  $i = 1$  to  $N_p$  do                                 $\triangleright$  Normalize weights
15:      $w_k^{(i)} = \frac{\tilde{w}_k^{(i)}}{\sum_{i=1}^{N_p} \tilde{w}_k^{(i)}}$ 
16:   if  $\text{mod}(k, T_s) = 0$  then                                 $\triangleright$  Cluster, resample, reassign
17:     Cluster and resample  $\{\mathbf{x}_k^{(i)}, w_k^{(i)}, c_k^{(i)} \mid i = 1, 2, \dots, N_p\}$ 
     to obtain  $\{\hat{\mathbf{x}}_k^{(i)}, \hat{w}_k^{(i)} = 1/N_p, \hat{c}_k^{(i)}\}$ 
18:     for  $i = 1$  to  $N_p$  do
19:        $\mathbf{x}_k^{(i)} = \hat{\mathbf{x}}_k^{(i)}, w_k^{(i)} = \hat{w}_k^{(i)}, c_k^{(i)} = \hat{c}_k^{(i)}$ 
20:     Sample  $\dot{\phi}_k^{(i)} \sim P(\dot{\phi}_k^{(i)} \mid \dot{\phi}_{k-1}^{(i)})$            $\triangleright$  Sample tempo
21:   Compute  $\mathbf{x}_{1:K}^* = \mathbf{x}_{1:K}^{(i*)} \mid i^* = \text{argmax}_i w_K^{(i)}$      $\triangleright$  MAP sequence

```

---

### 5.3.3 Inference extensions

The inference extensions proposed aim for better approximate inference with the bar pointer model, either by making it faster, or by

improving approximate inference.

### End-of-bar pattern sampling

**clean up this section based on comments from Andre** The use of bar (cycle) length rhythmic patterns for meter analysis in BP-model is well motivated. When there are multiple rhythmic patterns being tracked, we can theoretically infer the rhythmic pattern that occurred in the current bar only after observing the features corresponding to the whole bar. However, in the  $\text{AMPF}_0$  algorithm with the BP-model, at the beginning of every bar, the pattern transition matrix  $\mathbb{A}$  is used to sample a pattern for the current bar. The rhythmic pattern so sampled is fixed for the whole bar, which is suboptimal. This is contrary to intuition, in which we need the whole bar to see and infer which pattern occurred, a decision that can only be made at the end of the bar, not the beginning. An extension to  $\text{AMPF}_0$  algorithm is proposed to address this limitation.

The extension, called end-of-bar pattern sampling extension to AMPF (called  $\text{AMPF}_e$  in short), defers a decision of sampling the pattern in the current bar to the end of the bar. In the current bar that is being tracked, the algorithm accumulates likelihood over all the patterns being tracked, and uses this likelihood to choose the most likely pattern at the end of the bar. The particle weights are updated at the end of the bar based on such an accumulated likelihood.

The proposed enhancement can be formulated in a particle system using two different clustering steps and resampling steps. In addition to AMPF clustering based on metrical position and tempo, an additional grouping is achieved with the rhythmic patterns for each particle, each of which interact within the groups during resampling. Hence within a single system of particles, we can defer the inference of patterns till the end of a bar, as outlined in detail below.

We first start by rewriting the particle system of Eq. 5.12 as,

$$P(\mathbf{x}_{1:K} \mid \mathbf{y}_{1:K}) \approx \sum_{i=1}^{N_p} \sum_{j=1}^R w_K^{(i,j)} \delta(\mathbf{x}_{1:K} - \mathbf{x}_{1:K}^{(i,j)}) \quad (5.26)$$

where  $\mathbf{x}_{1:K}^{(i,j)}$  are particle trajectories with weights  $w_K^{(i,j)}$ , both indexed by  $i$  and  $j$ . Compared to the particle system in Eq. 5.12, the

additional index  $j$  is used to index the rhythmic patterns for each particle. The weights are two dimensional, one dimension denotes the subset of hidden variables  $\beta = [\phi, \dot{\phi}]$ , and the other dimension stores the weights of all patterns for each value of  $\beta$ . With a suitable proposal density, these weights can be computed recursively as,

$$w_k^{(i,j)} \propto w_{k-1}^{(i,j)} \frac{P(\mathbf{y}_k \mid \mathbf{x}_k^{(i,j)}) P(\mathbf{x}_k^{(i,j)} \mid \mathbf{x}_{k-1}^{(i,j)})}{Q(\mathbf{x}_k^{(i,j)} \mid \mathbf{x}_{k-1}^{(i,j)}, \mathbf{y}_k)} \quad (5.27)$$

As before, we choose to sample from the transition probability  $Q(\mathbf{x}_k^{(i,j)} \mid \mathbf{x}_{k-1}^{(i,j)}, \mathbf{y}_k) = P(\mathbf{x}_k^{(i,j)} \mid \mathbf{x}_{k-1}^{(i,j)})$ , which reduces weight update to,

$$w_k^{(i,j)} \propto w_{k-1}^{(i,j)} P(\mathbf{y}_k \mid \mathbf{x}_k^{(i,j)}) = w_{k-1}^{(i,j)} P(\mathbf{y}_k \mid \beta_k^{(i)}, r_k = j) \quad (5.28)$$

Let us define the following terms:

$$\mathbf{w}_k^{(i,:)} = [w_k^{(i,1)}, w_k^{(i,2)}, \dots, w_k^{(i,R)}] \quad (5.29)$$

$$\Omega_k^{(i)} = \sum_{j=1}^R w_k^{(i,j)} \quad (5.30)$$

Here,  $\mathbf{w}_k^{(i,:)}$  stores the weights of a particle for each rhythmic pattern and  $\Omega_k^{(i)}$  denotes the marginal of a particle trajectory over all rhythmic patterns.

The **AMPE** is outlined in Algorithm 4. The algorithm can be interpreted to have two groups of particles in the particle system, one grouped based on  $\beta$  and the other group based on rhythm patterns. These two groups are sampled in two different sampling steps, one every  $T_s$  with the  $\beta$ , and one at the end of the bar with the rhythm pattern group of particles for a specific value of  $\beta$ . After each of the two resampling steps, the weights are redistributed to maintain a valid probability distribution over the particle system. Since all rhythmic patterns at a specific value of  $\beta$  are to be resampled together, it is necessary that all patterns be of equal size, and hence the **AMPE** algorithm can only be used in the task of meter tracking.

## Faster Inference

The MO-model presented in Section 5.3.2 simplifies the BP-model and makes inference faster. Inference in BP-model can also be

---

**Algorithm 4** Outline of the  $\text{AMPF}_e$  algorithm (AMPF inference in BP-model with end-of-bar pattern sampling)

---

```

1: for  $i = 1$  to  $N_p$  do
2:   Sample  $\beta_0^{(i)} \sim P(\phi_0)P(\dot{\phi}_0)$ ,  $(r_0^{(i)}) \sim P(r_0)$   $\triangleright \beta_k = [\phi_k, \dot{\phi}_k]$ 
3:   Set  $\mathbf{w}_0^{(i,:)} = 1/(N_p \cdot R)$ ,  $\Omega_k^{(i)} = 1/N_p$ ,  $\psi^{(i)} = 0$ 
4: Cluster  $\{\beta_0^{(i)} | i = 1, 2, \dots, N_p\}$ , get cluster assignments  $\{c_0^{(i)}\}$ 
5: for  $k = 1$  to  $K$  do
6:   for  $i = 1$  to  $N_p$  do  $\triangleright \phi$ : Proposal and weights
7:     Sample  $\phi_k^{(i)} \sim P(\phi_k^{(i)} | \phi_{k-1}^{(i)}, \dot{\phi}_{k-1}^{(i)})$ , Set  $c_k^{(i)} = c_{k-1}^{(i)}$ 
8:     if  $\phi_k^{(i)} < \phi_{k-1}^{(i)}$  then  $\triangleright$  Bar crossed
9:        $j^* = \text{argmax}_j(\mathbf{w}_k^{(i,:)})$ ; Set  $r_{\psi^{(i)}:k-1}^{(i)} = j^*$ ,  $\psi^{(i)} = k$ 
10:      for  $j = 1$  to  $R$  do
11:         $w_k^{(i,j)} = \mathbb{A}(j^*, j) \cdot \Omega_k^{(i)}$   $\triangleright$  Weights redistributed
12:      else
13:         $r_k^{(i)} = r_{k-1}^{(i)}$ 
14:      for  $j = 1$  to  $R$  do
15:         $\tilde{w}_k^{(i,j)} = w_k^{(i,j)} \cdot P(\mathbf{y}_k | \phi_k^{(i)}, r = j)$ 
16: for  $i = 1$  to  $N_p$  do  $\triangleright$  Normalize weights
17:   for  $j = 1$  to  $R$  do
18:      $w_k^{(i,j)} = \frac{\tilde{w}_k^{(i,j)}}{\sum_{i=1}^{N_p} \sum_{j=1}^R \tilde{w}_k^{(i,j)}}$ 
19:   if  $\text{mod}(k, T_s) = 0$  then  $\triangleright$  Cluster, resample, reassign
20:     Cluster and resample  $\{\beta_k^{(i)}, \Omega_k^{(i)}, c_k^{(i)} | i = 1, 2, \dots, N_p\}$ 
21:     to obtain  $\{\hat{\beta}_k^{(i)}, \hat{\Omega}_k^{(i)} = 1/N_p, \hat{c}_k^{(i)}\}$ 
22:     for  $i = 1$  to  $N_p$  do
23:       Set  $\beta_k^{(i)} = \hat{\beta}_k^{(i)}$ 
24:       for  $j = 1$  to  $R$  do  $\triangleright$  Weights redistributed
25:          $w_k^{(i,j)} = w_k^{(i,j)} \cdot \frac{\hat{\Omega}_k^{(i)}}{\Omega_k^{(i)}}$ 
26:       Sample  $\dot{\phi}_k^{(i)} \sim P(\dot{\phi}_k^{(i)} | \dot{\phi}_{k-1}^{(i)})$ 
27:      $\beta_{1:K}^* = \beta_{1:K}^{(i^*)} | i^* = \text{argmax}_i \Omega_K^{(i)}$   $\triangleright$  MAP sequence
In the algorithm,  $\Omega_k^{(i)} = \sum_{j=1}^R w_k^{(i,j)}$ 

```

---

made faster by utilizing the time sparsity of onsets to make inference faster, using what we propose as hop inference. The idea of hop inference is that instead of performing inference at every time frame, we do inference only at specific frames that are associated with rhythmic events such as onsets. The motivation for such a hop inference is that the onsets might just be sufficient to infer metrical structures. This makes inference faster by skipping likelihood computation and sampling steps and can speed up by inference by a factor as large as 10.

Two different hop inference algorithms extensions are proposed for AMPF with BP-model in this work:

**Peak Hop Inference ( $\text{AMPF}_p$ )** : The peaks of the spectral flux feature sequence is an indicator of events such as onsets. Using a peak finding algorithm, the peak frames are estimated. The particles are sampled and their weights are updated only at these peak frames. The transition model updates Eq. 5.4-5.7 are to be redefined accordingly. In particular, the position variable update shown in Eq. 5.5 scales the instantaneous tempo by the number of frames hopped from the previous peak in order to maintain the same tempo even with a peak hop inference. Peak hop inference can speed up inference by up to a factor of 10.

**Onset gated weight update ( $\text{AMPF}_g$ )** : Despite the advantage of a faster inference, peak hop inference can lead to sharp discontinuities in  $\phi$  and tempo values due to large jumps in their values since they are sampled after significant number of frames. An improvement to peak hop while maintaining continuity is the gated weight update, where  $\phi$  and  $\dot{\phi}$  are updated every frame to maintain continuity, while the observation model and weights of particles are updated only at frames where there is a peak in the feature, indicating an event. The basic premise is to maintain the continuity in tracking the  $\phi$  and  $\dot{\phi}$  variables, while retaining the principle of peak hop. Gated weight update needs an observation likelihood computation only at peak frames, and hence speeds up inference.

The different meter tracking models were presented in detail in this section can be summarized in Table 5.3. We now present the ex-

Acronym	Model	Inference algorithm	Meter Analysis	
			Inference	Tracking
$\S HMM_0$	BP-model <sup>†</sup>	Viterbi algorithm	✓	✓
$\S AMPF_0$	BP-model	AMPF	✓	✓
$*HMM_m$	MO-model <sup>†</sup>	Viterbi algorithm	✗	✓
$*AMPF_m$	MO-model	AMPF	✗	✓
$*HMM_s$	SP-model <sup>†</sup>	Viterbi algorithm	✓	✓
$*AMPF_s$	SP-model	AMPF	✓	✓
$*AMPF_e$	BP-model	AMPF with end-of-bar pattern sampling	✓	✓
$*AMPF_p$	BP-model	Peak hop inference with AMPF	✓	✓
$*AMPF_g$	BP-model	Onset gated weight update in AMPF	✓	✓

**Table 5.3:** A summary of the meter analysis models and inference algorithms presented in this section. The symbol  $\S$  indicates an existing state of the art algorithm while the symbol  $*$  is used to denote an algorithm proposed in this dissertation. The symbol  $†$  indicates that a discretized counterpart of the model is used. The last two columns show the applicability of the algorithm in the meter analysis tasks of meter inference and meter tracking. ✓ indicates applicable, ✗ indicates not applicable.

periments and results of evaluation of these models and extensions on the annotated datasets.

## 5.4 Experiments and results

This section comprehensively presents the experiments and results of meter analysis with different algorithms and approaches described in the chapter. The goals of the experiments presented in the section are:

- To evaluate different meter analysis tasks: Meter inference, meter tracking and informed meter tracking on both Carnatic

and Hindustani music datasets. The main focus is on evaluation of meter tracking with different algorithms discussed for the task.

- To compare performance across different approaches to meter analysis. To compares and discuss the performance of different models and inference algorithms - the BP-model with Viterbi and particle filter inference, model extensions (MO-model, SP-model) and the inference extensions (end-of-bar pattern sampling, peak hop, onset gated weight update).
- To compare performance of these approaches across different datasets and music cultures, with a baseline comparison with the ballroom dataset.
- To further identify challenges to meter analysis in Indian art music and identify the limitations of these approaches to suggest further improvements.

### 5.4.1 Experiment parameters

Several results of different algorithms are presented. Unless otherwise specified, the following global settings for the experiments is used. The results are the mean performance over three runs in a two fold cross validation experiment. The results on the Carnatic dataset (**CMR**) and Hindustani music data subsets **HMR<sub>s</sub>** and **HMR<sub>l</sub>** are focused on. The datasets **CMR** and **CMR<sub>f</sub>** have equivalent content and show equivalent results. Hence only the results on the **CMR** dataset are reported in the dissertation. As discussed earlier in Section 4.2.2, Hindustani music divides tempo into three main tempo classes (**lay**): slow (**vilambit**, 10-60 MPM), medium (**madhya**, 60-150 MPM), and fast (**drt**, > 150 MPM). In our experiments, we will examine how the tempo class affects the tracking accuracy. Hence for Hindustani music, results are presented for **HMR<sub>l</sub>** (**vilambit** pieces) and **HMR<sub>s</sub>** (**madhya** and **drt** pieces) datasets separately to assess performance individually on long and short cycle pieces. The results are presented for each dataset as an average over the pieces in all the **tāla** (or meters), while specific comments on performance on each **tāla** is discussed when needed. Performance on

ballroom dataset is reported for meter inference and tracking tasks for comparison.

The performance of algorithms is presented for both beat and sama (downbeat) tracking. For beat tracking, we use the evaluation measures f-measure ( $f_b$ ), AML<sub>t</sub>(AML<sub>t,b</sub>) and information gain ( $\mathcal{I}_b$ ). The subscript  $b$  indicates that the measure refers to beat tracking. Sama tracking is measured using f-measure ( $f_s$ ). For evaluation in this paper, we used the evaluation toolkit developed by Matthew Davies using the code available at <http://code.soundsoftware.ac.uk/projects/beat-evaluation/>. To compute the f-measure in CMR, HMR<sub>s</sub>, and Ballroom datasets, an error tolerance window of 70 ms is used between the annotation and the estimated beat/sama. For other evaluation measures, we use default parameters in the evaluation toolbox.

**to be explained better: suggest newer measures of evaluation for long cycles** However, for computing f-measure with HMR<sub>l</sub> dataset, a bigger margin window is allowed. Since cycles are long in HMR<sub>l</sub> dataset and current evaluation approaches were not designed with such long cycles in mind, an allowance of 70ms is very tight. To account for the length of the cycle in the margin, a 6.25% median inter annotation interval is used as margin, as used in many other beat tracking evaluations. This choice of a larger allowance window also corroborates well with the observation that in vilambit pieces of the HMR<sub>l</sub> dataset, there can be significant freedom in pulsation and that larger errors go unnoticed since the pieces are not rhythmically dense. Arguably, the pulsation in vilambit pieces is also beyond the duration of what is called the perceptual present (Clarke, 1999).

For meter inference and tracking, we additionally report the results of median tempo estimation as estimated from the estimated beats. For evaluating median tempo estimation, we compare the median estimated tempo and the median annotated ground truth tempo with a 5% error margin. In addition, to understand a metrical ambiguities in tempo estimation, we compute both CML and AML tempo estimation accuracy. In addition to the correct metrical level, AML assumes a tempo scaling by factors of 0.25, 0.5, 1 (correct metrical level), 2, 4 to be correct. For meter inference, the algorithms also detect the rhythm class and hence the accuracy of tāla recognition is also reported for the task.

Most experiments are conducted for rhythmic patterns per rhythm class  $R = 1$  and  $R = 2$ , but the results are presented only for  $R = 1$ . Experiments with  $R = 2$  does not show any significant improvement/change. When necessary, performance with  $R = 2$  is indicated. It is to be noted that with  $R = 1$ , model extension  $\text{AMPF}_m$  and inference extension  $\text{AMPF}_e$  are equivalent to the baseline  $\text{AMPF}_0$ .

The goal of the experiments is to use as much prior information on the metrical structures being tracked. All experiments are done on the dataset from each music culture separately, to capture the specificities of each music culture. In meter inference experiments, the total set of *tālas* being tracked is known, along with their structure. The training dataset contains pieces from all the *tāla* contained in the dataset. In meter tracking experiments, the specific *tāla* being tracked and its structure is known, and the training data contains pieces from the specific *tāla* only. In all the meter tracking experiments on Hindustani music, experiments are done separately on the two subsets  $\text{HMR}_s$  and  $\text{HMR}_l$ . Hence the meter tracking experiments on Hindustani music are not only just *tāl* informed, but also *lay* (tempo class) informed, i.e. the algorithm knows if it is tracking long cycles or short cycles. For informed tracking, additional information is provided to the tracking algorithm on tempo and the first instance of downbeat, as discussed.

The tempo ranges are learned from training data of each fold, with 20% margin allowed on learned ranges for unseen data. However, a minimum and maximum tempo is set for each music culture independently, and if the learned tempo ranges lie outside that range, they are set to the these preset min and max values. The minimum and maximum tempo range for Carnatic music is set as [140, 520] *akṣaras* per minute, that for Hindustani music is set as [10, 370] *mātrās* per minute, and for ballroom dataset as [60, 230] BPM.

We use the number of bar positions,  $M_r = 1600$  for the longest rhythmic pattern we encounter in the dataset and scale all other pattern lengths accordingly. As indicated in Section 5.3.1, for meter tracking experiments,  $M_r = M = 1600$  is set for the longest pattern being tracked. The maximum  $M = 1600$  corresponds to *ādi tāla* in Carnatic music (8 beats and 32 *akṣaras*) and *tīntāl* (16 *mātrās*) in Hindustani music. If a different *tāla* is being tracked,

we set the value of  $M$  accordingly, e.g.  $M = 600$  for tracking the three beat *rūpaka tāla* in Carnatic music. For Ballroom dataset, we used  $M = 1600$  and  $M = 1200$  for tracking time signatures 4/4 and 3/4, respectively.

The number of beats  $B$  and the number of sections is set accordingly depending on the dataset and the *tāla*/s being tracked from Table 2.1 and Table 2.3. When  $R > 1$ , the transition probabilities of patterns are also learned from training data from the clustered bar/section length patterns.

For meter inference and tracking, we use uniform priors on all hidden variables within the allowed range of values. For informed tracking, priors on tempo and the position variables are set according to the prior information we have available on the tempo and the *sama* instances. The observation model uses a two dimensional spectral flux feature computed at a hop size  $h = 0.02$  seconds, as described in Figure 4.7. The bar is discretized into 64<sup>th</sup> note cells within which the observation probability is assumed to be constant.

For the HMM based Viterbi inference, the tempo state transition probability in Eq. 5.11 is set to  $n_p = 0.02$ , as used by Krebs et al. (2013), allowing a small probability of change of tempo.

For the AMPF, the number of particles is chosen as  $N_p = 1500 \times R$ . We set the user parameter that controls tempo variance in Eq. 5.6 to  $\sigma_n = 0.02$  and the maximum number of clusters in the MPF to 200. The resampling interval is set to  $T_s = 30$  frames, which corresponds to a resampling step every 0.6 seconds of audio. The other AMPF parameters are identical to the values used by Krebs, Holzapfel, et al. (2015).

There are several combinations of datasets (and their subsets), algorithms, evaluation measures and parameter settings for which the results can be reported. While the experimentation was comprehensive, only a selected set of relevant results are presented in the dissertation for brevity and conciseness. We first present results of meter inference with the bar pointer model as a baseline, followed by meter tracking for different variants, model and inference extensions. Informed meter tracking is discussed at the end. A final summary of results over all the Indian art music datasets is also presented for comparison of approaches.

	Algo.	$f_b$	$AML_{t,b}$	$\mathfrak{I}_b$	$f_s$	Tempo		Tāla %
				Bits		CML	AML	
CMR	$HMM_0$	0.718	0.722	1.44	0.440	0.718	0.938	64
	$AMPF_0$	0.825	0.906	2.17	0.574	0.802	1.000	68
$HMR_s$	$HMM_0$	0.759	0.698	1.21	0.551	0.533	0.721	60
	$AMPF_0$	0.828	0.834	1.54	0.569	0.714	0.946	63
$HMR_l$	$HMM_0$	0.338	0.225	0.77	0.280	0.119	0.350	37
	$AMPF_0$	0.390	0.427	1.35	0.268	0.350	0.740	27
Blrm.	$HMM_0$	0.853	0.910	2.52	0.666	0.755	0.988	91
	$AMPF_0$	0.813	0.850	2.15	0.529	0.709	0.957	89

**Table 5.4:** Results of meter inference with the bar pointer model ( $HMM_0$  and  $AMPF_0$ ) on different datasets. The first column indicates the dataset, with Blrm. denoting the Ballroom dataset. The last column of the table shows the tāla recognition (or time signature estimation for Ballroom dataset) accuracy. The table also reports tempo estimation performance (at both CML and AML), beat and sama (downbeat) tracking performance with different measures.

### 5.4.2 Meter inference

The results of meter inference provides a baseline for meter analysis algorithms when the underlying metrical structure is unknown. It is the hardest task, estimating the tāla, the tempo, the beats and sama. The results are presented for inference on the BP-model on CMR,  $HMR_s$ , and  $HMR_l$  and Ballroom datasets for both  $HMM$  variant sampling from pattern transition priors ( $HMM_0$ ) and  $AMPF$  variant sampling from pattern transition priors ( $AMPF_0$ ) algorithms in Table 5.4. The model training uses pooled data from all the rhythm classes within a particular dataset. The results are presented for  $R = 1$  per rhythm class, without any improvement for  $R = 2$ .

At a broad level, we see that the performance on Ballroom dataset is better than that for the Indian music datasets. The performance on long cycle pieces in  $HMR_l$  dataset is poor, showing the challenges in tracking long metrical cycle durations. The performance with  $HMM_0$  is marginal poorer than  $AMPF_0$  for Indian music datasets. Since metrical cycles in Indian music are longer in duration, it is neces-

sary to have a finer discretization grid. The poorer performance is largely attributed to the coarse grain discretization of the state space that is used.

From Table 5.4, from the last column that indicates *tāla* recognition accuracy, we see that the *tāla* recognition is better with short metrical cycle duration pieces in CMR and  $\text{HMR}_s$  dataset with an accuracy between 60-70%. For long cycle duration pieces in  $\text{HMR}_l$  dataset, the *tāla* recognition accuracy drops significantly (to less than 40%) indicating the difficulty in tracking long duration cycles. The time signature recognition performance in Ballroom dataset is also higher than that for Indian music datasets (about 90%).

We further can observe that the f-measure for *sama*/downbeat tracking (indicated by  $f_s$ ) is significantly poorer than beat tracking performance (indicated by  $f_b$ ), showing that while beat tracking is still possible without the knowledge of underlying metrical structures, estimating the downbeats is poor. Beat AMLt measure  $\text{AML}_{t,b}$  is comparable to beat f-measure. It was reported by Holzapfel et al. (2012) that an information gain of 1.5 beats is acceptable to users as satisfactory beat tracking. Such an acceptable beat tracking is achieved in many cases.

Median tempo estimation performance is poorer for meter inference at CML. The large difference in CML and AML tempo tracking performance shows that there are significant metrical level estimation errors in meter inference. This further contributes to poorer beat and downbeat tracking performance.

There is a large performance difference between  $\text{HMR}_s$  and  $\text{HMR}_l$  datasets, further emphasizing the difficulties of tracking long duration cycles. The tempo estimation at CML with  $\text{HMR}_l$  dataset is as low at 12% with  $\text{HMM}_0$  showing that the correct metrical level of tracking is achieved in very small number of cases. Discretization of the tempo state space is one reason for the inability to track long cycles, where an extremely fine grid of variables is needed.  $\text{AMPF}_0$  has no such restrictions and hence performs better for this case.

Within the CMR dataset, the performance is poorer for longer cycle *ādi tāla* for both beat and *sama* estimation at  $f_s = 0.36$  and  $f_b = 0.67$  with  $\text{HMM}_0$ . In Hindustani music  $\text{HMR}_s$  dataset, the performance is best for *dṛ̥t ēktāl* pieces that tend have high tempo and short duration cycles. Both these observations indicate that short duration cycles are better tracked by the inference algorithm. More

	Algo.	$f_b$	$AML_{t,b}$	$\mathfrak{I}_b$ Bits	$f_s$	Tempo	
						CML	AML
CMR	$HMM_0$	0.784	0.771	1.59	0.624	0.890	0.915
	$AMPF_0$	0.827	0.840	1.97	0.671	0.955	0.997
$HMR_s$	$HMM_0$	0.835	0.796	1.39	0.733	0.663	0.830
	$AMPF_0$	0.884	0.858	1.64	0.772	0.844	0.964
$HMR_l$	$HMM_0$	0.353	0.305	0.86	0.429	0.294	0.435
	$AMPF_0$	0.374	0.513	1.40	0.396	0.390	0.610
Blrm.	$HMM_0$	0.929	0.921	2.78	0.821	0.987	0.989
	$AMPF_0$	0.909	0.895	2.56	0.735	0.98	0.98

**Table 5.5:** Results of meter tracking with the bar pointer model ( $HMM_0$  and  $AMPF_0$ ) on different datasets. The first column indicates the dataset, with Blrm. denoting the Ballroom dataset. The table shows the tempo estimation performance at CML and AML, beat and sama (downbeat) tracking performance with different measures.

on specific **tālas** and comment on performance ??

### 5.4.3 Meter tracking

Meter tracking is the most relevant task in the context of Indian art music and hence is the main focus of the experiments presented here. Meter tracking experiments assume that the **tāla** is known, and hence meter tracking is done for each **tāla** in the datasets separately. The training data also includes pieces from the specific **tāla** being tracked. Some of the results presented in this section are published results from previous publications by Holzapfel et al. (2014); Srinivasamurthy et al. (2015, 2016).

Before presenting the results for model and inference extensions, we tabulate the performance of meter tracking with the bar pointer model for the Indian music datasets and Ballroom dataset for both  $HMM_0$  and  $AMPF_0$  algorithms in Table 5.5. This provides another baseline performance to compare with meter inference performance and also for all the extensions discussed in the thesis.

At a broad level, Table 5.5 shows an improvement in perfor-

mance with meter tracking compared to meter inference (Table 5.4). In addition, we see a lower difference between beat and downbeat tracking f-measure values, indicating a larger improvement in downbeat estimation when the underlying metrical structure is known i.e. knowing the *tāla* improves the *sama* tracking performance. Similar to meter inference, the performance on short duration cycle datasets CMR,  $HMR_s$ , and Ballroom datasets is better than that on  $HMR_1$  dataset. Similar to meter inference, the poorer performance of  $HMM_0$  compared to  $AMPF_0$  is largely attributed to the coarse grain discretization of the state space.

The median tempo estimation performance with meter tracking is better than meter inference since a more narrow range of tempo is estimated due to the presence only one *tāla* in the training dataset. The difference between CML and AML performance is significantly lower in Carnatic music, showing that most pieces have been tracked at the correct metrical level. A similar trend can be observed with  $HMR_1$  dataset, while there is still scope for improvement in CML accuracy. With the Ballroom dataset, tempo estimation is accurate for most pieces, with a high accuracy.

In Carnatic music, with an  $f_s = 0.41$  and  $0.39$  for  $HMM_0$  and  $AMPF_0$ , *ādi tāla* has a significantly lower *sama* tracking performance compared to the other *tālas*.

With such a baseline with meter tracking with the bar pointer model with  $AMPF_0$  showing an equivalent or better performance than  $HMM_0$ , we report the results for all further model and inference extension experiments for particle filter inference only with AMPF. The results on model extensions MO-model and SP-model are presented next.

### Mixture observation model (MO-model)

The results of meter tracking with bar pointer model using a mixture observation model MO-model is shown in Table 5.6, for  $R = 1$ . With  $R = 1$ , the  $AMPF_m$  algorithm is equivalent to  $AMPF_0$  algorithm. Contrary to expectation, it is also observed that there was no significant improvement for  $AMPF_m$  from  $R = 1$  to  $R = 2$ . Further analysis and comparison of MO-model with  $AMPF_m$  algorithm is presented at the end of this section along with comparisons between all the approaches.

Dataset	$f_b$	$AML_{t,b}$	$\mathfrak{I}_b$ Bits	$f_s$	Tempo	
					CML	AML
CMR	0.838	0.840	1.96	0.671	0.958	1.00
HMR <sub>s</sub>	0.886	0.864	1.66	0.783	0.837	0.942
HMR <sub>I</sub>	0.364	0.506	1.39	0.455	0.401	0.554
Ballroom	0.907	0.895	2.56	0.730	0.981	0.981

**Table 5.6:** Results of meter tracking with the bar pointer model using a mixture observation model ( $\text{AMPF}_m$  algorithm) on different datasets. The table shows the tempo estimation performance at CML and AML, beat and *sama* (downbeat) tracking performance with different measures. The table shows results with  $R = 1$ , which is equivalent to  $\text{AMPF}_0$  algorithm.

Dataset	$f_b$	$AML_{t,b}$	$\mathfrak{I}_b$ Bits	$f_s$	Tempo	
					CML	AML
CMR	0.868	0.879	2.16	0.717	0.958	1.00
HMR <sub>s</sub>	0.924	0.890	1.88	0.850	0.855	0.971
HMR <sub>I</sub>	0.414	0.590	1.63	0.509	0.458	0.644

**Table 5.7:** Results of meter tracking with the section pointer model (SP-model with  $\text{AMPF}_s$  algorithm) on Indian music datasets. The table shows the tempo estimation performance at CML and AML, beat and *sama* tracking performance with different measures.

### Section pointer model

The experiments aim to compare the performance of meter tracking using bar length (BP-model) and the proposed section length (SP-model) patterns. The BP-model applies the position variable  $\phi$  to the whole *tāla* cycle, while the proposed SP-model applies  $\phi$  to the sections (*vibhāg/aṅga*) and imposes a sequential structure as described in Section 5.3.2. It is hypothesized that section pointer model would be useful for tracking long duration metrical cycles often encountered in Indian art music. Since sections are not musically well defined for the music styles in the Ballroom dataset, an evaluation of SP-model is limited to the Indian music datasets.

The results of meter tracking with the SP-model and  $\text{AMPF}_s$  is

	Dataset	$f_b$	$AML_{t,b}$	$\mathfrak{I}_b$ Bits	$f_s$	Tempo	
						CML	AML
CMR	$\text{AMPF}_e$	0.826	0.842	1.97	0.668	0.958	0.997
	$\text{AMPF}_p$	0.519	0.561	0.67	0.213	0.927	0.969
	$\text{AMPF}_g$	0.756	0.756	1.51	0.580	0.938	0.98
$HMR_s$	$\text{AMPF}_e$	0.882	0.858	1.64	0.777	0.833	0.935
	$\text{AMPF}_p$	0.655	0.572	0.59	0.273	0.768	0.822
	$\text{AMPF}_g$	0.821	0.653	1.25	0.653	0.743	0.895
Blrm.	$\text{AMPF}_e$	0.908	0.895	2.56	0.734	0.98	0.98
	$\text{AMPF}_p$	0.631	0.694	1.49	0.322	0.922	0.923
	$\text{AMPF}_g$	0.831	0.815	2.13	0.579	0.939	0.943

**Table 5.8:** Results of meter tracking with inference extensions to the bar pointer model on different datasets. The first column indicates the dataset, with Blrm. denoting the Ballroom dataset. The table shows the tempo estimation performance at CML and AML, beat and sama (downbeat) tracking performance with different measures.

shown in Table 5.7. It shows a significant improvement in sama tracking f-measure compared to  $\text{AMPF}_0$  with bar pointer model. The improvement with the long cycle duration pieces in  $HMR_1$  dataset is further encouraging to use shorter section length patterns to track longer cycles.

A significant improvement is also observed in Carnatic music with  $\bar{\text{adi tāla}}$  with the sama tracking f-measure of  $f_s = 0.46$  from  $f_s = 0.39$  for  $\text{AMPF}_0$ .

### Inference extensions

After model extensions, we now present the results for inference extensions to meter tracking. We present results for three different inference extensions - end of bar sampling ( $\text{AMPF}_e$ ), peak hop inference ( $\text{AMPF}_p$ ), and onset gated weight update ( $\text{AMPF}_g$ ). The goal of the experiments in this section is to compare the performance of the inference extensions with  $\text{AMPF}_0$  algorithm. The long cycle duration  $HMR_1$  dataset is excluded from evaluation of the inference ex-

tensions. Inference extensions are only evaluated on CMR and HMR<sub>s</sub> datasets (Ballroom dataset shown for reference) and compared with  $\text{AMPF}_0$ . The results are shown in Table 5.8. The table shows results only for  $R = 1$ , which means that  $\text{AMPF}_e$  is equivalent to  $\text{AMPF}_0$ . It is important to note that  $\text{AMPF}_e$  does not show any significant improvement from  $R = 1$  to  $R = 2$ .

For both  $\text{AMPF}_p$  and  $\text{AMPF}_g$  algorithms, a peak picker is used to choose the frames at which inference is done. A peak picking threshold of 5% of the maximum value of the spectral flux sequence is used to select peaks. Further if two peaks are within three frames of each other, then only the largest peak is chosen into the peak sequence.

From the table, we see that  $\text{AMPF}_e$  has equivalent performance to  $\text{AMPF}_0$ . Though peak hop inference provides a significant boost in inference time (up to 10x faster), the performance is very poor. By tracking and doing inference only at peaks, continuity of tracking meter is lost and leads to poor performance. In most cases, the continuity in tracking is necessary, and hop inference with large hops loses on tempo continuity. Further, in many cases, the beats and downbeats do not always occur not only at the peaks of the spectral flux sequence. Doing inference only at peaks misses on these events, and leads to an unstable tempo and beat/downbeat tracking leading to poor performance.

Onset gated weight update overcomes this limitation by inducing the tempo and position variables of the bar pointer model every frame and hence maintaining continuity. Though it also speeds up inference, its performance is poorer since it fails to model the rhythmic events that happen between the two peaks, since the observation probability is updated only at peaks in observation feature sequence. These extensions show the importance of non-peak values in the observations and the importance of continuity in the task of meter tracking. Further analysis and comparison of these extensions with  $\text{AMPF}_0$  is presented at the end of this section.

#### 5.4.4 Informed meter tracking

Informed meter tracking aims to incorporate additional information into meter tracking and meter tracking improve performance. The results for informed meter tracking with BP-model(with  $\text{AMPF}_0$ )

Dataset	Algo.	$f_b$	$AML_{t,b}$	$\mathfrak{I}_b$	$f_s$
CMR	$\text{AMPF}_0$	0.899	0.952	2.35	0.792
	$\text{AMPF}_s$	0.898	0.950	2.38	0.814
$HMR_1$	$\text{AMPF}_0$	0.425	0.959	2.76	0.786
	$\text{AMPF}_s$	0.439	0.979	2.83	0.848
$HMR_s$	$\text{AMPF}_0$	0.939	0.941	1.99	0.882
	$\text{AMPF}_s$	0.943	0.943	2.00	0.918

**Table 5.9:** Results of tempo informed meter tracking with  $\text{AMPF}_0$  and  $\text{AMPF}_s$  on Indian music datasets. The table shows beat and sama tracking performance with different measures.

and SP-model(with  $\text{AMPF}_s$ ) is presented here to evaluate if providing additional information is useful for meter tracking. If it improves performance, then several semi-automatic automatic rhythm annotation applications can benefit from this, utilizing varying levels of additional prior information to improve meter tracking. Informed meter tracking is evaluated only within the context of Indian art music and hence only on Indian music datasets.

We will present results for two different informed meter tracking schemes as discussed at the beginning of the chapter: tempo-informed meter tracking, and tempo-sama-informed meter tracking. For tempo-informed meter tracking, we use the median ground truth tempo of the music piece being tracked and initialize the tempo variable  $\phi$  within a tight bound allowing for 10% variation in tempo around the median value. This enables the tracking algorithm to restrict the tempo variable within this allowed tight tempo range and track the correct tempo at the right metrical level. For tempo-sama informed meter tracking, we assume that in addition to the true median tempo, we have the first instance of the downbeat in the music piece being tracked. We use the ground truth median tempo to initialize the  $\phi$  within a tight range as before, and further use the first instance of sama to initialize the  $\phi$  variable to zero at that sama instant. The tracking algorithm hence knows the tempo and the beginning of the cycle in the piece, tracking the remaining beats and downbeats.

The results of tempo-informed meter tracking is shown with

Dataset	Algo.	$f_b$	$AML_{t,b}$	$\mathfrak{I}_b$	$f_s$
CMR	$\text{AMPF}_0$	0.880	0.943	2.37	0.834
	$\text{AMPF}_s$	0.917	0.946	2.40	0.901
$HMR_1$	$\text{AMPF}_0$	0.530	0.978	2.84	0.99
	$\text{AMPF}_s$	0.542	0.98	2.83	0.99
$HMR_s$	$\text{AMPF}_0$	0.959	0.920	2.02	0.911
	$\text{AMPF}_s$	0.958	0.915	2.01	0.933

**Table 5.10:** Results of tempo-sama informed meter tracking with  $\text{AMPF}_0$  and  $\text{AMPF}_s$  on Indian music datasets. The table shows beat and sama tracking performance with different measures.

the bar and section pointer model (BP-model and SP-model, with  $\text{AMPF}_0$  and  $\text{AMPF}_s$ , respectively) on Indian art music datasets is shown in Table 5.9. A similar set of results for tempo-sama-informed tracking is shown in Table 5.10. The tables show that  $\text{AMPF}_s$  marginally performs better than  $\text{AMPF}_0$  in informed tracking, while including sama information in tempo-sama-informed meter tracking further improves sama tracking performance with a marginal improvement in beat tracking performance.

We see from these tables that we can achieve a high f-measure of 0.9 for both sama tracking in informed tracking. A similar beat tracking can be achieved in CMR and  $HMR_s$  datasets too. With significant allowance on tempo variation allowed in Hindustani music vilambit pieces, the beat tracking performance with informed tracking is poorer since we use the median tempo and tight bounds. However, sam tracking with vilambit pieces is good showing that the algorithm is capable of recovering from these local tempo changes and track the sam accurately.

Though tested with a small number of pieces within the context of Indian art music, it is encouraging to observe that easily obtainable additional prior information can be used to improve meter tracking performance, and that the Bayesian models and inference algorithms allow for incorporating such prior information seamlessly for tracking.

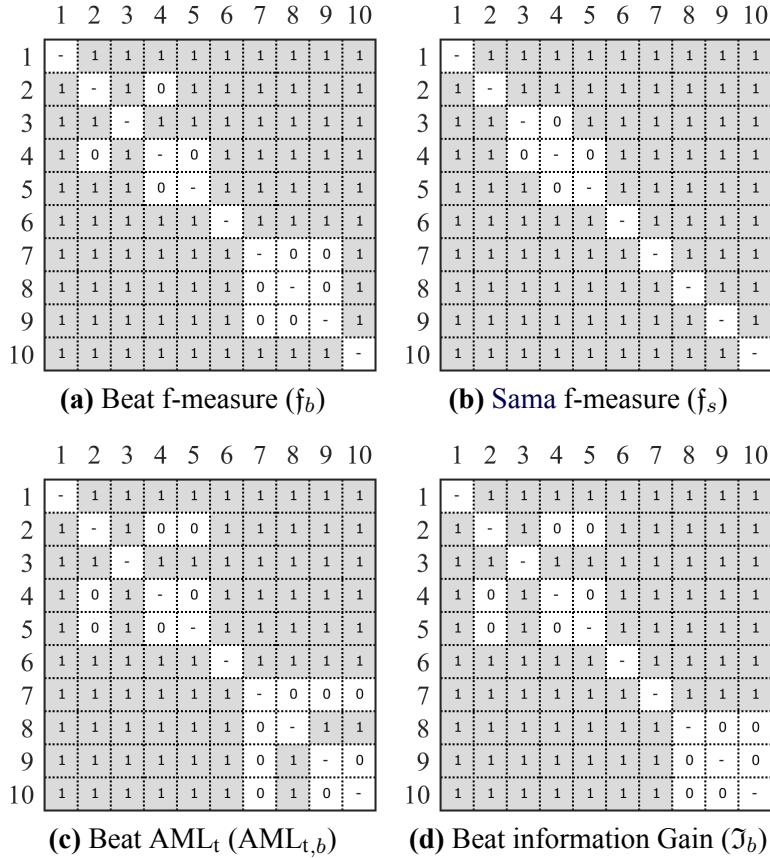
	Algo.	ID	$f_b$	$AML_{t,b}$	$\mathfrak{I}_b$ Bits	$f_s$	Tempo	
							CML	AML
Inf.	$HMM_0$	1	0.648	0.605	1.21	0.443	0.51	0.72
	$AMPF_0$	2	0.730	0.776	1.77	0.505	0.67	0.92
Track	$HMM_0$	3	0.707	0.677	1.36	0.618	0.67	0.77
	$AMPF_0$	4	0.747	0.774	1.73	0.645	0.79	0.89
	$AMPF_m$	5	0.750	0.775	1.73	0.662	0.79	0.88
	$AMPF_s$	6	0.779	0.817	1.92	0.704	0.81	0.91
t-Tr.	$AMPF_0$	7	0.809	0.950	2.32	0.822	0.99	0.99
	$AMPF_s$	8	0.813	0.954	2.35	0.857	1.00	1.00
ts-Tr.	$AMPF_0$	9	0.830	0.943	2.35	0.896	1.00	1.00
	$AMPF_s$	10	0.849	0.943	2.36	0.931	1.00	1.00

**Table 5.11:** Summary of meter analysis results on Indian music datasets. The meter analysis tasks are shown in the first column - with Inf., Track, t-Tr., and ts-Tr. referring to meter inference, meter tracking, tempo-informed meter tracking, and tempo-sama-informed meter tracking, respectively. The second column shows the different algorithms and the corresponding models. The table shows the tempo estimation performance at CML and AML, beat and sama (downbeat) tracking performance with different measures. The column ID on third column corresponds to the labels used in Figure 5.7, which shows the results of statistical significance tests on these results.

#### 5.4.5 Summary of results

A summary of the results to compare the performance of different algorithms is presented now. To compare results across algorithms, we pool the results from all the relevant Indian music datasets together and present the mean performance for the algorithm. It is to be noted that though a mean over all datasets is presented, the training and testing are separate for each dataset (for meter inference) and even for each *tāla* within a dataset (for meter tracking).

A paired sample t-test with  $p = 0.05$  is used to assess statistically significant differences between the performances of al-



**Figure 5.7:** Results of statistical significance testing of meter analysis results on Indian art music datasets. The figure shows the results for the four different performance measures: Beat f-measure ( $f_b$ ), Sama f-measure ( $f_s$ ), Beat AML<sub>t</sub> (AML<sub>t,b</sub>) and Beat information Gain ( $\mathcal{I}_b$ ) in panels (a), (b), (c), and (d), respectively. For each measure, the figure shows the results of a pairwise statistical test between methods (algorithms) numbered 1-10 as a matrix. A gray box with numeral 1 indicates a statistically significant difference (at  $p = 0.05$ ) while a white box with numeral 0 indicates a difference that is not statistically significant. The methods 1-10 map to the ID shown in Table 5.11.

gorithms. Statistical significance tests are done for the meter inference, meter tracking (model extensions), and informed tracking methods by pooling the results over all Indian music datasets

(CMR, HMR<sub>s</sub>, HMR<sub>I</sub> datasets - 269 pieces in total). Statistical significance tests are done for bar pointer model inference extensions ( $\text{AMPF}_e$ ,  $\text{AMPF}_p$ ,  $\text{AMPF}_g$ ) by pooling the results over CMR and HMR<sub>s</sub> (210 pieces in total) to compare with  $\text{AMPF}_0$ .

We pool the results of meter inference, meter tracking (model extensions), tempo-informed tracking, and tempo-sama-informed tracking on all the Indian music datasets and present it in Table 5.11. The results of statistical significance tests between these approaches is presented in Figure 5.7. Table 5.11 and Figure 5.7 are to be analyzed in conjunction. In both the table and the figure, since  $R = 1$ , note that  $\text{AMPF}_m$  is equivalent to  $\text{AMPF}_0$ .

From Table 5.11, we see a consistent increase over the rows of the table across different meter analysis experiments (inference, tracking and informed tracking) indicating that incorporating additional prior information can lead to improved meter analysis. Informed meter tracking has the best performance, while we see that meter tracking performance is mid-way between inference and informed meter tracking.

The Figure 5.7 shows that  $\text{AMPF}_0$  and  $\text{AMPF}_m$  are equivalent and produces results that are not statistically significantly different for all performance measures. The panel (a) in the figure for beat f-measure ( $f_b$ ) shows that  $\text{AMPF}_0$  algorithm in inference and tracking have statistically insignificant differences. In addition,  $\text{AMPF}_0$  and  $\text{AMPF}_s$  in tempo-informed tracking have insignificant differences with  $\text{AMPF}_0$  in tempo-sama-informed tracking. Sama f-measure ( $f_s$ ) shown in panel (b) indicates statistically insignificant differences between  $\text{HMM}_0$  and  $\text{AMPF}_0$ .

The SP-model shows statistically significant improvement over the methods that use BP-model indicating the use of section length shorter patterns for tracking downbeats. The beat AML<sub>t,b</sub> measure in panel (c) is comparable to beat f-measure. Informed tracking methods have several statistically insignificant differences with the beat AML<sub>t,b</sub> measure since the correct metrical level is already provided to the algorithm. An acceptable beat information gain ( $\mathcal{I}_b > 1.5$  bits) is obtained in most cases, with several statistically insignificant differences in performance in informed tracking.

To summarize the results from Table 5.11 and Figure 5.7, we see that informed tracking and algorithms using SP-model improve sama tracking performance significantly, while beat tracking per-

Algo.	ID	$f_b$	AML <sub>t,b</sub>	$\mathfrak{I}_b$ Bits	$f_s$	Tempo	
						CML	AML
$\text{AMPF}_0$	1	0.852	0.848	1.83	0.715	0.90	0.97
$\text{AMPF}_e$	2	0.850	0.849	1.82	0.716	0.90	0.97
$\text{AMPF}_p$	3	0.579	0.566	0.63	0.240	0.85	0.93
$\text{AMPF}_g$	4	0.784	0.761	1.40	0.612	0.85	0.94

**Table 5.12:** Summary of meter tracking performance of inference extensions on CMR and HMR<sub>s</sub> datasets. The second column shows the different algorithms. The table shows the tempo estimation performance at CML and AML, beat and sama (downbeat) tracking performance with different measures.

formance also improves to a lesser extent.

For an analysis and comparsion of inference extensions, we pool the results on Indian music datasets CMR and HMR<sub>s</sub> and present it in Table 5.12. We compare the extensions  $\text{AMPF}_e$ ,  $\text{AMPF}_p$ , and  $\text{AMPF}_g$  with the baseline meter tracking with  $\text{AMPF}_0$ . In the table, since  $R = 1$ , note that  $\text{AMPF}_e$  is equivalent to  $\text{AMPF}_0$ . Statistical tests indicate that for all measures,  $\text{AMPF}_0$  and  $\text{AMPF}_e$  are equivalent and show statistically insignificant difference in performance. In addition, for all measures,  $\text{AMPF}_p$  and  $\text{AMPF}_g$  both give significantly lower performance compared to  $\text{AMPF}_0$ . The hop inference extensions need further improvement and do not match up to the performance of inference at every frame.

With that summary of results, we now focus on some more analysis with  $R = 2$  comparing meter tracking performance of  $\text{AMPF}_0$  with  $\text{AMPF}_m$  and  $\text{AMPF}_e$ . Table 5.13 shows a summary of results over all the Indian music datasets for meter tracking with  $\text{AMPF}_0$  and  $\text{AMPF}_m$  and  $R = 2$ . An analysis showed that there is no statistically significant difference in results between  $R = 1$  and  $R = 2$  for either  $\text{AMPF}_0$  or  $\text{AMPF}_m$  (for all measures). However, for  $R = 2$ , the beat tracking measures show an improvement for  $\text{AMPF}_m$  over  $\text{AMPF}_0$ . For the case of  $\text{AMPF}_e$  compared with  $\text{AMPF}_0$  however, there was no statistically significant improvement with more patterns.

Algo.	$f_b$	$AML_{t,b}$	$\mathfrak{I}_b$ Bits	$f_s$	Tempo	
					CML	AML
<b>AMPF<sub>0</sub></b>	0.735	0.751	1.68	0.641	0.77	0.88
<b>AMPF<sub>m</sub></b>	<b>0.749</b>	<b>0.773</b>	<b>1.75</b>	0.660	0.78	0.89

**Table 5.13:** Comparing the meter tracking performance of **AMPF<sub>0</sub>** and **AMPF<sub>m</sub>** algorithms for  $R = 2$  patterns. Numbers in bold for the beat and sama tracking measures indicate a statistically significant improvement.

## 5.5 Conclusions

We defined different meter analysis tasks within the context of Indian art music, pointing out the distinctions between meter inference, meter tracking and informed meter tracking. A set of preliminary experiments on Carnatic music, we explored Bayesian models for jointly tracking several aspects of meter. The state of the art bar pointer model was presented and several model and inference extensions were presented to improve meter analysis.

An extensive evaluation of different meter analysis models and algorithms was provided for different Indian art music datasets, with Ballroom dataset results reported for comparison. Indian art music, with complex metrical structures is an ideal case to study the performance of novel methods for meter analysis and hence such an evaluation is valuable to improve state of the art in meter tracking in MIR. To the best of our knowledge, the work in this chapter is the first collective and comprehensive work on meter analysis in Carnatic and Hindustani music.

The algorithms explicitly considered all the possible musically relevant information for meter analysis, leading to culture specific algorithms. However, the algorithms and models are flexible and can easily adapt to cyclical metrical structures in other music cultures such as Turkish makam music (with *usul* as the cyclical metrical structures) and to Arab-andalusian music. Such Bayesian machine learning models require a small amount of beat and downbeat annotated training data from which we can learn these models and build specific algorithms. Exploring such extensions to different music cultures is one of the goals of future work in the area.

The SP-model shows significant promise in automatic meter analysis. It is a flexible model that can track any cyclical metrical structure by tracking smaller meaningful sub-patterns of the cycle. It provides a significant improvement with Indian music, and it would be fruitful to explore it further to other music cultures. It further goes on to show that tracking shorter length patterns is useful for tracking long duration metrical structures, an intuitive conclusion considering that several additive meters are tracked that way.

The results were mostly reported on the CMR, HMR<sub>s</sub>, and HMR<sub>I</sub> datasets that consist of 2 minute long pieces. However, as reported by Srinivasamurthy et al. (2015) and as seen from additional experiments, these algorithms extend to full length pieces in Carnatic music, showing an equivalent performance. While computational complexity is one factor for meter analysis in full length pieces, there can be several ways in which it can be reduced and the approaches described in the chapter can be applied. An future evaluation on a larger dataset with full length pieces, such as the rhythm annotated pieces in HMD<sub>o</sub> and CMD<sub>o</sub> collections, will further boost such a claim.

The approaches in the chapter utilized bar/section length rhythm patterns for meter tracking. Indian art music is replete with several rhythmic patterns and hence should benefit algorithms that use multiple patterns to model a cycle. However, the experiments did not show such an improvement. There was no statistically significant improvement observed with additional rhythmic patterns ( $R > 1$ ). This can be primarily attributed the simpler GMM based observation model and the spectral flux based feature that fail to capture nuances from mulitple patterns and model them effectively. Better features that can capture nuances and a better observation model need to explored to utilize the variability in patterns we encounter in Indian art music and use them for meter analysis.

Vilāṁbit (slow tempo) pieces in Hindustani music are significant challenge for meter tracking. They are further a challenge for evaluating a meter tracker output. During the rendering of metrical cycles as long as a minute, the mātrās within the cycle are quite flexibly rendered with expressive timing. In addition, given the large inter-mātrā interval, larger errors in tracking are acceptable for listeners. However, the mātrā at the beginning and end of the cycle are more important to keep the time and hence have to be

more accurate. An evaluation measure that treats all the beats of an output as the same is not the best evaluation measure for such a case. The standard evaluation measures considered in the thesis, including the continuity measures  $CML_t$  and  $AML_t$ , cannot handle such cases where there needs to different weights on errors depending on metrical position and tempo. In the evaluation of  $HMR_1$  pieces in this chapter, we used 6.25% of the median inter-*mātrā* interval as the error window for all *mātrā* of the piece. Though it allows for more flexibility in evaluation of long duration metrical cycles, better measures that can consider the metrical position might be more meaningful. Such measures are to be further developed and tested to reflect a more accurate performance of the meter tracking algorithms from a listener's perspective.

Apart from the percussion patterns played on mridangam and tabla, as the case may be, are indicative of the position in cycle, melodic patterns also in many cases indicate the position in cycle. This is further true in composition renderings, which are composed in a specific *tāla*. Melody also can be used to track the progression through a *tāla* with several melodic and lyrical markers indicating the *sama*. Incorporating melodic features into the observation model is hence hypothesized to additionally help to improve meter analysis performance. Recent approaches that use deep learning to build observation models have also seen some success(Böck & Schedl, 2011).

The presented Bayesian models can be further improved to incorporate other structures and priors that can be utilized for improving meter tracking - such as tigher bounds on tempo, tighter restriction on continuity, and allowance for errors such a skipping a beat. This is in addition to the ideas explored already: such as hop inference that aims to track meter only at specific event cues. Such models need to be further explored, with suitable and efficient inference algorithms. The computationally efficient mixture observation model and the inference extensions presented in the chapter show some promise, but need further improvement.

The meter analysis algorithms discussed in the chapter were developed within the context of CompMusic project aligned with its goals. Automatic meter analysis provides valuable content based metadata for a piece of music. Several useful applications of automatic meter analysis exist, a few of them are detailed in Chapter 7.

# Percussion pattern transcription and discovery

The metaphorical usage of ‘language’ for a musical system is paralleled by a literal usage that refers to the ways in which many drum musics may be represented with spoken syllables.

---

Kippen and Bel (1989)

The chapter presents the second problem addressed in this thesis - percussion pattern transcription and discovery. The work presented in the chapter is basic and exploratory, and only for demonstrating the utility of the onomatopoeic mnemonic syllabic percussion system in percussion pattern transcription and discovery. Most experiments presented contain preliminary results, needing further work. The goals of the chapter are:

1. To discuss two main broad approaches to percussion transcription. To focus of timbre based transcription, and to present how meaningful representations can be obtained for overall timbres of percussion strokes in syllabic percussion systems.

2. To present an approach to percussion pattern transcription and discovery in syllabic percussion systems based on a speech recognition framework.
3. To present experiments on *jīngjù* percussion patterns, as a simpler useful example test case of percussion pattern transcription and pattern classification.
4. To extend the approach to Indian art music, and evaluate it on mridangam and tabla solo datasets.
5. To identify the advantages and shortcomings of such an approach, with possible future research directions to pursue.

We first describe two broad approaches that can be taken for the task of percussion pattern transcription.

## 6.1 Approaches

Percussion transcription aims to transcribe an audio recording of a percussion solo into a time aligned sequence of symbols. Depending on symbol set chosen and acoustic events being modeled, transcription can be approached from two different broad ways. The two approaches differ mainly in the way they define, transcribe and discover percussion patterns in audio.

### Transcription based on individual instruments

An audio recording can be transcribed into time aligned sequence of percussion instrument stroke onsets, e.g. transcribing a drum solo into a sequence of bass, hi-hats and snare drum onsets. The output of such a task is a drum score showing all the drums, and their onset times. A percussion pattern hence would be a part of that score, and needs onset information from all instruments. Such an approach is useful for transcription from drum mixtures, when percussion patterns are defined as a sequential combination of different instruments.

Since most percussion solos have simultaneous strokes from multiple instruments, such an approach needs a decomposition of

the audio containing drum mixtures into individual drum components. As a preprocessing step, an instrument-wise onset detection might be needed. As discussed before in Section 2.3.11, event-based systems (Gillet & Richard, 2004b; Gouyon et al., 2002; Goto & Muraoka, 1994; Gillet & Richard, 2008) segment the input signal based on percussion events then extract and classify features from these segments to uncover its musically meaningful content, such as onsets.

Source separation based methods (Paulus & Virtanen, 2005; Smaragdis, 2004a; Abdallah & Plumley, 2003) decompose the input audio signal containing drum mixtures into basis functions capturing spectral characteristics of the sources (ideally, individual percussion instruments). Tian et al. (2014) present an exploratory study on the use of NMF-based source separation techniques for instrument specific onset detection in *jīngjù* percussion instrument ensembles, aiming at providing a baseline for further research in *jīngjù* percussion transcription.

### Transcription based on overall timbre

A contrasting approach is to consider the overall timbre sequence of a percussion solo, without any regard to individual instruments. A percussion pattern is then defined as a sequence of combined instrument timbres and the goal is to transcribe an audio recording into a sequence of such combined timbres. Such an approach is useful when percussion patterns can be defined based on timbral sequences e.g. in syllabic percussion systems. A notable limitation of the approach however is when a pattern cannot be accurately defined by the overall timbre, and individual instrument timbres are necessary, as often is the case with a drum set.

In syllabic percussion systems, patterns can be defined using syllables, which is both musically meaningful and accurate in representation. In certain percussion systems such as in *jīngjù* percussion, the syllables represent the overall combined timbre of a percussion ensemble, while in Indian art music (both Carnatic and Hindustani music), the syllables represent the different timbres that can be produced by the (often) single percussion instrument used, either tabla or mridangam. In either case, the overall timbre represented by the syllable is sufficient to define a percussion pattern.

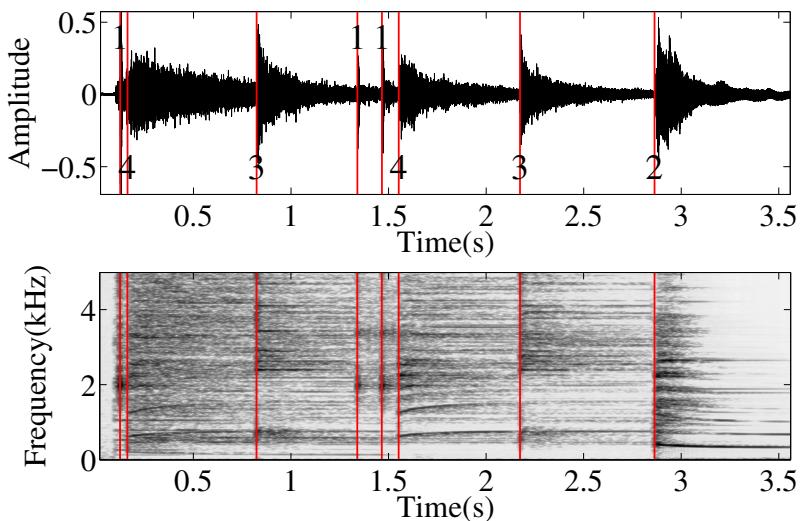
In this chapter, we explore only the second approach, using syllables to define overall timbres of percussion strokes and use them to define, transcribe and discover percussion patterns. In the remainder of the chapter, the goal is to test the effectiveness and relevance of percussion syllables in representation and modeling of percussion patterns for automatic transcription and discovery. Since these syllables have a clear analogy to speech and language, the transcription task has a definite analogy to speech recognition and we can apply several tools and knowledge from this well explored research area with many state of the art algorithms and systems (Huang & Deng, 2010).

The final goal is to automatically discover percussion patterns from segmented percussion solo audio recordings of Indian art music, and we take the transcription+search approach as briefly discussed in Section 3.3.2. The task however is challenging since it requires a concrete definition of a percussion pattern, while a concrete definition of what constitutes a percussion pattern is ambiguous in Indian art music. Further, this ambiguity also leads to possibly a large number of percussion patterns that can be played on the mridangam and tabla.

Hence as an initial test case for our hypothesis and methods, we consider the case of percussion patterns in *jīngjù*, which overcomes both these challenges. Percussion patterns are well defined and limited in number in *jīngjù*. Once we verify our hypothesis, we then extend the method to Indian art music, with a data driven but extendable definition of a percussion pattern. Since there are a limited number of *jīngjù* percussion patterns, the problem of pattern discovery is simplified into a pattern classification task, as is described further in the following section.

## 6.2 The case of Beijing Opera (*Jīngjù*)

To recall from Section 2.2.5, percussion ensemble in *jīngjù* consists of five instruments played by four musicians, and can be grouped based on timbre into four instrument groups - *bǎngǔ* (clapper-drum), *xiǎoluó* (small gong), *dàluó* (big gong) and *náobó* (cymbals). The waveform and spectrogram of an audio example with all four of

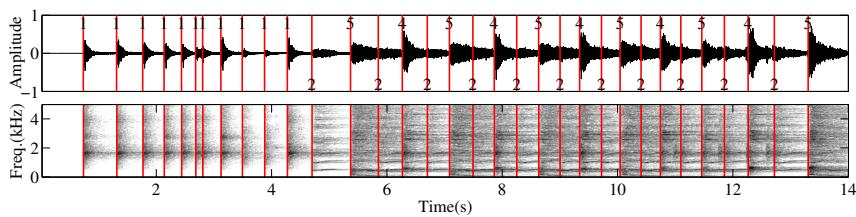


**Figure 6.1:** The waveform and spectrogram of an audio example containing all four instrument groups of *jīngjù*. The top panel shows the waveform and the bottom panel is the spectrogram, the x-axis for both panels is time (in seconds). The vertical lines (in red) mark the onsets of the instruments. The onsets are labeled to indicate the specific instrument onset: *bāngǔ*-1, *dàluó*-2, *náobó*-3, *xiǎoluó*-4.

these instrument classes is shown in Figure 6.1<sup>1</sup>. We can see the amplitude dynamics and spectral shapes for each instrument. *Dàluó* has a falling pitch, while *xiǎoluó* has a rising pitch profile. *Náobó* has a broadband spectrum with significant energy in higher frequencies, as is characteristic of cymbals. Onsets generated by *bāngǔ* are sharp, have much lower amplitude and shorter transient time and happen in higher density than those generated by the cymbal instruments, and hence the *bāngǔ* onsets are easily masked by the cymbals and gongs. We can also see how the bang stroke is masked by an adjoining *xiaoluó* stroke (0-0.5s in Figure 6.1).

Using combinations of these instruments, several different combined percussion strokes are produced, each of which is labeled through an onomatopoeic oral mnemonic syllable. Since there are

<sup>1</sup>The audio example is available here: <http://www.freesound.org/people/ajaysm/sounds/205971/>



**Figure 6.2:** The waveform and spectrogram of an audio example of the pattern shǎnchuí. The top panel shows the waveform and the bottom panel is the spectrogram. The vertical lines (in red) mark the onsets of the syllables. The onsets are labeled to indicate the specific syllable group: DA-1, TAI-2, QI-3, QIE-4, and CANG-5 (QI is not present in this pattern). The score for the pattern is shown in Figure 2.5e. Notice that the audio example has two additional repetitions of the sub-sequence CANG-TAI-QIE-TAI in the pattern.

many syllables that map to a single timbre, we reduced the complete set of syllables into five syllable groups - DA, TAI, QI, QIE, and CANG, as listed in Table 2.6. The use of these oral syllabic sequences simplify and unify the representation of these patterns played by an ensemble.

To further recall Section 2.2.5, the percussion patterns in *jīngjù* music are sequences of strokes played by different combinations of the percussion instruments, and the resulting variety of timbres are transmitted using oral syllables as mnemonics. Each percussion pattern is a sequence of syllables in their pre-established order, along with their specific rhythmic structure and dynamic features. Each particular pattern has a single unique syllabic representation shared by all the performers. Hence, the use of these oral syllabic sequences simplify and unify the representation of these patterns played by an ensemble, making them optimal for the transcription and automatic classification of the patterns.

A performance starts and ends with percussion patterns, they generally introduce and conclude arias, and mark transition points within them. The patterns accompany the actors' movements on stage and set the mood of the play, the scene, the aria or a section of the aria. An automatic description of these percussion patterns is

thus quite important in providing the overall description of the aria. Therefore, the detection and characterization of percussion patterns is a fundamental task for the description of the music dimension in Beijing opera. In practice, there is a limited set of named patterns that are played in performance.

Though the patterns are limited in number and predefined, there are several challenges to the problem of percussion pattern transcription and classification. Being an oral tradition, the syllables used for the representation of the patterns lacks full consistency and general agreement. The result being that one particular timbre might be represented by more than one syllable. Furthermore, the syllabic representation conveys information for the conjoint timbre of the ensemble, so only the main structural sounds are represented. In an actual performance, a particular syllable might be performed by different combinations of instruments - e.g. in Figure 2.5e, the first occurrence of the syllable TAI is played just by the xiǎoluó, but in the rest of the pattern is played by xiǎoluó and the bǎngǔ together. In fact, generally speaking, the strokes of the bǎngǔ are seldom conveyed in the syllabic sequence (as can be seen in the third measure in Figure 2.5e for the second sixteenth-note of the bǎngǔ), except for the introductions and other structural points played by the drum alone. As indicated in Table 2.6, CANG is mostly a combination of all the three metallophones, but in some cases, CANG can be played with just the dàluó, or just the dàluó+náobó combination.

In the cases where the percussion pattern is to accompany the movements of actors on stage, certain syllable sub-sequences in the pattern are repeated indefinitely. This causes the same pattern in different performances to have variable lengths, and these repetitions need to be explicitly handled. The timing of these patterns is expressive and matches the acting in the scene, and hence we consider only the sequence of syllables and do not consider timing relationships between the syllables to define patterns. Finally, although the patterns are usually played in isolation, in many cases the string instruments or even the vocals can start playing before the patterns end, presenting challenges in identification and classification. Figure 6.2 shows an audio example of the pattern shānchuí, along with time aligned markers to indicate the syllable onsets. The spectrogram also shows the timbral characteristics of the percus-

sion instruments *xiaoluo* (increasing pitch) and *daluo* (decreasing pitch). Some variation to the notated score can also be seen, such as expressive timing and additional insertion of syllables.

At the outset, it is clear that *jīngjù* percussion patterns are well defined and limited in number. Further, in Beijing opera, the recognition of the pattern as a whole is more important than an accurate syllabic transcription of the pattern. Due to the limited set of pattern classes and owing to all the variations possible in a pattern, we are primarily interested in classifying an audio pattern into one of the possible pattern classes. Syllabic transcription is only considered as an intermediate step towards pattern classification. The named patterns can be used to build a library of patterns. The patterns can be referred to as “pattern classes” for the purpose of classification, and classifying an instance of a pattern occurring in the audio recording of an aria into one of these pattern classes is thus a primary task.

### 6.2.1 Percussion pattern classification

Beijing Opera percussion pattern transcription and classification is a first test case for percussion pattern discovery approaches. In this work, we restrict to five predominantly used percussion patterns in *jīngjù* - *dǎobǎn tóu*, *màn chángchuí*, *duótóu*, *xiǎoluó duótóu*, and *shǎnchuí* (pattern scores provided in Figure 2.5). Further, we restrict ourselves to percussion patterns that occur at the introduction of the aria, since they convey significant information about the structure of the aria that follows it, e.g. *dǎobǎn tóu* pattern is always followed by an aria in *banshi dǎobǎn*. **Check with Rafa, XX**

We now present a formulation for transcription and recognition of syllable based audio percussion patterns, and evaluate it on the JPP dataset (see Section 4.2.6). The dataset is a collection of 133 audio percussion patterns spanning five different pattern classes, with over 2200 syllables in total. There is a significant analogy of this task to connected word speech recognition using word models. Syllables are analogous to words and a percussion pattern to a sentence - a sequence of words. There are language rules to form a sentence using a vocabulary, just as each percussion pattern is formed with a defined sequence of syllables from a vocabulary. However unlike in the case of speech recognition where infinitely

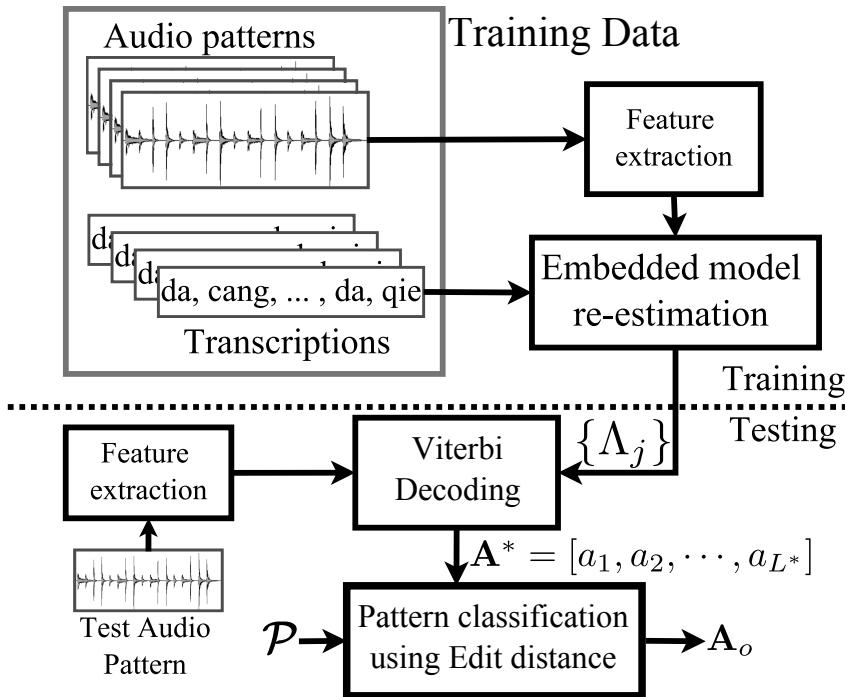
many sentences are possible, in our case we have a small number of percussion patterns to be recognized.

Similar to the work by Nakano et al. (2004), we explore a speech recognition based framework in this study. This approach is different to ours in the sense that these onomatopoeic representations were created by the authors, while we are relying on already existing oral traditions. To the best of our knowledge, Srinivasamurthy, Caro, et al. (2014) presented the first work to explore transcription and classification of syllable based percussion patterns, as applied to Beijing opera. The method and results presented in this section are from that work.

Following the notation presented in Section 3.3.2, consider a set of  $N_a$  pattern classes  $\mathcal{P} = \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_{N_a}\}$ , each of which is a sequence of syllables from the set of  $N_s$  syllables  $\mathcal{A} = \{A_1, A_2, \dots, A_{N_s}\}$ . Hence,  $\mathbf{A}_j = [a_1, a_2, \dots, a_{L_j}]$  where  $a_i \in \mathcal{A}$  and  $L_j$  is the length of  $\mathbf{A}_j$ . Given a test audio pattern  $f[n]$ , the transcription task aims to obtain a syllable sequence  $\mathbf{A}^* = [a_1, a_2, \dots, a_{L^*}]$  and the classification task aims to assign  $\mathbf{A}^*$  into one of the patterns in the set  $\mathcal{P}$ .

The syllables are non-stationary signals and to model their timbral dynamics, we build an HMM for each syllable (analogous to a word-HMM). Using these syllable HMMs and a language model, an input audio pattern is transcribed into a sequence of syllables using Viterbi decoding, and then classified to a pattern class in the library using a measure of distance.

A block diagram of the approach is shown in Figure 6.3. We first build syllable level HMMs  $\{\Lambda_i\}$ ,  $1 \leq i \leq N_s (= 5)$ , for each syllable  $a_i$  using features extracted from the training audio patterns. We use the MFCC features to model the timbre of the syllables. To capture the temporal dynamics of syllables, we add the velocity and the acceleration coefficients of the MFCC. The stereo audio is converted to mono, since there is no additional information in stereo channels. The 13 dimensional (including the 0<sup>th</sup> coefficient) MFCC features are computed from audio patterns with a frame size of 23.2 ms and a shift of 5.8 ms. We also explore the use of energy (as measured by the 0<sup>th</sup> MFCC coefficient) in classification performance. Hence we have two sets of features, **MFCC\_0\_D\_A**, the 39 dimensional feature including the 0<sup>th</sup>, delta (velocity) and double-delta (acceleration) coefficients, and **MFCC\_D\_A**, the 36 dimensional



**Figure 6.3:** The block diagram jīngjù percussion pattern classification approach

vector without the 0<sup>th</sup> coefficient.

We model each syllable using a 5-state left-to-right HMM including an entry and an exit non-emitting states. The emission densities for each state is modeled with a four component GMM to capture the timbral variability in syllables. We experimented with eight and sixteen component GMM, but with little performance improvement. Since we do not have time aligned transcriptions in the Jingju percussion pattern dataset (JPP) dataset, an isolated HMM training for each syllable is not possible. Hence we use an embedded model Baum-Welch re-estimation to train the HMMs using just the syllable sequence corresponding to each feature sequence. The HMMs are initialized with a flat start using all of the training data. All the experiments were done using the HMM Toolkit (HTK) (Young et al., 2006).

Given a test audio pattern, using these syllable HMMs and a basic language model, we obtain a rough syllabic transcription of the test pattern. We then classify the test pattern into one of the pat-

tern classes in the library based on a measure of distance between the test pattern and the pattern classes. For testing, since we only need a rough syllabic transcription independent of the pattern class, we treat the test pattern as a first order time-homogenous discrete Markov chain, which can consist of any finite length sequence of syllables, with uniform unigram and bigram (transition) probabilities, i.e.  $P(a_1 = A_i) = 1/N_s$  and  $P(a_{k+1} = A_j | a_k = A_i) = 1/N_s$ ,  $1 \leq i, j \leq N_s$ , with  $k$  being the sequence index. This also forms the language model for forming the percussion patterns using syllables. Given the feature sequence extracted from test audio pattern, we use the HMMs  $\{\Lambda_i\}$  to do a Viterbi (forced) alignment, which aims to provide the best sequence of syllables  $\mathbf{A}^*$ , given a syllable network constructed from the language model.

Given the decoded syllable sequence  $\mathbf{A}^*$ , we compute the string edit distance (Navarro, 2001) between  $\mathbf{A}^*$  and elements in the set  $\mathcal{P}$ . The use of edit distance is motivated by two factors. First, due to errors in Viterbi alignment,  $\mathbf{A}^*$  can have insertions (I), deletions (D), substitutions (S), and transposition (T) of syllables compared to the ground truth. Secondly, to handle the allowed variations in patterns, an edit distance is preferred over an exact match to the sequences in  $\mathcal{P}$ . We explore the use of two different string edit distance measures, Levenshtein distance ( $d_1$ ) that considers I, D, S errors and the Damerau–Levenshtein distance ( $d_2$ ) that considers I, D, S, and also T errors.

As discussed earlier, there can be repetitions of a sub-sequence in some patterns. Though the number of repetitions is indefinite, we observed in the dataset that there are at most two repetitions in a majority of pattern instances. Hence for the pattern classes that allow repetition of a sub-sequence, we compute the edit distance for the cases of zero, one and two repetitions and then take the minimum distance obtained among the three cases. This way, we can handle repeated parts in a pattern. Finally, the  $\mathbf{A}^*$  is assigned to the pattern class  $\mathbf{A}_o \in \mathcal{P}$  for which the edit distance  $d$  (either  $d_1$  or  $d_2$ ) is minimum, as in Eq. 6.1.

$$\mathbf{A}_o = \operatorname{argmin}_{1 \leq j \leq N_a} d(\mathbf{A}^*, \mathbf{A}_j) \quad (6.1)$$

Feature	Syllable		Pattern	
	$\mathfrak{C}$	$\mathfrak{A}$	$d_1$	$d_2$
MFCC_D_A	78.14	26.32	93.23	89.47
MFCC_0_D_A	84.98	39.63	91.73	89.47

**Table 6.1:** Syllable transcription and pattern classification performance on JPP dataset, with Correctness ( $\mathfrak{C}$ ) and Accuracy ( $\mathfrak{A}$ ) measures for syllable transcription. Pattern classification results are shown for both distance measures  $d_1$  and  $d_2$ . All values are in percentage.

### 6.2.2 Results and discussion

We present the syllable transcription and pattern classification results on the JPP dataset described in Section 4.2.6. The results shown in Table 6.1 are the mean values in a leave-one-out cross validation. We report the syllable transcription performance using the measures of Correctness ( $\mathfrak{C}$ ) and Accuracy ( $\mathfrak{A}$ ). If  $L$  is the length of the ground truth sequence, then the two measures are defined as,

$$\mathfrak{C} = \frac{L - D - S}{L} \quad (6.2)$$

$$\mathfrak{A} = \frac{L - D - S - I}{L} \quad (6.3)$$

The Correctness measure penalizes deletions and substitutions, while Accuracy measure additionally penalizes insertions too. The pattern classification performance is shown for both edit distance measures  $d_1$  and  $d_2$  in Table 6.1. All the results are reported for both the features, MFCC\_0\_D\_A and MFCC\_D\_A. The difference in performance between the two features was found to be statistically significant for both Correctness and Accuracy measures in a Mann-Whitney U test at  $p = 0.05$ , assuming an asymptotic normal distribution (Mann & Whitney, 1947).

In general, we see a good pattern classification performance while syllable transcription accuracy is poor. We see that MFCC\_0\_D\_A has a better performance with syllable transcription, while both kinds of features provide a comparable performance for pattern classification. Though syllable transcription is not the primary task we focus on, an analysis of its performance provides several in-

sights. The set of percussion instruments in Beijing Opera is fixed, but there can be slight variations across different instruments of the same kind. The training examples are varied and representative, and models built can be presumed to be source independent. Nevertheless, there can be unrepresented syllable timbres in test data leading to a poorer transcription performance. A bigger training dataset can improve the performance in such a case. The energy co-efficient provides significant information about the kind of syllables and hence gives a better syllable transcription performance.

We see that the Correctness is higher than Accuracy showing that the exact sequence of syllables, as indicated in the score was never achieved in a majority of the cases, with several insertion errors. This is due to the combined effect of errors in decoding and allowed variations in patterns. An edit distance based distance measure for classification is quite robust in the present five class problem and provides a good classification performance, despite the low transcription accuracy. Both distance measures provide comparable performance, indicating that the number of transposition errors are low. To see if there are any systematic classification errors, we build a confusion matrix (Table 6.2) with one of the well performing configurations: **MFCC\_0\_D\_A** with  $d_1$  distance. We see that **duótóu** (ID = 3) has a low recall, and gets confused with **shānchuí** (ID = 5) often. A close examination of the scores showed that a part of the pattern *duotou* is contained within *shanchui*, which explains source of confusion. Such confusions can be handled with better language models, which need further exploration.

## Conclusions and Summary

We presented a formulation based on connected-word speech recognition for transcription and classification of syllabic percussion patterns on Beijing Opera, as a initial study case. On a representative collection of Beijing opera percussion patterns, the presented approach provides a good classification performance, despite a simplistic language model and inadequate syllabic transcription accuracy. The approach is promising, however, the evaluation using a small dataset necessitates a further assessment of the generalization capabilities. Better language models can be explored, that use sequence and rhythmic information more effectively, and the task can

Pattern class	Total	ID	1	2	3	4	5
dǎobǎn tóu	62	1	100				
màn chángchuí	33	2		93.9			6.1
duótóu	19	3	10.5		68.4		21.1
xiǎoluó duótóu	11	4			18.2	81.8	
shānchuí	8	5		12.5			87.5

**Table 6.2:** The confusion matrix for pattern classification in JPP dataset, using the feature MFCC\_0\_D\_A with  $d_1$  distance measure. The first and second column show the pattern class label and the total examples in each class, and class labels correspond to the ID in Table 4.19. The rows and column headers represent the True Class and Assigned Class, respectively. All other values are in percentage and the empty blocks are zeros (omitted for clarity).

be extended to a much larger dataset spanning more pattern classes. We used isolated patterns in this study, assuming segmented audio patterns. But an automatic segmentation of patterns from audio is a good direction for future work.

Given the effectiveness of the approach using syllables for percussion pattern representation, we can now do similar formulation for Indian art music - for both tabla and mridangam solo recordings. The percussion system in Indian art music is more complex than *jīngjù*, with a larger variety of syllables and ill defined larger number of patterns, while it still has a syllabic percussion system.

### 6.3 The case of Indian art music

The syllabic percussion system in both Carnatic and Hindustani music provides a musically relevant representation system for percussion patterns. In the remainder of the chapter, we describe an approach for percussion pattern discovery from audio recordings of percussion solos. We define percussion patterns using a reduced set of syllable groups (instead of the inconvenient term of syllable group, we call the syllable groups as just syllables) using the mapping described for Carnatic music in Table 2.2 and for Hindustani music Table 2.4. To address the problem of percussion pat-

tern discovery in Indian art music percussion solo recordings, we follow a data driven transcription + search approach. The approach mainly has three components:

**Pattern library generation:** Creating a library of characteristic percussion patterns (query patterns) from a corpus of syllabic percussion pattern scores of solos.

**Automatic transcription:** Transcribe a given percussion solo audio recording into a time aligned sequence of syllables using syllable timbre models.

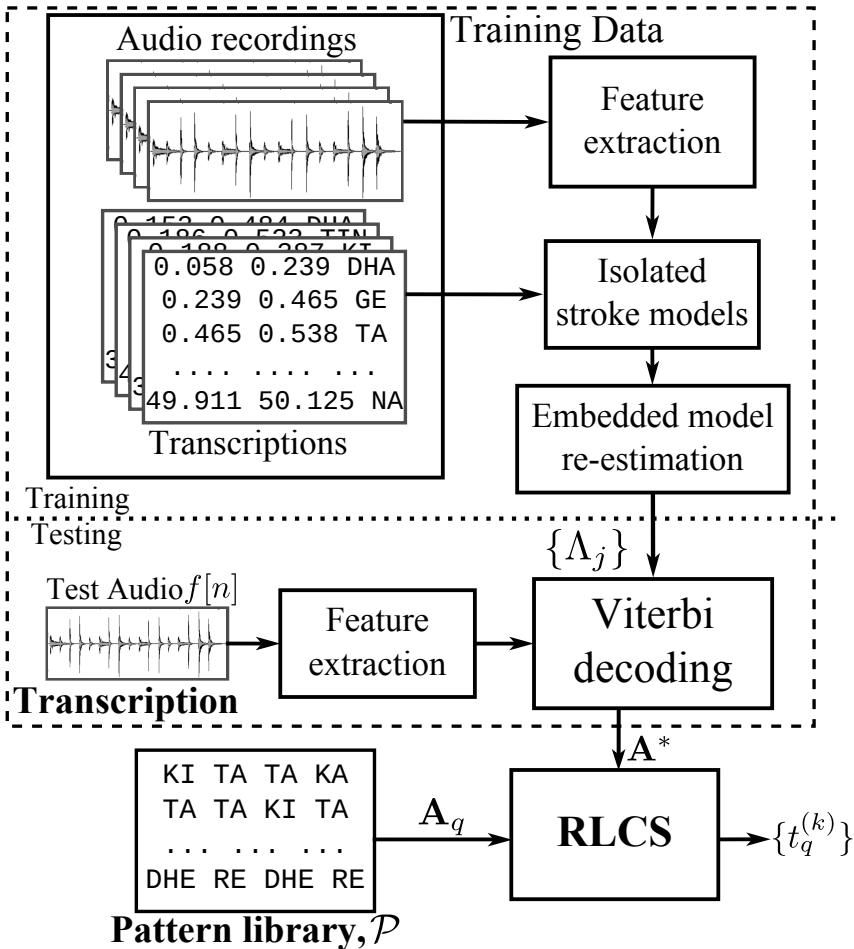
**Approximate pattern search:** Search for the query patterns in the transcribed output syllable sequence using approximate string search algorithms.

We describe each of these components in greater detail. Despite involving a search for a known query pattern in transcribed scores, since the query patterns are also discovered automatically from a collection of scores, this method is different from a supervised pattern search. The approach addresses a discovery problem that can automatically find audio percussion patterns from a corpus of percussion solo audio recordings.

The framework is similar for both Hindustani and Carnatic music, and hence we describe the approach together for both tabla and mridangam solos together. The approach is evaluated on a collection of tabla solo recordings (the [MTS](#) dataset) and mridangam solo recordings (the [UMS](#) dataset). A block diagram of the approach is shown in Figure 6.4. Some of the work described in this section has been discussed by [Gupta et al. \(2015\)](#) and [Gupta \(2015\)](#).

### 6.3.1 Pattern library generation

Percussion patterns are built hierarchically in Indian art music, with smaller standard phrases used to build longer sequences of percussion patterns. With a limited set of syllables, there are smaller patterns that are played very often, which are grouped in different combinations to create larger patterns. A library of such patterns can hence be obtained from music scores of percussion solos - we use the accompanying scores in [MTS](#) dataset and [UMS](#) dataset for tabla and mridangam, respectively to build such pattern libraries.



**Figure 6.4:** A block diagram of percussion pattern discovery approach in Indian art music. The figure considers the example of tabla solos for illustration.

In tabla solos, despite the differences across gharanas, there are also many similarities due to the fact that the same forms and standard phrases reappear across these repertoires (Gottlieb, 1993, p. 52). This enables in creation of a library of standard phrases or patterns across compositions of different *gharānās* present in MTS dataset.

From Section 3.3.2, we recall the use of a simplistic definition of a pattern as a sequence of syllables, without considering the relative and absolute durations of the constituent syllables, as well as the metrical position of the pattern in the *tāla*. In this dissertation,

we take a data driven approach to build a set of  $N_a$  query patterns,  $\mathcal{P} = \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_{N_a}\}$ . In addition, we assume that the most often played patterns are the most characteristic. Without any prior knowledge, such an assumption enables us to create a library of valid set of patterns with an objective criterion and further allows for a better evaluation since there are several examples of those patterns in the test datasets. It is however to be noted that discovery approaches need not make this assumption, and any other criteria for automatic discovery of patterns can be used.

Using the simple definition of a pattern as a sequence of syllables, we use the scores of the compositions in the **MTS** dataset (for tabla) and **UMS** dataset (for mridangam) to generate all the  $L$  length patterns that occur in the score collection. We sort them by their frequency of occurrence to get an ordered set of patterns for each stated length. We then manually choose musically representative patterns from this ordered set of most commonly occurring patterns to form a set of query patterns  $\mathcal{P}$ . We create a set of query patterns of length  $L = 4, 6, 8, 16$  (Beronja, 2008, p. 126). These lengths were chosen based on the structure of **tālas** in the score collections (**ādi** and **rūpaka tāla** in mridangam solo dataset and **tīntāl** in tabla solo dataset).

Table 6.3 shows the query tabla patterns used in this work obtained from the **MTS** dataset. The table also shows their length and their count in the dataset, leading to a total of 1425 instances. We want a diverse collection of patterns to test if the algorithms generalize. Hence we choose patterns that have a varied set of syllables with different timbral characteristics, like syllables that are harmonic (DHA), syllables played with a flam (DHE, RE) and syllables having bass (GE).

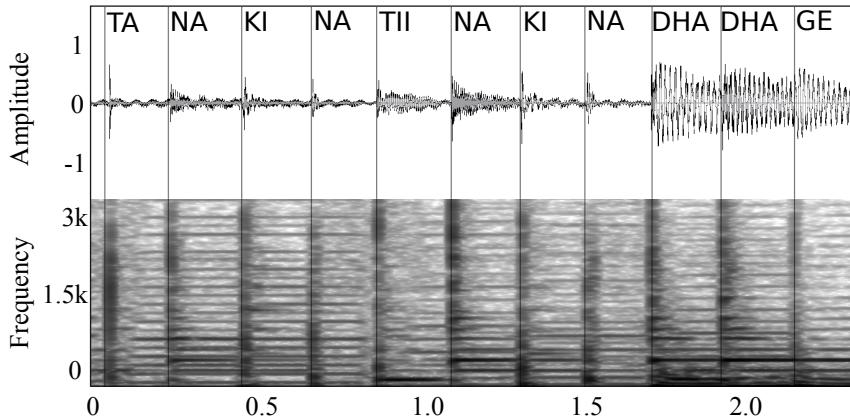
Table 6.4 shows the query mridangam patterns used in this work obtained from the **UMS** dataset. The table also shows their length and their count in the dataset, leading to a total of 976 instances. As confirmed by a carnatic percussionist, these patterns are very commonly played in practice and hence are a good set of candidates to evaluate pattern discovery methodologies.

ID	Pattern	<i>L</i>	Count
A <sub>1</sub>	DHE, RE, DHE, RE, KI, TA, TA, KI, NA, TA, TA, KI, TA, TA, KI, NA	16	47
A <sub>2</sub>	TA, TA, KI, TA, TA, KI, TA, TA, KI, TA, TA, KI, TA, TA, KI, TA	16	10
A <sub>3</sub>	TA, KI, TA, TA, KI, TA, TA, KI	8	61
A <sub>4</sub>	TA, TA, KI, TA, TA, KI	6	214
A <sub>5</sub>	TA, TA, KI, TA	4	379
A <sub>6</sub>	KI, TA, TA, KI	4	450
A <sub>7</sub>	TA, TA, KI, NA	4	167
A <sub>8</sub>	DHA, GE, TA, TA	4	97

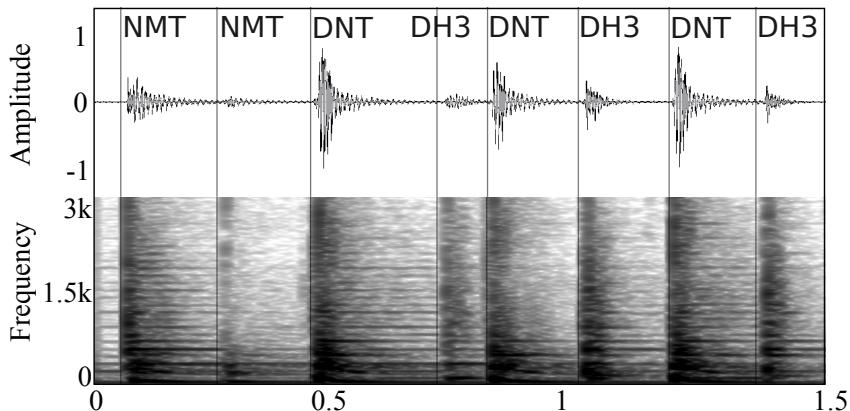
**Table 6.3:** Query tabla percussion patterns, their ID (*j*), length (*L*) and the number of instances in the **MTS** dataset (Total instances: 1425).

ID	Pattern	<i>L</i>	Count
A <sub>1</sub>	DH3, TA, DH3, TA, TH, DH3, TH, TA	8	70
A <sub>2</sub>	TA, DH3, TA, TH, DH3, TH, TA, TM	8	69
A <sub>3</sub>	DH3, TA, DH3, TA, TH, DH3	6	89
A <sub>4</sub>	DH3, TA, TH, DH3, TH, TA	6	70
A <sub>5</sub>	TA, TH, DH3, TH, TA, TM	6	69
A <sub>6</sub>	DH3, TA, TH, DH3	4	291
A <sub>7</sub>	DH3, TA, DH3, TA	4	114
A <sub>8</sub>	TH, DH3, TA, TH	4	102
A <sub>9</sub>	TA, TH, DH3, TH	4	102

**Table 6.4:** Query mridangam percussion patterns, their ID (*j*), length (*L*) and the number of instances in the **UMS** dataset (Total instances: 976).



**Figure 6.5:** The waveform and spectrogram of an audio example of a tabla percussion pattern shown with the onsets and the mapped syllable names from Table 4.15.



**Figure 6.6:** The waveform and spectrogram of an audio example of a mridangam percussion pattern shown with the onsets and the mapped syllable names from Table 4.17.

### 6.3.2 Automatic transcription

Given an audio recording of percussion solo, automatic transcription refers to transcribing a given percussion solo audio recording into a time aligned sequence of syllables using syllable timbre models. An audio example of a percussion pattern is shown in Figure 6.5 for tabla, and in Figure 6.6 for mridangam. In the figures, we can see the pitched nature of some of the strokes, with clear onsets in many cases, but an overlap between adjacent strokes of the pattern.

This needs a modeling of timbre, along with modeling of sequential information in syllables.

Some *bōls* of tabla may be pronounced with a different vowel or consonant depending on the context, without altering the drum stroke (Chandola, 1988). Furthermore, the *bōls* and the strokes vary across different *gharānās*, making the task of transcription of tabla solos challenging. Mridangam syllables are further less specific as discussed earlier, and using the timbral grouping aims to address this challenge. To model the timbral dynamics of syllables, we build an **HMM** for each syllable (analogous to a word-**HMM**). We use these **HMMs** along with a language model to transcribe an input audio solo recording into a sequence of syllables.

The stereo audio is converted to mono, since there is no additional information in stereo channels. We use the **MFCC** features to model the timbre of the syllables. To capture the temporal dynamics of syllables, we add the velocity and the acceleration coefficients of the **MFCC**. The 13 dimensional **MFCC** features (including the zeroth coefficient) are computed from the audio with a frame size of 23.2 ms and a shift of 5.8 ms. We also explore the use of energy (as measured by the zeroth **MFCC** coefficient) in transcription performance. Hence we have two sets of features, **MFCC** features with energy, delta and double-delta coefficients (**MFCC\_0\_D\_A**), the 39 dimensional feature including the zeroth, delta and double-delta coefficients, and **MFCC** features without energy but with delta and double-delta coefficients (**MFCC\_D\_A**), the 36 dimensional vector without the zeroth coefficient.

Using the features extracted from training audio recordings, we model each syllable  $A_j$  using a 7-state left-to-right **HMM**  $\{\Lambda_j\}$ ,  $1 \leq j \leq N_s$ , including an entry and an exit non-emitting states. For tabla solo transcription,  $N_s = 18$  while for mridangam solo transcription task,  $N_s = 21$ . The emission density of each emitting state is modeled with a three component Gaussian Mixture Model (GMM) to capture the timbral variability in syllables. We experimented with higher number of components in the **GMMs**, but with little performance improvement.

The **UMS** mridangam solo dataset lacks such time aligned transcriptions and hence all syllables are initialized with a flat start **HMM** using all the data in the dataset. The **MTS** tabla solo dataset is a parallel corpus of audio and time aligned syllabic transcrip-

tions, each syllable HMM is initialized through an isolated HMM training of each syllable. Additionally for comparison, we report results with a flat start on MTS tabla dataset too. The initialized HMMs are then trained further in an embedded model Baum-Welch re-estimation to get the final syllable HMM.

Percussion solos in Indian art music are built hierarchically using short phrases, and hence some bōls/solkattus tend to follow a bōl/solkattu more often than others. In such a scenario, a language model can improve transcription. In addition to a flat language model with uniform unigram and transition probabilities, i.e.  $P(a_1 = A_j) = 1/N_s$  and  $P(a_{k+1} = A_j | a_k = A_i) = 1/N_s$ , with  $1 \leq i, j \leq N_s$  and  $k$  being the sequence index, we explore the use of a bigram language model learned from data. The bigram language model is learned from all the scores in the training data.

For testing, we treat the feature sequence extracted from test audio file to have been generated from a first order time-homogeneous discrete Markov chain, which can consist of any finite length sequence of syllables. From the extracted feature sequence, we use the HMMs  $\{\Lambda_j\}$  and a syllable network constructed from the language model to do a Viterbi (forced) alignment, which aims to provide the best sequence of syllables and their onset timestamps, given as  $\mathbf{A}^* = [(t_1, a_1), (t_2, a_2), \dots, (t_{L^*}, a_{L^*})]$ , where  $t_i$  is the onset time of  $a_i$  and  $L^*$  is the length of the transcribed sequence. All the transcription experiments were done using the HMM Toolkit (HTK) (Young et al., 2006).

### 6.3.3 Approximate pattern search

The automatically transcribed output syllable sequence  $\mathbf{A}^*$  is used to search for the query patterns. Transcription is often inaccurate in both the sequence of syllables and in the exact onset times of the transcribed syllables. We need to handle both these errors in a pattern search task from audio. We primarily focus on the errors in syllabic transcription in this work. We use the syllable boundaries output by the Viterbi algorithm, without any additional post processing. We can improve the output syllable boundaries using an onset detector (Bello et al., 2005), but we leave this task to future work.

Searching of a query syllable sequence in a transcribed sequence of syllables is akin to string search. As discussed in the case of *jīngjù* percussion pattern transcription task, errors in transcription are mainly insertions (I), deletions (D), substitutions (S), and transpositions (T). Further, the query pattern is to be searched in the whole transcribed composition, where several instances of the query can occur. With both these issues, the problem of pattern search can be addressed as a subsequence search. **RLCS** method is a suitable choice for such a case. **RLCS** is a subsequence search method that searches for roughly matched subsequences while retaining the local similarity (Lin et al., 2011). We make further enhancements to **RLCS** to handle the I, D and S errors in transcription.

We use a modified version of the **RLCS** approach as proposed by Lin et al. (2011) with changes proposed by S. Dutta and Murthy (2014) to handle substitution errors. We propose a further enhancement to handle insertions and deletions, and explore its use in the current task. S. Dutta and Murthy (2014) used a modified version of **RLCS** for motif spotting in *ālāpanas* of Carnatic music. We propose to use a similar approach with minor modifications to suit the symbolic domain specific to our use case. We first present a general form of **RLCS** and then discuss different variants of the algorithm.

Given a query pattern  $\mathbf{A}_q \in \mathcal{P}$  of length  $L_q$  and a reference sequence (transcribed syllable sequence)  $\mathbf{A}^*$  of length  $L_*$ , **RLCS** uses a dynamic programming approach to compute a score matrix (of size  $L_* \times L_q$ ) between the reference and the query with a rough length of match. We can use a threshold on the score matrix to obtain the instances of the query occurring in the reference. We can then use the syllable boundaries in the output transcription and retrieve the audio segment corresponding to the match.

For the ease of notation, we index the transcribed syllable sequence  $\mathbf{A}^*$  with  $i$  and the query syllable sequence  $\mathbf{A}_q$  with  $j$ . We compute the rough and actual length of the subsequence matches similar to the way computed by S. Dutta and Murthy (2014). At every position  $(i, j)$ , a syllable is included into the matched subsequence if  $d(a_i, a_j) \leq \delta$ , where  $d(a_i, a_j)$  is the timbral distance between the syllables at positions  $i$  and  $j$  in the transcription and query, respectively.  $\delta$  is the threshold distance below which the two syllables are said to be equivalent. The matrices of rough length of

match ( $\mathbb{H}$ ) and the actual length of match ( $\mathbb{H}^a$ ) are updated as,

$$\mathbb{H}(i, j) = \mathbb{H}(i - 1, j - 1) + (1 - d(a_i, a_j)) \cdot \mathbb{1}_d \quad (6.4)$$

$$\mathbb{H}^a(i, j) = \mathbb{H}^a(i - 1, j - 1) + \mathbb{1}_d \quad (6.5)$$

where,  $\mathbb{1}_d$  is an indicator function that takes a value of 1 if  $d(a_i, a_j) \leq \delta$ , else 0. The matrix  $\mathbb{H}$  thus contains the length of rough matches ending at all combinations of the syllable positions in reference and the query. The rough length and an appropriate distance measure handles the substitution errors during transcription.

To penalize insertion and deletion errors, we compute a “density” of match using two measures called the Width Across Reference (WAR) and Width Across Query (WAQ), respectively. The WAR ( $\mathbb{R}$ ) and WAQ ( $\mathbb{Q}$ ) matrices are initialized to  $\mathbb{R}_{i,j} = \mathbb{Q}_{i,j} = 0$  when  $i, j = 0$ , and propagated as,

$$\mathbb{R}_{i,j} = \begin{cases} \mathbb{R}_{i-1,j-1} + 1 & d(a_i, a_j) \leq \delta \\ \mathbb{R}_{i-1,j} + 1 & d(a_i, a_j) > \delta, \mathbb{H}_{i-1,j} \geq \mathbb{H}_{i,j-1} \\ \mathbb{R}_{i,j-1} & d(a_i, a_j) > \delta, \mathbb{H}_{i-1,j} < \mathbb{H}_{i,j-1} \end{cases} \quad (6.6)$$

$$\mathbb{Q}_{i,j} = \begin{cases} \mathbb{Q}_{i-1,j-1} + 1 & d(a_i, a_j) \leq \delta \\ \mathbb{Q}_{i-1,j} & d(a_i, a_j) > \delta, \mathbb{H}_{i-1,j} \geq \mathbb{H}_{i,j-1} \\ \mathbb{Q}_{i,j-1} + 1 & d(a_i, a_j) > \delta, \mathbb{H}_{i-1,j} < \mathbb{H}_{i,j-1} \end{cases} \quad (6.7)$$

Here,  $\mathbb{R}_{i,j}$  is the length of sub-string containing the subsequence match ending at the  $i^{\text{th}}$  and the  $j^{\text{th}}$  position of the reference and the query, respectively.  $\mathbb{Q}_{i,j}$  represents a similar measure in the query. When incremented,  $\mathbb{R}_{i,j}$  and  $\mathbb{Q}_{i,j}$  are incremented by 1 similar to the way formulated by Lin et al. (2011). At the same time, the increment is done based on the conditions formulated by S. Dutta and Murthy (2014).

Using the rough length of match ( $\mathbb{H}$ ), actual length of match ( $\mathbb{H}^a$ ), and width measures ( $\mathbb{R}$  and  $\mathbb{Q}$ ), we compute a score matrix  $\sigma$  that incorporates penalties for substitutions, insertions, deletions, and additionally, the fraction of the query matched as,

$$\sigma_{i,j} = \begin{cases} \left[ \eta \cdot f\left(\frac{\mathbb{H}_{i,j}}{\mathbb{R}_{i,j}}\right) + (1 - \eta) \cdot f\left(\frac{\mathbb{H}_{i,j}}{\mathbb{Q}_{i,j}}\right) \right] \cdot \frac{\mathbb{H}_{i,j}}{L_q} & \text{if } \frac{\mathbb{H}_{i,j}^a}{L_q} \geq \rho \\ 0 & \text{otherwise} \end{cases} \quad (6.8)$$

where  $\sigma_{i,j}$  is the score for the match ending at the  $i^{\text{th}}$  and the  $j^{\text{th}}$  position of the reference and the query, respectively.  $f$  is a warping function for the rough match length densities  $\frac{\mathbb{H}_{i,j}}{\mathbb{R}_{i,j}}$  in the reference and  $\frac{\mathbb{H}_{i,j}}{\mathbb{Q}_{i,j}}$  in the query. The parameter  $\eta$  controls their weights in the convex combination for score computation. The term  $\frac{\mathbb{H}_{i,j}^a}{L_q}$  is the fraction of the query length matched and is used for thresholding the minimum fraction of the query to be matched. The parameter  $\rho$  is the threshold for the minimum fraction that contributes to the score. Starting with all combinations of  $i$  and  $j$  as the end points of the match in the reference and the query, respectively, we perform a traceback to get the starting points of the match.

**RLCS** algorithm outputs a match when the score is more than a score threshold  $\xi$ . However, with a simple score thresholding, we get multiple overlapping matches, from which we select the match with the highest score. If the scores of multiple overlapping matches are equal, we select the ones that have the lowest width (WAR). This way, we obtain a match that has the highest score density. We use these non-overlapping matches and the corresponding syllable boundaries to retrieve the audio patterns.

## Variants of RLCS

The generalized **RLCS** provides a framework for subsequence search. The parameters  $\rho$ ,  $\eta$ ,  $\xi$  and  $\delta$  can be tuned to make the algorithm more sensitive to different kinds of transcription errors. The variants we consider here use different distance measures  $d(a_i, a_j)$  in Eq. 6.4 to handle substitutions and different functions  $f(\cdot)$  in Eq. 6.8 to handle insertions and deletions. We explore these variants for the current task and evaluate their performance.

In a default **RLCS** configuration ( $\text{RLCS}_0$ ), we only consider exact syllable matches. We set  $\delta = 0$  and use a binary distance metric based on the syllable label, i.e.  $d(a_i, a_j) = 0$  if  $a_i = a_j$ , and 1 otherwise. Further, an identity warping function,  $f(y) = y$  is used. The rough length match densities can be transformed using a non-linear warping function to penalize low density values more than the higher ones, leading to another variant of **RLCS** ( $\text{RLCS}_s$ ). In

this dissertation, we only explore warping functions of the form,

$$f(y) = \frac{e^{\kappa y} - 1}{e^\kappa - 1} \quad (6.9)$$

where  $\kappa > 0$  is a parameter to control warping, larger values of  $\kappa$  lead to more deviation from an identity transformation.  $\text{RLCS}_0$  is a limiting case of  $\text{RLCS}_s$  when  $\kappa \rightarrow 0$ .

We hypothesize that the substitution errors in transcription are due to the confusion between timbrally similar syllables. A timbral similarity (distance) measure between the syllables can thus be used to make an **RLCS** algorithm robust to specific kinds of substitution errors. In essence, we want to disregard and give a greater allowance for substitutions between timbrally similar syllables during RLCS matching. Computing timbral similarity is a wide area of research and has many different proposed methods (Pachet & Aucouturier, 2004), but we restrict ourselves to a basic timbral distance measure: the Mahalanobis distance between the cluster centers obtained using a K-means clustering of MFCC features (with 3 clusters) from isolated audio examples of each syllable (Aucouturier & Pachet, 2002). We call this variant of RLCS as  $\text{RLCS}_d$  and experiment with different thresholds  $\delta$ .

### 6.3.4 Results and discussion

Similar to the results in Section 6.2.2, we present an evaluation of percussion pattern transcription and discovery for both tabla and mridangam solo datasets. The results of automatic transcription and those of approximate pattern search are presented separately in each case. We first present it for the tabla solo dataset (**MTS** dataset), followed by the mridangam solo dataset (**UMS** dataset). It is important to note the contrast between the two datasets being evaluated: the recordings in **UMS** dataset consists of short phrases with the query patterns being the same order of length, while the recordings in **MTS** dataset are full length compositions spanning multiple **tāl** cycles. We will also analyze the effect of this difference in datasets on the results of pattern search.

LM	Feature	Training		Test	
		C	A	C	A
Flat	MFCC_D_A	67.82	46.05	<u>64.21</u>	37.94
	MFCC_0_D_A	70.63	51.78	<u>66.30</u>	<u>43.86</u>
Bigram	MFCC_D_A	<b>68.50</b>	<b>50.48</b>	<u>65.33</u>	<b>44.10</b>
	MFCC_0_D_A	69.33	46.72	<u>64.49</u>	39.48

**Table 6.5:** Automatic transcription results on the MTS dataset using HMMs initialized with a flat start for each syllable. The table shows both training and test performance, for both a flat and a bigram language model, using the Correctness (C) and Accuracy (A) performance measures. The best performing combination with highest test Accuracy is shown in bold. For test data performance, the values underlined in each column are statistically equivalent to the best result (in a paired-sample t-test at 5% significance levels).

LM	Feature	Training		Test	
		C	A	C	A
Flat	MFCC_D_A	68.42	52.69	64.07	45.01
	MFCC_0_D_A	68.91	56.78	64.26	49.27
Bigram	MFCC_D_A	70.16	57.83	<u>65.53</u>	49.97
	MFCC_0_D_A	<b>70.71</b>	<b>60.77</b>	<u>66.23</u>	<b>53.13</b>

**Table 6.6:** Automatic transcription results on the MTS dataset using HMMs initialized using isolated stroke examples for each syllable. The table shows both training and test performance, for both a flat and a bigram language model, using the Correctness (C) and Accuracy (A) performance measures. The best performing combination with highest test Accuracy is shown in bold. For test data performance, the values underlined in each column are statistically equivalent to the best result (in a paired-sample t-test at 5% significance levels).

### Results on tabla solo dataset

The tabla solo dataset (MTS dataset) described in Section 4.2.3 is used to evaluate the performance of transcription and discovery in tabla percussion solo recordings. The results of automatic tran-

scription is first presented, and the best performing transcription system is used to present the results of approximate pattern search using different variants of **RLCS**.

The performance of automatic transcription is shown in Table 6.5-6.6 as the mean value over the whole dataset in a leave-one-piece out cross validation experiment. The performance measures are Correctness ( $\mathfrak{C}$ ) and Accuracy ( $\mathfrak{A}$ ) as defined in Eq. 6.2. We experimented with the two different MFCC features (**MFCC\_D\_A** and **MFCC\_0\_D\_A**), two different initializations of HMMs (an isolated training and a flat start, both followed by embedded reestimation training) and two language models (a flat model and a bigram learnt from data).

With the parallel time aligned transcriptions in the dataset, we experiment with both a flat initialization of syllables with an isolated training initialization of syllable **HMMs**, followed by embedded training. The results with flat start **HMMs** is shown in Table 6.5 and the results for **HMMs** initialized with isolated stroke models is shown in Table 6.6. In each table, the results are shown for both a flat (uniform) language model that assumes equal unigram and bigram probabilities, and for a bigram language model learned from training data. The tables also show both training accuracy (measured on training data) and test accuracy (measured on test data).

Overall, we see a best test Accuracy of 53.13% for isolated stroke initialization with a bigram language model and **MFCC\_0\_D\_A** feature, which justifies the use of a robust approximate string search algorithm for pattern retrieval. We see that the Accuracy measure for all cases is lower than the Correctness measure, which shows that there are a significant number of insertion errors in transcription. Training Accuracy is higher than test Accuracy, but with a small margin showing that there is some difficulty in modeling unseen data. Isolated stroke **HMM** initialization improves performance, and hence its useful to work with time aligned transcriptions. The use of a bigram language model learned from data improves the transcription performance when using isolated stroke **HMM** initialization. With the features, when using isolated stroke **HMM** initialization, we see that the use of the energy co-efficient in **MFCC\_0\_D\_A** performs better when compared to the feature **MFCC\_D\_A**, which shows the the use of relative volume dynamics between strokes improves transcription performance.

Variant	Parameter	Precision ( $\text{p}$ )	Recall ( $\text{r}$ )	f-measure ( $\text{f}$ )
Baseline	-	0.479	0.254	0.332
$\text{RLCS}_0$	$\delta = 0$	0.384	0.395	0.389
$\text{RLCS}_d$	$\delta = 0.3$	0.139	0.466	0.214
$\text{RLCS}_d$	$\delta = 0.6$	0.084	0.558	0.145
$\text{RLCS}_s$	$\kappa = 1$	0.412	0.350	0.378
$\text{RLCS}_s$	$\kappa = 4$	0.473	0.268	0.342
$\text{RLCS}_s$	$\kappa = 7$	0.482	0.259	0.336
$\text{RLCS}_s$	$\kappa = 9$	0.481	0.258	0.335

**Table 6.7:** Performance of approximate pattern search on tabla solo dataset using different **RLCS** variants using the best performing parameter settings for  $\text{RLCS}_0$  ( $\rho = 0.875$ ,  $\eta = 0.76$  and  $\xi = 0.6$ ).

We use the output transcriptions from the best performing combination (**MFCC\_0\_D\_A** and a bigram language model) to report the performance of approximate string matching with different **RLCS** variants for the query patterns shown in Table 6.3. For pattern retrieval, we don't evaluate the accuracy of boundary segmentation. However, we call a retrieved pattern from **RLCS** as *correctly retrieved* if it has at least a 70% overlap with the pattern instance in ground truth. To evaluate pattern search performance, we use the standard information retrieval measures precision (the ratio between the number of correctly retrieved patterns and all retrieved patterns) and recall (the ratio between number of correctly retrieved patterns and the patterns in the ground truth). The harmonic mean of precision and recall, the f-measure is also reported. To form a baseline for string search performance with the output transcriptions, we used an exact string search algorithm and report its performance in Table 6.7 (shown as Baseline). We see that the baseline has a precision that is similar to transcription performance, but a very poor recall leading to a poor f-measure.

To establish the optimum parameter settings for **RLCS**, we performed a grid search over the values of  $\eta$ ,  $\rho$  and  $\xi$  with  $\text{RLCS}_0$ . The parameters  $\eta$  and  $\xi$  are varied in the range 0 to 1. To ensure that the minimum length of the pattern matched is at least 2, we varied  $\rho$  in the range,  $1.1/\min(L_q) \leq 1$ .

The parameter  $\eta$  is the convex sum parameter for the contri-

bution of the rough match length density of the reference and the query towards the final score. With increasing  $\eta$ , we give more weight to the reference length ratio, allowing more insertions. We observed a poor true positive rate with larger  $\eta$ , and hence we validate the observation that insertion errors contribute to a majority of transcription errors.

The best average f-measure over all the query patterns in an experiment using  $\text{RLCS}_0$  is reported in Table 6.7. We see that  $\text{RLCS}_0$  improves the recall, but with a lower precision and an improved f-measure, showing that the flexibility in approximate matching provided by  $\text{RLCS}$  comes at the cost of additional false positives. The values of  $\rho$ ,  $\eta$  and  $\xi$  that give the best f-measure are then fixed for all subsequent experiments to compare the performance of the proposed  $\text{RLCS}$  variants.

It is observed that the patterns composed of smaller repetitive patterns (and hence having ambiguous boundaries) result in a poor precision (e.g.  $\mathbf{A}_2$  and  $\mathbf{A}_3$  in Table 6.3 with a precision of 0.108 and 0.239, respectively). Both are commonly played patterns with several repetitions and have a poor precision due to incorrect segmentation.  $\mathbf{A}_1$  in Table 6.3, on the contrary, has non-ambiguous boundaries leading to a good precision of 0.692. The effect of the length of a pattern on precision is also evident. Small patterns (with  $L = 4$ ) that have non-ambiguous boundaries (e.g.  $\mathbf{A}_8$  in Table 6.3 with a precision of 0.384) have a poor precision as compared to longer patterns with non-ambiguous boundaries (e.g.  $\mathbf{A}_1$ ). The reason for this is that the smaller patterns are more prone to errors as the search algorithm has to match a lower number of syllables.

The results with other variants of  $\text{RLCS}$  are also reported in Table 6.3. The results from  $\text{RLCS}_d$  show that the use of a timbral syllable distance measure with higher threshold  $\delta$  further improves the recall, but with a much lower precision and f-measure. Although we find matches that have substitution errors using the distance measure, we retrieve additional matches that do not have substitution errors contributing to additional false positives. On the contrary, using a non-linear warping function  $f(\cdot)$  in  $\text{RLCS}_s$  improves the precision with a higher value of  $\kappa$ . The penalties on matches with higher number of insertions and deletions is high and they are left out, leading to good precision at the cost of recall. We observe that both the above mentioned variants improve either pre-

LM	Feature	Training		Test	
		$\mathfrak{C}$	$\mathfrak{A}$	$\mathfrak{C}$	$\mathfrak{A}$
Flat	MFCC_D_A	76.66	59.43	<u>74.08</u>	55.64
	MFCC_0_D_A	<b>76.63</b>	<b>63.79</b>	<u>74.13</u>	<b>60.23</b>
Bigram	MFCC_D_A	78.12	57.69	75.90	54.02
	MFCC_0_D_A	78.78	60.54	76.50	57.38

**Table 6.8:** Automatic transcription results on the [UMS](#) dataset using HMMs trained using a flat start for each syllable. The table shows both training and test performance, for both a flat and a bigram language model, using the Correctness ( $\mathfrak{C}$ ) and Accuracy ( $\mathfrak{A}$ ) performance measures. The best performing combination with highest test Accuracy is shown in bold. For test data performance, the values underlined in each column are statistically equivalent to the best result (in a paired-sample t-test at 5% significance levels).

cision or recall at the cost of the other measure. They need further exploration with better timbral similarity measures to be combined in an effective way to improve the search performance.

### Results on mridangam solo dataset

Similar to an evaluation on the tabla solo dataset, we present a parallel evaluation with the mridangam solo dataset ([UMS](#) dataset). Unlike the tabla solo dataset, since the mridangam solo dataset does not have time aligned ground truth transcriptions, we report automatic transcription results only for embedded [HMM](#) training. With the best performing combination, we then report results of pattern search using different [RLCS](#) variants using the query patterns in Table 6.4. We use identical definitions of performance measures as used while reporting results for tabla solo dataset.

The results of automatic transcription are shown in Table 6.8 for all the combinations of conditions. Overall, we see a best test Accuracy of 60.23% for [MFCC\\_0\\_D\\_A](#). Similar to results on tabla dataset, we see that the Accuracy measure for all cases is lower than the Correctness measure, which shows that there are a significant number of insertion errors in transcription. Training Accuracy is higher than test Accuracy, but with a small margin showing that

Variant	Parameter	Precision ( $\rho$ )	Recall ( $\tau$ )	f-measure ( $f$ )
Baseline	-	XX	XX	XX
$\text{RLCS}_0$	$\delta = 0$	0.902	0.492	0.637

**Table 6.9:** Performance of approximate pattern search using different **RLCS** variants using the best performing parameter settings for  $\text{RLCS}_0$  ( $\rho = 0.275$ ,  $\eta = 0.36$  and  $\xi = 0.95$ ).

there is some difficulty in modeling unseen data. With the features, we see that the use of the energy co-efficient in **MFCC\_0\_D\_A** performs better when compared to the feature **MFCC\_D\_A**, which shows the the use of relative volume dynamics between strokes improves transcription performance.

Contrary to results on tabla dataset, the use of a bigram language model learned from data does not improve the transcription performance. The better performing combination uses a flat language model. We hypothesize that it is because there is much more variety in mridangam stroke playing in the dataset and a bigram language model learned from training data restricts the possibility of unseen stroke sequences adversely. It also hints towards the use of better language models that can incorporate longer contexts than a simplistic bigram language model.

Using the output transcriptions from the best performing combination (**MFCC\_0\_D\_A** and a flat language model), we report the performance of approximate string matching with **RLCS** algorithm for the query patterns shown in Table 6.4. Table 6.9 shows the average results of pattern search with the mridangam solo dataset with the  $\text{RLCS}_0$  algorithm, with an exact string search baseline also shown. We further establish the optimum parameter settings for **RLCS** using a grid search similar to experiments with tabla solo dataset. The numbers in the table show the results for the best performing parameter settings. Since  $\text{RLCS}_d$  and  $\text{RLCS}_s$  algorithms did not show any improvement in f-measure for tabla solos, only the results of  $\text{RLCS}_0$  are reported for mridangam solos. **Explanation of Table 6.9 to be completed.**

To summarize, the goal of these evaluation on tabla and mridangam solo datasets is to present a methodology for transcription and discovery of percussion patterns in syllabic percussion sys-

tems. The work presented is preliminary and not comprehensive, and can be significantly improved. However, the basic idea of using a musically meaningful representation system to define and describe patterns is valid and useful.

We addressed the unexplored problem of discovering syllabic percussion patterns in Tabla solo recordings. The presented formulation used a parallel corpus of audio recordings and syllabic scores to create a set of query patterns, that were searched in an automatically transcribed (into syllables) piece of audio. We used a simplistic definition of a pattern and explored RLCS based subsequence search algorithm, using an HMM based automatic transcription. Compared to a baseline, we showed that the use of approximate string search algorithms improved the recall at the cost of precision. Additionally, proposed variants evaluated on the MTS dataset improved either the precision or recall, but do not provide a significant improvement in the f-measure over the basic RLCS.

For future work, we aim to improve syllable boundaries output by transcription using onset detection. Inclusion of the rhythmic information can be an interesting aspect in defining and discovering percussion patterns, and will help in comprehensively evaluating the task of pattern discovery. The next steps would be to incorporate better timbral similarity measures and inclusion of segment boundaries into the RLCS algorithm that effectively combines the proposed variants.

---

# Applications, Summary and Conclusions

The outcome of any serious research can only be  
to make two questions grow where only one grew  
before.

---

Thorstein Veblen (1908)

The concluding chapter of the dissertation aims to present some concrete applications of the rhythm analysis approaches and results presented in the previous chapters. A summary of the work presented in the dissertation, along with some key results and conclusions then follow. The thesis opens up a host of open problems and pointers and directions for future work based on the thesis form the last part of the chapter.

## 7.1 Applications

There are several applications for the research work presented in the dissertation. Some of these applications have already been identified in Chapter 1 and Chapter 3. The goal of this section is to present concrete examples of such applications, and further suggest other applications that might be built or get benefited from the work presented here. The section describes some of the applications that

have already been resulted from the work in CompMusic so far, and those that have been planned within the efforts of CompMusic. Possible future applications with the data and methods are also discussed briefly. At the outset, the primary objective and application of automatic rhythm analysis is to use them to define rhythm similarity measures between music pieces (and within excerpts within them) on large corpora of music. While this has not been addressed in this thesis, there are several ways in which the research discussed in the thesis can be used to define similarity measures and use them for various applications.

The main application area for automatic rhythm analysis algorithms in the thesis is enriched listening, with additional rhythm related metadata along with the music recording. Meter analysis outputs can be further used to improve higher level MIR tasks analyzing the musical structure. The tools can also aid in corpus level musicological studies for analysis of both music theory and performance.

Enriched music listening is a primary application of the meter analysis methods presented in the thesis. With rich additional rhythm related information such as the *tāla*, time varying tempo, beats and the *sama* can all enrich the music listening experience. It finds audience both for serious music listeners and also music students who wish to understand more about the underlying rhythmic structures. Large archives of music can be organized with added rhythm metadata and presented to listeners. Semi-automatic rhythm annotation applications can be built with these analysis tools. For a music expert (such as a musician, musicologist, an expert listener, or even a music student) curating these collections, it might be possible to tap some instances of the *tāla* for a piece and an informed meter tracking algorithm can track the rest of the piece using that initialization. Such a semi-automatic annotation tool significantly enhances the accuracy of *tāla* tracking and hence is practical for real world applications.

Percussion pattern transcription and discovery finds its application in helping navigate through percussion solo recordings (*tani* recordings in Carnatic music) in a more meaningful way. Applications such as search by patterns can be conceived, such as query by example, query by drumming, or in the case of Indian art music, query by vocals. The syllabic system in Indian art music enables

us to build a query system where the query is the vocalized pattern of syllables, which can then be searched in a corpus of percussion solos. An improvement of such a query by vocals system is a machine improvisation system.

During Hindustani music concerts, it is common to have a call-response improvisatory passages between musicians, called a *sawaal-jawaab* (literally, question-answer). It is also common in tabla solos to have a *sawaal-jawaab* between a musician reciting vocal syllables and a response by the musician playing the tabla. A basic prototype system called Sawaal-Jawaab<sup>1</sup> has been built with this idea, with the call being the vocal recitation of syllables. The response is an improvisation of the call built using timing, rhythmic and timbral features from the call, exploiting the onomatopeioc nature of the tabla bols. Such an improvisation is done within the framework of a specific *tāl*.

Musicologists working with rhythm would benefit from the corpora and tools developed in the thesis. Musicological applications include tools for analysis of large corpora. The CompMusic corpora and datasets are representative well curated with several useful information that can be used to derive valid musicological findings. Semi-automatic meter analysis tools can lead to complete accurate meter tracking and hence be used to analyze larger corpora of recordings, which would be a tedious time consuming task if done manually by musicologists. Percussion pattern discovery is useful for style analysis of different tabla *gharānās* and mridangam style schools of teaching. Though it would require larger corpora and significant musicological intervention, automatic pattern discovery framework would aid such a task.

To conclude, two specific applications built within CompMusic are described below: Dunya and Sarāga. Both these applications are collaborative efforts of the CompMusic team. A brief introduction to the tools is provided for a better understanding, and then we emphasize on how the rhythm analysis methods developed in thesis apply and integrate into these tools.

---

<sup>1</sup>Further details and a demo available at [http://labrosa.ee.columbia.edu/hamr\\_ismir2015/proceedings/doku.php?id=sawaal-jawaab](http://labrosa.ee.columbia.edu/hamr_ismir2015/proceedings/doku.php?id=sawaal-jawaab) or <http://compmusic.upf.edu/node/283>

### 7.1.1 Dunya

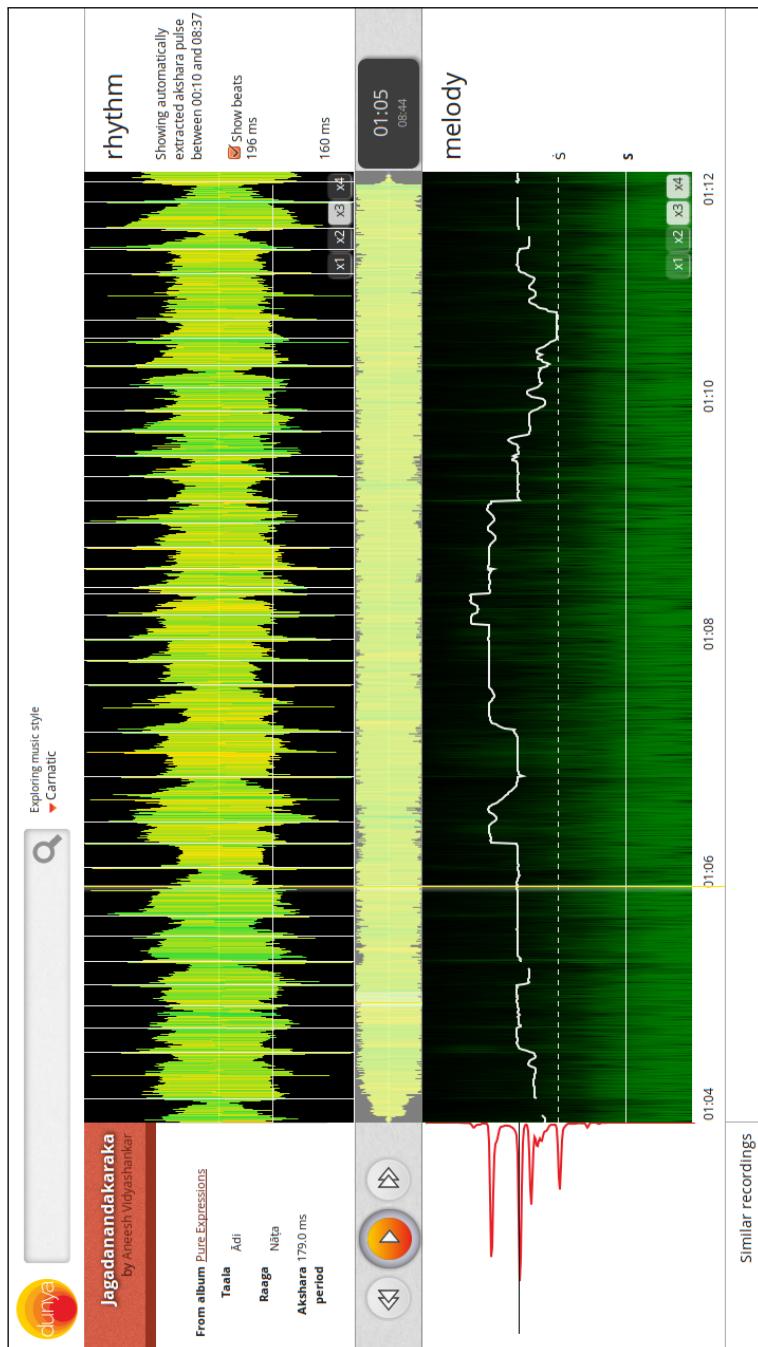
As described earlier, Dunya<sup>2</sup> comprises a set of cross-platform open source tools for navigating through music collections in a culture aware and musically relevant way. It is also a test platform to evaluate the research results of CompMusic where users can interact with the music collections under study and with which we can evaluate most of the research results from a user perspective. Dunya is aimed at music connoisseurs of the particular music traditions. It uses the technologies developed for melodic and rhythmic description to navigate through the audio recordings and through the other information items available in a particular music collection. This navigation promotes the discovery of relationships between the different information entities.

Dunya aggregates music and related metadata from various sources such as music archives, Wikipedia and MusicBrainz and makes it available to users for an enriching listening experience with music. The automatically and manually extracted audio and other features, and curated metadata can also be accessed through Dunya. Dunya has a front end interactive web based tool where users can interactively browse through these music collections. Dunya provides an interface for music similarity based navigation through music collections, and has a detailed recording page that will provide an interactive interface with a visualization of automatically extracted metadata. It will also provide an interface for navigating through the main musically meaningful entities of the specific music culture using characteristic rhythmic and melodic patterns. It also has a back end along with an API that provides access to all these data. Dunya hence acts as the central permanent online repository to store the metadata, audio, annotations and research results.

The research results from the presented work on rhythm analysis are partly integrated into Dunya, and further integration is in progress. The rhythm analysis tools developed will be a part of the suite of MIR tools integrated into Dunya. Essentia is an audio analysis and audio based MIR toolkit (Bogdanov et al., 2013a, 2013b). The Dunya backend uses Essentia to extract features, and hence where possible, specific rhythm extractors from the developed algorithms is also be added to Essentia.

---

<sup>2</sup><https://dunya.compmusic.upf.edu>



**Figure 7.1:** A screenshot of the recording page of Dunya showing rhythm related metadata in the top panel superimposed on top of the waveform.

Drawing information from various data sources and relating them with ontologies, Dunya is the best showcase of the tools and algorithms developed as a part of the thesis. A screenshot of the Dunya recording page interface for a Carnatic music recording is shown in Figure 7.1. The recording page shows important rhythm related metadata related to the recording<sup>3</sup> *tāla* editorial metadata, and automatically extracted median *akṣara* pulse period ( $\tau_o$ ). In addition, the waveform panel on the top shows the time varying  $\tau_o$  curve along with *akṣara* markers extracted automatically using the approach presented by Srinivasamurthy and Serra (2014). All of these editorial, automatically extracted, and manually annotated rhythm metadata can also be accessed from the Dunya API.

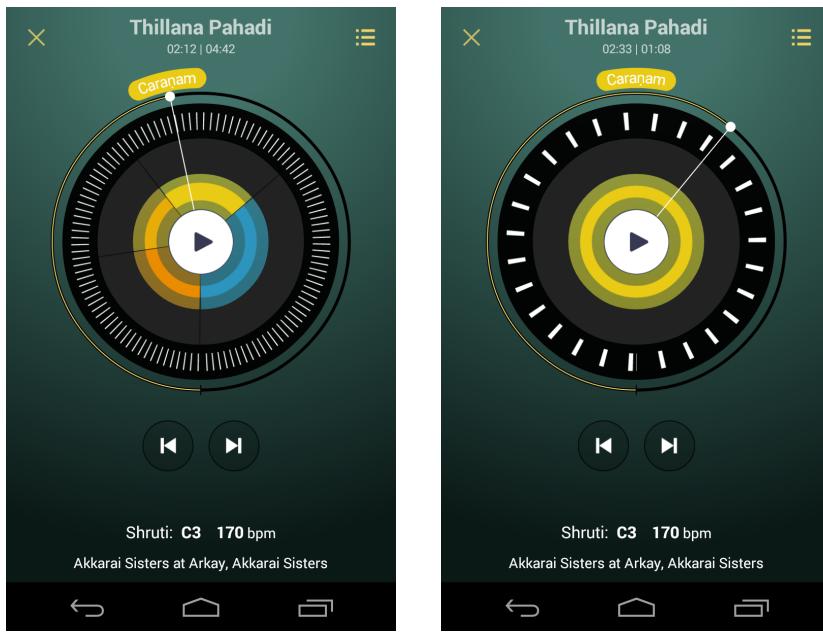
### 7.1.2 Sarāga

Culture-aware music technologies (CAMUT) is a project that aims to take the research results of CompMusic to practical real world possibly commercial applications, aiming to build technologies to foster learning and teaching of Indian music forms. Sarāga<sup>4</sup> is a music appreciation and infotainment application for students and listeners developed as a part of CAMUT. Sarāga is an android application that provides an enriched listening atmosphere over the open collection of Carnatic ( $CMD_o$ ) and Hindustani ( $HMD_o$ ) music. It allows Indian art music connoisseurs and casual listeners to navigate, discover and listen to these music traditions using familiar, relevant and culturally grounded concepts. Sarāga includes innovative visualizations and inter and intra-song navigation patterns that present musically rich information to the user. These time synchronized visualizations of musically relevant facets such as melodic patterns, sama locations and sections provides a user with better understanding and appreciation of these music traditions. It additionally features unique compound filters over *rāgas*, *tālas*, instruments and artists for finding songs.

---

<sup>3</sup>The recording shown in Figure 7.1 is a violin rendering of the composition Jagadanandakaraka (<http://musicbrainz.org/recording/de94ed93-7399-47e3-aa8e-d77b49d94bd3>) from the album Pure expressions (<http://musicbrainz.org/release/bcb30e6f-bb13-499d-8e0f-9447af8555a3>) by Aneesh Vidyashankar

<sup>4</sup>Application summary paraphrased from <http://musicmuni.com/>



(a) The entire music piece

(b) The carana section zoomed

**Figure 7.2:** Screenshots of the mobile application Sarāga for a recording. Panel (a) shows the entire music piece with all the sections, while panel (b) shows the carana section zoomed. The sama markers can be seen as white colored ticks on the outer circle. The tempo of the piece is displayed at the bottom of the screen.

A screenshot of the application in Figure 7.2 shows the rich and novel visualization of a music recording<sup>5</sup> including several different associated metadata. Figure 7.2a shows all the sections of the piece while fig:saraga:charana shows only the carana (also called caranam) section. The median tempo of the piece is shown as 170 BPM at the bottom of the panel. The whole piece (or a section when zoomed in) is summarized in concentric circles, with white colored time ticks on the outer circle indicating the location of samas. Both the tempo and the samas shown on recordings in Sarāga have been

<sup>5</sup>The screenshot shows the recording of a tillāna in rāga Pahādi (<http://musicbrainz.org/recording/50c2fea1-d267-4506-a155-73bbefd5da27>) from the album Akkarai Sisters in Arkay (<http://musicbrainz.org/release/513e205a-8d71-4d4a-95f7-96d131fa15bc>)

semi-automatically extracted from audio using  $\text{AMPF}_0$  algorithm using the bar pointer model, and then corrected for any errors manually.

## 7.2 Contributions

A summary of the specific contributions from the work presented in the dissertation are listed below.

### Contributions to creating Research Corpora and datasets

Building research corpora for MIR research is one of the primary tasks of CompMusic project. Significant collaborative efforts have been put into building research corpora and datasets, and relevant datasets that have a major contribution by the authors are listed below. The links to all the datasets are provided in Appendix B.

- CompMusic Carnatic Music Rhythm ( $\text{CMR}_f$ ) dataset:  $\text{Tāla}$ , beat and  $\text{sama}$  annotated collection of 176 Carnatic music pieces, built with the support of Vignesh Ishwar, a professional Carnatic musician who also verified the annotations. The dataset has about 16.6 hours of audio spanning pieces from four popular  $\text{tālas}$ . A representative subset of the dataset with 118 pieces ( $\text{CMR}$  dataset) was also built.
- CompMusic Hindustani Music Rhythm ( $\text{HMR}_f$ ) dataset:  $\text{Tāl}$ ,  $\text{mātrā}$  and  $\text{sam}$  annotated collection of 151 Hindustani music excerpts, built with the support of Kaustuv Kanti Ganguli, a professional Hindustani musician who also verified the annotations. The dataset has about 5 hours of audio spanning pieces from four popular  $\text{tāls}$  and three different  $\text{lay}$  (tempo classes). Two subsets of the dataset grouped based on  $\text{lay}$ :  $\text{HMR}_l$  and  $\text{HMR}_s$  were also built.
- CompMusic Carnatic open music collection ( $\text{CMD}_o$  collection): The  $\text{sama}$  annotations in the  $\text{CMD}_o$  collection in collaboration with Vignesh Ishwar. The dataset contains over 190 pieces spanning  $\text{XX}$  hours of music and  $\text{XX}$   $\text{sama}$  annotations.
- CompMusic Hindustani open music collection ( $\text{CMD}_o$  collection): The  $\text{sam}$  and section annotations in the  $\text{HMD}_o$  collection in collab-

oration with Kaustuv Kanti Ganguli. The dataset contains over **XX** hours of music, **XX** sam and **XX** section annotations.

- CompMusic Mulgaonkar Tabla Solo dataset ([MTS](#)) dataset: The tabla solo dataset consisting of audio recordings of 38 tabla solo compositions from different *gharānās* with time aligned syllabic transcription was built with Swapnil Gupta.
- CompMusic *Jīngjù* percussion pattern ([JPP](#)) dataset: The [JPP](#) dataset was built with Rafael Caro, a musicologist, and consists of 133 audio percussion patterns spanning five different pattern classes, with about 22 minutes of audio and over 2200 percussion syllables.
- *Jīngjù* percussion instrument ([JP1](#)) dataset: Built with Mi Tian at Centre for Digital Music, Queen Mary University of London, the dataset has over 3000 audio samples for four different percussion instrument classes in *jīngjù*.

### Technical and Scientific Contributions

- Identification of challenges, opportunities and applications of automatic rhythm analysis of Indian art music.
- Identification of several interesting automatic rhythm analysis problems in Indian art music, along with a review and evaluation of the state of art for some of the tasks, establishing the need for joint estimation methods that can incorporate higher level music information.
- Engineering formulation of meter analysis (meter inference, meter tracking and informed meter tracking) and percussion pattern discovery (transcription and search) in Indian art music.
- An illustrative evaluation of the Carnatic and Hindustani music research corpora based on the methodology by Serra (2014).
- An illustrative demonstration of the utility of corpora and rhythm analysis tools for a corpus level musicological analysis, as exemplified by rhythmic pattern analysis in Carnatic and Hindustani music.

- Demonstration of the utility of percussion syllables in representation, transcription and discovery of percussion patterns, using a syllabic mapping and grouping system for syllables based on timbral similarity for both tabla and mridangam percussion syllables (joint work with Swapnil Gupta and Akshay Anantapadmanabhan).
- Bayesian methods for meter analysis in Indian art music: The task of meter analysis addressed for the first time in Carnatic and Hindustani music, developing approaches that are aware of underlying metrical structures and utilize them explicitly. Novel meter analysis model extensions (MO-model and SP-model) and inference extensions are proposed. The novel SP-model shows improvement in tracking long metrical cycles, a task which has been addressed for the first time in MIR.
- Approaches for percussion solo transcription and discovery in syllabic percussion systems, applied on Beijing Opera as a test case and then extended to tabla and mridangam percussion solos in Indian art music.

### 7.3 Conclusions and Summary

We present a summary, conclusions and key results from the thesis, organized based on the chapters of the dissertation. Broadly, the dissertation aimed to build culture-aware and domain specific data driven MIR approaches using Bayesian models for automatic rhythm analysis in Indian art music, focusing mainly on the tasks of meter analysis and percussion pattern discovery, with the eventual goal of developing rhythm similarity measures. Such approaches would lead to tools and technologies that can improve our experience with music helping us to navigate through large music collections in a musically meaningful way, all of it within the sociocultural context of the music culture. The applications lie in enriched music listening, music archival, music learning, musicology, and as pre-processing steps in MIR for higher level information such as structure and style analysis.

The dissertation focused on rhythm analysis tasks within the purview of CompMusic project. The scope of the thesis was lim-

ited to rhythm analysis in audio collections of Indian art music using Bayesian approaches, emphasizing on data and models. The thesis aimed to address the question of utility of culture specific data driven approaches to rhythm analysis and their applications. A summary of rhythm in Indian art music was discussed in Chapter 2 to provide a background to music concepts encountered in the thesis, showing the contrasting differences between several rhythm concepts in eurogenetic popular music and Indian art music. *Jīngjù* (Beijing Opera) percussion provides a suitable case to study percussion patterns and hence a basic introduction to *jīngjù* was also provided. A review of the state of the art in rhythm analysis tasks in MIR provided a basis for understanding relevant rhythm analysis tasks in Indian art music.

Chapter 3 identified some of the unique challenges and opportunities to rhythm analysis in Indian art music. The complexities and characteristics of rhythm in Indian art music make it an ideal candidate for analysis and push the boundaries of the state of the art in rhythm analysis in MIR. Important and relevant rhythm analysis problems within the context of Indian art music were identified and described. An evaluation of the state of the art with some of these problems indicated the need for culture-aware domain specific methods to address these tasks. The set of tasks identified in the chapter will be useful for a researcher looking to solve relevant problems in this new area of research. Definitions relating to rhythm in Indian art music suffered inconsistencies, which were addressed by formulating the research problems of meter analysis and percussion pattern discovery more accurately.

The problem of creating research corpora and datasets for data-driven MIR research, the content of Chapter 4, shows that significant efforts are needed to build relevant datasets for research. It is possible to build relevant datasets, using a combination of criteria that are used to continuously evaluate the datasets and corpora for their suitability to the tasks at hand. Further, research corpora can be used for corpus level analysis to draw several inferences from it. The dissertation is aimed to be a comprehensive resource for the rhythm related datasets developed as a part of CompMusic, with suggestions on tasks where each of those datasets can be useful.

Meter analysis was one of the main problems addressed in the thesis. Chapter 5 presented a comprehensive analysis of meter in-

ference, meter tracking, and informed meter tracking in Indian art music. Preliminary experiments showed poor performance with *sama* tracking in Carnatic music, indicating the need for meter analysis methods that can utilize metrical structure information. The bar pointer model is one such Bayesian model that allows for a joint estimation of components of meter and incorporates the metrical structure explicitly incorporates the underlying metrical structure. An evaluation of the state of the art BP-model on Indian art music showed the utility of Bayesian models for meter analysis. Novel model extensions (MO-model and SP-model) to improve on BP-model were proposed. In addition, novel inference extensions based on particle filters for faster inference from these models were also proposed. The algorithms were evaluated in the settings of meter inference, meter tracking, tempo-informed meter tracking, and tempo-sama-informed meter tracking. The experiments clearly show that incorporating additional *tāla* information and making the algorithms more “informed” improves performance of algorithms. Further, the SP-model shows improvement in tracking long metrical cycles, a task which has been addressed for the first time in MIR. The inference extensions did not show much improvement in performance, but have promising ideas to make inference faster with these Bayesian models. The models and inference extensions are capable of generalizing to other music cultures, as was demonstrated with an evaluation of Ballroom dataset.

A framework for percussion pattern discovery from solo recordings, along with some exploratory experiments for the task was the subject matter of Chapter 6. Utilizing the syllabic percussion system in Indian art music, we used the onomatopoeic oral-mnemonic syllables to represent, transcribe and search for percussion patterns from audio recordings of percussion solos. A syllabic representation is suitable for timbrally similar percussion patterns and a transcription+search framework was explored for discovery of patterns. Preliminary experiments on percussion pattern transcription and classification were presented on *jīngjù* percussion patterns, and then extended to pattern discovery by transcription followed by approximate search on tabla and mridangam solo recordings. The transcription was based on a speech recognition framework using timbral syllable models with a language model. A set of query patterns were automatically derived from data. Transcriptions have

errors and hence an approximate search algorithm (such as **RLCS**) was used to search for these query patterns in transcribed recording. Preliminary experiments on the **MTS** and **UMS** datasets show that there are several insertion errors and there is a need for approximate search algorithms. Exact string search algorithm gives a better precision but a poor recall due to all the transcription errors. An approximate string search algorithm **RLCS** showed improvement in pattern recall with the **MTS** dataset that included full length compositions. With the **UMS** dataset consists of short audio files segmented by phrases and hence **RLCS** does not show much improvement in pattern recall over an exact string search. The variants of **RLCS** need to be further explored and improved. For the task, the combination of transcription+search for the problem is a promising approach, while further experiments with a comprehensive evaluation is needed to make stronger claims on performance and suggest improvements.

## 7.4 Future directions

There are several directions for future work based on the thesis. One of the goals of the dissertation was to present relevant research problems in rhythm analysis of Indian art music. Some of these problems presented in Chapter 3 are a good start to extend the work presented in the dissertation. Several rhythm tasks for Indian art music were proposed in Chapter 3, while only a few of them were addressed in the thesis. The problems such as building rhythm ontologies and rhythm based segmentation have received no attention from the research community so far. Both are relevant areas of research with potential to be explored in the future.

The goal of automatic rhythm analysis is to define relevant rhythm similarity measures, a topic that has not been addressed in the dissertation. Using both the rhythmic structures and patterns extracted from audio to define better measures is an important part of future work. In addition, the work in the thesis used only audio recordings to extract meaningful rhythm information. However, using additional metadata (such as lyrics, scores, editorial metadata) along with audio features and combining them with suitable rhythm ontologies can lead to better similarity measures, which is to be ex-

plored as a part of future work.

The sizeable curated research corpora and datasets have immense opportunity for be utilized for a variety of research problems in the future. The availability of the datasets now opens up the possibility of significant data driven automatic rhythm analysis research in these music cultures. The problems that can use these datasets were detailed in Chapter 4. In addition, the Creative Commons music collections developed for Carnatic and Hindustani music can be used to build open data and algorithms without restrictive copyright issues. The research corpora can evolve over time and that would be an important task for future, to improve the research corpora and build additional datasets for rhythm analysis tasks. The use of these datasets for musicological research was hinted in our experiments, but a rigorous study of suitability of the corpora and datasets for musicology, and its adoption for musicological studies is one direction to pursue.

Meter analysis tasks such as meter inference and tracking were addressed in detail in the dissertation. However, there are several open questions that still have to be more rigorously answered. The experiments need to be extended to full length music pieces from the present experiments on shprter excerpts, to make it more practical. A part of immediate future work would be to evaluate it further on larger Indian music datasets. The SP-model for meter analysis showed significant promise in tracking a wide range of metrical structures in both Carnatic and Hindustani music. Further formal evaluations on its extendability to other music cultures is an important future work. The model and inference extensions need to be analyzed further to improve their performance. The use of spectral flux feature for meter analysis is limiting. Developing additional audio features that can perhaps also include melodic features are to be further explored.

Percussion pattern discovery was a problem that was addressed to a lesser extent in the dissertation with only preliminary results presented on a small datasets. While the framework using syllabic representation for percussion patterns shows promise, extensive evaluation on larger datasets spanning different playing styles, schools, instrument timbres, and variability in patterns is to be done in the future. Better transcription can be achieved in real world scenarios with such diverse data to build acoustic and language

models. Approximate string search using RLCS shows promise in searching for short query patterns in longer transcribed audio recordings, while some improvements suggested to it need further exploration to make the search algorithm robust to transcription errors.

Integration of these algorithms and methods into practical applications requires additional effort to understand the gaps between research tasks and practical needs. Tools such as Dunya aim to bridge that gap by providing an evaluation framework for research algorithms. In the future, an integration of all the described rhythm analysis approaches into Dunya is important and helps improve the algorithms, while aiming to provide users with better experience with large collections of music.

We sincerely believe that the dissertation has opened up the new area of research in automatic rhythm analysis of Indian art music. With several challenging and interesting research problems in the area, there is significant scope and potential for novel approaches and methodologies to solve these problems. The future directions discussed here provide pointers for new research in the field, and this is perhaps where a new PhD thesis can begin!



# Appendix A

---

## List of Publications by the author

To be completed ...



# Appendix B

---

## Resources

To be completed ...

Links to all resources: code, datasets, examples, papers, hacks

<http://compmusic.upf.edu/phd-thesis-ajay>



# Glossary

## C.1 Carnatic Music

- akṣara** The lowest metrical pulse
- ālāpana** An unmetered melodic improvisation
- aṅga** The sections of a tāla
- āvartana** One complete cycle of a tāla
- caraṇa** The end section of a Carnatic music composition
- kachēri** A concert of Carnatic music
- edupu** The phase/offset of the composition relative to the sama
- konnakōl** The art form of reciting the percussion syllables in Carnatic music
- kṛti** A common compositional form in Carnatic music
- mṛdaṅgam** The primary percussion accompaniment
- muttusvāmi dīkṣitar** A prominent Carnatic music composer
- naḍe** The subdivision structure within a beat
- caturaśra** A naḍe with 2 or 4 akṣaras per beat
  - tiśra** A naḍe with 3 or 6 akṣaras per beat
- rāga** The melodic framework of Carnatic music
- sama** The first akṣara of āvartana
- śyāmā śāstri** A prominent Carnatic music composer
- solkaṭṭu** The onomatopoeic rhythmic syllables of Carnatic music used in oral percussion
- tāla** The rhythmic framework of Carnatic music
- ādi** A tāla with 32 akṣaras in a cycle

- khaṇḍa chāpu** A tāla with 10 akṣaras in a cycle  
**mīśra chāpu** A tāla with 14 akṣaras in a cycle  
**rūpaka** A tāla with 12 akṣaras in a cycle  
**tambūra** The drone instrument used in Carnatic music  
**tani-āvartana** The solo performance of a percussion ensemble  
**tani** Short for tani-āvartana  
**tillāna** A rhythmic piece in Carnatic music widely used in dance performance  
**tyāgarāja** A prominent Carnatic music composer

## C.2 Hindustani Music

- āvart** One complete cycle of a tāl  
**ālāp** An unmetered melodic improvisation  
**āmad** literally approach, a phrase leading to a sam  
**bandiś** A fixed, melodic composition in Hindustani vocal/instrumental music  
**bāyān** The left drum  
**bōl** The onomatopoeic rhythmic syllables  
**dāyān** The right drum  
**dhrupad** A music style in Hindustani music  
**diggā** Another name for the left drum  
**gharānā** The stylistic schools of Hindustani music  
**khālī** A clap in the tāl cycle  
**khyāl** A music style in Hindustani music  
**lay** The tempo class  
  **dṛt** Fast tempo class  
  **madhya** Medium tempo class  
  **vilābit** Slow tempo class  
**mātrā** The lowest metrical pulse  
**pakhāvaj** A double barrel drum used as rhythm accompaniment in Hindustani music  
**rāg** The melodic framework of Hindustani music  
**sam** The first mātrā of an āvart  
**sāraṅgi** A bowed music instrument used in Hindustani music  
**tāl** The rhythmic framework of Hindustani music  
  **ēktāl** A tāl with 12 mātrās in a cycle  
  **jhaptāl** A tāl with 10 mātrās in a cycle

- rūpak tāl** A tāl with 7 mātrās in a cycle  
**tīntāl** A tāl with 16 mātrās in a cycle  
**thālī** A clap in the tāl cycle  
**gaṭ** A composition type in tabla  
**kāyadā** A composition type in tabla  
**palaṭā** A composition type in tabla  
**pēśkār** A composition type in tabla  
**rēlā** A composition type in tabla  
**thēkā** The basic bōl pattern associated with a tāl  
**vibhāg** The sections of a tāl

## C.3 Acronyms

- AMS** Anantapadmanabhan Mridangam Strokes dataset  
**JPI** Jingju percussion instrument dataset  
**JPP** Jingju percussion pattern dataset  
**CMR<sub>f</sub>** Carnatic music rhythm dataset  
**CMD<sub>o</sub>** Creative Commons Carnatic corpus  
**CMR** Carnatic music rhythm dataset (subset)  
**HMR<sub>f</sub>** Hindustani music rhythm dataset  
**HMR<sub>l</sub>** Hindustani music rhythm dataset (subset with *vilāmbit* and *madhya lay* pieces)  
**HMD<sub>o</sub>** Creative Commons Hindustani corpus  
**HMR<sub>s</sub>** Hindustani music rhythm dataset (subset with *dṛ̥t* lay pieces)  
**HMM<sub>m</sub>** HMM mix obs model of ISMIR 2015  
**HMM<sub>0</sub>** HMM variant sampling from pattern transition priors  
**HMM<sub>s</sub>** HMM section pointer model of ICASSP 2016  
**AMPF<sub>e</sub>** AMPF with end of bar sampling likelihood  
**AMPF<sub>m</sub>** AMPF mix obs model of ISMIR 2015  
**AMPF<sub>g</sub>** AMPF with gated observation likelihood update  
**AMPF<sub>p</sub>** AMPF with peak hop inference  
**AMPF<sub>0</sub>** AMPF variant sampling from pattern transition priors  
**AMPF<sub>s</sub>** AMPF section pointer model of ICASSP 2016  
**DAV** The algorithm by Davies and Plumley (2007)  
**GUL** The algorithm by Gulati et al. (2012)  
**HOC-SVM** The algorithm by Hockman et al. (2012)  
**HOC** The algorithm by Hockman et al. (2012)  
**KLA** The algorithm by Klapuri et al. (2006)

- OP** The algorithm by Pohle et al. (2009)
- PIK** The algorithm by Pikrakis et al. (2004)
- SRI** The algorithm by Srinivasamurthy et al. (2012)
- STM** The algorithm by Holzapfel and Stylianou (2011)
- MTS** Mulgaonkar Tabla Solo dataset
- UMS** UKS Mridangam Solo dataset
- DBN** Dynamic Bayesian network
- GMM** Gaussian mixture model
- HMM** Hidden Markov model
- HTK** Hidden Markov model Toolkit
- ICT** Information and Communication Technologies
- MFCC** Mel-Frequency cepstral co-efficients
- MFCC\_0\_D\_A** MFCC features with energy, delta and double-delta coefficients
- MFCC\_D\_A** MFCC features without energy but with delta and double-delta coefficients
- MIR** Music Information Research
- MIREX** Music Information Retrieval EXchange
- MMA** Madras Music Academy
- RLCS** Rough Longest Common Subsequence

---

# Bibliography

- Abdallah, S. A., & Plumbley, M. D. (2003, April). Probability as metadata: event detection in music using ICA as a conditional density model. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Signal Separation (ICA 2003)* (pp. 233–238). Nara, Japan. (pg. 53, 233)
- Anantapadmanabhan, A., Bello, J., Krishnan, R., & Murthy, H. (2014, January). Tonic-Independent Stroke Transcription of the Mridangam. In *Proceedings of the 53rd AES International Conference on Semantic Audio*. (pg. 77)
- Anantapadmanabhan, A., Bellur, A., & Murthy, H. A. (2013, May). Modal analysis and transcription of strokes of the mridangam using non-negative matrix factorization. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)* (p. 181-185). Vancouver, Canada. (pg. 77, 160)
- Atlı, H. S., Uyar, B., Şentürk, S., Bozkurt, B., & Serra, X. (2014, November). Audio Feature Extraction for Exploring Turkish Makam Music. In *Proceedings of the 3rd International Conference on Audio Technologies for Music and Media*. Ankara, Turkey: Bilkent University. Department of Communication and Design. (pg. 110)
- Aucouturier, J. J., & Pachet, F. (2002). Music similarity measures: What's the use? In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)* (pp.

- 157–163). Paris, France. (pg. 255)
- Bello, J. P., Daudet, L., Abdullah, S., Duxbury, C., Davies, M., & Sandler, M. (2005, September). A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, 13(5), 1035-1047. (pg. 40, 42, 159, 174, 176, 251)
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 34–43. (pg. 79)
- Beronja, S. (2008). *The Art of the Indian tabla*. Rupa and Co. New Delhi. (pg. 30, 32, 247)
- Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011, October). The Million Song Dataset. In *Proc. of the 12th International Conference on Music Information Retrieval (ISMIR 2011)* (p. 591-596). Miami, USA. (pg. 108)
- Bhatkhande, V. N. (1990). *Hindustani Sangeet Paddhati: Kramik Pustak Maalika Vol. I-VI*. Sangeet Karyalaya. (pg. 61, 120)
- Böck, S., Krebs, F., & Schedl, M. (2012). Evaluating the Online Capabilities of Onset Detection Methods. In *Proceedings of the 13th International Society for Music Information Retrieval Conference* (pp. 49–54). (pg. 42, 133)
- Böck, S., Krebs, F., & Widmer, G. (2014). A Multi-model Approach to Beat Tracking Considering Heterogenous Music Styles. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (p. 602-607). Taipei, Taiwan. (pg. 48, 183, 200)
- Böck, S., & Schedl, M. (2011). Enhanced Beat Tracking with Context-aware Neural Networks. In *14th International Conference on Digital Audio Effects (DAFx-11)* (pp. 135–139). Paris, France. (pg. 45, 230)
- Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., ... Serra, X. (2013a, November). ESSENTIA: an Audio Analysis Library for Music Information Retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)* (pp. 493–498). Curitiba, Brazil. (pg. 178, 266)
- Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., ... Serra, X. (2013b, October). ESSENTIA: an Open-Source Library for Sound and Music Analysis. In *Proceedings of the 21st ACM International Conference on Multimedia*

- (*MM'13*) (pp. 855–858). Barcelona, Spain. (pg. 266)
- Brachman, R., & Levesque, H. (2004). *Knowledge Representation and Reasoning (The Morgan Kaufmann Series in Artificial Intelligence)*. Morgan Kaufmann. Hardcover. (pg. 79)
- Cannam, C., Landone, C., & Sandler, M. (2010, October). Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files. In *Proceedings of the ACM Multimedia 2010 International Conference* (pp. 1467–1468). Florence, Italy. (pg. 93, 126, 163)
- Caro, R., & Serra, X. (2014, October). Creating a Corpus of Jingju (Beijing Opera) Music and Possibilities for Melodic Analysis. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (p. 313-318). Taipei, Taiwan. (pg. 110)
- Chandola, A. (1988). *Music as Speech: An Ethnomusicological Study of India*. Navrang. (pg. 250)
- Chen, R., Shen, W., Srinivasamurthy, A., & Chordia, P. (2012, October). Chord Recognition Using Duration-explicit Hidden Markov Models. In *Proceedings of the 13th International Society for Music Information Retrieval Conference* (p. 445-450). Porto, Portugal. (pg. 52)
- Chordia, P. (n.d.). Automatic transcription and representation of solo tabla music. *Computing in Musicology*, 13. (pg. 61)
- Chordia, P. (2005a). *Automatic Transcription of Solo Tabla Music* (Unpublished doctoral dissertation). Stanford University. (pg. 77)
- Chordia, P. (2005b). Segmentation and recognition of tabla strokes. In *Proceedings of 6th International Society for Music Information Retrieval Conference (ISMIR 2005)* (pp. 107–114). (pg. 77)
- Chordia, P., Albin, A., & Sastry, A. (2010). Evaluating Multiple Viewpoints Models of Tabla Sequences. In *Proceedings of acm multimedia workshop on music and machine learning*. (pg. 77, 82)
- Chordia, P., Sastry, A., & Şentürk, S. (2011). Predictive Tabla Modelling Using Variable length Markov and Hidden Markov Models. *Journal of New Music Research*, 40(2), 105–118. (pg. 77)
- Chordia, P., Sastry, A., Mallikarjuna, T., & Albin, A. (2010). Mul-

- tiple viewpoints modeling of tabla sequences. In *Proceedings of International Conference on Music Information Retrieval*. (pg. 77, 82)
- Clarke, E. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *The Psychology of Music* (2nd ed., p. 473-500). Academic Press, San Diego. (pg. 60, 200, 212)
- Clayton, M. (1996). Free Rhythm: Ethnomusicology and the Study of Music Without Metre. *Bulletin of the School of Oriental and African Studies*, 59, 323–332. (pg. 34)
- Clayton, M. (2000). *Time in Indian Music : Rhythm, Metre and Form in North Indian Rag Performance*. Oxford University Press. (pg. 19, 27, 32, 60, 68, 72, 73, 169, 200, 201)
- Cohen, K. B., Ogren, P. V., Fox, L., & Hunter, L. (2005). Empirical data on corpus design and usage in biomedical natural language processing. In *AMIA Annual Symposium Proceedings* (pp. 156–160). (pg. 108)
- Cooper, G., & Meyer, L. B. (1960). *The rhythmic structure of music*. University of Chicago Press. (pg. 18)
- Davies, M. E. P., Degara, N., & Plumley, M. D. (2009, October). Evaluation Methods for Musical Audio Beat Tracking Algorithms. *Technical Report C4DM-TR-09-06, Queen Mary University of London*. (pg. 48)
- Davies, M. E. P., & Plumley, M. D. (2006, September). A spectral difference approach to downbeat extraction in musical audio. In *Proceedings of the 14th European Signal Processing Conference (EUSIPCO 2006)*. Florence, Italy. (pg. 47, 102, 173)
- Davies, M. E. P., & Plumley, M. D. (2007, March). Context-Dependent Beat Tracking of Musical Audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3), 1009–1020. (pg. 44, 69, 94, 285)
- Dixon, S. (2006, September). Onset Detection Revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx '06)* (pp. 133–137). Montreal, Canada. (pg. 43)
- Dixon, S. (2007). Evaluation of The Audio Beat Tracking System Beatroot. *Journal of New Music Research*, 36(1), 39–50. (pg. 43, 44)
- Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Non-linear Filtering*. (pg. 193, 194)

- Dutta, A. E. (1995). *Tabla: Lessons and Practice*. Ali Akbar College. (pg. 32)
- Dutta, S., & Murthy, H. A. (2014, February). A modified rough longest common subsequence algorithm for motif spotting in an Alapana of Carnatic Music. In *Proceedings of the 20th National Conference on Communications (NCC)* (p. 1-6). Kanpur, India. (pg. 252, 253)
- Ellis, D. P. W. (2007). Beat Tracking by Dynamic Programming. *Journal of New Music Research*, 36(1), 51–60. (pg. 44, 69, 100)
- Fitzgerald, D. (2010). Harmonic/percussive separation using median filtering. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*. Graz, Austria. (pg. 43)
- Fitzgerald, D., & Paulus, J. (2006). Unpitched Percussion Transcription. In A. Klapuri & M. Davy (Eds.), *Signal Processing Methods for Music Transcription* (pp. 131–162). Springer US. (pg. 52)
- Fletcher, N. H., & Rossing, T. D. (1998). *The Physics of Musical Instruments*. Springer. (pg. 53)
- Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo 2000* (Vol. 1, pp. 452 – 455). Tokyo, Japan. (pg. 44, 52, 178)
- Foote, J., & Uchihashi, S. (2001). The Beat Spectrum: a new approach to rhythm analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo 2001* (pp. 881 – 884). Tokyo, Japan. (pg. 46)
- Fouloulis, A., Papadelis, G., Pastiadis, K., & Papanikolaou, G. (2010). Estimating Similarity of Musical Rhythm Patterns through the use of a Neural Network Model. In *German Annual Conference on Acoustics (DAGA)* (pp. 199–200). Berlin, Germany. (pg. 51)
- Gainza, M. (2009, April). Automatic musical meter detection. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2009)* (pp. 329–332). Taipei, Taiwan. (pg. 46)
- Gillet, O., & Richard, G. (2004a). Automatic Labelling of Tabla Signals. In *Proceedings of the 4th International Conference*

- on Music Information Retrieval (ISMIR 2004)*. (pg. 77)
- Gillet, O., & Richard, G. (2004b, May). Automatic transcription of drum loops. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)* (Vol. 4, pp. 269–272). Montreal, Canada. (pg. 52, 53, 233)
- Gillet, O., & Richard, G. (2007). Supervised and unsupervised sequence modeling for drum transcription. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 2007)* (pp. 219–224). (pg. 77)
- Gillet, O., & Richard, G. (2008, March). Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3), 529 - 540. (pg. 53, 233)
- Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological Engineering with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. Series: Advanced Information and Knowledge Processing* (1st ed.). Springer. (pg. 79)
- Goto, M., & Muraoka, Y. (1994, May). A Sound Source Separation System for Percussion Instruments. *Transactions of the Institute of Electronics, Information and Communication Engineers D-II, J77-D-II(5)*, 901-911. (pg. 53, 233)
- Gottlieb, R. S. (1993). *Solo Tabla Drumming of North India: Its Repertoire, Styles, and Performance Practices*. Motilal BanarsiDass Publishers. (pg. 30, 32, 246)
- Gouyon, F., & Dixon, S. (2005). A Review of Automatic Rhythm Description Systems. *Computer Music Journal*, 29(1), 34-54. (pg. 40)
- Gouyon, F., Herrera, P., & Cano, P. (2002). Pulse-dependent Analyses of Percussive Music. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. (pg. 53, 233)
- Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C., & Cano, P. (2006). An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5), 1832-1844. (pg. 166)
- Grosche, P., & Müller, M. (2011a). Extracting Predominant Lo-

- cal Pulse Information from Music Recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6), 1688–1701. (pg. 44)
- Grosche, P., & Müller, M. (2011b). Tempogram Toolbox: MATLAB tempo and pulse analysis of music recordings. In *12th International Conference on Music Information Retrieval (ISMIR), late-breaking contribution*). Miami, USA. (pg. 44, 52, 174, 176)
- Gulati, S., Rao, V., & Rao, P. (2011, March). Meter detection from audio for Indian music. In *Proc. of 8th International Symposium on Computer Music Modeling and Retrieval (CMMR)* (pp. 34–43). Bhubaneswar, India. (pg. 69)
- Gulati, S., Rao, V., & Rao, P. (2012). Meter Detection from Audio for Indian Music. In S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen, & S. Mohanty (Eds.), *Speech, Sound and Music Processing: Embracing Research in India: 8th International Symposium, CMMR 2011, 20th International Symposium, FRSM 2011, Bhubaneswar, India, March 9-12, 2011, Revised Selected Papers, Lecture Notes in Computer Science, vol. 7172* (pp. 34–43). Springer: Berlin Heidelberg. (pg. 69, 93, 285)
- Guoyon, F. (2005). *A Computational Approach to Rhythm Description* (Unpublished doctoral dissertation). Universitat Pompeu Fabra, Barcelona, Spain. (pg. 40, 41)
- Gupta, S. (2015). *Discovery of Percussion Patterns from Tabla Solo Recordings* (Master's Thesis). Universitat Pompeu Fabra, Barcelona, Spain. (pg. 159, 245)
- Gupta, S., Srinivasamurthy, A., Kumar, M., Murthy, H., & Serra, X. (2015, October). Discovery of Syllabic Percussion Patterns in Tabla Solo Recordings. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (pp. 385–391). Malaga, Spain. (pg. 16, 159, 245)
- Hainsworth, S., & Macleod, M. (2003, October). Beat tracking with particle filtering algorithms. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (p. 91–94). New Paltz, New York. (pg. 48)
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. (pg. 45)

- Hockman, J. A., Davies, M. E. P., & Fujinaga, I. (2012, October). One in the Jungle: Downbeat Detection in Hardcore, Jungle, and Drum and Bass. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)* (p. 169-174). Porto, Portugal. (pg. 47, 102, 103, 173, 285)
- Holzapfel, A., Davies, M., Zapata, J., Oliveira, J., & Gouyon, F. (2012, November). Selective Sampling for Beat Tracking Evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9), 2539-2548. (pg. 44, 45, 46, 60, 92, 216)
- Holzapfel, A., Flexer, A., & Widmer, G. (2011, July). Improving Tempo-sensitive and Tempo-robust Descriptors for Rhythmic Similarity. In *Proceedings of the Conference on Sound and Music Computing*. Padova, Italy. (pg. 50, 51)
- Holzapfel, A., Krebs, F., & Srinivasamurthy, A. (2014). Tracking the “odd”: Meter inference in a culturally diverse music corpus. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (p. 425-430). Taipei, Taiwan. (pg. 15, 132, 183, 197, 200, 217)
- Holzapfel, A., & Stylianou, Y. (2009, October). Rhythmic Similarity in Traditional Turkish Music. In *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR 2009)* (p. 99-104). Kobe, Japan. (pg. 51)
- Holzapfel, A., & Stylianou, Y. (2011). Scale transform in rhythmic similarity of music. *IEEE Transactions on Speech and Audio Processing*, 19(1), 176-185. (pg. 50, 51, 95, 286)
- Huang, X., & Deng, L. (2010, feb). An Overview of Modern Speech Recognition. In N. Indurkhy & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed., p. 339-366). Chapman and Hall/CRC. (pg. 234)
- Hughes, A., & Gerson-Kiwi, E. (accessed April 28, 2014). Solmization. In *Grove music online. oxford music online*. Oxford University Press. Retrieved from <http://www.oxfordmusiconline.com/subscriber/article/grove/music/26154> (pg. 75)
- Hughes, D. (2000). No nonsense: the logic and power of acoustic-iconic mnemonic systems. *British Journal of Ethnomusicology*, 9(2), 93–120. (pg. 75, 76)

- Huron, D. (2002). Music information processing using the humdrum toolkit: Concepts, examples, and lessons. *Computer Music Journal*, 26(2), 11–26. (pg. 77)
- ISO/TC. (2001). *ISO/TC 15919:2001 Information and documentation – Transliteration of Devanagari and related Indic scripts into Latin characters*. Geneva, Switzerland: International Organization for Standardization. (pg. 90)
- Jehan, T. (2005). *Creating music by listening* (Unpublished doctoral dissertation). Massachusetts Institute of Technology. (pg. 47)
- Jha, R. (2001). *Abhinav Geetanjali Vol. I-V*. Sangeet Sadan Prakashan. (pg. 61, 120)
- Johansen, A., & Doucet, A. (2008). A note on auxiliary particle filters. *Statistics and Probability Letters*, 78(12), 1498–1504. (pg. 194)
- Kapur, A., Benning, M., & Tzanetakis, G. (2004, October). Query by beatboxing: Music information retrieval for the dj. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 2004)*. Barcelona, Spain. (pg. 53)
- Kippen, J., & Bel, B. (1989, June). The identification and modelling of a percussion ‘language,’ and the Emergence of Musical Concepts in a machine-learning experimental set-up. *Computers and the Humanities*, 23(3), 199–214. (pg. 231)
- Klapuri, A. (1999, March). Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999)* (Vol. 6, pp. 3089–3092). Phoenix, USA. (pg. 43)
- Klapuri, A., & Davy, M. (2006). *Signal Processing Methods for Music Transcription*. Springer. (pg. 52)
- Klapuri, A., Eronen, A. J., & Astola, J. T. (2006). Analysis of the Meter of Acoustic Musical Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), 342–355. (pg. 44, 45, 47, 92, 94, 99, 285)
- Koduri, G. K. (2014, November). Culture-aware approaches to modeling and description of intonation using multimodal data. In *International Conference on Knowledge Engineering and Knowledge Management (EKAW)*. Linkoping, Sweden. (pg. 80)

- Koduri, G. K., Ishwar, V., Serrá, J., & Serra, X. (2014). Intonation analysis of ragas in Carnatic music. *Journal of New Music Research*, 43(1), 73–94. (pg. 82)
- Koduri, G. K., Miron, M., Serra, J., & Serra, X. (2011, October). Computational approaches for the understanding of melody in Carnatic Music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)* (pp. 263 – 268). Miami, USA. (pg. 65)
- Koduri, G. K., & Serra, X. (2013, October). A knowledge-based approach to computational analysis of melody in Indian art music. In *International Workshop on Semantic Music and Media, International Semantic Web Conference* (p. 1-10). Sydney, Australia. (pg. 80)
- Kolinski, M. (1973). A cross-cultural approach to metro-rhythmic patterns. *Ethnomusicology*, 17(3), 494-506. (pg. 18)
- Krebs, F., Böck, S., & Widmer, G. (2013, November). Rhythmic Pattern Modeling for Beat- and Downbeat Tracking in Musical Audio. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2014)*. Curitiba, Brazil. (pg. 48, 133, 166, 183, 189, 214)
- Krebs, F., Böck, S., & Widmer, G. (2015, October). An efficient state-space model for joint tempo and meter tracking. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*. Malaga, Spain. (pg. 183, 191)
- Krebs, F., Holzapfel, A., Cemgil, A. T., & Widmer, G. (2015, May). Inferring Metrical Structure in Music Using Particle Filters. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(5), 817-827. (pg. 48, 183, 187, 191, 193, 194, 195, 214)
- Kuriakose, J., Kumar, J. C., Sarala, P., Murthy, H. A., & Sivaraman, U. K. (2015, February). Akshara Transcription of Mru-dangam Strokes in Carnatic Music. In *Proceedings of the 21st National Conference on Communication (NCC)*. Mumbai, India. (pg. 77, 160, 162)
- Lee, Y.-Y., & Shen, S.-Y. (1999). *Chinese Musical Instruments (Chinese Music Monograph Series)*. Chinese Music Society of North America Press. (pg. 36)
- Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal*

- music*. MIT Press Cambridge. (pg. 18, 60)
- Liberman, M., & Cieri, C. (1998, May). The Creation, Distribution and Use of Linguistic Data: The Case of the Linguistic Data Consortium. In *1st International Conference on Language Resources and Evaluation*. Granada, Spain. (pg. 108)
- Ligeti, G. (2007). Brief an Kai Jakobs. In M. Lichtenfeld (Ed.), *Gesammelte Schriften*. Paul Sacher Stiftung. (pg. 18)
- Lin, H., Wu, H., & Wang, C. (2011). Music Matching Based on Rough Longest Common Subsequence. *Journal Information Science and Engineering*, 27(1), 95–110. (pg. 54, 252, 253)
- London, J. (2004). *Hearing in time: Psychological aspects of musical meter*. Oxford: Oxford University Press. (pg. 18)
- London, J. (accessed 19.12.2012). Metre. In L. Macy (Ed.), *Grove music online*. <http://www.grovemusic.com>. (pg. 18)
- Mann, H. B., & Whitney, D. R. (1947, March). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 18(1), 50–60. (pg. 242)
- McKinney, M. F., Moelants, D., Davies, M. E. P., & Klapuri, A. (2007). Evaluation of Audio Beat Tracking and Music Tempo Extraction Algorithms. *Journal of New Music Research*, 36(1), 1–16. (pg. 48, 181)
- Miron, M. (2011). *Automatic Detection of Hindustani Talas* (Master's Thesis). Universitat Pompeu Fabra, Barcelona, Spain. (pg. 32, 69, 77)
- Moelants, D., & McKinney, M. F. (2004). Tempo Perception and Musical Content: What makes a piece fast, slow or temporally ambiguous ? In *Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC)* (p. 558-562). Evanston, IL, USA. (pg. 18)
- Mu(穆文义), W. (2007). *Jingju dajiyue jiqiao yulianxi: yan zou jiaocheng* 京剧打击乐技巧与练习: 演奏教程 (*Technique and practice of Beijing opera percussion music: a performance course*). Beijing: Renmin yinyue chubanshe. (pg. 36, 39)
- Murphy, K. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning* (Unpublished doctoral dissertation). UC Berkeley, Computer Science Division. (pg. 52, 183)
- Naimpalli, S. (2005). *Theory and practice of Tabla*. Popular

- Prakashan. (pg. 32)
- Nakano, T., Ogata, J., Goto, M., & Hiraga, Y. (2004, October). A Drum Pattern Retrieval Method by Voice Percussion. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)* (p. 550-553). (pg. 53, 239)
- Navarro, G. (2001, March). A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, 33(1), 31–88. (pg. 241)
- Ono, N., Miyamoto, K., Le Roux, J., Kameoka, H., & Sagayama, S. (2008, August). Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram. In *16th European Signal Processing Conference (EUSIPCO 2008)*. Lausanne, Switzerland. (pg. 43)
- Pachet, F., & Aucouturier, J. J. (2004). Improving timbre similarity: How high is the sky. *Journal of negative results in speech and audio sciences*, 1(1), 1–13. (pg. 255)
- Pan, S., & Weng, W. (2002). Designing a speech corpus for instance-based spoken language generation. In *Proceedings of the 2nd International Conference on Natural Language Generation* (pp. 49–56). (pg. 108)
- Parncutt, R. (1994). A Perceptual Model of Pulse Salience and Metrical Accent in Musical Rhythms. *Music Perception*, 11(4), 409–464. (pg. 18)
- Parry, M., & Essa, I. (2003). Rhythmic Similarity through Elaboration. In *Proceedings of 4th International Conference on Music Information Retrieval (ISMIR 2003)*. Baltimore, USA. (pg. 51)
- Paulus, J., & Klapuri, A. (2009). Drum Sound Detection in Polyphonic Music with Hidden Markov Models. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009, 497292. (pg. 53)
- Paulus, J., & Virtanen, T. (2005, September). Drum transcription with non-negative spectrogram factorisation. In *Proceedings of the 13th European Signal Processing Conference (EUSIPCO)* (pp. 4–8). Antalya, Turkey. (pg. 52, 53, 233)
- Peeters, G., & Fort, K. (2012, October). Towards a (Better) Definition of the Description of Annotated MIR Corpora. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)* (pp. 25 – 30). Porto,

- Portugal. (pg. 89, 108)
- Peeters, G., & Papadopoulos, H. (2011). Simultaneous Beat and Downbeat-Tracking Using a Probabilistic Framework: Theory and Large-Scale Evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 19(6), 1754-1769. (pg. 45, 47)
- Pikrakis, A., Antonopoulos, I., & Theodoridis, S. (2004, October). Music meter and tempo tracking from raw polyphonic audio. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*. Barcelona, Spain. (pg. 46, 94, 97, 286)
- Pohle, T., Schnitzer, D., Schedl, M., & Knees, P. (2009, October). On rhythm and general music similarity. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)* (pp. 525–530). Kobe, Japan. (pg. 50, 95, 286)
- Porter, A., Sordo, M., & Serra, X. (2013, November). Dunya: A System for Browsing Audio Music Collections Exploiting Cultural Context. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)* (p. 101-106). Curitiba, Brazil. (pg. 4, 83, 123)
- Rabiner, L. R. (1989, February). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE* (Vol. 77, pp. 257–286). (pg. 191)
- Rae, A., & Chordia, P. (2010). Tabla Gyan: An Artificial Tabla Improviser. In *Proceedings of the First International Conference on Computational Creativity (ICCCX)*. (pg. 77)
- Raimond, Y. (2008). *A Distributed Music Information System* (Unpublished doctoral dissertation). University of London. (pg. 80)
- Raman, C. V. (1934). The Indian Musical Drums. *Journal of Mathematical Sciences*, 1(3), 179–188. (pg. 77)
- Raman, C. V., & Kumar, S. (1920). Musical Drums with Harmonic Overtones. *Nature*, 104. (pg. 77)
- Ranjani, H., & Sreenivas, T. (2013, November). Grouping carnatic music notes using a multi- gram language model. In *Proceedings of Acoustics 2013*. New Delhi, India. (pg. 82)
- Ranjani, H., & Sreenivas, T. (2015, April). Multi-instrument detection in polyphonic music using Gaussian Mixture based facto-

- rial HMM. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (p. 191-195). Brisbane, Australia. (pg. 78)
- Ravikiran, C. N. (2008). *Perfecting Carnatic Music, Vol. I-II*. The International Foundation for Carnatic Music, www.ravikiranmusic.com. (pg. 61)
- Sachs, C. (1953). *Rhythm and tempo*. W. W. Norton & Co. (pg. 17)
- Salamon, J., & Gómez, E. (2012, August). Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6), 1759–1770. (pg. 175)
- Sambamoorthy, P. (1998). *South Indian Music Vol. I-VI*. The Indian Music Publishing House. (pg. 21, 23, 24)
- Sarala, P., & Murthy, H. A. (2013, November). Inter and Intra Item Segmentation of Continuous Audio Recordings of Carnatic Music for Archival. In *Proceedings of the 14th International Society for Music Information Retrieval (ISMIR) Conference* (pp. 487–492). Curitiba, Brazil. (pg. 78, 87)
- Sastry, A. (2012). *N-gram Modeling of Tabla Sequences using Variable-length Hidden Markov Models for Improvisation and Composition* (Master's Thesis). Georgia Institute of Technology, Atlanta, USA. (pg. 77)
- Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition* (PhD Thesis). Stanford University. (pg. 43)
- Serra, X. (1997). Musical Sound Modeling with Sinusoids plus Noise. In C. Roads, S. T. Pope, A. Picialli, & G. De Poli (Eds.), *Musical Signal Processing* (pp. 91–122). Swets & Zeitlinger. (pg. 175)
- Serra, X. (2011, October). A multicultural approach in Music Information Research. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)* (pp. 151–156). Miami, USA. (pg. 3, 110)
- Serra, X. (2014, January). Creating Research Corpora for the Computational Study of Music: the case of the CompMusic Project. In *Proceedings of the 53rd AES International Conference on Semantic Audio*. London. (pg. 5, 89, 108, 271)
- Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer,

- A., ... Widmer, G. (2013). *Roadmap for Music Information ReSearch*. Retrieved from [http://www.mires.cc/sites/default/files/MIRES\\_Roadmap\\_ver\\_1.0.0.pdf](http://www.mires.cc/sites/default/files/MIRES_Roadmap_ver_1.0.0.pdf) (pg. 5)
- Shankar, V. (1999). *The art and science of carnatic music*. Parampara. (pg. 17)
- Smaragdis, P. (2004a). Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs. In C. G. Puntonet & A. Prieto (Eds.), *Independent Component Analysis and Blind Signal Separation* (Vol. 3195, p. 494-499). Springer Berlin Heidelberg. (pg. 53, 233)
- Smaragdis, P. (2004b, September). Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs. *Technical Report TR2004-104, Mitsubishi Electric Research Laboratories*. (pg. 53)
- Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 2, pp. 629–633). Washington, DC, USA. (pg. 158)
- Srinivasamurthy, A., Caro, R., Sundar, H., & Serra, X. (2014, October). Transcription and Recognition of Syllable based Percussion Patterns: The Case of Beijing Opera. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (pp. 431–436). Taipei, Taiwan. (pg. 16, 165, 239)
- Srinivasamurthy, A., & Chordia, P. (2012a, June). Multiple Viewpoint Modeling of North Indian Classical Vocal Compositions. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)* (p. 344-356). London, UK. (pg. 82)
- Srinivasamurthy, A., & Chordia, P. (2012b, July). A Unified System for Analysis and Representation of Indian Classical Music using Humdrum Syntax. In *Proceedings of the 2nd Comp-Music Workshop* (p. 38-42). Istanbul, Turkey. (pg. 61)
- Srinivasamurthy, A., Holzapfel, A., Cemgil, A. T., & Serra, X. (2015, October). Particle Filters for Efficient Meter Tracking with Dynamic Bayesian Networks. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (pp. 197–203). Malaga, Spain. (pg. 15, 183, 217, 229)

- Srinivasamurthy, A., Holzapfel, A., Cemgil, A. T., & Serra, X. (2016, March). A generalized Bayesian model for tracking long metrical cycles in acoustic music signals. In *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*. Shanghai, China. (pg. 15, 183, 200, 204, 217)

Srinivasamurthy, A., Holzapfel, A., & Serra, X. (2014). In Search of Automatic Rhythm Analysis Methods for Turkish and Indian Art Music. *Journal of New Music Research*, 43(1), 97–117. (pg. 3, 7, 14, 71, 91, 165)

Srinivasamurthy, A., Koduri, G. K., Gulati, S., Ishwar, V., & Serra, X. (2014, September). Corpora for Music Information Research in Indian Art Music. In *Proceedings of Joint International Computer Music Conference/Sound and Music Computing Conference*. Greece. (pg. 14)

Srinivasamurthy, A., & Serra, X. (2014, May). A Supervised Approach to Hierarchical Metrical Cycle Tracking from Audio Music Recordings. In *Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)* (pp. 5237–5241). Florence, Italy. (pg. xxvi, 15, 174, 268)

Srinivasamurthy, A., Subramanian, S., Tronel, G., & Chordia, P. (2012, July). A Beat Tracking Approach to Complete Description of Rhythm in Indian Classical Music. In *Proceedings of the 2nd CompMusic Workshop* (pp. 72–78). Istanbul, Turkey. (pg. 69, 94, 286)

Sundberg, J., Gu, L., Huang, Q., & Huang, P. (2012, March). Acoustical study of classical Peking Opera singing. *Journal of Voice*, 26(2), 137–143. (pg. 84)

Swartz, A. (2002, January). MusicBrainz: a semantic Web service. *IEEE Intelligent Systems*, 17(1), 76–77. (pg. 80)

T. K. Govinda Rao. (2003a). *Compositions of Muddusvāmi Dīkshitar*. Ganamandir Publications. (pg. 114)

T. K. Govinda Rao. (2003b). *Compositions of Śyāmā Śāstri, Subbaraya Śāstri and Anṇasvāmi Śāstri*. Ganamandir Publications. (pg. 114)

T. K. Govinda Rao. (2009). *Compositions of Tyāgarāja*. Ganamandir Publications. (pg. 113, 114)

Thoshkahna, B., & Ramakrishnan, K. R. (2011, November). A Probabilistic Model for Tracking Metrical Cycles in Indian Classical Music. In *Proceedings of the 2011 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*. Prague, Czech Republic. (pg. 113, 114)

- ber). A Postprocessing Technique for Improved Harmonic/Percussion Separation for Polyphonic Music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)* (p. 251-256). Miami, USA. (pg. 43)
- Tian, M., Srinivasamurthy, A., Sandler, M., & Serra, X. (2014, May). A Study of Instrument-wise Onset Detection in Beijing Opera Percussion Ensembles. In *Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)* (pp. 2174–2178). Florence, Italy. (pg. 37, 164, 233)
- Toussaint, G. T. (2004). A comparison of rhythmic similarity measures. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*. Barcelona, Spain. (pg. 50)
- Typke, R., Wiering, F., & Veltkamp, R. C. (2005, September). A Survey of Music Information Retrieval Systems. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)* (pp. 153–160). London, UK. (pg. 54)
- Uhle, C., & Herre, J. (2003, September). Estimation of Tempo, Micro Time and Time Signature from Percussive Music. In *Proceedings of 6th International Conference on Digital Audio Effects (DAFX-03)*. London, UK. (pg. 46)
- Verma, P., Vinutha, T. P., Pandit, P., & Rao, P. (2015, April). Structural segmentation of Hindustani concert audio with posterior features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (p. 136-140). Brisbane, Australia. (pg. 78)
- Vermaak, J., Doucet, A., & Pérez, P. (2003, October). Maintaining multimodality through mixture tracking. In *Proceedings of the 9th IEEE International Conference on Computer Vision* (pp. 1110–1116). Nice, France. (pg. 194)
- Vinutha, T. P., & Rao, P. (2014, March). Audio Segmentation of Hindustani Music Concert Recordings. In *Proceedings of the International Symposium, Frontiers of Research on Speech and Music (FRSM)*. Mysore, India. (pg. 78)
- Vinutha, T. P., Sankagiri, S., & Rao, P. (2016, March). Reliable Tempo Detection for Structural Segmentation in Sarod Concerts. In *Proceedings of the National Conference on Commu-*

- nications.* Guwahati, India. (pg. 78)
- Viswanathan, T., & Allen, M. H. (2004). *Music in South India*. Oxford University Press. (pg. 21)
- Vos, J., & Rasch, R. (1981). The perceptual onset of musical tones. *Perception & Psychophysics*, 29(4), 323–335. (pg. 42)
- Whiteley, N., Cemgil, A. T., & Godsill, S. (2006, October). Bayesian modelling of temporal structure in musical audio. In *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)*. Victoria, Canada. (pg. 183)
- Whiteley, N., Cemgil, A. T., & Godsill, S. (2007, April). Sequential Inference of Rhythmic Structure in Musical Audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)* (Vol. 4, pp. 1321–1325). Honolulu, USA. (pg. 183)
- Wichmann, E. (1991). *Listening to Theatre: The Aural Dimension of Beijing Opera*. University of Hawaii Press, Honolulu. (pg. 36)
- Widdess, R. (1994). Involving the Performers in Transcription and Analysis: A Collaborative Approach to Dhrupad. *Ethnomusicology*, 38(1), 59–79. (pg. 34)
- Wu, F.-H. F., Lee, T.-C., Jang, J.-S. R., Chang, K. K., Lu, C.-H., & Wang, W.-N. (2011, October). A Two-Fold Dynamic Programming Approach to Beat Tracking for Audio Music with Time-Varying Tempo. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)* (pp. 191–196). Miami, USA. (pg. 44, 177)
- Wynne, M. (Ed.). (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. (pg. 108)
- Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., ... Woodland, P. C. (2006). *The HTK book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department. (pg. 240, 251)
- Zapata, J., & Gómez, E. (2013, May). Using Voice Suppression Algorithms to improve Beat Tracking in the Presence Of Highly Predominant Vocals. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)* (pp. 51–55). Vancouver, Canada. (pg. 46, 175)
- Zapata, J. R., Holzapfel, A., Davies, M. E. P., Oliveira, J. L., &

- Gouyon, F. (2012, October). Assigning a Confidence Threshold on Automatic Beat Annotation in Large Datasets. In *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR 2012)* (pp. 157–162). (pg. 46, 181)
- Zhang, Y., & Zhou, J. (2003, July). A study on content-based music classification. In *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications* (Vol. 2, pp. 113–116). Paris, France. (pg. 84)



---

# Index

Bayesian models, 54

CompMusic, 3

Eurogenetic music, 3

Meter, 18

Meter analysis, 169