

A Data-driven Bayesian Approach to Automatic Rhythm Analysis of Indian Art Music

Ajay Srinivasamurthy

TESI DOCTORAL UPF / 2016

Director de la tesi

Dr. Xavier Serra Casals

Music Technology Group

Dept. of Information and Communication Technologies



Copyright © 2016 by Ajay Srinivasamurthy

<http://www.ajaysrinivasamurthy.in/phd-thesis>

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

Licensed under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0



You are free to share – to copy and redistribute the material in any medium or format under the following conditions:

- **Attribution** – You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial** – You may not use the material for commercial purposes.
- **NoDerivatives** – If you remix, transform, or build upon the material, you may not distribute the modified material.

The doctoral defense was held on at the Universitat Pompeu Fabra and scored as

Dr. Xavier Serra Casals
(Thesis Supervisor)
Universitat Pompeu Fabra (UPF), Barcelona

Dr. Simon Dixon
(Thesis Committee Member)
Queen Mary University of London (QMUL), London

Dr. Geoffroy Peeters
(Thesis Committee Member)
Institut de Recherche et Coordination Acoustique/Musique
(IRCAM), Paris

Dr. Juan Pablo Bello
(Thesis Committee Member)
New York University (NYU), New York

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

This thesis has been carried out between Oct. 2012 and Jun. 2016 at the Music Technology Group (MTG) of Universitat Pompeu Fabra (UPF) in Barcelona (Spain), supervised by Dr. Xavier Serra Casals. The work in Chapter 3 and Chapter 5 has been conducted in collaboration with Dr. Andre Holzapfel (The Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria) and Dr. Ali Taylan Cemgil (Boğaziçi University, Istanbul, Turkey). The work in Chapter 4 and Chapter 6 has been conducted in collaboration with the Comp-Music team at MTG and research partners Dr. Hema Murthy (Indian Institute of Technology Madras, Chennai, India) and Dr. Preeti Rao (Indian Institute of Technology Bombay, Mumbai, India). This work has been supported by the Dept. of Information and Communication Technologies (DTIC) PhD fellowship (2012-16), Universitat Pompeu Fabra and the European Research Council under the European Union's Seventh Framework Program, as part of the [CompMusic](#) project (ERC grant agreement 267583).

Acknowledgements

Someone said four years is a long time, but for me it has gone past in a blink and has left me wanting more! The main reason I believe for that is the group of people I have been fortunate to be a part of. I am extremely grateful to my advisor Xavier Serra for the direction, guidance, mentoring, support and freedom he has given me. His research vision, initiative, perspective and far-sightedness are some qualities that leave me in awe every time. I also thank him for his initiative of conceiving and leading the CompMusic project, and the opportunity he gave me to be a part of it.

I wish to express a deep sense of gratitude to Andre Holzapfel, who has guided me throughout our continuous collaboration. I have greatly benefited from his multidisciplinary experience and a deep understanding of rhythm in music cultures around the world. It is a great joy to work with him and learn from his meticulous approach to research spanning diverse areas. When I was stuck in a research problem, he would surely know what was to be done next! His second PhD is a sign of his steadfastness and is a great inspiration for me! I would like to thank Taylan Cemgil for his crucially important ideas stemming from his vast experience with Bayesian methods and inference. I also thank Taylan Cemgil and Andre Holzapfel for inviting me for a short research stay at Boğaziçi University to work with them.

I thank Hema Murthy and Preeti Rao for their pioneering contributions in CompMusic, and for their guidance, support and collaboration on several research topics. I also thank Barış Bozkurt

for his ideas and fruitful discussions. I sincerely thank all my collaborators and co-authors Mark Sandler, Mi Tian, Manoj Kumar and Harshavardhan for their contribution. Special thanks to Martin Clayton for his interest in technology tools for rhythm analysis, guidance on musicological aspects of the work, and a review of parts of the dissertation to provide musicological insights.

I have had the best experience working with the CompMusic team, my culturally diverse research family that supported me both professionally and personally. My personal thanks go to Sankalp Gulati (the dependable guy next seat who can make even a block of wood sound good and even a lump of salt taste good), Gopala Krishna Koduri (the guru of discipline and planning), Sertan Şentürk (Otomobil), Vignesh Ishwar (cooking and singing, up for both), Kaustuv Kanti Ganguli (best Kedar and Kesaribhath ever), Rafael Caro (who spreads laughter contagiously), Swapnil Gupta (the fellow *tāl* companion), Mohamed Sordo (the walking-talking experience), Rong Gong (chef de cuisine), Georgi Dzhambazov (the fellow graphical models explorer), Alastair Porter (if he doesn't know, nobody else will), Andres Ferraro, Yile Yang, Vinutha Prasad, Shuo Zhang, Padi Sarala, Hasan Sercan Atlı, Joe Cheri Ross, Sridharan Sankaran, Akshay Anantapadmanabhan, Jom Kuriakose, Jilt Sebastian and Amruta Vidwans. I also thank the collaborators of CompMusic T M Krishna, Amin Chachoo, Suvarnalata Rao for helping us formulate research problems and correcting us with feedback.

MTG has been an incubator for several of my thoughts and ideas, fueled by encouraging conversations and discussions with many of its past and present members: Sergi Jordà, Emilia Gómez, Rafael Ramírez, Jordi Bonada, Perfecto Herrera, Agustín Martorell, Marius Miron, Oscar Mayor, Frederic Font, Sebastián Maella, Panos Papiotis, Sergio Giraldo, Sergio Oramas, Jordi Pons, Dmitry Bogdanov, Esteban Maestre, Carles Julià, Juan José Bosch, Nadine Kroher, Cárthach Ó Nuanáin, Álvaro Sarasúa, Zacharias Vamvakousis, Julio Carabias, Giuseppe Bandiera, Martí Umbert and Daniel Gomez. Many thanks to all mentioned here and others I might have missed. Special thanks to Rafael and Jordi for their help in translating the abstract and showing how much my Catalan and Spanish can be improved! Thanks to Cristina Garrido, Sonia

Espí, Alba Rosado, Vanessa Jimenez, Jana Safrankova and Lydia García for their reassuring support in hooping through piles of paperwork.

I gratefully acknowledge Florian Krebs, Sebastian Böck and Gerhard Widmer for their inputs on this work and for providing me access to their code and data. Special thanks to Juan Bello and Carlos Guedes for the opportunity to attend the rhythm workshop at NYUAD (2014) and broaden my horizons with a multitude of perspectives on rhythm. On my journey so far, I have interacted and learned from several more researchers who have been eager to give their inputs and share ideas - (ordered alphabetically by last name) Jean-Julian Aucouturier, Emmanouil Benetos, Matthew Davies, Simon Dixon, Simon Durand, Gyorgy Fazekas, Fabien Gouyon, Shantala Hegde, Robert Kaye, Andres Lewin-Richter, Meinard Müller, Gautham Mysore, Luiz Naveda, Uri Nieto, Maria Panteli, Geoffroy Peeters, Colin Raffel, Ranjani H G, Robert Reigle, Martin Rocamora, Gerard Roma, Justin Salamon, Joan Serrà, Bill Sethares, Bob Sturm, Julian Urbano, Anja Volk, Frans Weiring and Jose Zapata. My work has been influenced by them in more ways than one, many thanks to all of them and others in the MIR community. I am specially grateful to Dan Ellis, Anssi Klapuri, Aggelos Pikrakis, Matthew Davies and Jason Hockman for providing their code and implementation of the algorithms for evaluation.

A note of thanks to ERC, DTIC-UPF and Erasmus+ for funding parts of the thesis work. Humble thanks to the maestros of music who have kept the Indian music cultures alive through generations, and the present day music community for bring receptive and providing a context for this work. My sincere thanks go to Vid. G S Nagaraj, to whom I owe all my rhythm learning.

I wholeheartedly thank my flatmates over the years at Mallorca 529: Alastair, Gopal, Sankalp and Shefali (a life coach whose words are collectables), all of whom can be nominated to the best flatmate awards. Special thanks to Eva, whose sense of care and concern goes beyond my comprehension. Thanks to Ratheesh, Geordie, Manoj, Kalpani, Windhya, Waqas, Praveen, Princy, Pal-labi and Srinibas for making sunny Barcelona sunnier.

I wish to express a deep sense of gratitude to my parents in In-

dia for their unflinching support, daily reminders towards overall well-being, and for bearing with me even during the weeks-long communication blackouts during submission deadlines. Thanks to Sunanda, Veena (the benevolent devil's advocate of my work), Rashmi and friend-philosopher Srinivas for their patient effort to understand my work. Special thanks to my guide-friend-brother Sharath, who stood 7 hours ahead of me to shield and guide me through matters small and big. And to Soumya, who will be a co-author on all our life's publications in the time to come!

Unlike some journeys that end, the research journey never does . . .

Ajay Srinivasamurthy

29th June 2016

Abstract

Large and growing collections of a wide variety of music are now available on demand to music listeners, necessitating novel ways of automatically structuring these collections using different dimensions of music. Rhythm is one of the basic music dimensions and its automatic analysis, which aims to extract musically meaningful rhythm related information from music, is a core task in Music Information Research (MIR).

Musical rhythm, similar to most musical dimensions, is culture-specific and hence its analysis require culture-aware approaches. Indian art music is one of the major music traditions of the world and has complexities in rhythm that have not been addressed by the current state of the art in MIR, motivating us to choose it as the primary music tradition for study. Our intent is to address unexplored rhythm analysis problems of Indian art music to push the boundaries of the current MIR approaches by making them culture-aware and generalizable to other music traditions.

The thesis aims to build data-driven signal processing and machine learning approaches for automatic analysis, description and discovery of rhythmic structures and patterns in audio music collections of Indian art music. After identifying challenges and opportunities, we present several relevant research tasks that open up the field of automatic rhythm analysis of Indian art music. Data-driven approaches require well curated data corpora for research and efforts towards creating such corpora and datasets are documented in detail. We then focus on the topics of meter analysis and

percussion pattern discovery in Indian art music.

Meter analysis aims to align several hierarchical metrical events with an audio recording. Meter analysis tasks such as meter inference, meter tracking and informed meter tracking are formulated for Indian art music. Different Bayesian models that can explicitly incorporate higher level metrical structure information are evaluated for the tasks and novel extensions are proposed. The proposed methods overcome the limitations of existing approaches and their performance indicate the effectiveness of informed meter analysis.

Percussion in Indian art music uses onomatopoeic oral mnemonic syllables for the transmission of repertoire and technique, providing a language for percussion. We use these percussion syllables to define, represent and discover percussion patterns in audio recordings of percussion solos. We approach the problem of percussion pattern discovery using hidden Markov model based automatic transcription followed by an approximate string search using a data derived percussion pattern library. Preliminary experiments on Beijing Opera percussion patterns, and on both tabla and mridangam solo recordings in Indian art music demonstrate the utility of percussion syllables, identifying further challenges to building practical discovery systems.

The technologies resulting from the research in the thesis are a part of the complete set of tools being developed within the Comp-Music project for a better understanding and organization of Indian art music, aimed at providing an enriched experience with listening and discovery of music. The data and tools should also be relevant for data-driven musicological studies and other MIR tasks that can benefit from automatic rhythm analysis.

Resum

Les col·leccions de música són cada vegada més grans i variades, fet que fa necessari buscar noves fòrmules per a organitzar automàticament aquestes col·leccions. El ritme és una de les dimensions bàsiques de la música, i el seu anàlisi automàtic és una de les principals àrees d'investigació en la disciplina de l'recupерació de la informació musical (MIR, acrònim de la traducció a l'anglès).

El ritme, com la majoria de les dimensions musicals, és específic per a cada cultura i per tant, el seu anàlisi requereix de mètodes que incloguin el context cultural. La complexitat rítmica de la música clàssica de l'Índia, una de les tradicions musicals més grans al món, no ha estat encara treballada en el camp d'investigació de MIR - motiu pel qual l'escollim com a principal material d'estudi. La nostra intenció és abordar les problemàtiques que presenta l'anàlisi rítmic de la música clàssica de l'Índia, encara no tractades en MIR, amb la finalitat de contribuir en la disciplina amb nous models sensibles al context cultural i generalitzables a altres tradicions musicals.

L'objectiu de la tesi consisteix en desenvolupar tècniques de processament de senyal i d'aprenentatge automàtic per a l'anàlisi, descripció i descobriment automàtic d'estructures i patrons rítmics en col·leccions de música clàssica de l'Índia. Després d'identificar els reptes i les oportunitats, així com les diverses tasques d'investigació rellevants per a aquest objectiu, detallarem el procés d'elaboració del corpus de dades, fonamentals per als mètodes basats en dades. A continuació, ens centrem en les tasques

d'anàlisis mètric i descobriment de patrons de percussió.

L'anàlisi mètric consisteix en alinear els diversos esdeveniments mètrics -a diferents nivells- que es produueixen en una gravació d'àudio. En aquesta tesi formulem les tasques de deducció, seguiment i seguiment informat de la mètrica. D'acord amb la tradició musical estudiada, s'avaluen diferents models bayesianos que poden incorporar explícitament estructures mètriques d'alt nivell i es proposen noves extensions per al mètode. Els mètodes proposats superen les limitacions dels mètodes ja existents i el seu rendiment indica l'efectivitat dels mètodes informats d'anàlisis mètric.

La percussió en la música clàssica de l'Índia utilitz a onomatopeies per a la transmissió del repertori i de la tècnica, fet que construeix un llenguatge per a la percussió. Utilitzem aquestes sílabes percussives per a definir, representar i descobrir patrons en enregistraments de solos de percussió. Enfoquem el problema del descobriment de patrons percussius amb un model de transcripció automàtica basat en models ocults de Markov, seguida d'una recerca aproximada de strings utilitzant una llibreria de patrons de percussions derivada de dades. Experiments preliminars amb patrons de percussió d'òpera de Pequín, i amb gravacions de solos de tabla i mridangam, demostren la utilitat de les sílabes percussives. Identificant, així, nous horitzons per al desenvolupament de sistemes pràctics de descobriment.

Les tecnologies resultants d'aquesta recerca són part de les eines desenvolupades dins el projecte de CompMusic, que té com a objectiu millorar l'experiència d'escoltar i descobrir música per a la millor comprensió i organització de la música clàssica de l'Índia, entre d'altres. Aquestes dades i eines poden ser rellevants per a estudis musicològics basats en dades i, també, altres tasques MIR poden beneficiar-se de l'anàlisi automàtic del ritme.

Resumen

Las colecciones de música son cada vez mayores y más variadas, haciendo necesarias nuevas fórmulas para su organización automática. El análisis automático del ritmo tiene como fin la extracción de información rítmica de grabaciones musicales y es una de las principales áreas de investigación en la disciplina de recuperación de la información musical (MIR por sus siglas en inglés).

La dimensión rítmica de la música es específica a una cultura y por tanto su análisis requiere métodos que incluyan el contexto cultural. Las complejidades rítmicas de la música clásica de la India, una de las mayores tradiciones musicales del mundo, no han sido tratadas hasta la fecha en MIR, motivo por el cual la elegimos como nuestro principal objeto de estudio. Nuestra intención es abordar cuestiones de análisis rítmico aún no tratadas en MIR con el fin de contribuir a la disciplina con nuevos métodos sensibles al contexto cultural y generalizables a otras tradiciones musicales.

El objetivo de la tesis es el desarrollo de técnicas de procesamiento de señales y aprendizaje automático dirigidas por datos para el análisis, descripción y descubrimiento automáticos de estructuras y patrones rítmicos en colecciones de audio de música clásica de la India. Tras identificar retos y posibilidades, así como varias tareas de investigación relevantes para este objetivo, detallamos la elaboración del corpus de estudio y conjuntos de datos, fundamentales para métodos dirigidos por datos. A continuación, nos centramos en las tareas de análisis métrico y descubrimiento de patrones de percusión.

El análisis métrico consiste en la alineación de eventos métricos a diferentes niveles con una grabación de audio. En la tesis formulamos las tareas de deducción de metro, seguimiento de metro y seguimiento informado de metro de acuerdo a la tradición estudiada, se evalúan diferentes modelos bayesianos capaces de incorporar explícitamente información de estructuras métricas de niveles superiores y se proponen nuevas extensiones. Los métodos propuestos superan las limitaciones de las propuestas existentes y los resultados indican la efectividad del análisis informado de metro.

La percusión en la música clásica de la India utiliza onomatopeyas para la transmisión del repertorio y la técnica. Utilizamos estas sílabas para definir, representar y descubrir patrones en grabaciones de solos de percusión. A tal fin generamos una transcripción automática basada en un modelo oculto de Márkov, seguida de una búsqueda aproximada de subcadenas usando una biblioteca de patrones de percusión derivada de datos. Experimentos preliminares en patrones de percusión de ópera de Pekín, y en grabaciones de solos de tabla y mridangam, demuestran la utilidad de estas sílabas, identificando nuevos retos para el desarrollo de sistemas prácticos de descubrimiento.

Las tecnologías resultantes de esta investigación son parte de un conjunto de herramientas desarrollado en el proyecto CompMusic para el mejor entendimiento y organización de la música clásica de la India, con el objetivo de proveer una experiencia mejorada de escucha y descubrimiento de música. Estos datos y herramientas pueden ser también relevantes para estudios musicológicos dirigidos por datos y otras tareas de MIR que puedan beneficiarse de análisis automáticos de ritmo.

Contents

Abstract	xi
Contents	xvii
List of Symbols	xxi
List of Figures	xxv
List of Tables	xxix
1 Introduction	1
1.1 Context and relevance	3
1.2 Motivation	5
1.3 Scope and objectives	8
1.4 Organization and thesis outline	13
2 Background	19
2.1 Rhythm: terminology	19
2.2 Music background	21
2.2.1 Indian art music	22
2.2.2 Rhythm and percussion in Carnatic music .	24
2.2.3 Rhythm and percussion in Hindustani music	29
2.2.4 Carnatic and Hindustani music: Comparison	36
2.2.5 Percussion in Beijing opera	38
2.3 A review of automatic rhythm analysis	43

2.3.1	Onset detection	43
2.3.2	Tempo estimation	47
2.3.3	Beat tracking	48
2.3.4	Time signature estimation	50
2.3.5	Downbeat tracking	51
2.3.6	Meter tracking	52
2.3.7	Evaluation measures	52
2.3.8	Rhythm similarity measures	54
2.3.9	Domain-specific approaches	56
2.3.10	Percussion pattern analysis	56
2.4	Relevant technical concepts	58
2.4.1	Bayesian models	58
2.4.2	Speech recognition technologies and tools .	60
3	Automatic rhythm analysis of Indian art music	63
3.1	Challenges and Opportunities	64
3.1.1	Challenges	64
3.1.2	Opportunities	68
3.1.3	Characteristics of Indian Art Music	70
3.2	Research problems in rhythm analysis	71
3.2.1	Building data corpora	73
3.2.2	Automatic rhythm annotation	74
3.2.3	Rhythm and percussion pattern analysis	81
3.2.4	Rhythm based audio segmentation	86
3.2.5	Ontologies for rhythm concepts	88
3.2.6	Rhythm similarity measures	88
3.2.7	Symbolic music analysis	90
3.2.8	Evaluation and Integration	91
3.2.9	Extensions to other music cultures	92
3.3	Thesis problems: A formulation	93
3.3.1	Meter inference and tracking	93
3.3.2	Percussion pattern transcription and discovery	95
3.3.3	Datasets for research	97
3.4	In search of automatic rhythm analysis methods	99
3.4.1	Cycle length estimation	101
3.4.2	Downbeat tracking	109
3.4.3	Discussion	111

4 Data corpora for research	115
4.1 CompMusic research corpora	118
4.1.1 Criteria for creation of research corpora	118
4.1.2 Carnatic music research corpus	120
4.1.3 Hindustani music research corpus	128
4.1.4 Creative Commons music collections	131
4.2 Test datasets	133
4.2.1 Carnatic music rhythm dataset	133
4.2.2 Hindustani music rhythm dataset	147
4.2.3 Tabla solo dataset	167
4.2.4 Mridangam datasets	168
4.2.5 Jingju percussion instrument dataset	171
4.2.6 Jingju percussion pattern dataset	172
4.2.7 Other evaluation datasets	174
5 Meter inference and tracking	177
5.1 The meter analysis tasks	178
5.2 Preliminary experiments	182
5.2.1 Meter tracking using dynamic programming	182
5.3 Bayesian models for meter analysis	191
5.3.1 The bar pointer model	192
5.3.2 Model extensions	205
5.3.3 Inference extensions	213
5.4 Experiments and results	218
5.4.1 Experimental setup	220
5.4.2 Meter inference	224
5.4.3 Meter tracking	226
5.4.4 Informed meter tracking	231
5.4.5 Summary of results	234
5.5 Conclusions	238
6 Percussion pattern transcription and discovery	243
6.1 Approaches	244
6.2 The case of Beijing opera	247
6.2.1 Percussion pattern classification	250
6.2.2 Results and discussion	254
6.3 The case of Indian art music	257
6.3.1 Pattern library generation	258
6.3.2 Automatic transcription	260

6.3.3	Approximate pattern search	263
6.3.4	Results and discussion	267
6.4	Conclusions	275
7	Applications, Summary and Conclusions	277
7.1	Applications	277
7.1.1	Dunya	280
7.1.2	Sarāga	282
7.2	Contributions	284
7.3	Conclusions and Summary	286
7.4	Future directions	289
Appendix A	List of Publications	293
Appendix B	Resources	297
Appendix C	Glossary	303
Bibliography		309
Index		331

List of Symbols

The following is a list of different symbols used in the dissertation along with a short description of each symbol.

Symbol	Description
Audio Signal and Music	
t	Time variable (in seconds)
k	Time frame index
$f[n]$	Discrete audio signal
$f_p[n]$	Percussion enhanced audio signal
T_f	Time duration of an audio recording
h	Hop size used for audio analysis
K	Number of frames in an audio recording
z	Audio music recording index
\mathcal{B}_z	Set of all beats in an audio music recording z
\mathcal{S}_z	Set of all <i>samas</i> /downbeats in an audio music recording z
\mathcal{O}_z	Set of all <i>akşaras</i> /tatum pulses in an audio music recording z
τ_s	Inter- <i>sama</i> interval
τ_b	Inter-beat/ <i>mātrā</i> interval
τ_o	Inter- <i>akşara</i> interval or <i>akşara</i> pulse period

Symbol	Description
F_0	Fundamental frequency
BPM	beats per minute
MPM	matras per minute
B	Number of beats in a bar/ tāla cycle
Meter Analysis	
\mathbf{y}_k	Audio feature vector at frame k
\mathbf{x}_k	A vector of latent variables at frame k
\mathbf{G}	The tempogram matrix
M	Total metrical position states
ϕ	Position in bar/section/metrical cycle
m	Position in bar/section/metrical cycle (discrete)
N	Total tempo states
$\dot{\phi}$	Rate of change of position in bar/section/metrical cycle
n	Rate of change of position in bar/section/metrical cycle (discrete)
R	Number of rhythm patterns
r	Rhythm pattern indicator variable
V	Number of metrical sections
v	Section (vibhāg) indicator variable
\mathbb{A}	Rhythmic pattern transition matrix
\mathbb{B}	Section transition matrix
$\mathbb{1}$	Indicator function
N_p	Number of particles in the particle filter
w	The weight of a particle
\mathbf{w}	A vector of particle weights
c	Cluster assignment variable in mixture particle filters
T_s	Resampling interval for systematic resampling
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

Symbol	Description
Percussion Pattern Analysis	
\mathcal{A}	The set of all percussion syllables
N_s	The total number of percussion syllables
A	A percussion syllable in the set \mathcal{A}
a	An instance of a percussion syllable
\mathbf{A}	A syllabic percussion pattern
L	Length of a percussion pattern (in syllables)
\mathcal{P}	The set of all syllabic percussion patterns
N_a	The total number of percussion patterns
Λ	The set of all HMM parameters
\mathbb{R}	RLCS Width Across Reference (WAR) matrix
\mathbb{Q}	RLCS Width Across Query (WAQ) matrix
\mathbb{H}	RLCS length of match matrix
σ	RLCS weight matrix
Evaluation measures	
O	The big O notation (Order of the function)
Θ	The <i>overlap</i> measure for data corpora
\mathfrak{p}	Precision
\mathfrak{r}	Recall
\mathfrak{f}	f-measure
\mathfrak{I}	Information Gain
CML_t	Correct Metrical Level (No continuity) measure
AML_t	All Metrical Levels (No continuity) measure
\mathfrak{C}	Correctness measure for automatic transcription
\mathfrak{A}	Accuracy measure for automatic transcription
I	String insertion error
D	String deletion error
S	String substitution error
T	String transposition error

List of Figures

1.1	Automatic rhythm analysis from audio	10
1.2	Dependencies between the chapters of the dissertation	16
2.1	Four popular Carnatic tālas	25
2.2	Structure of tiśra nađe ādi tāla	27
2.3	Four popular Hindustani tāls	31
2.4	Ēktāl in dṛt lay	32
2.5	Percussion patterns in jingju	41
2.6	Functional units of a rhythm description system	44
3.1	Audio signal characteristics of Indian art music	70
3.2	Relevant automatic rhythm analysis problems in Indian art music	72
4.1	The number of artists by the number of their performances in the Carnatic music corpus	124
4.2	Coverage of the Carnatic artists	126
4.3	Histogram of τ_s in the CMR _f dataset	136
4.4	Histogram of τ_b in the CMR _f dataset	137
4.5	Histogram of median normalized τ_s in the CMR _f dataset	138
4.6	Histogram of median normalized τ_b in the CMR _f dataset	139
4.7	Computation of the spectral flux feature	141
4.8	Rhythm patterns in ādi tāla learned from CMR _f dataset .	143
4.9	Rhythm patterns in ādi tāla learned from CMR dataset .	143
4.10	Rhythm patterns in rūpaka tāla learned from CMR _f dataset	144

4.11	Rhythm patterns in rūpaka tāla learned from CMR dataset	144
4.12	Rhythm patterns in miśra chāpu tāla learned from CMR _f dataset	145
4.13	Rhythm patterns in miśra chāpu tāla learned from CMR dataset	145
4.14	Rhythm patterns in khaṇḍa chāpu tāla learned from CMR _f dataset	146
4.15	Rhythm patterns in khaṇḍa chāpu tāla learned from CMR dataset	146
4.16	Histogram of τ_s in the HMR _I dataset	151
4.17	Histogram of τ_b in the HMR _I dataset	152
4.18	Histogram of τ_s in the HMR _s dataset	153
4.19	Histogram of τ_b in the HMR _s dataset	154
4.20	Histogram of median normalized τ_s in the HMR _I dataset	155
4.21	Histogram of median normalized τ_b in the HMR _I dataset	156
4.22	Histogram of median normalized τ_s in the HMR _s dataset	157
4.23	Histogram of median normalized τ_b in the HMR _s dataset	158
4.24	Rhythm patterns in tīntāl learned from HMR _I dataset	161
4.25	Rhythm patterns in tīntāl learned from HMR _s dataset	161
4.26	Rhythm patterns in ēktāl learned from HMR _I dataset	163
4.27	Rhythm patterns in ēktāl learned from HMR _s dataset	163
4.28	Rhythm patterns in jhaptāl learned from HMR _I dataset	164
4.29	Rhythm patterns in jhaptāl learned from HMR _s dataset	164
4.30	Rhythm patterns in rūpak tāl learned from HMR _I dataset	165
4.31	Rhythm patterns in rūpak tāl learned from HMR _s dataset	165
5.1	Block diagram of the tāla tracking algorithm proposed by Srinivasamurthy and Serra (2014)	183
5.2	An illustration of percussion enhancement	184
5.3	Estimated time varying tempo curve with tempogram	186
5.4	The meter analysis models used in the dissertation	193
5.5	An illustration of the bar pointer model	195
5.6	An illustration of SP-model transition matrices	212
5.7	Results of statistical significance testing of meter analysis results on Indian art music datasets	235
6.1	Waveform and spectrogram of jingju percussion strokes	248
6.2	Waveform and spectrogram of the pattern shanchui	249
6.3	Block diagram: Jingju percussion pattern classification	252

6.4	Block diagram: Percussion pattern discovery in Indian art music	259
6.5	An example of a tabla percussion pattern	262
6.6	An example of a mridangam percussion pattern	262
7.1	A screenshot of the recording page of Dunya	281
7.2	Screenshots of Sarāga	283

List of Tables

2.1	Structure of Carnatic tālas	25
2.2	The syllables used in Carnatic music percussion	28
2.3	Structure of Hindustani tāls	30
2.4	The tabla bōls used in Hindustani music	33
2.5	The ṭhēkās for popular Hindustani tāls	35
2.6	Syllables used in Beijing opera percussion	39
3.1	Performance of meter estimation using GUL algorithm . .	104
3.2	Performance of cycle length estimation using PIK al- gorithm	105
3.3	Accuracy of cycle length recognition using KLA algorithm	106
3.4	Accuracy of cycle length recognition using SRI algorithm	107
3.5	Accuracy of cycle length recognition using compara- tive approaches	109
3.6	Accuracy of downbeat tracking in Carnatic music subset	110
4.1	Coverage of the Carnatic music corpus	123
4.2	Completeness of the Carnatic music corpus	127
4.3	Coverage of the Hindustani music corpus	129
4.4	Completeness of the Hindustani music corpus	130
4.5	CMR _f dataset description	134
4.6	Tāla cycle length indicators for CMR _f dataset	134
4.7	CMR dataset description	135
4.8	Tāla cycle length indicators for CMR dataset	135
4.9	HMR _f dataset description	148

4.10	Tāl cycle length indicators for HMR _f dataset	148
4.11	HMR ₁ dataset description	149
4.12	Tāl cycle length indicators for HMR ₁ dataset	149
4.13	HMR _s dataset description	150
4.14	Tāl cycle length indicators for HMR _s dataset	150
4.15	The tabla solo dataset	167
4.16	The mridangam strokes dataset	169
4.17	The mridangam solo dataset	170
4.18	The jingju percussion instrument dataset	171
4.19	The jingju percussion pattern dataset	173
5.1	Results of akṣara period tracking on CMR _f dataset	188
5.2	Results of sama tracking on CMR _f dataset	189
5.3	Summary of the meter analysis models and inference algorithms	219
5.4	Results of meter inference with the bar pointer model	225
5.5	Results of meter tracking with the bar pointer model	227
5.6	Results of meter tracking with the mixture observation model	228
5.7	Results of meter tracking with the section pointer model	229
5.8	Results of meter tracking with inference extensions to the bar pointer model	230
5.9	Tempo informed meter tracking results on Indian music datasets	232
5.10	Tempo-sama-informed meter tracking results on Indian music datasets	233
5.11	Summary of meter analysis performance on Indian art music datasets	234
5.12	Summary of meter tracking performance of inference extensions	237
5.13	Comparing the meter tracking performance of AMPF ₀ and AMPF _m algorithms	237
6.1	Transcription and classification results on JPP dataset	255
6.2	Confusion matrix for percussion pattern classification in JPP dataset	255
6.3	Query tabla percussion patterns	260
6.4	Query mridangam percussion patterns	261

6.5	Automatic transcription results on tabla solo dataset (Flat start HMM)	269
6.6	Automatic transcription results on tabla solo dataset (Isolated start HMM)	269
6.7	Performance of approximate pattern search on tabla solo dataset	270
6.8	Automatic transcription results on the mridangam solo dataset	272
6.9	Performance of approximate pattern search on mridan- gam solo dataset	273

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, which keep increasing each 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, enhancing our experience with the data, information and knowledge of the world. The technology needs in a multicultural world 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, necessitating novel ways of automatically structuring these collections. The interaction with music has grown beyond just listening into an enriching and engaging experience with the music content. Such a scenario provides a great opportunity to enhance our experience interacting with music.

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) is a specialized area of research within music technology that 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.

Music is multi-dimensional and rhythm is one of the most basic dimensions of music. Rhythm can also be studied from many different perspectives (Bello, Rowe, Guedes, & Toussaint, 2015), and this work takes an **MIR** viewpoint. Music manifests as musical events unfolding in time, and the arrangement of these events constitutes the rhythm of a music piece. These events can be grouped and organized in several layers into rhythmic structures and patterns. Automatic rhythm analysis aims to estimate and characterize these rhythmic structures and patterns from music to extract musically meaningful rhythm related information 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 now delve into the context and motivation for the thesis. The scope and objectives of the thesis are then clearly identified. The concluding section of the chapter 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) music¹ 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 to address these limitations.

CompMusic² (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),

¹The 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.

²<http://compmusic.upf.edu>

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, 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 categorized. 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.

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 Dunya.

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 has been conducted in 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 imbues and inherits all the goals and context of CompMusic project as applied to rhythm analysis. A meaningful automatic rhythm analysis hence 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 dimension 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.

Musical events are often organized in several hierarchical layers leading to metrical structures. The metrical structures provide a fundamental framework in time to organize events and hence play a pivotal role in music. Most melodic and rhythmic phrases, lyrical lines, harmonic changes are organized around metrical structures and hence the estimation of different aspects of the meter 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 structures. These are structural components (e.g. verse, chorus, bridge, intro, outro, solo) and 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 a song.

Music is also replete with rhythmic patterns at several different levels. Music is expressed through a grouping of events into rhythmic patterns and hence these patterns are 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 can be further 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. The target audience for such tools span across serious music listeners who wish for an enhanced experience 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 these tools in their work, and music collectors and record labels who can use these tools to 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 a 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 music content of the audio recordings). Further, navigating within a recording would also mean that these similarity measures are additionally needed intra-piece i.e. for different parts of a single piece.

As specified earlier, a meaningful navigation and retrieval can be better achieved 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 in Indian art music

can help to identify the limitations of current approaches to improve their performance and make them better and more general. As emphasized earlier, there is a 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 culture-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-aware methods that are also generalizable and adaptive to other contexts and musics.

The motivation for culture-aware 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 with many different open problems, it is important and necessary to clearly identify the scope and objectives of this dissertation.

1.3 Scope and objectives

The work presented in the dissertation on automatic rhythm analysis stands at the intersection of audio music processing, machine learning, music theory, musicology, and the application of enriched music listening. 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 challenges and opportunities in automatic rhythm analysis of Indian art music and formulate relevant automatic rhythm

analysis problems. Convert musical definitions into engineering formulations amenable to quantitative analysis using signal processing and machine learning approaches.

- To build useful annotated music 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-aware computational rhythm models for Hindustani and Carnatic music
- To develop novel signal processing and machine learning methods for rhythm analysis of Indian art music
- To devise and develop specific rhythm similarity measures for a better structuring and discovery of Indian art music music collections
- 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 data-driven signal processing and machine learning approaches for automatic analysis, description and discovery of rhythmic structures and patterns in audio music 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 main 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. An audio recording is hence at the centre of analysis, with several types of rhythm related information extracted from a recording.

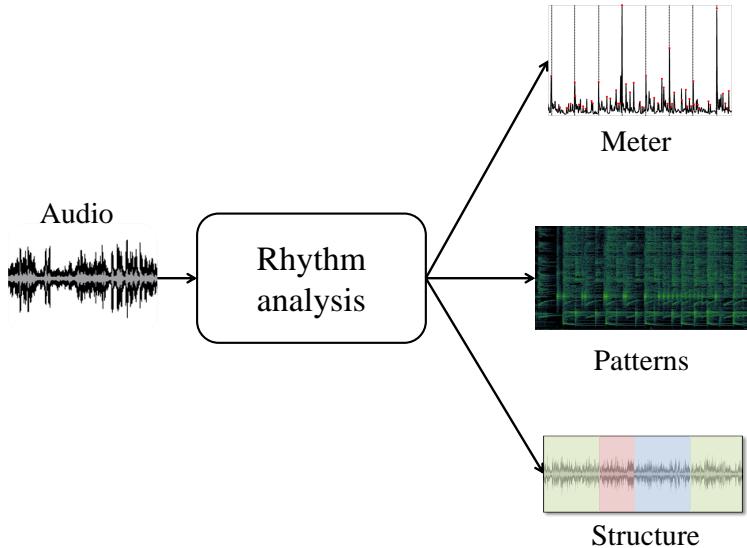


Figure 1.1: Example of automatic rhythm analysis from audio recordings estimating meter, rhythmic patterns and structure from audio recordings. The approaches in the dissertation follow a similar flow, with an audio recording being at the centre of analysis.

Figure 1.1 shows an example of such a paradigm, showing 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 are considered secondary sources in the scope of the thesis, though used in some tasks. 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 models and methods provide an effective framework to bring in higher level music knowledge into models, in terms of model structure and priors. Bayesian models are hence a central theme for the analysis

approaches in the thesis.

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 are built to work on real world representative music collections - 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 aid musicologists further in their work 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 CompMusic project. The analysis methods developed in the thesis do not aim to replace expert musician opinions, but only work within the framework provided by musicians, musicologists, listeners and learners 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 musicologist community.

In addition to exploring novel approaches to automatic rhythm analysis, the thesis aims to answer the following research 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 rhythmic struc-

- tures. To what extent does incorporating higher level knowledge affect the performance of automatic analysis? What kinds of higher level information are useful and lead 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. The methods can then utilize the interplay between the components for better estimation. Does a holistic approach work better or is it better to estimate individual components separately? Which components of rhythm is better estimated jointly, and which components 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?
 6. Given the availability of useful annotated datasets, one of the questions to ask is if the annotations and the data can be used for a corpus level analysis leading to meaningful and valid musicalological 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 rhythm analysis tools and methods to be a 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.

A 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. An open dissemination strategy is one of the 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. The writing style followed in the thesis is a mixture of both active and passive voice. Most of the dissertation derives content from published research papers describing the work done by collaborative teams. Hence the word ‘we’ refers to the author and in cases additionally includes the co-authors and collaborators in research papers. When presenting results and making observations, the word ‘we’ further includes the reader who can equally make such an observation from the results. However, the main original contributions by the author of the thesis are emphasized appropriately, wherever needed.

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 prob-

lems in Indian art music. Chapter 4 presents the research corpora and all the rhythm related datasets compiled as a part of Comp-Music 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 analysis and percussion pattern discovery, respectively. Chapter 7 presents some of the applications and conclusions 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 dissertation including unfamiliar terms related to Indian art music which are all listed with a short description in a glossary in Appendix C. The glossary also lists the acronyms used for datasets and analysis methods in the dissertation.

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 for 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 of the chapter is compiled and presented from several external sources cited appropriately when necessary.

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 are 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 of the dissertation and presents an evaluation of the state of the art for some of these tasks on Indian art music. A large part of the content of the chapter is derived from several discussions of with collaborators of Comp-Music, musicians, musicologists and listeners on what they consider are important rhythm analysis problems, with some content and results from our published journal article (Srinivasamurthy, Holzapfel, & Serra, 2014).

Chapter 4 describes the efforts of CompMusic in compiling and annotating the research corpora and test datasets. The chapter presents a systematic framework to elucidate a set of design principles to build data corpora for research in Indian art music (Serra,

2014). 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 an illustrative corpus level statistical analysis of relevant test datasets is also presented to point out some interesting observations. The rhythm related datasets described in the chapter have a major contribution from the author, but still are collective efforts of the CompMusic team, as indicated with each dataset. Some of the content of the chapter is from our published papers (Srinivasamurthy, Koduri, Gulati, Ishwar, & Serra, 2014; Srinivasamurthy & Serra, 2014; Tian, Srinivasamurthy, Sandler, & Serra, 2014; Srinivasamurthy, Caro, Sundar, & Serra, 2014; Gupta, Srinivasamurthy, Kumar, Murthy, & Serra, 2015; Srinivasamurthy, Holzapfel, Cemgil, & Serra, 2016) while some of it is unpublished content.

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 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 beat and downbeat tracking (Krebs, Böck, & Widmer, 2013; Krebs, Holzapfel, Cemgil, & Widmer, 2015), along with some of our recently published papers (Srinivasamurthy & Serra, 2014; Srinivasamurthy, Holzapfel, Cemgil, & Serra, 2015; Srinivasamurthy et al., 2016; Holzapfel, Krebs, & Srinivasamurthy, 2014) and from latest unpublished work.

Chapter 6 presents the other important contribution of the thesis and describes the task of percussion pattern 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 search of percussion patterns in percussion solo recordings. As a preliminary test case, experiments on percussion pattern classification in Beijing opera are presented. The approaches are extended to Indian art

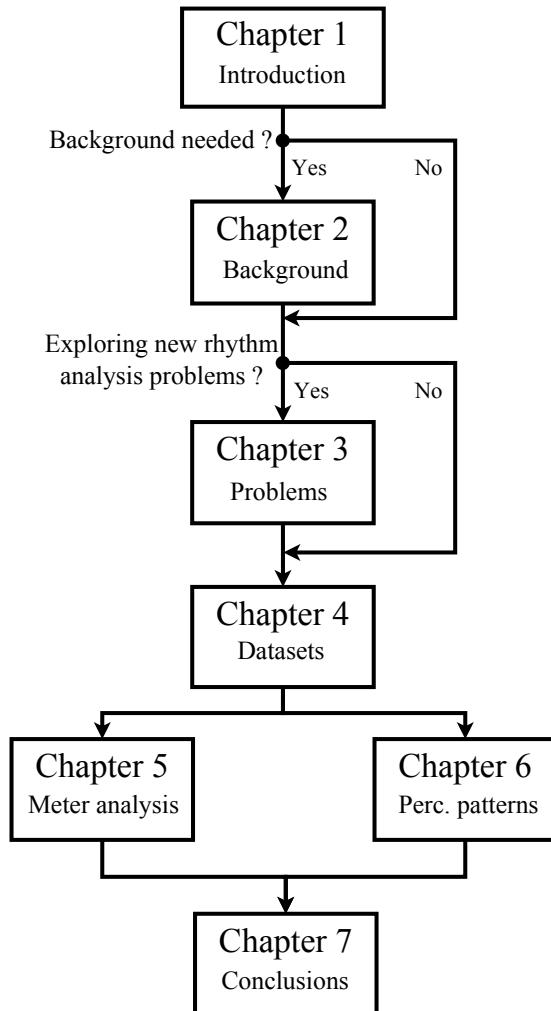


Figure 1.2: Dependencies between the chapters of the dissertation, also indicating a suggested reading order for the chapters

music and an evaluation is provided on both tabla and mridangam drum solo datasets. A part of the results presented in the chapter are derived from our published papers (Gupta et al., 2015; Srinivasamurthy, Caro, et al., 2014) while many results are yet unpublished. 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 of the dissertation is self contained and can be read

in isolation with sufficient background. However, the following dependencies exist between chapters, which is a possible indicator of the recommended order for reading and is further summarized in Figure 1.2. Starting with Chapter 1, if the reader has sufficient background on rhythm in Indian music and rhythm analysis problems in MIR, Chapter 2 can be skipped. For a researcher starting out and exploring new problems and resources in Indian art music, 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. 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 at rhythm analysis of 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 a music 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.

Background

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

Shankar (1999, p. 61)

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

1. To establish a consistent terminology for several rhythm related music concepts
2. To describe relevant rhythm related concepts in Indian art music and Beijing opera
3. To present an overview of the state of the art in automatic rhythm analysis problems that will be addressed in this dissertation
4. To briefly describe the relevant technical concepts necessary to understand the algorithms and methods presented in the 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 subdivision (also called the *tatum*), beat (also called the *tactus*), and measure (or bar): from short to long time-span (London, 2001, accessed June 2016). 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 (Parncutt, 1994; Müller, Ellis, Klapuri, Richard, & Sagayama, 2011), 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). A shortcoming of most attempts to describe rhythm and meter is the assumption about pulsation to consist of recurring, precisely equivalent and equally-spaced isochronous 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 art music. In addition, in Indian art 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. Hence, additional references that have an in depth discussion of the presented concepts are provided when necessary.

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 melody and rhythm 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.

There are several unfamiliar terms that are introduced in this section and hence before presenting those terms, a note on disambiguating terminology is in order. The latin transliteration of the Indian art music terms are according to the ISO 15919 (Transliteration of Devanagari and related Indic scripts into Latin characters) standard (ISO/TC, 2001). 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 latin transliteration of terms change accordingly.

We will use the term Indian art music (or even Indian 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 collec-

tively 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, called the *gāyaki* (singing) style (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 spread over the Indian subcontinent, it is diverse with several different music styles falling under its gamut. Several different styles of Hindustani music exist, 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 *sitār*, *sarōd*, flute, *santūr*), 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*, or 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 (Mehta, 2008).

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 provided in the following section, highlighting these similarities and differences.

2.2.2 Rhythm and percussion in Carnatic music

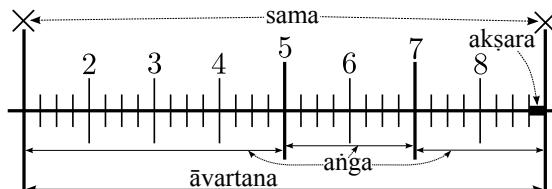
Sambamoorthy (1998) provides a comprehensive description of *tālas* in Carnatic music. In Carnatic music, the *tāla* provides a broad structure for repetition of music phrases, motifs and improvisations. It consists of fixed length time cycles called *āvartana* which can be referred to as the *tāla* cycle. In an *āvartana* of a *tāla*, phrase refrains, melodic and rhythmic changes occur usually at the beginning of the cycle. An *āvartana* is divided into basic equidistant time units called *akṣaras*. The first *akṣara* pulse of each *āvartana* is called the *sama*, which marks the beginning of the cycle (or the end of the previous cycle, due to the cyclic nature of the *tāla*). 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 use an adapted definition of a beat for the purpose of a common ground, defined as a uniform pulsation. It is to be noted 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 as the beats of the *tāla*. The perceptual foot tapping time “beats” are a subset of this uniform beat pulsation.

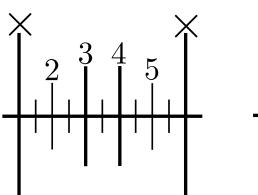
The subdivision grouping structure of the *akṣaras* within a beat is called the *naḍe* (also spelled *nadai*) or *gati*. The most common *naḍe* is *caturaśra*, in which a beat is divided into 4 *akṣaras*. Another important aspect of rhythm in Carnatic music is the *edupu*, the “phase” or offset of the 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

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

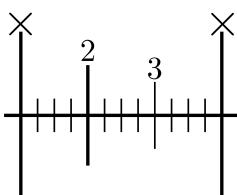
Table 2.1: Structure of Carnatic tālas, showing the akṣaras per beat (an indicator of nađe), the number of beats and akṣaras in each cycle



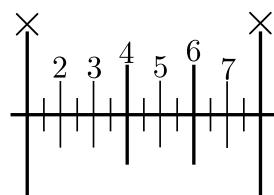
(a) Ādi tāla, illustrated



(b) Khaṇḍa chāpu tāla



(c) Rūpaka tāla



(d) Miśra chāpu tāla

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

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 that incorporates

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 *naḍe*, 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 are composed in four popular *tālas* - *ādi*, *rūpaka*, *miśra chāpu*, and *khaṇḍa 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 naḍe* (division of a beat into two or four *akṣaras*). The different concepts related to the *tālas* of Carnatic music are also illustrated¹ 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 which are in *caturaśra naḍe*. But for completeness, 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.

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

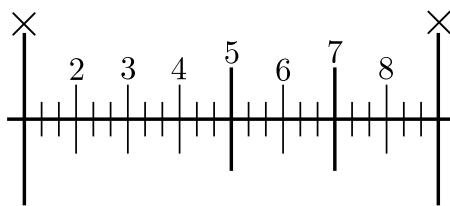


Figure 2.2: The structure of *tiśra nađe ādi tāla*, to contrast with the popularly used *caturaśra nađe*. Note the three *akṣara* beats, and only 24 *akṣara* cycle, as compared to the 32 *akṣara* cycle in its *caturaśra nađe* counterpart.

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 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 *tala*. 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

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 (*solkattus*) used in mridangam, grouped based on timbre along with the symbol we use for the syllable group in this dissertation. The last column also provides a short description. Most strokes are combinations of left+right strokes on the mridangam.

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 oral mnemonic syllables (called *solkattus*) representative of the different strokes of the mridangam. An oral recitation of these syllables is itself an art form called *konnakol*, and is often a part of a *tani-āvartana*. The syllables used be-

long to mridangam, but is widely used with other percussion instruments used in 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 syllable groups, assuming that such a timbral grouping is sufficient for discovery of timbrally similar percussion patterns. This timbre based grouping further 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. For simplicity and brevity, we will refer to the syllable groups as just syllables in this work when there is no ambiguity. Finally however, we carefully note that the syllables also have a loosely defined functional role, and such a timbre based grouping used in the thesis 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**, marking the end of the previous cycle and the beginning of the next cycle. The first **mātrā** of the cycle (**sam**) is highly significant structurally, with many important

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.

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-vilambit (very slow), *vilambit* (slow), *madhya* (medium), *dṛt* (fast) to ati-dhṛt (very fast). Depending on the *lay*, the mātrā may be further subdivided into shorter time-span pulsations, indicated through additional filler strokes of the tabla. However, since these pulses are not well defined in music theory, 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 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 tāls defined, while about 15 tāls are performed in practice. Figure 2.3 shows four popular Hindustani tāls - tīntāl, ēktāl, jhaptāl, and rūpak tāl. The structure

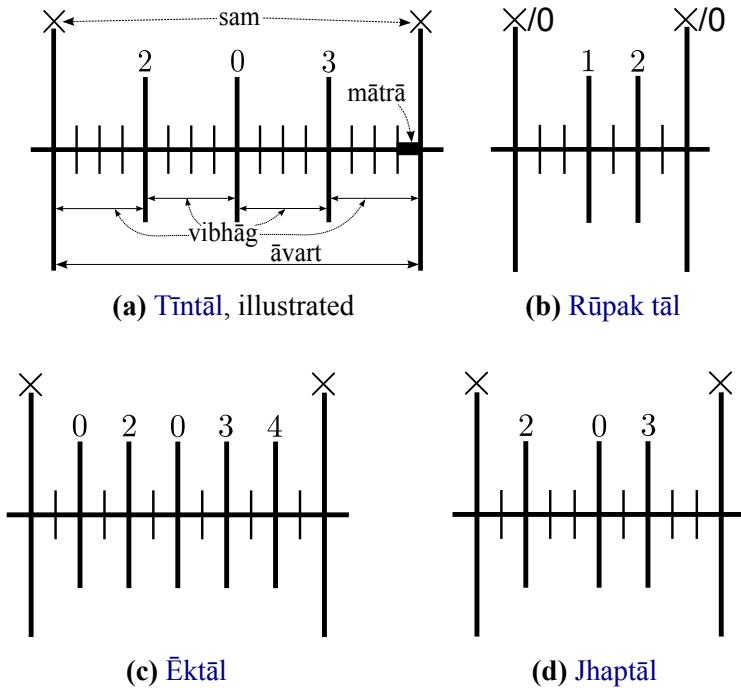


Figure 2.3: An *āvart* of four popular Hindustani *tāls*, showing the *mātrās* (all time ticks), *vibhāgs* (long and bold time ticks) and the *sam* (\times). *Tintāl* is also illustrated using the terminology used in this article.

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 *tintāl* *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, which is represented as four equal duration *vibhāgs* of three *mātrās*

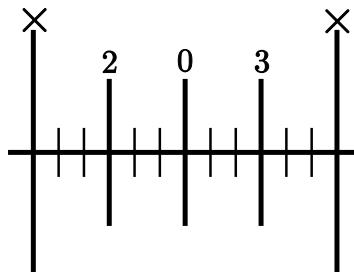


Figure 2.4: An alternative structure of Ektāl in dṛt lay

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 accompaniment. 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 lead 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).

The **bōls** of the tabla vary marginally within and across **gharānās**, and 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 functional 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

ID	Bōls	Symbol	Description
1	D, DA, DAA	DA	A closed stroke on the <i>dāyān</i> (right drum)
2	N, NA, TAA	NA	A ringing stroke on the <i>dāyān</i>
3	DI, DIN, DING	DIN	An open stroke on the <i>dāyān</i>
4	KA, KAT, KE, KI, KII	KI	A closed stroke on the <i>bāyān</i> (left drum)
5	GA, GHE, GE, GHI, GI	GE	A modulated stroke on the left drum
6	KDA, KRA, KRI, KRU	KDA	Two quick successive strokes (played as a flam), one each on <i>dāyān</i> and <i>bāyān</i>
7	TA, TI, RA	TA	A closed stroke on the <i>dāyān</i>
8	CHAP, TIT	TIT	A closed stroke on the <i>dāyān</i>
9	DHA	DHA	A resonant combined stroke with NA and GE
10	DHE	DHE	A closed combined stroke played with the full palm on the <i>dāyān</i> with a closed GE

Table 2.4: (1/2) The *bōls* used in tabla are shown in the second column, grouped by similarity of timbre. The symbol we use for the syllable group in the dissertation is shown in the third column and a short description of the timbre (Beronja, 2008) is shown in the fourth column. Combined stroke has strokes on left and right drum played together simultaneously (**contd...**).

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 improvisatory patterns. Miron (2011), Clayton (2000), A. E. Dutta (1995), Beronja (2008), and Naimpalli (2005) provide a more detailed discussion of *tāl* in Hin-

ID	Bōls	Symbol	Description
11	DHET	DHET	A combined stroke played with a closed stroke on <i>dāyān</i> with GE
12	DHI	DHI	A closed combined stroke with GE and a soft resonant stroke on <i>dāyān</i>
13	DHIN	DHIN	An open combined stroke with GE and a soft resonant stroke on <i>dāyān</i>
14	RE	RE	A closed stroke on the <i>dāyān</i> played with the palm
15	TE	TE	A closed stroke on the <i>dāyān</i> played with the palm
16	TII	TII	A combined stroke with KI and a soft closed resonant stroke on <i>dāyān</i>
17	TIN	TIN	A combined stroke with KI and a soft open resonant stroke on <i>dāyān</i>
18	TRA	TRA	Two quick successive closed strokes on <i>dāyān</i> (played as a flam)

Table 2.4: (2/2) The *bōls* used in tabla are shown in second column, grouped by similarity of timbre. The symbol we use for the syllable group in the dissertation is shown in the third column and a short description of the timbre (Beronja, 2008) is shown in the fourth column. Combined stroke has strokes on left and right drum played together simultaneously.

dustani music including *ṭhēkās* for commonly used *tāls*².

To showcase the nuances of the *tāl* 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*, *palatā*, *rēlā*, *pēskār* and *gaṭ* 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

²Some audio examples illustrating the *tāls* at <http://compmusic.upf.edu/examples-taal-hindustani>

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

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.

taken up by the lead musician during the solo.

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 (*dṛ̥ṣṭ*) 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 ranges for these classes: *vilambit lay* for a median tempo between 10-60 MPM, *madhya lay* for 60-150 MPM, and *drt lay* for >150 MPM. This large range of allowed tempi means that the duration of a *tāl* cycle in Hindustani music ranges from less than 2 seconds to over a minute. A *mātrā* in *vilambit lay* hence can last about 6 seconds, 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 allowed 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 a 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 large (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 musician. 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 oral syllables that provide a language for percussion, which even have evolved into art forms 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 musicological 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 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 percussion pattern analysis are first tested on Beijing opera and then extended to Indian art music.

Beijing opera (Jingju, 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 recitatives 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 fulfil this task is a set of predefined and labeled percussion patterns.

The percussion ensemble in jingju 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 predefined, 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

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

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.

The percussion patterns in jingju 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, *bangu* (clapper-drum). The other three instruments are idiophones: the *xiaoluo* (small gong), the *daluo* (big gong) and the *naobo* (cymbals) (Lee & Shen, 1999; Wichmann, 1991).

Bangu has a high pitched drum-like sound while the rest of three instruments are metallophones with distinct timbres³. 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 jingju, several syllables can

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

be mapped to a single timbre. This many-syllable to one-timbre mapping is useful to reduce the syllable space for computational analysis of percussion patterns.

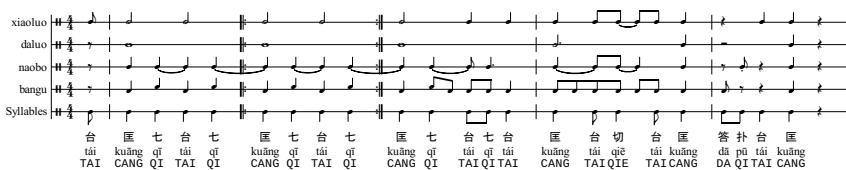
We first mapped each syllable to one or several of the instrument categories considered for analysis, as explained by Tian et al. (2014), without considering differences in playing technique or dynamics. Based on inputs from expert musicologists, we then grouped the syllables with similar timbres into five syllable groups - DA, TAI, QI, QIE, and CANG, as shown in Table 2.6. Every individual stroke of the [bangu](#), both drum and clappers, have been grouped as DA. In the rest of the syllable groups, the [bangu](#) can be played simultaneously or not. The single strokes of the [xiaoluo](#) and the [naobo](#) are called TAI and QI respectively, and the combined stroke of these two instruments together is the syllable QIE. Finally, any stroke of the [daluo](#) or any combination that includes [daluo](#) 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.

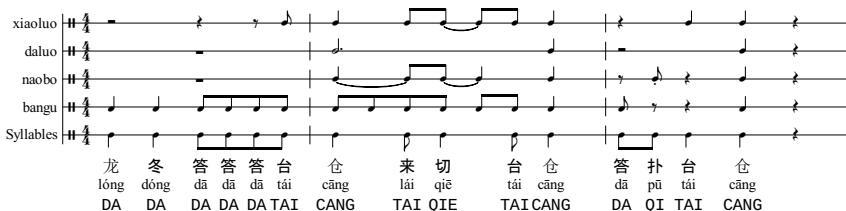
In practice, there is a library of limited set of named patterns (called [luogu jing](#), 锣鼓经) that are played in a performance, with each of these having a specific role in the arias. Although a definite agreed number for the total number of these patterns is lacking, some estimations, e.g. by Mu(穆文义) (2007), suggest the existence of around ninety of them. Figure 2.5 shows the scores for five predominantly used percussion patterns in jingju - daoban tou,



(a) daoban tou 【导板头】



(b) man changchui 【慢长锤】



(c) duotou 【夺头】

Figure 2.5: Scores for percussion patterns in jingju, showing the instruments and a syllabic representation of the pattern using the unmapped syllables, and the mapped syllable groups shown in Table 2.6 (contd...)

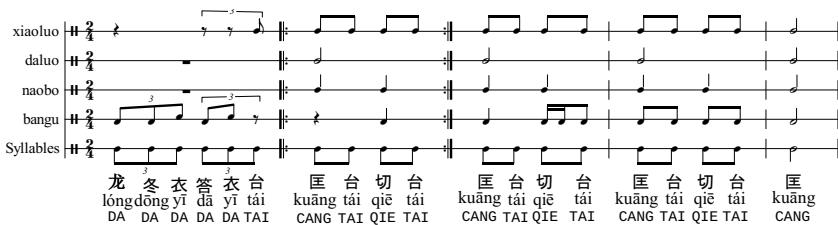
man changchui, duotou, xiaolu duotou, and shanchui⁴. The figure also shows how a possible transcription in staff notation, adapted from the scores provided by Mu(穆文义) (2007), can be simplified in a single line by the oral syllabic system. Hence, the use of these oral syllabic sequences simplify and unify 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

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



(d) xiaolu duotou 【小锣夺头】



(e) shanchui 【闪锤】

Figure 2.5: Scores for percussion patterns in jingju, showing the instruments and a syllabic representation of the pattern using the unmapped syllables, and the mapped syllable groups shown in Table 2.6

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 shanchui shown in Figure 2.5e) accompany scenic movements of undefined duration. In these cases, certain syllable subsequences in the pattern are repeated indefinitely, e.g. the subsequence cāng-tái-qiē-tái in the pattern shanchui can be repeated indefinitely until the scene completes.

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

2.3 A review of automatic rhythm analysis

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 of the relevant rhythm research problems is presented to provide a basis for further work proposed in the dissertation. The review of previous works in this section is generic 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 automatic rhythm description into complementary modules, each considering a specific task, and possibly using information and outputs from other modules (Gouyon, 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 research 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 an 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 on-

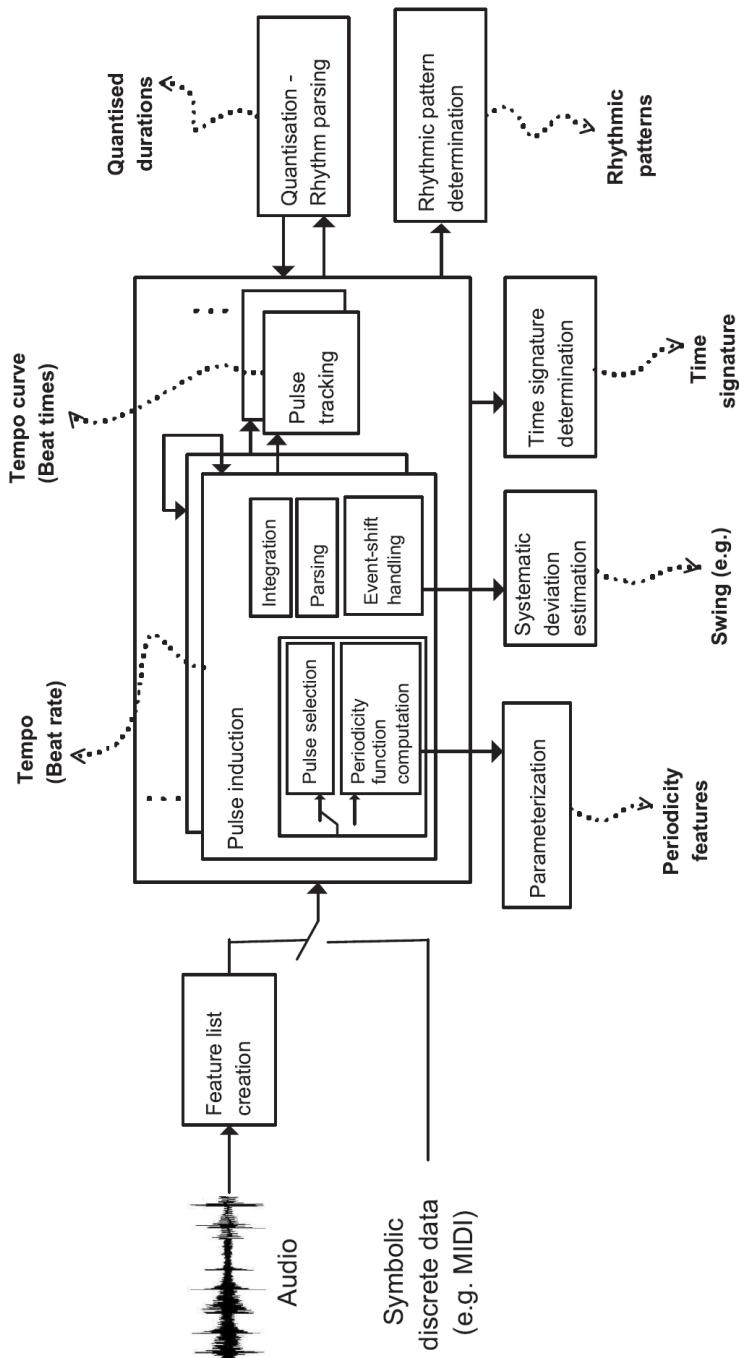


Figure 2.6: Functional units of a rhythm description system as described by Gouyon (2005) (Figure reproduced with permission)

set events is an essential part of many music signal analysis algorithms and has various applications 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), with additional improvements suggested by Dixon (2006). 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 frame-wise Short-Time Fourier Transform (STFT) 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 instruments having longer transient times without sharp bursts of energy rises. Vos and Rasch (1981) approached this issue by in-

troducing 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.

Instrument-wise onset detection

Instrument-wise onset detection refers to detecting the onsets of specific instruments from an audio signal that is a mixture of many music instruments. By detecting onsets of specific instruments, we can focus on characterizing the aspect the music that the instrument dominates, e.g. the onsets of the percussion instruments might be more useful for rhythmic analysis. One approach to instrument-wise onset detection is to separate out the part of audio that contains the information from the specific instrument we wish to extract onsets from. To detect percussion onsets, it might be advantageous to enhance the percussive parts of the signal and suppress the harmonic component.

Harmonic-Percussive Source Separation (HPSS) aims to separate an audio signal into harmonic (melodic instruments) and percussive (drums) components. Though an accurate source separation for human listening is a difficult task, for further computational analysis is simpler. Most approaches to HPSS process the spectrogram using characteristics shown by melodic and percussive music sources. The basic idea is that melodic sources show up as horizontal lines (owing to all their harmonics) in the spectrogram while percussive sources (due to their sharp attacks and broadband spectra) as vertical lines (Fitzgerald, 2010). This enables us to enhance the vertical spectral lines (or suppress the horizontal spectral lines) to enhance the percussive component of the audio, and vice versa to enhance the melodic component, such as the approaches by Thoskhahna and Ramakrishnan (2011) and Ono, Miyamoto, Le Roux, Kameoka, and Sagayama (2008).

A more informed approach uses a predominant melody extraction algorithm, which tracks the fundamental frequency (F_0) of the audio signal to uses signal analysis tools to suppress all the harmonics of the melodic source. The residual left in the spectrogram

corresponds to the percussive component of the signal. The Harmonic plus residual model by Serra (1989) along with a melody extraction algorithm e.g. by Salamon and Gómez (2012) are some of the tools that could be used for the task. The onsets from the percussion enhanced signal would give us the percussion onsets.

2.3.2 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 (or at a different metrical level, if musically well defined). 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 through the piece instead of single tempo estimate for a music piece. Despite the metrical ambiguity, tempo estimation is a useful task for further analysis tasks such as 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 BeatRoot system proposed by Dixon (2007) are pulse selection methods that measure the **Inter-Onset Interval (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, significant metrical ambiguity exists 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 audio features have been proposed 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).

2.3.3 Beat tracking

In the context of MIR, beat tracking is commonly defined as determining the time instances in the audio recording where a human listener is likely to tap his/her foot to the music. Being an important and relevant MIR task, 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 Plumbley (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. Peeters and Papadopoulos (2011) present another approach for simultaneous beat and downbeat tracking using a probabilistic framework using beat templates built using linear discriminant analysis and an algorithm that estimates the beat positions within a bar, with an evaluation on six different datasets. Some early approaches have explored particle filtering and approximate inference for beat tracking task (Hainsworth & Macleod, 2003), while probabilistic graphical models (see Section 2.4.1) have also been explored for the task of beat tracking (Lang, 2004; Lang & Freitas, 2005).

Böck and Schedl (2011) proposed a data driven approach to beat tracking using context-aware neural networks. A Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cells (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. Recently, neural network beat trackers have significantly improved beat tracking state of the art and have aimed towards joint beat and downbeat tracking.

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. The disagreement between beat trackers indicates that a piece is difficult to track (for the automatic beat trackers) and hence such pieces can further be used to improve beat tracking performance with better beat trackers.

Beat tracking is an important MIR task and has been a part of Music Information Retrieval EXchange (MIREX) challenge since its inception. There are also several datasets that have been used for evaluating beat tracking algorithms, such as the SMC dataset

(Holzapfel et al., 2012), Ballroom dataset (Gouyon et al., 2006; Dixon, Guoyon, & Widmer, 2004; Böck & Schedl, 2011), RWC database (Goto, 2006), Hainsworth dataset (Hainsworth & Macleod, 2003), McKinney dataset (Moelants & McKinney, 2004), and more recently, the GTZAN-Rhythm dataset (Marchand, Fresnel, & Peeters, 2015) that adds the beat, downbeat and swing annotations to the GTZAN dataset (Tzanetakis & Cook, 2002). We use the Ballroom dataset to evaluate 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. 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; Zapata, Holzapfel, Davies, Oliveira, & Gouyon, 2012).

2.3.4 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 Co-efficients (MFCC) extracted from 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.

2.3.5 Downbeat tracking

Downbeat tracking is the estimation of the instant of beginning of the bar. 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.

2.3.6 Meter tracking

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 learned from onset detection function as the features (Krebs et al., 2013; Böck, Krebs, & Widmer, 2014; Krebs, Holzapfel, et al., 2015). Recent interest has also been to explore deep neural networks for meter tracking, where multiple musically inspired features capturing different aspects of music have been used (Durand, Bello, David, & Richard, 2015), with extensions that used feature adapted convolutional neural networks (Durand, Bello, & David, 2016).

2.3.7 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 measures the extent of correlation between the annotations and the estimated beat locations. McKinney, Moe-lants, Davies, and Klapuri (2007) present a survey of the performance of several beat tracking algorithms using multiple accuracy measures⁵. For our evaluation, we use the f-measure, Information Gain, CML_t, and AML_t measures. These measures are character-

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

ized by a set of diverse properties and are often used in beat tracking evaluations in MIREX⁶. The measures are now explained for the task of beat tracking, but extend to downbeat tracking as well, with the same tolerances.

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 (denoted 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 (p) is the ratio between the number of correctly detected beats and all detected beats, while recall (r) 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, no continuity) is a number between 0 and 1, is the 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 misalignments in the estimated beats are allowed for by applying tolerance windows before computing the CML_t and AML_t measures.

Information Gain (\mathfrak{I}) aims at determining if there exists any kind of relation between the estimated beats and the annotations, and indicates how much information the estimated beats provide

⁶e.g. MIREX 2012, http://www.music-ir.org/mirex/wiki/2012:Audio_Beat_Tracking

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 window 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.

2.3.8 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 rhythm 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 [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 (Ballroom dataset).

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, Pastiadiis, 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.

2.3.9 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. As mentioned earlier, 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). 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 \(HMM\)](#). 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 2.3.3, might also be useful to bring the modeling capabilities of neural networks to modeling structured data such as music.

2.3.10 Percussion pattern analysis

One of the main percussion pattern analysis tasks is percussion transcription from audio. 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 growing 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 that 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. 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).

A parallel to natural language can be drawn with percussion/drum patterns where patterns composed from a small alphabet can be analogous to words. A corpus-wide analysis of rhythm patterns using a data-driven natural language processing approach was presented by (Mauch & Dixon, 2012), identifying the analogy between rhythm patterns and natural language.

Nakano, Ogata, Goto, and Hiraga (2004) explored drum pattern retrieval using vocal percussion, using a 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 of Indian art music percussion.

More recently, Paulus and Klapuri (2009) have proposed the use of connected HMMs for drum transcription in polyphonic music. Thompson, Dixon, and Mauch (2014) explore the task of drum transcription by classifying bar length rhythm patterns, utilizing low level timbre features and long term statistics from rhythm patterns. Both these approaches aim to transcribe individual drums 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 symbolic representations for drum patterns.

2.4 Relevant technical concepts

The thesis uses several well established and well studied signal processing and machine learning algorithms and techniques to address automatic rhythm analysis problems. There are excellent resources available to study and understand those models and approaches in depth, and hence only a brief mention of those methods along with references to the resources are provided in this section. The purpose is to list the algorithms and techniques and provide adequate references for a background study, and hence the section is not comprehensive in description.

2.4.1 Bayesian models

A probabilistic graphical model is a probabilistic model that expresses conditional dependence between random variables using a graph. A Bayesian model (or a Bayesian network) is a probabilistic graphical model that represents a set of random variables along and their (conditional) dependencies with a directed acyclic graph. Graphical models are generic models and most of the classical multivariate probabilistic systems (e.g. mixture models, hidden Markov models and Kalman filters) are special cases of graphical models.

With a fundamental dependence on time, any model that aims to accurately represent rhythm and 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.

DBNs hence provide an effective and explicit way to encode dependence relationships between different components of rhythm in music, and explored for meter analysis problems. The **HMM** is a special case of a **DBN** that is also used for modeling percussion syllables in the thesis. The book by Koller and Friedman (2009) is a comprehensive guide for probabilistic graphical models including practical applications. **HMMs** have been extensively used in machine learning research, and basics of an **HMM** along with the core problems are discussed in a beginner friendly tutorial by Rabiner (1989). A comprehensive resource to understand and apply **DBNs** in theory and to practical applications is written by Murphy (2002).

Inference in Bayesian models

Inference with a built Bayesian model refers to the operation in which we estimate the probability distribution of one or more unknown variables (or attributes), given that we know the values of other variables. In the context of rhythm analysis, inference can refer to using the “observed” audio features extracted from music to estimate the unknown rhythm and meter related variables.

Exact inference in Bayesian models, in its simplest form involves marginalizing over variables to obtain the distribution over the required set of variables, achieved by direct marginalization, factoring, variable elimination and other techniques (D’Ambrosio, 1999). With the exception of some toy examples, exact inference is complex without closed form solutions for real world Bayesian models and hence we resort to approximate methods.

There are efficient inference algorithms for specific inference problems in **HMM**, one such being the Viterbi algorithm (Rabiner, 1989) to decode the most likely sequence of hidden states given a observed feature sequence. In more complex **DBNs**, **Sequential Monte Carlo (SMC)** methods are often effective. **SMC** methods (also called particle filters) are a class of approximate inference algorithms that have been effectively applied for estimating posterior densities in **DBN**. Particle filters are used for efficient inference from Bayesian models for meter analysis and are described in detail in the tutorial by Doucet and Johansen (2009).

2.4.2 Speech recognition technologies and tools

Automatic speech recognition aims to build tools and techniques for automatic understanding of speech, primarily focusing on transcribing spoken utterances into written words. Being an important **ICT** for natural language processing, it has received attention from a large research community over the past many decades and has evolved into a mature research area with state of the art methods for the task (Huang & Deng, 2010). While research continues in speech recognition on several open problems, there is a potential to utilize some of its proven technologies and extend them to analogous tasks in **MIR**. There is extensive literature on speech recognition, with comprehensive description provided by, e.g. Rabiner and Juang (1993) and recently Huang and Deng (2010). Percussion syllables provide a language for percussion, with a significant analogy to speech. Hence, we explore the use of speech recognition tools for percussion pattern transcription and discovery.

Percussion transcription follows the standard data flow of speech recognition, aiming to transcribe an audio recording into a sequence of syllables. A string search on transcribed sequence can then be sued to search for patterns, using string search methods. However, transcription is often inaccurate with many errors, and any pattern search on transcribed data needs to use approximate search algorithms (Navarro, 2001). There are several other attempts to deal with search in symbolic sequences, many are described in detail by Typke, Wiering, and Veltkamp (2005). Well explored techniques formulate pattern search as **Longest Common Subsequence (LCS)** problem. However, **LCS** does 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, that we adapt for the problem of approximate string search in the thesis.

Automatic rhythm analysis of 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

Automatic rhythm analysis of Indian art music has not been explored systematically, which further means that the challenges, opportunities and relevant research problems have not been formally studied. The chapter presents the efforts to open up this research area by introducing several relevant research problems, with a review of the state of the art in these problems for Indian art music. With the background from all the relevant research problems, we define and formulate the thesis problems of meter analysis and percussion pattern discovery. The main objectives 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. Some of these challenges will help to rethink and reformulate the existing problems to be more inclusive, while improving their performance. The opportunities in turn help to pursue novel directions of research in MIR.

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 still use the commonly used terms, but it is 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 *nāde*, 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 metrical 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, the grouping 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 approaches.

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, *avadhāna 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 *tāla* 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 tempo 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 different 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, 2006; Srinivasamurthy & Chordia, 2012b). However, the use of scores itself is limited as 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 abso-

lute annotated tempo and expressive tempo are further challenging. The presence of additive meters is expected to pose challenges to existing analysis approaches, especially for Hindustani music. 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 (*solkat̄u*) 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 (loosely) 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 *solkaṭtu*). 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 significant 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

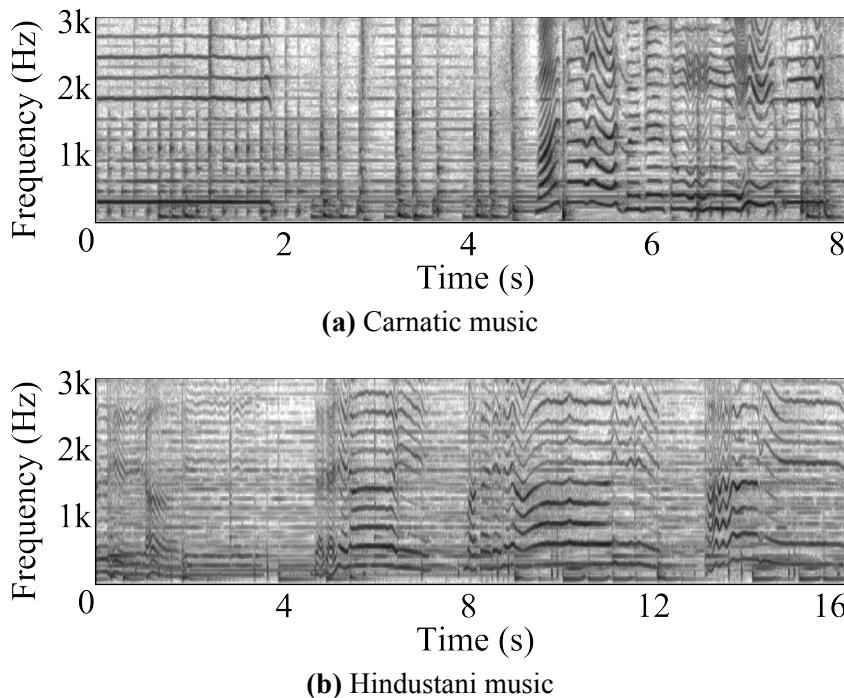


Figure 3.1: Audio signal characteristics of Indian art music. The figure shows the spectrogram of the audio excerpt of Carnatic and Hindustani music, showing the frequencies up to 3 kHz.

consistent analysis than by just taking into account the audio signal.

3.1.3 Characteristics of Indian Art Music

It is necessary to illustrate some basic signal characteristics of Indian art music that will be useful to extract relevant audio features for rhythm analysis. Figure 3.1 shows an illustrative example of the spectrogram of audio excerpts of both Carnatic music¹ (Figure 3.1a) and Hindustani music² (Figure 3.1b). To focus on melody and percussion, the spectrogram is plotted only up to a frequency

¹A short excerpt from Seethamma, a *kṛti* in *rāga* Vasanta and *rūpaka tāla*, from the album *K P Nandini at Arkay* by K P Nandini: <http://musicbrainz.org/recording/559b19c7-2d30-47db-8aab-6e4448d867fb>

²A short excerpt from Bhor Hi Aheerin, a *bandiś* in *rāg* Āhir Bhairon in *jhaptāl*, from the album *Geetinandan : Part 3* by Pt. Ajoy Chakrabarty: <http://musicbrainz.org/recording/51656b20-295c-40f9-8dab-005b9b90fa98>

of 3 kHz, though there is information in higher harmonics and percussion strokes at higher frequencies. The time durations of the excerpts are also different, with the Carnatic music example being shorter in duration.

Indian art music is predominantly melodic and heterophonic, with usually two (sometimes more) simultaneous melodic voices - a lead melody and a melodic accompaniment. The lead melody and its harmonics can be clearly seen in both the figures, along with the continuously changing fundamental frequency (F_0). There is a drone in the background that provides the tonic for the performance, provided by the [tānpura](#) (Hindustani music) or [tambūra](#) (Carnatic music). The drone can also be seen as the unchanging set of spectral frequencies in both the spectrograms. The accompaniment in Carnatic music (mainly the violin) closely follows the lead voice and can be seen with a lower amplitude in Figure 3.1a. Harmonium is the melodic accompaniment in the Hindustani music excerpt, which can also be seen with a lower amplitude in Figure 3.1b.

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. Both the figures show the percussion strokes as vertical lines, showing their broad spectrum. We can identify the strokes from the left and right drums distinctly in different frequency ranges in both cases. In addition, we can see the some of the harmonics of the pitched percussion strokes, and the quick decay of the percussion strokes. The Hindustani excerpt is taken from a [madhya lay](#) piece and we can see the longer notes and sparser tabla strokes, indicating lower rhythmic density due to lower tempo.

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

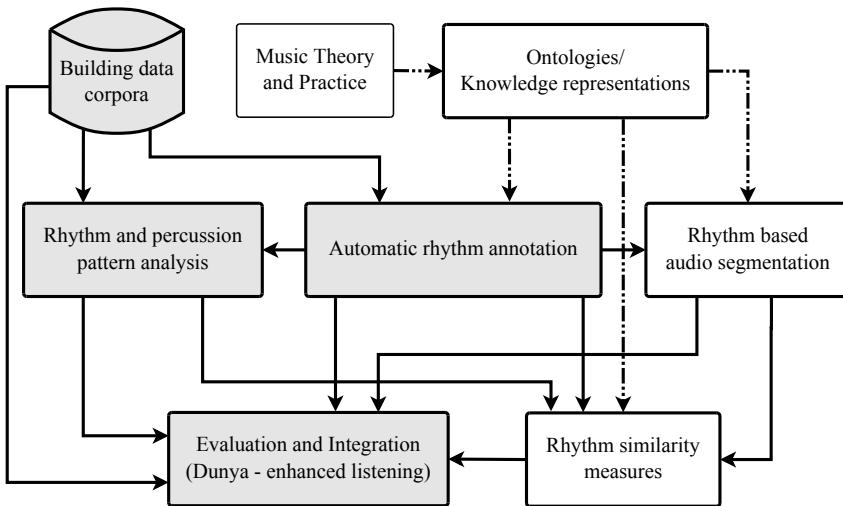


Figure 3.2: Relevant automatic rhythm analysis topics and applications in Indian art music. The solid lines indicate the flow of data and signals. The dot-dash lines indicate the flow of high level information such as parameters and priors. The subset of problems focused in this dissertation are shown in gray.

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 be 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 components 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 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 3.2 shows an overview of the rhythm related research topics and applications Indian art music. It also shows the flow of data and information across the important units. The set of topics addressed in the dissertation are shown in gray. Each of those topics uses information knowledge representations derived from music theory and practice, making them more informed and culture-aware. The data-driven approaches finally culminate with rhythm similarity measures computed from the data outputs from different automatic analyses. Each of the units are further described in detail, with sub-problems within them.

3.2.1 Building data corpora

A significant part of data-driven research with signal processing and machine learning approaches needs good quality data. Data corpora that are representative of the music culture 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 (semantic) data. 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 de-

veloping 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, beat and 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 characteristic components of the *tāla* from audio. 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 implicitly the underlying metrical structure), the *naḍe* (and hence the subdivision structure), and the *edupu* 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 (possibly more than one) *tāla* tag 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 has 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 esoteric importance. 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 *ṭhēkā*) 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, creating a circular dependency.

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 *tālas* 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 such other specific level of interest apart from the *tāla* cycle is the 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 *nāde* estimation. Further, we need to point out here that there is a clear definition of the subdivision meter in the form of *nāde* 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 dataset 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 *thēkā* 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) to be estimated. 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 IOI 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, personal music collections, or 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 and not relevant with these collections. Further, manually curating audio collections with **tāla** tags is less time consuming and less resource intensive than tasks such as tempo and beat tracking. Hence, 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 **dṛt** (and possibly even more 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 *edupu*, it is likely that the melodic changes tend to occur at the *edupu* 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. But when required, we will examine the effect of non-zero *edupu* on other rhythm analysis tasks.

Tāla tracking

Tāla tracking (or generally meter tracking) refers to a set of problems that aim to estimate 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 inference - 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 estimate the time-varying tempo over the audio recording of a music piece, measured in **akṣaras/beats/mātrās** per 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 **vi-lambit** 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. The possibly irregular pulse sequence is a subset of the uniform pulsation estimated from the algorithms. 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 **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 [aksara](#) 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 [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). 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 form 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 this thesis. 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 structures 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 differ-

ent percussion timbres. To contrast, 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 it is necessary to estimate onsets of various instruments in music, since that creates the time-aligned sequence of note/stroke events that can be further used to obtain both rhythmic and percussion patterns. Some important sub-problems within 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

The task of instrument-wise onset detection refers to detecting the onsets of specific instruments from an audio signal that is a mixture of many music instruments and was described in Section 2.3.1. For rhythm analysis in Indian art music, instrument-wise onset detection can help to obtain cues for both meter tracking and for analysis of percussion patterns. The onsets of percussion instruments *mridangam* and *tabla* provide cues to the *tāla* and delimit percussion patterns. A differentiation between the left (bass) and right drum onsets in both instruments is additionally insightful and useful. Instrument-specific onsets are often not estimated explicitly, but are estimated as a part of a bigger task, such as percussion transcription.

It is a difficult task to extract out percussion onsets with a great accuracy, given that the percussion in Indian art music is also tuned and shares the same frequency range as the melody. Within Comp-Music, instrument-wise onset detection has been explored for Beijing opera (Tian et al., 2014). HPSS can help to improve onset detection of percussion instruments (Section 2.3.1) and has been applied to Carnatic music by Srinivasamurthy and Serra (2014) with limited success. Given the complexity of the task, it is preferable to build models that are robust to errors in onset detection of specific instruments.

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 one full 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. However, there are long rhythmic patterns played on mridangam/*tabla* and accentuated through melody that can last many cycles. 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 larger corpora.

Rhythmic patterns are closely tied to the *tāla*, and further within a *tāla* cycle to the sections of 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 study of rhythm patterns in Indian art music is lacking, and an illustrative effort towards that is a part of the dissertation (Section 4.2.1-4.2.2).

Percussion pattern transcription and discovery

Percussion pattern transcription is mainly applied 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 percussion syllables. Percussion transcription is a sub-problem of the more general music transcription. 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 dis-

covered patterns.

The use of oral mnemonics is not unique to Indian art music. 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, 2001, accessed June 2016), like for instance the syllables of solfège, of the Indian svaras (notes) 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 many traditional musics. D. Hughes (2000) mentions 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. Such a link between drum patterns and natural language has been explored by Mauch and Dixon (2012). However, there are some challenges to percussion transcription in Indian art music. 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 as 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) are 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 longer 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), with their more recent work mainly on sequence modeling of rhythm sequences (Gillet & Richard, 2007). 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 and Chordia (2010) developed an automatic tabla improviser. Extending further, most of work using tabla sequences has been in a predictive modeling setup using the multiple viewpoint modeling framework (Chordia, Sastry, & Albin, 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 NMF based approach to decomposition of mridangam strokes into its modes, and used them for transcription. The work was further extended using cent-filterbank based features to make the transcription approach independent of the 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. Segmentation 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 recording), and to further develop similarity measures. Audio segmentation is not the main focus of the dissertation, but several rhythm related segmentation problems are briefly described for completeness.

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 aims to segment a piece into the different sections of the piece, and is useful for navigation and similarity. An applause based segmentation of Car-

natic music concerts was proposed by Sarala and Murthy (2013), which was also extended to intra-piece segmentation into sections of a 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 *sarōd* 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 from the ontology. In summary, ontologies can be built both for a direct use for navigation and inference, and for building domain specific machine learning algorithms.

Previous work on ontologies have been mainly on 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 similar-

ity 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 multicultural setting. This would necessitate the use of knowledge based approaches for similarity modeling. An in-depth study of rhythm similarity measures is not a part of the 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 that 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 a data based 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 then be 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*, *nāde*, *edupu*, *lay* classes. The measures obtained from data can be used to further 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 seldom used in performance, but used to a limited extent in music training and archival. There are no standard notation systems for melody or percussion, in both Hindustani and Carnatic music, which 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 *thēkās* 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. The syllabic scores of tabla and mridangam are useful to discover percussion patterns from symbolic data, a problem that is further addressed.

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.

Automatic score analysis research in the context of melody, rhythm, and percussion for Indian art music is scarce. Symbolic scores have been 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 melodic 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, Mallikarjuna, & Albin, 2010; Chordia, Sastry, & Albin, 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 a ground truth that musicians and listeners largely agree upon, 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, aiming 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 accent. An *usul* is analogous to [tāla](#), but is less complex than the [tāla](#) system. Hence, most of the algorithms developed for Indian music would 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 a task analogous to [tāla](#) tracking. Beijing opera percussion shares the concept of a syllabic percussion system, which is simpler and better defined than Indian art music. It is hence an ideal pilot case for percussion transcription with syllabic representation of percussion patterns, a topic of study in Chapter 6.

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 approaches. The problems across Hindustani and Carnatic music are quite 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 thesis is meter analysis of audio recordings. Meter analysis is an umbrella term used for the problems of meter inference and meter tracking. To the best of our knowledge, a comprehensive automatic 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 following research question: To what extent does building culture specific models of *tāla* and informed meter analysis (that provides additional information about the *tāla* *a priori* into algorithms) improve performance leading to more accurate tracking of the components of the *tāla* ?

To address the problem, we formulate tasks that can incorporate additional known rhythm information - informed meter analysis. The additional information provided to the algorithms is studied at various levels, from the least informed meter inference to the most informed tempo-sama-informed meter tracking. We then build Bayesian models that can explicitly incorporate higher level metrical information explicitly, and study their effectiveness and applicability for meter analysis in Indian art music. Finally, we use data-derived audio features indicative of rhythmic events in music. All together, these three focus points lead to data-driven informed Bayesian approaches for meter analysis.

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. Most commercial releases already are segmented into pieces and hence such an assumption a fair assumption. Even with large music collections, a manual or a semi-automatic segmentation of audio recordings into excerpts with a single *tāla* is less time consuming than meter tracking, and hence such an assumption is also relevant. 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*, *mīśra chāpu*, and *khaṇḍa chāpu tālas*, and for Hindustani music to *tīntāl*, *ēktāl*, *jhapitāl*, and *rūpak tāl*. Since our approaches are supervised and data-driven, this restriction is mainly due to the lack of availability of annotated training data in less popular *tālas* - the rare *tālas* have very few examples available even in large music archives. The performance of the approaches is likely to extend to other *tālas* as well, provided we have sufficient training data. From a practical standpoint, these four *tālas* will cover a majority of compositions in both Carnatic and Hindustani music, and hence such a restriction is justified.

Let a music recording z 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}_z , and the set of *sama/sam* time instants be denoted as \mathcal{S}_z . In addition, the set of *akṣara* pulses in a Carnatic music recording be denoted as \mathcal{O}_z . By this definition, we have $\mathcal{S}_z \subset \mathcal{B}_z \subset \mathcal{O}_z$. 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 time varying sequence of tempo value estimates for every frame k , called a tempo curve can be measured in inter-beat/*mātrā* interval $\tau_{b,k}$ (or equivalently $60/\tau_{b,k}$ measured in beats/*mātrās* per min), or as inter-*sama/sam* interval $\tau_{s,k}$. For Carnatic music, tempo can additionally be measured in inter-*akṣara* interval $\tau_{o,k}$ (or equivalently $60/\tau_{o,k}$ measured in *akṣaras* per minute). The approaches, experiments and results for meter inference and tracking problems 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. Per-

cussion solos with other instruments (e.g. *khañjira*, *ghaṭam*, and *mōrsiṅg*) 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, which 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 and hence can be trained with lower amount of 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 each A_j is a syllable in the set and 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 i is defined as $\mathbf{A}_i = [a_1, a_2, \dots, a_{L_i}]$, where $a_j \in \mathcal{A}$ and L_i is the length of \mathbf{A}_i . 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 library \mathcal{P} 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 to build 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_j is the onset time of a_j 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 $\{t_q^{(\cdot)}\}$ 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 algorithm. An analogous task for this search in speech recognition research is keyword spotting, where a known word (or a phrase) is searched in a longer piece of speech recording (Wilpon, Rabiner, Lee, & Goldman, 1990).

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 then be used to 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 representative rhythm annotated sub-collection with beat and *sama* level annotations. For rhythm based segmentation, 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 for experiments. The datasets built in the context of this thesis are further elaborated in Chapter 4. An interesting corpus level cycle length rhythmic pattern analysis using the rhythm annotated datasets is also presented in Section 4.2.1 for Carnatic music and in Section 4.2.2 for Hindustani music. Interesting musicological inferences can be drawn from such an analysis, showing the potential of such a methodology.

A note on terminology and style

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.

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. Most of the content of this section comes from the paper by Srinivasamurthy, Holzapfel, and 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 to further 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 data had to be adapted to a common ground in which an evaluation could be done. Hence, the evaluations are not strict and comprehensive, but still provide insights into the approaches.

We focus on the problems that are not explicitly addressed in subsequent 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 was 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 tālas*, 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 - *vi-*

laṁbit, *madhya* and *dṛt*. 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 music, with pieces from *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, which we refer to as 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 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 septuple 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 proposed 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 algorithm: The approach proposed by Pohle et al. (2009) that uses Onset Patterns (OP) as the rhythm similarity measure.

STM algorithm: The approach 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 the subdivision level. As the algorithms were 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, 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 1/2, 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

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

Table 3.1: Performance of meter estimation using [GUL](#) algorithm.

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 differentiate 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 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 two 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 tāla* is only 65.08%, which leads to considerable decrease 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

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.

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 [vilañbit 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.

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	25.40	-	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.

tions 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 for these as compared to τ_s , τ_b , and τ_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 [vilarābit](#) 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	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 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 [mīśra 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 \(2007\)](#) beat tracker, one being at the cycle level (bar length in beats), and the second at the beat level (subdivision meter, or [nāde](#)). 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.

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.

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.

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

Table 3.5: Accuracy of cycle length recognition using comparative approaches

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 provide 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 Plumley (2006) that assumes that percussive events and harmonic changes tend to be correlated with the downbeat.

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*. All values are in percentage.

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).

In Table 3.6, we depict the downbeat recognition accuracies (in percentage) 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 presented in this section. 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/āvartana* 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 need 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 into rhythm description of Indian art music through mutual reinforcement.

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 in themselves are difficult to estimate. Since downbeat information can help in estimating the cycle length and also beat tracking, 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](#) cycle 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 [tāl](#). 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, which can be adapted to a target music using machine learning methods that can adapt flexibly to the underlying rhythmic structures.

We discussed several important rhythm analysis tasks in the chapter, opening up the area of rhythm analysis research in Indian art music. While only meter analysis and percussion pattern discovery problems will be addressed in the subsequent chapters, the other relevant problems are yet unexplored. The evaluation of the state of art motivates us to explore culture-aware informed approaches to these relevant rhythm analysis tasks, using well curated datasets to test our approaches.

Data corpora for research

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

Tim Berners-Lee

Computational data-driven 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 is 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 datasets 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 datasets can even consist of synthetic data that can be used for testing. Unlike a 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 Consortium¹ (LDC) by Liberman and Cieri (1998) for language resources and PhysioBank² for physiological signals. Other open repositories of data such as MusicBrainz³ or Wikipedia⁴ 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 Dataset⁵ (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011) and AcousticBrainz⁶ (Porter, Bogdanov, Kaye, Tsukanov, & Serra, 2015) which are good research corpora 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-

¹<https://www.ldc.upenn.edu/>

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

³<http://musicbrainz.org/>

⁴<http://www.wikipedia.org/>

⁵<http://labrosa.ee.columbia.edu/millionsong/>

⁶<https://acousticbrainz.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 (built for automatic rhythm analysis) 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 rhythm annotated 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 and quantities presented for the research corpora in this dissertation 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. Finally, whenever possible, in the spirit of open research and data, the research corpora and the test datasets will be made accessible and available for further work on these music cultures.

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-aware 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 makam music research corpus has been presented in detail by Atlı, Uyar, Şentürk, Bozkurt, and Serra (2014), while Caro and Serra (2014) have described the Beijing opera (*jingju*) research corpus comprehensively. We first discuss the criteria for creating research corpora, and then describe the Carnatic and Hindustani music research corpora. Most of the content in this section is from papers by Serra (2014) and Srinivasamurthy, Koduri, et al. (2014), and describe the collective efforts of the CompMusic team in creating CompMusic research corpora. Due to continued efforts in building corpora, the numbers and analysis presented for the Indian art music research corpora are correct as of June 2014.

4.1.1 Criteria for creation of research corpora

Serra (2014) 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 de-

velop 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 approaches, 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 in 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, leading to better 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 to data for research.

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 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 audio recordings, 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 Krishna⁷ to arrive at a representative collection of Carnatic music audio. The main institutional reference for Carnatic music is the [Madras Music Academy \(MMA\)](#)⁸, 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

⁷<http://www.tmkrishna.com/>

⁸<http://musicacademymadras.in/>

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 to include their teachers, and popular musicians of their era. The record label *Charsur*⁹ 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 *vīṇā*, violin, flute, saxophone, and mridangam as lead instruments). The whole audio collection is commercial and hence 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 **MusicBrainz 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 Latin alphabet and a Latin transliteration was used when the language of the release was not English. The *rāga* and *tāla* information have been added as work attributes.

⁹<http://www.charsur.com/>

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*, *Śyāmā śāstri*, and *Muttusvāmi dīkṣitar*, in published compilations by T. K. Govinda Rao (2009), T. K. Govinda Rao (2003b) and T. K. Govinda Rao (2003a), respectively. However, these compilations are not machine readable and hence not amenable to computational analysis.

There are several good online open repositories for lyrics, such as sahityam.net¹⁰, 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. It 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. Shivkumar Kalyanaraman¹¹ is the main source of machine readable music 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.com¹² is an up-to-date directory of artist biographies, music venues, concerts and events. The category of Carnatic music on Wikipedia¹³ 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

¹⁰<http://www.sahityam.net>

¹¹<http://www.shivkumar.org>

¹²<http://www.kutcheris.com>

¹³http://en.wikipedia.org/wiki/Category:Carnatic_music

	Corpus	Raaga.com	Kutcheris	Charsur
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.

encyclopedia of music concepts providing linked information, online music forums with discussions provide opinions from which some of these links can be inferred. The rasikas.org¹⁴ 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 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.com¹⁵ 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

¹⁴<http://www.rasikas.org/>

¹⁵<http://www.raaga.com>

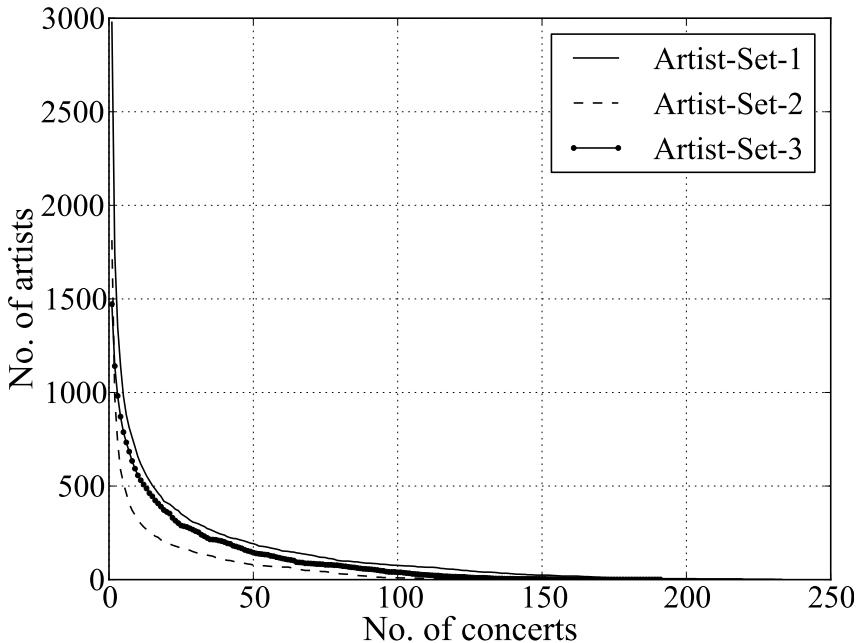


Figure 4.1: The number of artists by the number of their performances in the Carnatic music corpus

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 (Damerau, 1964).

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* (Θ) as,

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

where Θ_i^j is the *overlap* measure of the entity i with reference j , ς_i^c is the set of entities in the corpus, ς_i^j is the set of entities in the reference, and $|\varsigma|$ denotes the cardinality of a set ς . 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 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 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 dis-

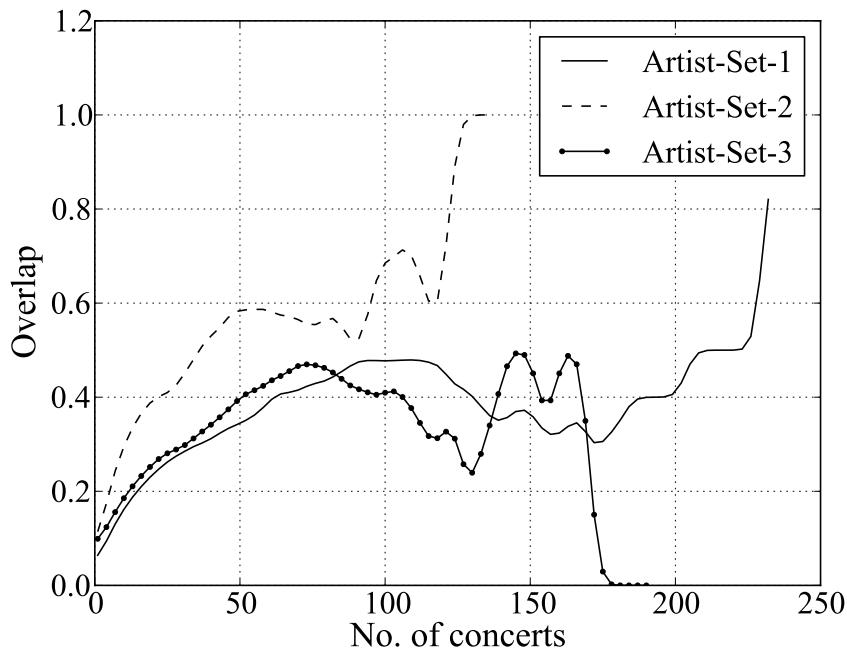


Figure 4.2: Coverage of Carnatic artists. The ordinate is the overlap value of the set of artists in corpus, compared against a set of artists in Kutcheris.com who have performed in at least as many concerts as the abscissa.

count 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

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.

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 (as of June 2014), 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 ac-

curately 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 public and private institutions that have compiled large 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 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 in-

	Corpus	ITC-SRA	Swarganga
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.

struments 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 (**bandiś**) 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 Latin alphabet, following a uniform transliteration scheme to maintain 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 **bandiś** using a standardized notation for Hindustani music. However, they are not available in a machine readable form, though a small collection of scores from these books are available in machine readable Humdrum format (Srinivasamurthy & Chordia, 2012b). Swarganga Music Foundation¹⁶ has a good archive of **rāgs**, **tāls** and **bandiś**. The category of Hindustani music on Wikipedia¹⁷ 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

¹⁶<http://www.swarganga.org/>

¹⁷http://en.wikipedia.org/wiki/Category:Hindustani_music

Accompanying metadata	# Recordings	% of total
Lead Artist	1096	100.0
Accompanying artist	658	60.0
Rāg	960	87.6
Tāl	627	57.2
Work (Bandiś)	576	52.5

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

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 subcontinent 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 (as of June 2014) 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 or compositions (*bandiś*) 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 repositories of information such as MusicBrainz and Wikipedia. The releases in the Carnatic¹⁸ and Hindustani¹⁹ 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 available²⁰. In CompMusic, have developed 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 Archive²¹,

¹⁸<http://musicbrainz.org/collection/f96e7215-b2bd-4962-b8c9-2b40c17a1ec6>

¹⁹<http://musicbrainz.org/collection/213347a9-e786-4297-8551-d61788c85c80>

²⁰<http://compmusic.upf.edu/corpora>

²¹www.archive.org

with both the audio and associated metadata and annotations available through the Dunya API. The open music collections are growing collections with new releases being added, and hence the numbers in this section (correct as of June 2016) are approximate and indicative.

The Carnatic Creative Commons music collection (CMD_o)²² is a collection of 20 vocal Carnatic concerts (with more releases being added) with 197 tracks and over 41 hours of music by professional Carnatic musicians. The audio in the CMD_o collection was professionally recorded in multi-track at 44.1kHz sampling rate at Arkay Convention Center, Chennai, India, and mastered professionally. The pieces from the concerts were split into individual recordings and released together 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 about 16880 *sama* annotations that can be used for meter analysis.

The Hindustani Creative Commons music collection (HMD_o)²³ is a collection of 36 vocal Hindustani music albums (with more releases being added) with 108 tracks and over 43 hours of music by professional Hindustani musicians, sourced from personal collections of musicians. The audio in the HMD_o collection are stereo mp3 tracks sampled at 44.1 kHz. The tracks procured from personal collections have been grouped into musically meaningful short compilations 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 about 11260 *sam* annotations that can be used for rhythm analysis.

²²<https://musicbrainz.org/collection/a163c8f2-b75f-4655-86be-1504ea2944c2>

²³<https://musicbrainz.org/collection/6adc54c6-6605-4e57-8230-b85f1de5be2b>

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 corpora (or 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. There are several test datasets for different music cultures built within CompMusic²⁴, while 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.

4.2.1 Carnatic music rhythm dataset

The Carnatic Music Rhythm dataset (CMR_f)²⁵ is a rhythm annotated test corpus for many automatic rhythm analysis tasks in Carnatic Music(Srinivasamurthy & Serra, 2014). 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* (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 down-mixed from stereo recordings, and sampled at 44.1 kHz. The audio is also available as downmixed mono WAV files for experiments.

²⁴<http://compmusic.upf.edu/datasets>

²⁵<http://compmusic.upf.edu/carnatic-rhythm-dataset>

Tāla	#Pieces	Total Duration hours (min)	\bar{T}_f	#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_f 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 (m and s indicate minutes and seconds, respectively)

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

Table 4.6: Tāla cycle length indicators for CMR_f dataset. $\bar{\tau}_s$ and σ_s indicate the mean and standard deviation of the median inter-sama interval of the pieces, respectively. $\bar{\tau}_o$ and σ_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 τ_s and hence the range of τ_s in the dataset. All values in the table are in seconds.

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

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 ± 0.868	0.17 ± 0.027	[2.88, 7.07]
Rūpaka	2.12 ± 0.225	0.18 ± 0.019	[1.40, 3.10]
Miśra chāpu	2.81 ± 0.272	0.20 ± 0.019	[2.03, 3.65]
Khaṇḍa chāpu	1.87 ± 0.290	0.19 ± 0.029	[1.00, 2.84]

Table 4.8: Tāla cycle length indicators for CMR dataset. $\bar{\tau}_s$ and σ_s indicate the mean and standard deviation of the median inter-sama interval of the pieces, respectively. $\bar{\tau}_o$ and σ_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 τ_s and hence the range of τ_s in the dataset. All values in the table are in seconds.

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.

CMR_f 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

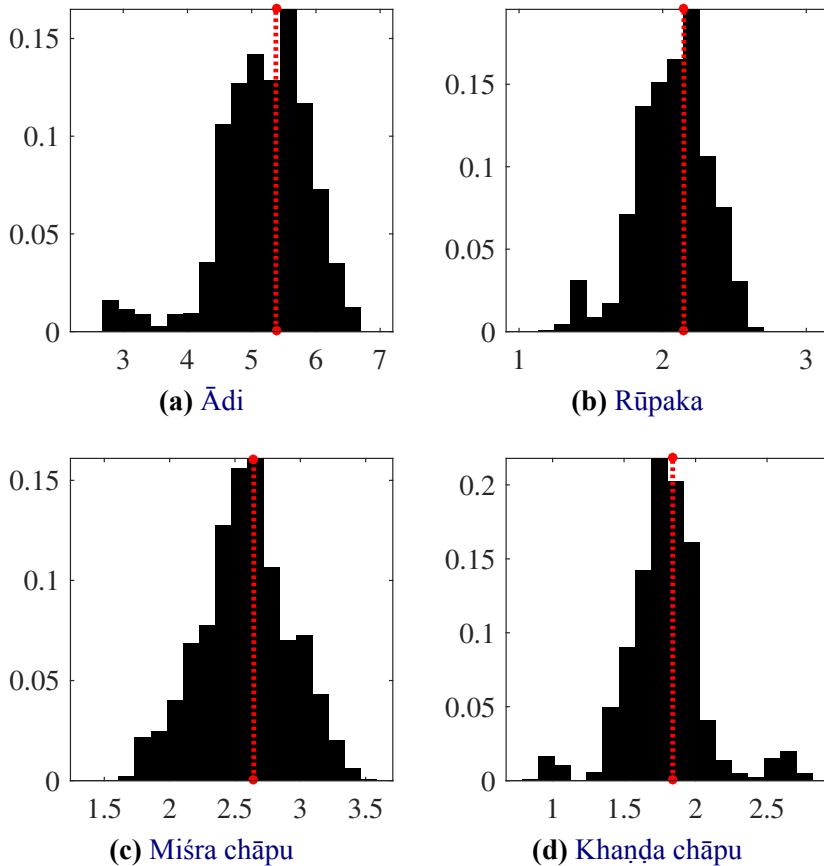


Figure 4.3: A histogram of the inter-sama interval τ_s in the CMR_f dataset for each tāla. The ordinate is the fraction of the total count corresponding to the τ_s value shown in abscissa. The median τ_s for each tāla is shown as a red dotted line.

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 cycle durations that are present in Carnatic music pieces. 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_f dataset is also compiled as the CMR dataset, with two minute excerpts of pieces in CMR_f (or the full piece if the piece is shorter than 2 minutes). These short ex-

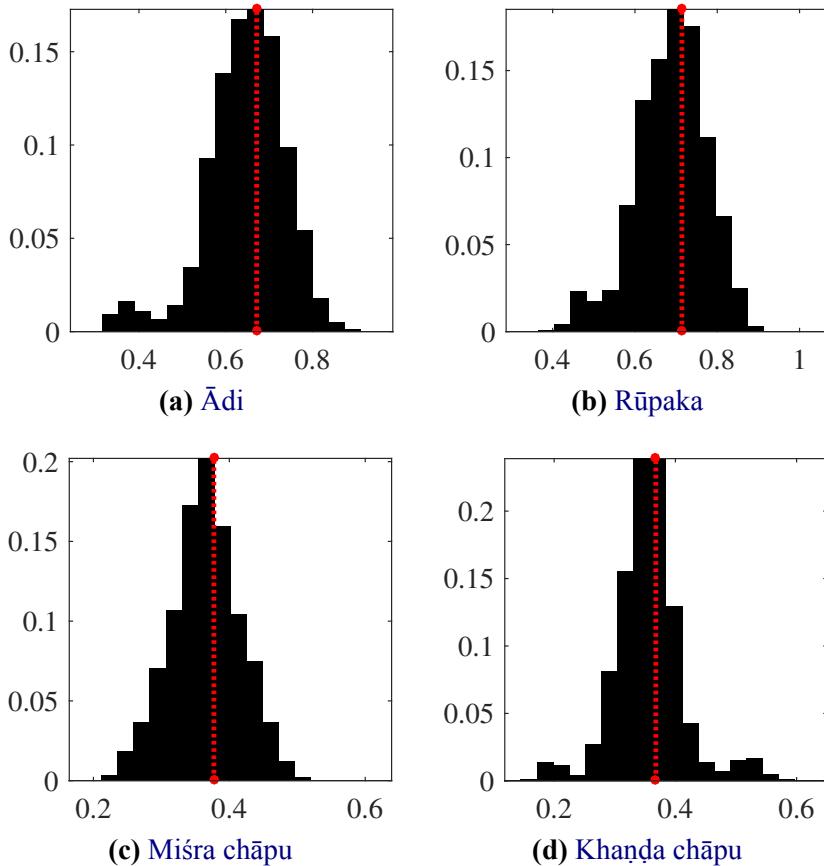


Figure 4.4: A histogram of the inter-beat interval τ_b in the CMR_f dataset for each tāla . The ordinate is the fraction of the total count corresponding to the τ_b value shown in abscissa. The median τ_b for each tāla is shown as a red dotted line.

cerpts additionally contain all the annotations of the full dataset, including time aligned *sama* and beat annotations. The smaller CMR dataset will be useful for faster testing of approaches and algorithms.

The CMR dataset is described in Table 4.7, showing the four tāla s 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_f dataset shown in Table 4.6, showing that CMR dataset

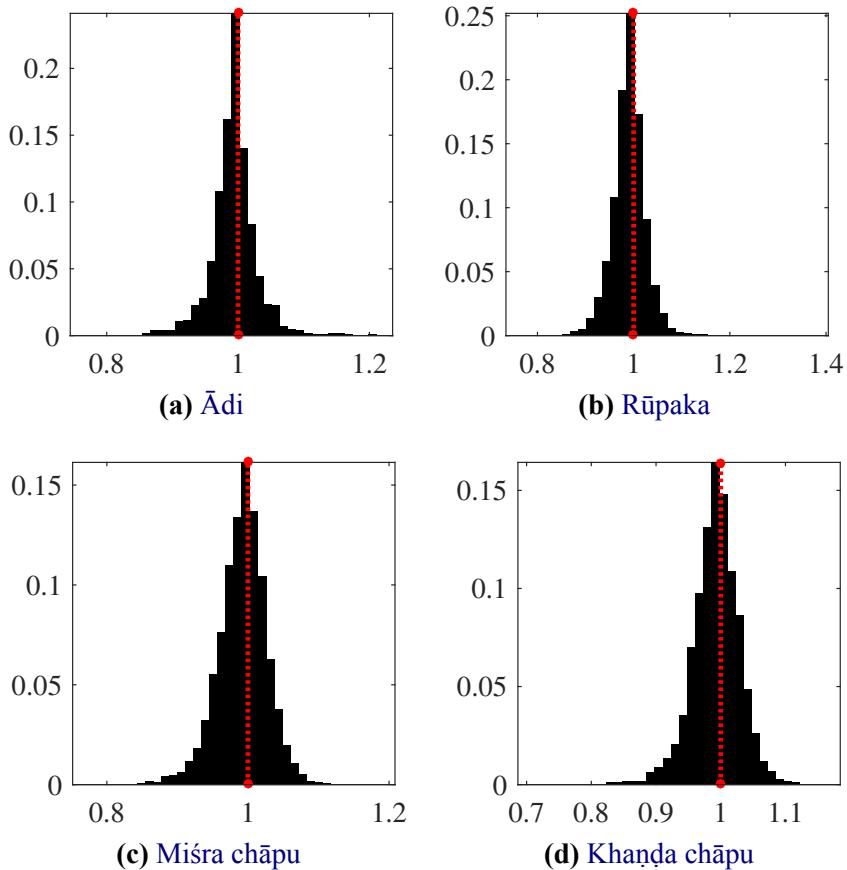


Figure 4.5: A histogram of the median normalized inter-sama interval τ_s in the CMR_f dataset for each tāla. The ordinate is the fraction of the total count corresponding to the normalized τ_s value shown in abscissa.

is a representative subset of CMR_f 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 Table 4.6 we present further analysis of the inter-sama interval (τ_s) and inter-beat interval (τ_b) for each **tāla** over the whole CMR_f dataset. A histogram of τ_s and τ_b for each **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 τ_s for each **tāla**, around the median value. Despite the large range of τ_s values, the distribution

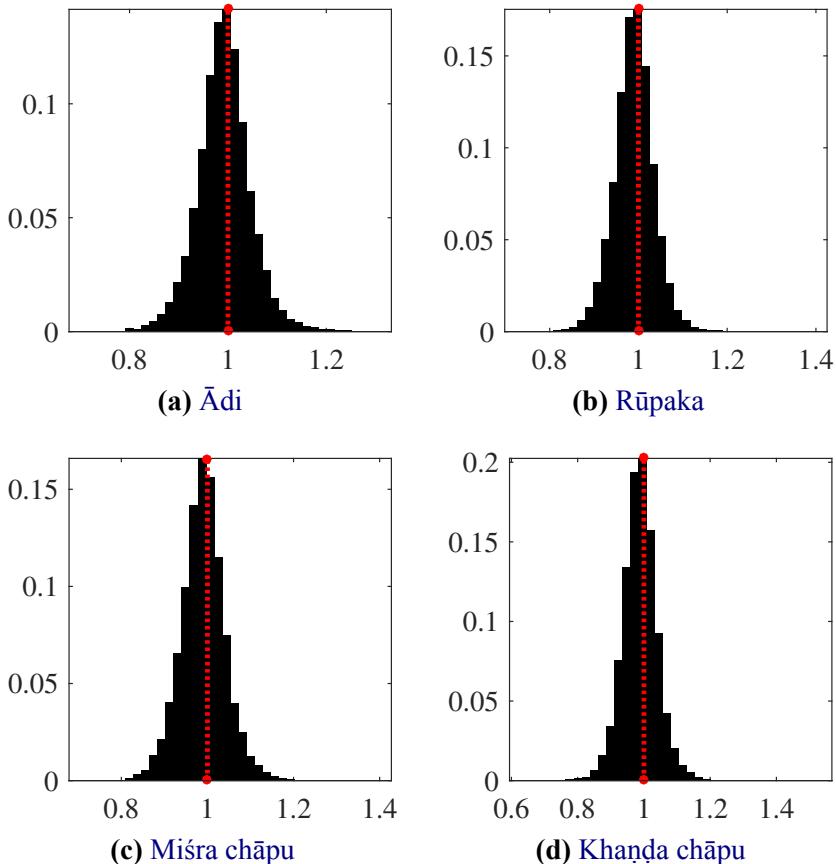


Figure 4.6: A histogram of the median normalized inter-beat interval τ_b in the CMR_f dataset for each tāla. The ordinate is the fraction of the total count corresponding to the normalized τ_b value shown in abscissa.

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.

To illustrate and measure the time varying tempo of music pieces in Carnatic music, we normalize all the τ_s and τ_b values in a piece by the median value of the piece to obtain median normalized τ_s and τ_b values, a histogram of which is shown for CMR_f dataset 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 median tempo of the piece for all *tālas*.

Rhythm patterns in CMR_f and 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 a potential application of corpus level analysis, while showing their utility for meter tracking in MIR, and for performance analysis and comparative analysis in musicology.

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 need 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 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), DFT 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 is then computed. Since the bass onsets have significant information about the rhythmic patterns, the features are computed in two frequency bands (Low: ≤ 250 Hz, High: > 250 Hz) to additionally consider

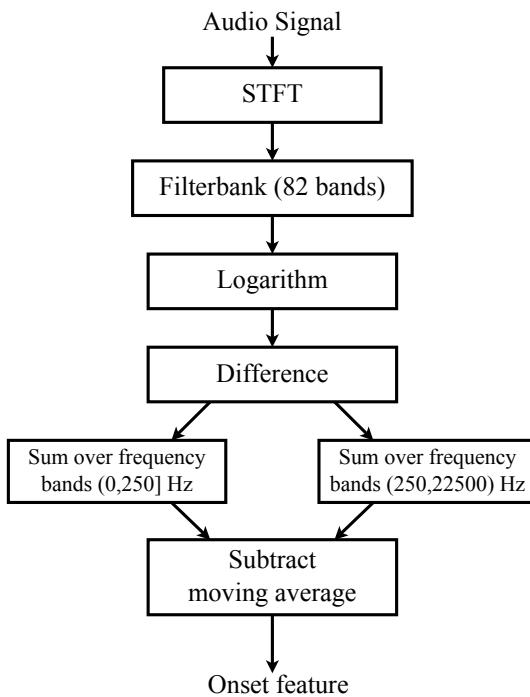


Figure 4.7: Computation of the spectral flux onset feature in two frequency bands, from Holzapfel et al. (2014).

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 such cycle length sequence instances for a specific *tāla* is computed in both the frequency bands and used as a representative rhythmic pattern illustrated here.

At the outset, it is necessary to note here that the patterns played in a *tāla* cycle are to be described using timing, energy and timbre descriptors. The rhythm patterns generated here using the spectral flux feature and can only explain timing and energy accents. A minor effect of timbre can be seen in these rhythm patterns, but are predominantly affected by the other two characteristics. These patterns are averaged over the whole dataset for a *tāla*, and hence cannot capture specific nuances of individual pieces, but only can give a broad perspective. The patterns here are indicative of the sur-

face rhythm present in the audio recordings, and hence completely reflect the underlying canonical metrical structures.

The rhythm patterns are roughly indicative of the energies 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 (and 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.

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.

We list down and discuss some salient qualitative observations from figures for each *tāla*, for both *CMR_f* dataset and its subset *CMR*. The Figures 4.8-4.15 show the cycle length rhythm patterns for all *tālas* for both *CMR_f* and *CMR* datasets. For each *tāla*, we plot the rhythm patterns together to compare patterns across the short excerpts in *CMR* dataset and full length pieces in *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 accents organized in three different strengths, reflecting the metrical levels of the *aṅga*, the beat and the *akṣara*. The two *akṣara* long beats in *miśra chāpu* and *khaṇḍa chāpu tālas*, and the four *akṣara* long beats in *ādi* and *rūpaka tālas* can be additionally seen. The patterns and *ṭhēkās* played in Carnatic music are quite diverse, and no obvious representative *tāla* pattern can be inferred, apart from the varied accents at three metrical levels.

The patterns illustrated here are average patterns that occur and

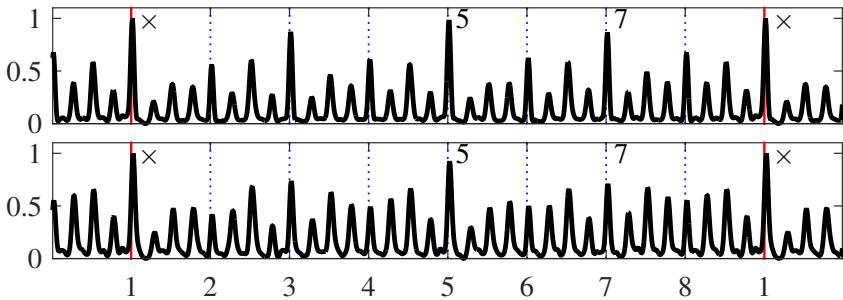


Figure 4.8: Cycle length rhythmic patterns learned from CMR_f dataset for ādi tāla. In each of the following Figures 4.8-4.15, the patterns are computed from spectral flux feature and averaged over all the pieces in the dataset. The bottom/top pane corresponds to the low/high frequency bands, respectively. 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 aṅga is indicated with beat numbers at the top of each pane (sama shown as x). The plot shows the cycle extended by a beat at the beginning and end to illustrate the cyclic nature of the tāla.

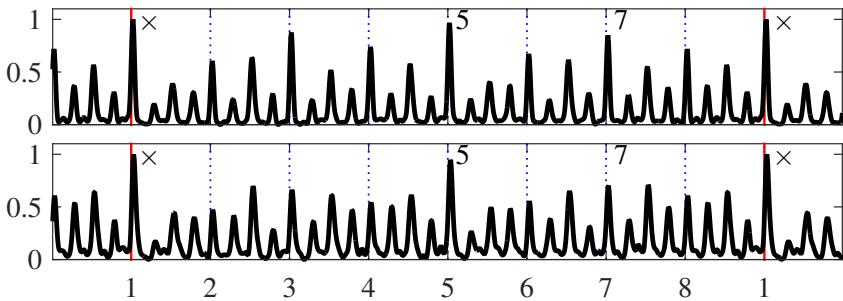


Figure 4.9: Cycle length rhythmic patterns learned from CMR dataset for ādi tāla.

do not tell us 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.

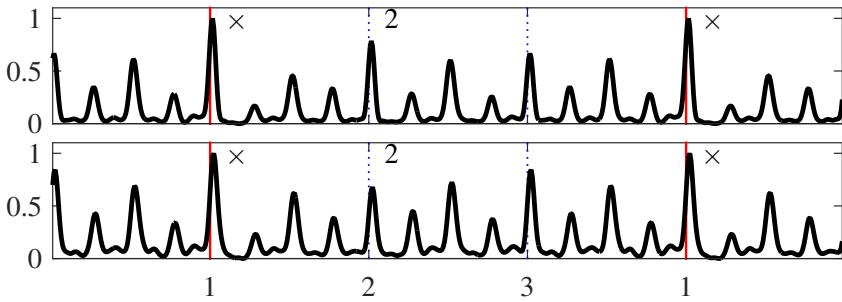


Figure 4.10: Cycle length rhythmic patterns learned from [CMR_f](#) dataset for [rūpaka tāla](#).

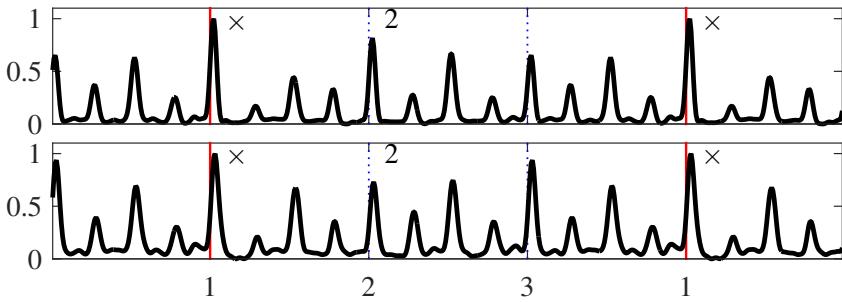


Figure 4.11: Cycle length rhythmic patterns learned from [CMR](#) dataset for [rūpaka tāla](#).

The rhythm patterns computed using [CMR](#) dataset are very similar to those computed using [CMR_f](#) dataset, showing that [CMR](#) is a good representative subset of the larger [CMR_f](#). Additionally, all the observations we make with patterns from [CMR_f](#) 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 - in-

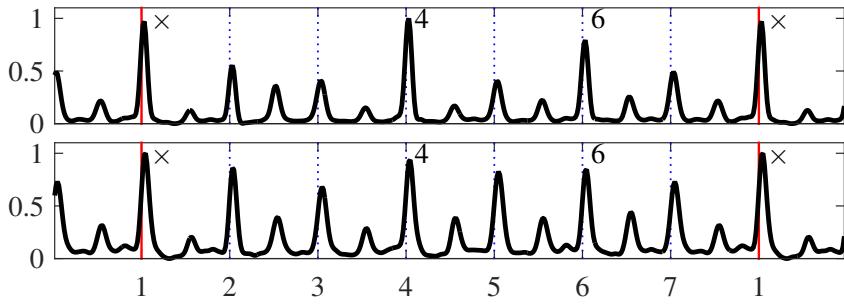


Figure 4.12: Cycle length rhythmic patterns learned from CMR_f dataset for miśra chāpu tāla.

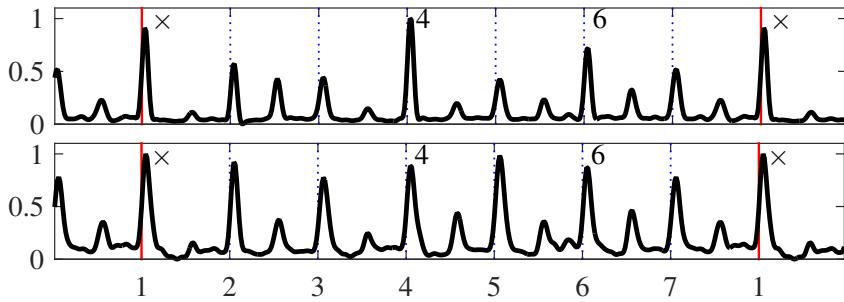


Figure 4.13: Cycle length rhythmic patterns learned from CMR dataset for miśra chāpu tāla.

dicating 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 anča 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.

The rhythm patterns of khaṇḍa chāpu tāla shown in Figures 4.14-4.15 have a strong left accent on beat 4, which is used an anchor within the cycle. A stronger right accent on beat 3 shows the progression through the unequal ančas. The 2+1+2 akṣara grouping

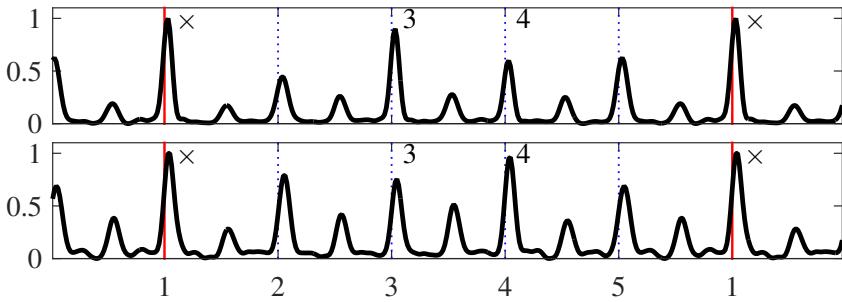


Figure 4.14: Cycle length rhythmic patterns learned from [CMR_f](#) dataset for [khanḍa chāpu tāla](#).

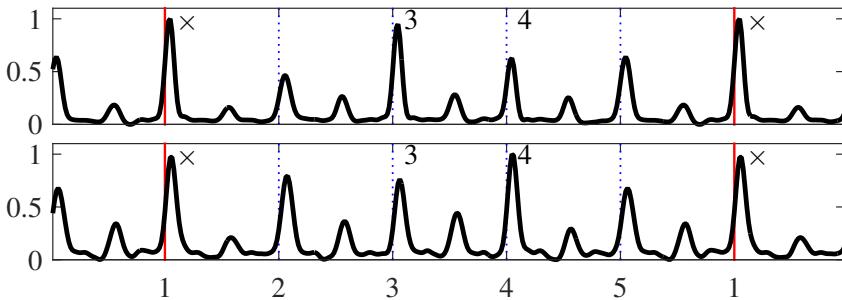


Figure 4.15: Cycle length rhythmic patterns learned from [CMR](#) dataset for [khanḍa chāpu tāla](#).

structure of [khanḍa chāpu](#) is often played out as 3+2 or 2+3, showing strong accents on beats 3 and 4.

These are some observations from rhythm patterns that have interesting musicological significance. A professional Carnatic musician has informally validated these observations, but they still have to be formally studied in depth to make valid musicological conclusions.

Applications of the Carnatic rhythm dataset

The [CMR_f](#) dataset and its subset [CMR](#) dataset are intended to be test corpora for several automatic rhythm analysis tasks in Carnatic music. Possible tasks include [sama](#) and beat tracking, tempo estimation and tracking, [tāla](#) recognition, rhythm based segmenta-

tion 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 with some experiments extended to the full CMR_f dataset to verify their applicability to larger datasets.

4.2.2 Hindustani music rhythm dataset

Hindustani Music Rhythm dataset (HMR_f)²⁶ is a rhythm annotated test corpus for automatic rhythm analysis tasks in Hindustani Music (Srinivasamurthy et al., 2016). The collection consists of audio excerpts from the CompMusic Hindustani research corpus, manually annotated time aligned markers indicating the progression through the *tāl* cycle, 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 research corpus. The pieces include a mix of vocal and instrumental recordings, new and old recordings, and to span three *lay*. For each *taal*, there are pieces in *dṛt* (fast), *madhya* (medium) and *vilambit lay*. 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 **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* cycle illustrated 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.

²⁶<http://compmusic.upf.edu/hindustani-rhythm-dataset>

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_f 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`).

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

Table 4.10: Tāl cycle length indicators for HMR_f dataset. $\bar{\tau}_s$ and σ_s indicate the mean and standard deviation of the median inter-`sam` interval of the pieces, respectively. $\bar{\tau}_b$ and σ_b 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 τ_s and hence the range of τ_s in the dataset. All values in the table are in seconds.

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_f dataset is described in Table 4.9, showing the four `tāls` and the number of pieces for each `tāl`, totaling to 151 pieces. The

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_I 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).

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

Table 4.12: Tāl cycle length indicators for HMR_I dataset. $\bar{\tau}_s$ and σ_s indicate the mean and standard deviation of the median inter-sam interval of the pieces, respectively. $\bar{\tau}_b$ and σ_b 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 τ_s and hence the range of τ_s in the dataset. All values in the table are in seconds.

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 than 150 ms to over 6 seconds. This huge range of cycle lengths and mātrā periods is a significant challenge in Hindustani music

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_s 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).

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

Table 4.14: Tāl cycle length indicators for HMR_s dataset. $\bar{\tau}_s$ and σ_s indicate the mean and standard deviation of the median inter-sam interval of the pieces, respectively. $\bar{\tau}_b$ and σ_b 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 τ_s and hence the range of τ_s in the dataset. All values in the table are in seconds.

automatic meter inference. Across different tāls, we see that tīntāl and ēktāl have the largest range of $\bar{\tau}_s$, since they are performed in all the lay classes, vilābit to dṝt. Jhaptāl and rūpak tāl have smaller $\bar{\tau}_s$ ranges.

The dataset consists of excerpts with a wide tempo range from 10 MPM (mātrās 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 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 (vilābit),

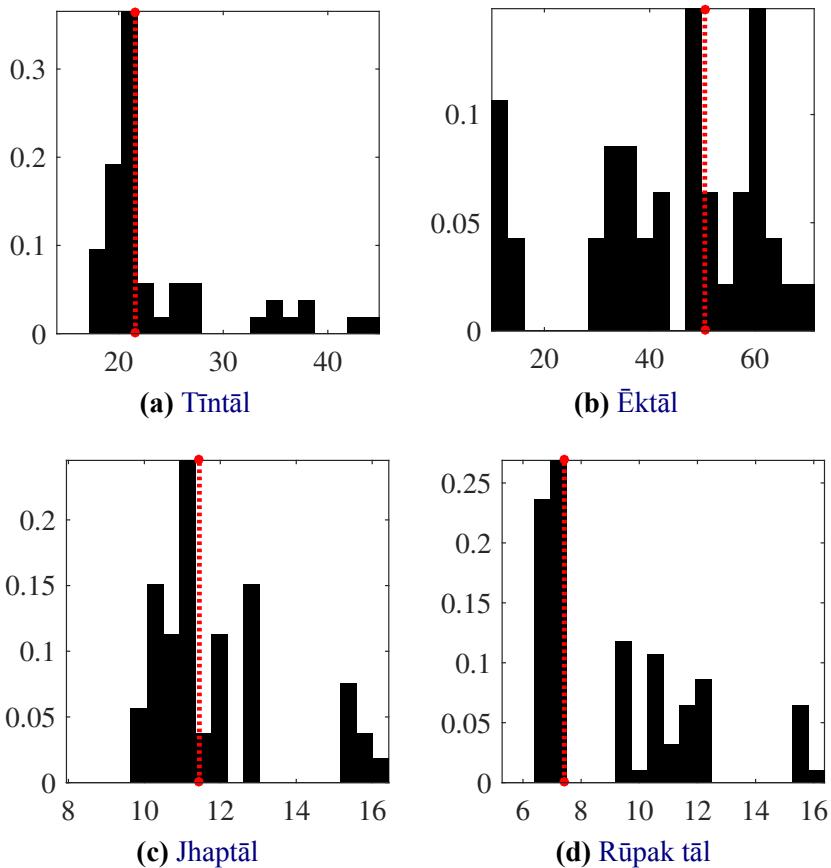


Figure 4.16: A histogram of the inter-sam interval τ_s in the HMR_1 dataset for each tāl . The ordinate is the fraction of the total count corresponding to the τ_b value shown in abscissa. The median τ_s for each tāl is shown as a red dotted line.

medium (**madhya**), and fast (**dṛt**) 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 duration subset called the HMR_1 dataset (shown in Table 4.11) consisting of **vilābit** pieces with a median tempo between 10-60 MPM, and the short cycle duration subset HMR_s dataset (shown in Table 4.13) with **madhya lay** (60-150 MPM) and the **dṛt lay** (150+ MPM) pieces.

HMR_1 dataset shown in Table 4.11 consists of 59 pieces in **vi-**

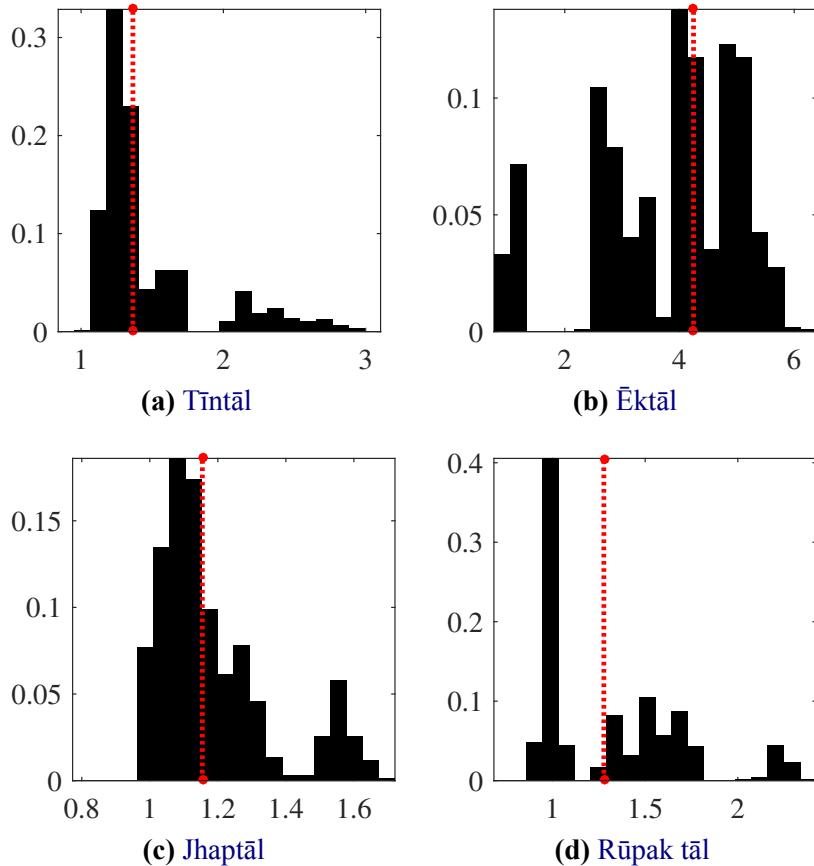


Figure 4.17: A histogram of the inter-mātrā interval τ_b in the HMR_1 dataset for each $tāl$. The ordinate is the fraction of the total count corresponding to the τ_b value shown in abscissa. The median τ_b for each $tāl$ is shown as a red dotted line.

lambit 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 **vilambit lay jhaptāl** and **rūpak tāl**, there are only 6 and 8 pieces for those $tāls$, respectively. As described with HMR_f , a basic statistical analysis of the $tāl$ cycle length indicators in Table 4.12 shows that the median inter-sam interval and its range for **jhaptāl** and **rūpak tāl** are less than that for **tīntāl** and **ēktāl**.

HMR_s dataset consists of 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 Ta-

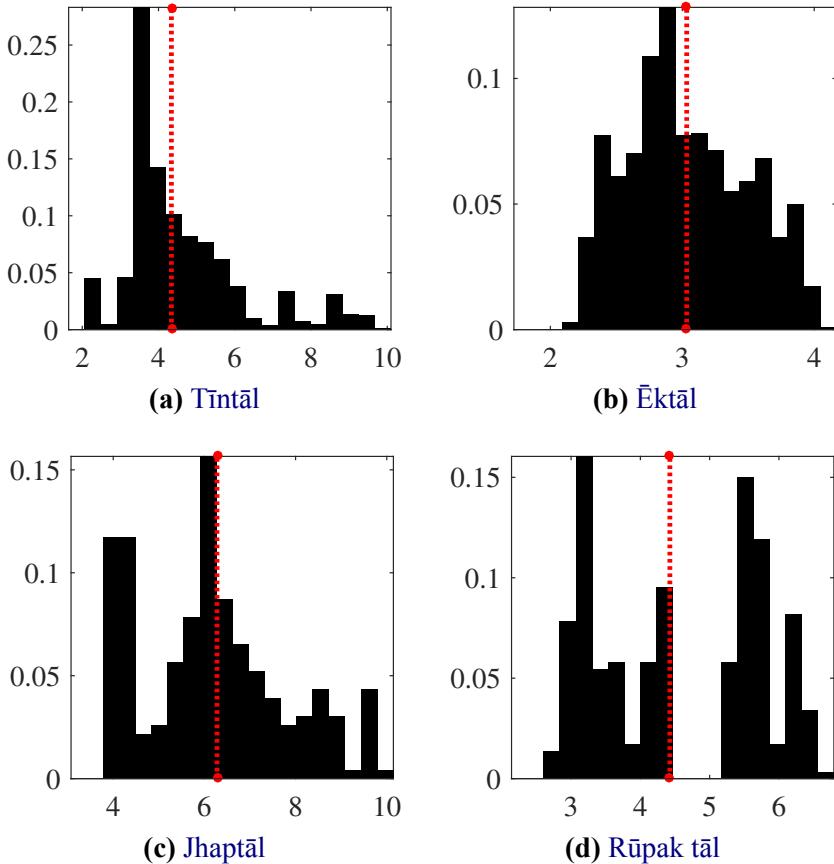


Figure 4.18: A histogram of the inter-sam interval τ_s in the HMR_s dataset for each tāl. The ordinate is the fraction of the total count corresponding to the τ_b value shown in abscissa. The median τ_s for each tāl is shown as a red dotted line.

ble 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 vilaṁbit or dṛt and its rare for a 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 inter-

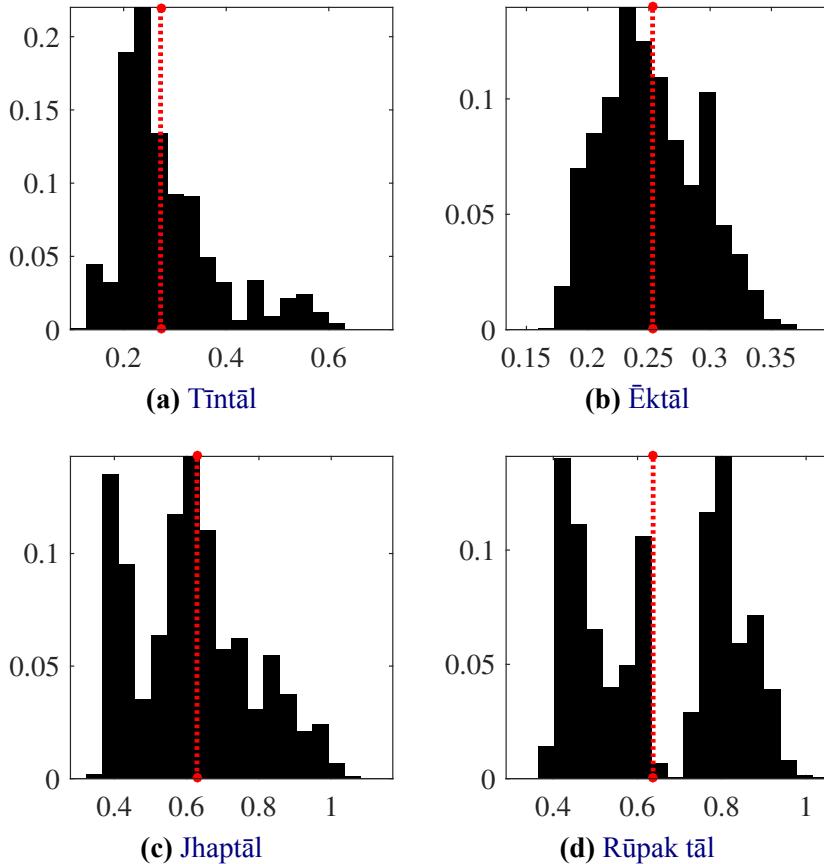


Figure 4.19: A histogram of the inter-mātrā interval τ_b in the HMR_s dataset for each tāl . The ordinate is the fraction of the total count corresponding to the τ_b value shown in abscissa. The median τ_b for each tāl is shown as a red dotted line.

val (τ_s) and inter-mātrā interval (τ_b) for each tāl . For better comparison, we present this analysis for each data subset HMR_I and HMR_s separately.

A histogram of τ_s and τ_b for each tāl for HMR_I dataset is shown in Figure 4.16 and Figure 4.17, respectively, and those for HMR_s 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 τ_s for each tāl , around the median value. The large range of τ_s and τ_b values and an irregular distribution spanning the whole range is seen with both datasets, unlike the Carnatic music

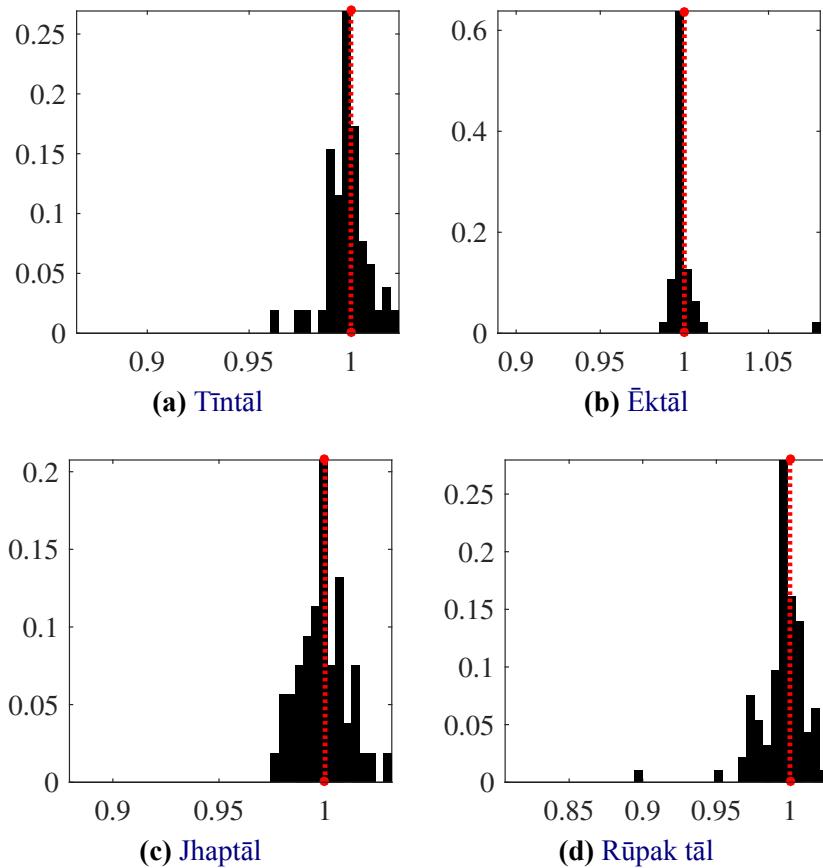


Figure 4.20: A histogram of the median normalized inter-sam interval τ_s in the HMR_1 dataset for each tāl . The ordinate is the fraction of the total count corresponding to the normalized τ_s value shown in abscissa.

CMR_f dataset with a smaller 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 τ_s and τ_b values in a piece by the median in the piece to obtain median normalized τ_s and τ_b values, a histogram of which is shown in Figure 4.20 and Figure 4.21, respectively for HMR_1 dataset and Figure 4.22 and Figure 4.23, respectively for HMR_s 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

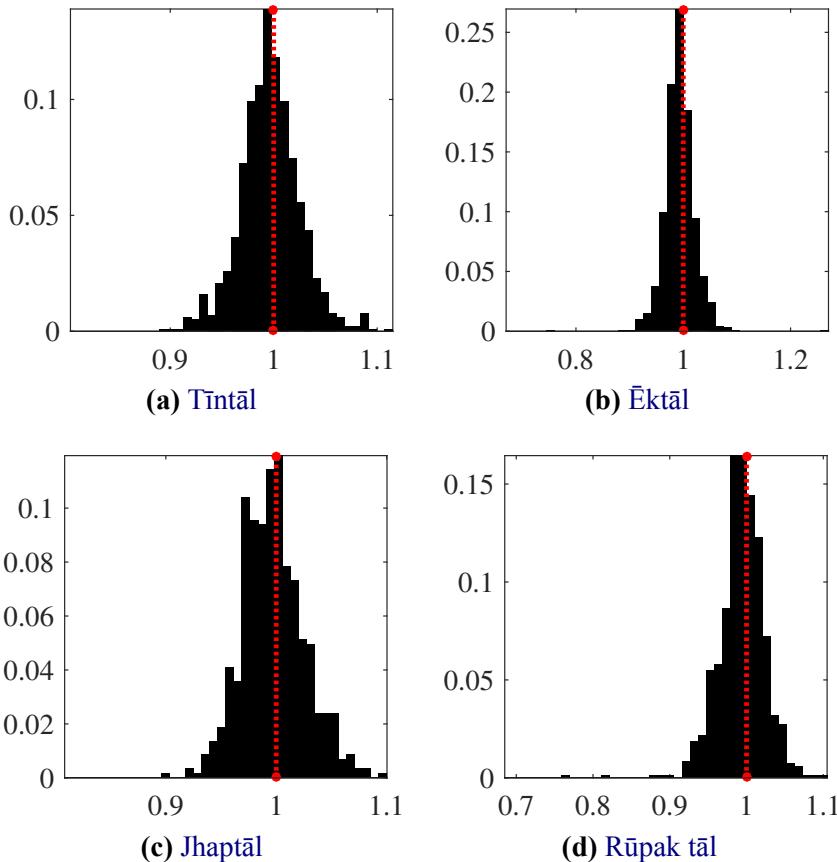


Figure 4.21: A histogram of the median normalized inter-*mātrā* interval τ_b in the HMR_1 dataset for each *tāl*. The ordinate is the fraction of the total count corresponding to the normalized τ_b value shown in abscissa.

tempo of the piece for all *tāls*. This is in contrast to Carnatic music where the median normalized tempo had a higher maximum deviation ($\sim 20\%$). One possible reason for this lower tempo deviation in Hindustani music compared to Carnatic music is because of lesser rhythmic improvisation, with the tabla acting as an accurate timekeeper. However, this could also be possibly due to the fact that the Hindustani pieces in the dataset are two minute short excerpts, compared to full length Carnatic pieces in the CMR_f dataset, and hence lesser tempo variability.

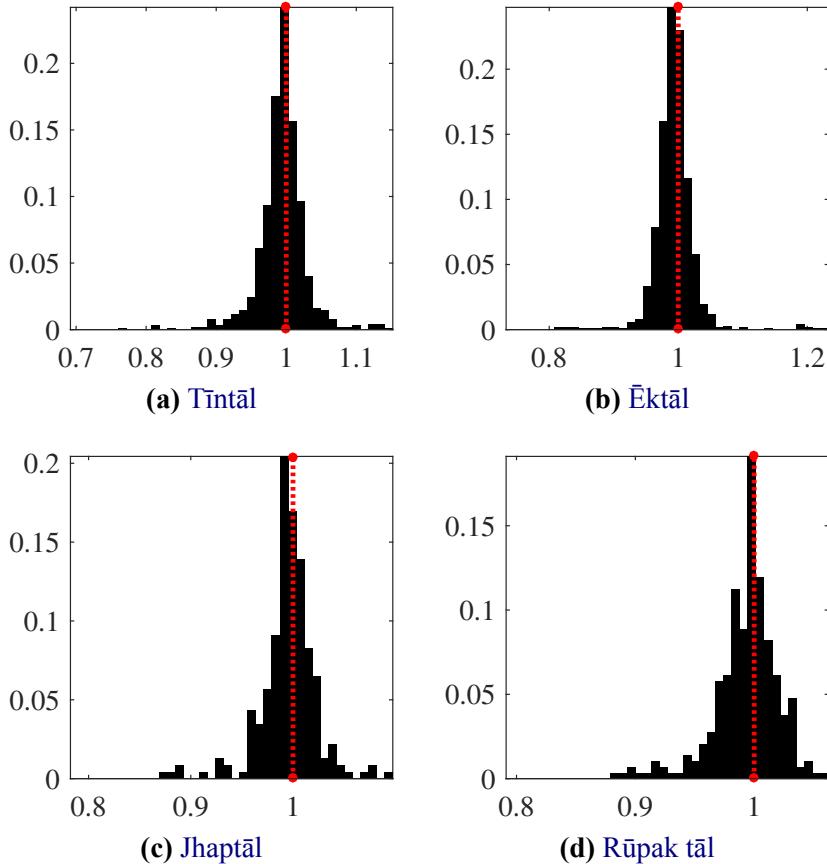


Figure 4.22: A histogram of the median normalized inter-sam interval τ_s in the HMR_s dataset for each tāl . The ordinate is the fraction of the total count corresponding to the normalized τ_s value shown in abscissa.

Rhythm patterns in Hindustani rhythm datasets

Similar to Carnatic music, we do corpora level analysis of rhythm patterns in Hindustani music and illustrate several musicological inferences and insights, and contrast if there are any differences between music theory and practice. The rhythm patterns described in this section were obtained using spectral flux, in an identical process as described for Carnatic music.

The Figures 4.24-4.31 show the cycle length rhythm patterns for all tāls for both HMR_1 and HMR_s datasets, using the spectral flux feature computed identically to the way it was computed for Car-

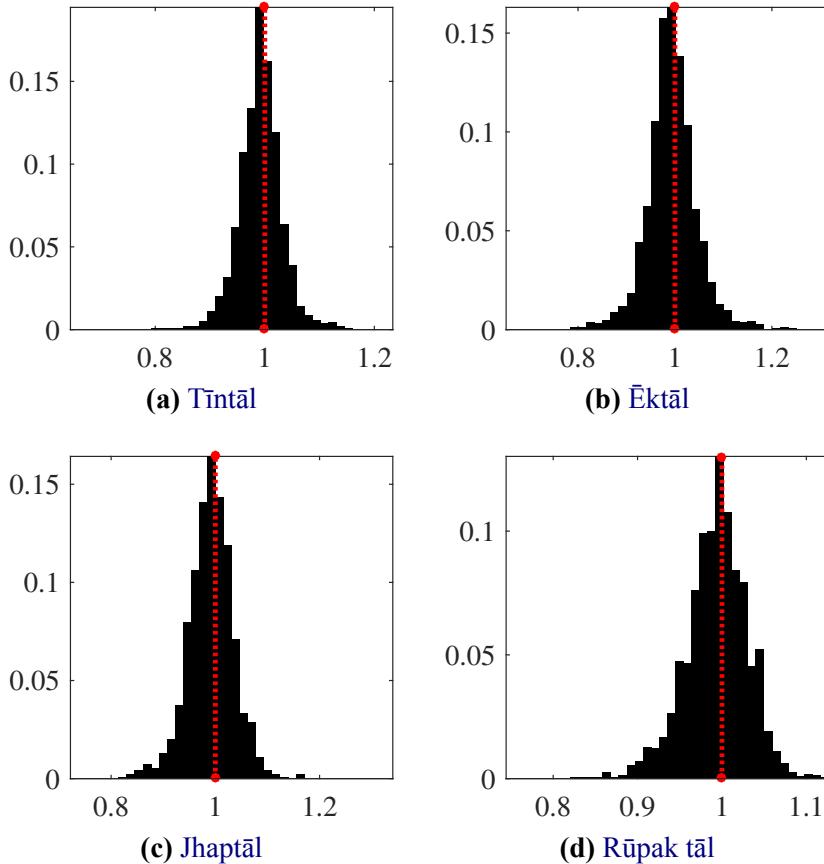


Figure 4.23: A histogram of the median normalized inter-*mātrā* interval τ_b in the *HMR_s* dataset for each *tāl*. The ordinate is the fraction of the total count corresponding to the normalized τ_b value shown in abscissa.

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_I** dataset (with **vilambit lay** pieces) and **HMR_s** 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 **vilambit lay** and **dṛ̥t lay**. The patterns are indicative of the surface rhythm present in these audio recordings.

There are several observations from the plotted rhythm patterns that have interesting musicological significance. A professional Hindustani musician has informally validated these observations, but they still have to be formally studied in depth to make valid musicological conclusions. 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**, **thēkās** 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 **vilambit** 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 concentrated towards the second half of the **mātrā**. The 1st **mātrā** (and often the 2nd **mātrā**) is quite empty with few accents, while the last few **mātrās** of the cycle have dense accents. This is to place a special emphasis on the **sam**, indicating the approaching of **sam** with fillers and dense stroke playing while there is a short recovery period after the **sam**.

with fewer strokes. 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 (τ_b) 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). Further, due to a large time interval available to play the **thēkā**, 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 have 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 the 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 phe-

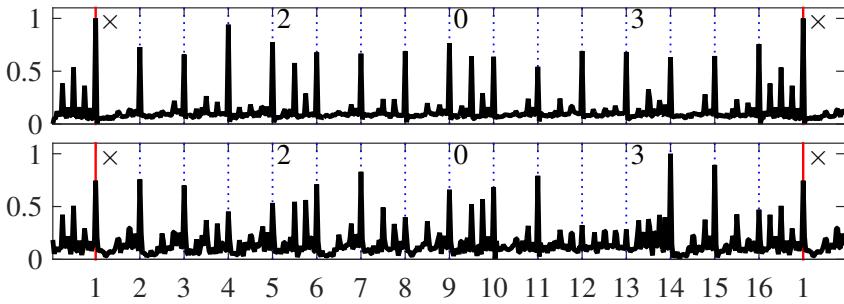


Figure 4.24: Cycle length rhythmic patterns learned from HMR_1 dataset for tintāl , computed from spectral flux feature and averaged over all the pieces in the dataset. The bottom/top pane corresponds to the low/high frequency bands, respectively. The abscissa is the mātrā number within the cycle (dotted 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 plot shows the cycle extended by a mātrā at the beginning and end to illustrate the cyclic nature of the tāl .

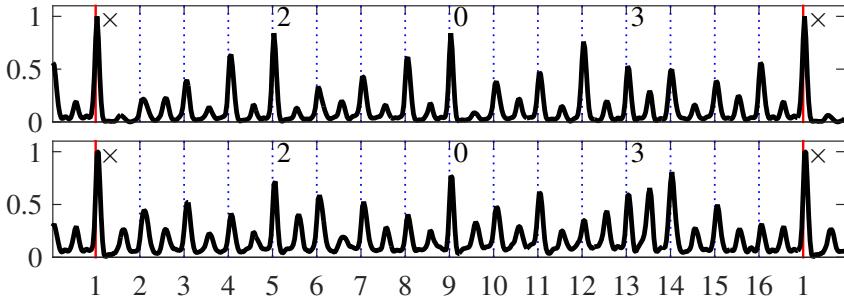


Figure 4.25: Cycle length rhythmic patterns learned from HMR_s dataset for tintāl .

nomenon known in music theory as **āmad** (literal meaning - the approach). A strong left accent on the 9th mātrā 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 known (to practising musicians) difference between theory and practice and can additionally be observed in the patterns too. As described earlier, the right stroke fillers are fewer in mātrās 1 and 2, and the left accents support the timekeeping task when the right accents are weaker there. 4th mātrā has a

strong right accent perhaps to indicate the end of the 1st *vibhāg*, after a filler-less *mātrās* 2 and 3. The beginning of the 2nd and 3rd *vibhāgs*, labeled 2 and 0 have higher number of fillers. The left accents between the 11th and the 14th matra are weak - with the 11th and 14th *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 *vilāmbit* 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 *vilāmbit*. The accents are more regular due to higher tempi associated. Similar to *vilāmbit*, the 9th matra has a strong left accent, which again is a well known difference between theory and practice. The 11th and 14th *mātrās* have strong left accents to support the build up of accents through *mātrās* 12-14 and indicate the arrival of *sam* (*āmad*). It is interesting to note that the right accent at *vibhāg* boundary (*mātrā* 13) is weaker than that at the previous *mātrā* 12. This is perhaps due to the stroke on *mātrā* 13 being skipped and a strong left stroke on *mātrā* 14 often played to indicate the approaching *sam*.

Vilāmbit 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, see Table 2.5) 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 distributed throughout 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

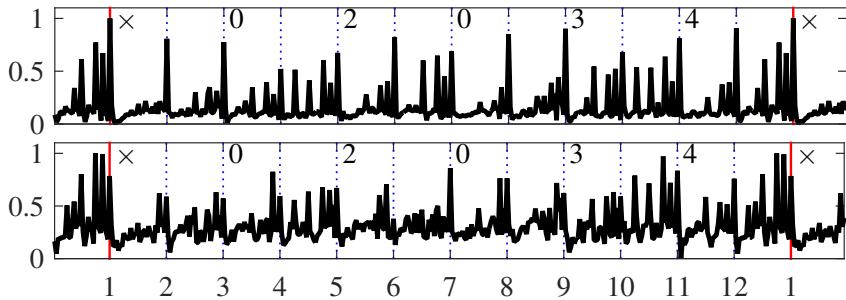


Figure 4.26: Cycle length rhythmic patterns learned from HMR_I dataset for ēktāl .

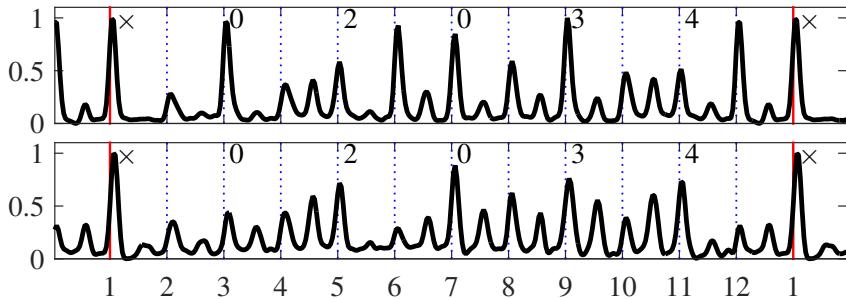


Figure 4.27: Cycle length rhythmic patterns learned from HMR_s dataset for ēktāl .

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*, which complement 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

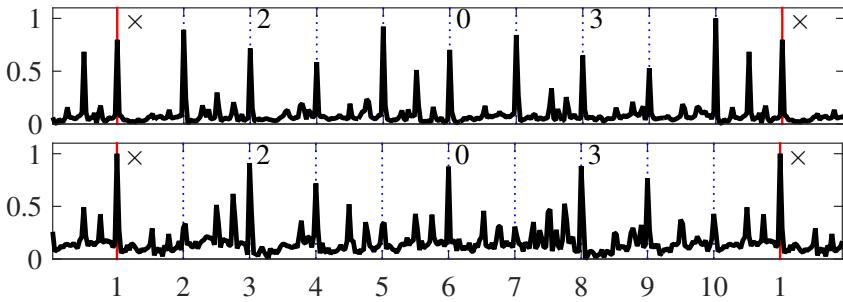


Figure 4.28: Cycle length rhythmic patterns learned from HMR_1 dataset for *jhaptāl*.

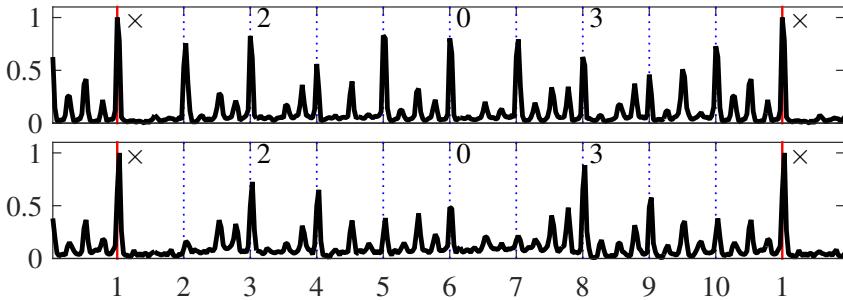


Figure 4.29: Cycle length rhythmic patterns learned from HMR_s dataset for *jhaptāl*.

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 TE 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 *vilambit 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.

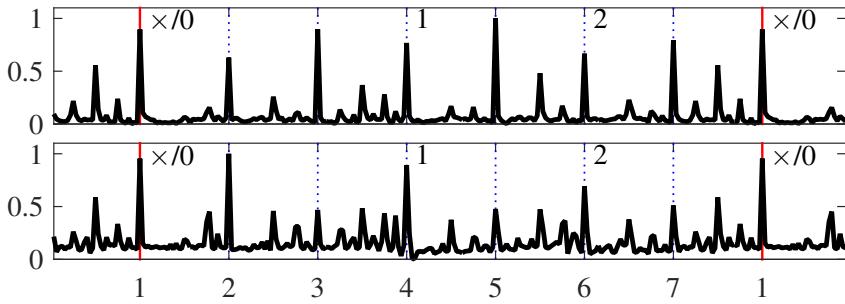


Figure 4.30: Cycle length rhythmic patterns learned from HMR_1 dataset for *rūpak tāl*.

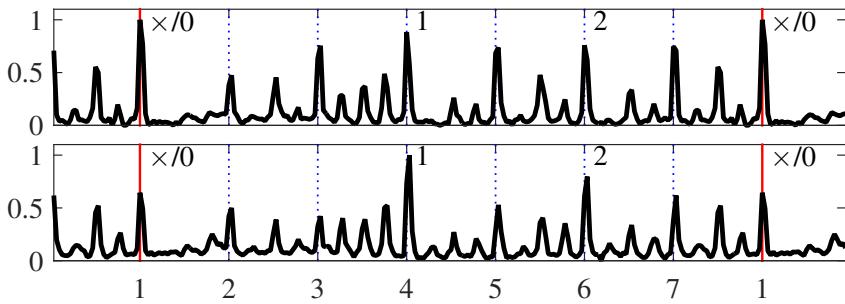


Figure 4.31: Cycle length rhythmic patterns learned from HMR_s dataset for *rūpak tāl*.

Vilambit 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 TE - KE DHI DHI NA | TI NA TE - 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 with basic *ṭhēkā* playing. The envelope of accents through the cycle is more regular than in *vilambit 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.

Vilambit 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 vilambit 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 stoke 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. In dṛt rūpak tāl, the accent on the second mātrā is softer than vilambit rūpak tāl, going back to its canonical 3+2+2 structure compared to 1+2+2+2 structure in vilambit rūpak tāl.

Applications of the HMR_f dataset

The HMR_f dataset and its subsets HMR_I and HMR_s datasets are intended to be test datasets for several automatic rhythm analysis tasks in Hindustani music. Possible tasks where the datasets 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 on the two subsets separately, to contrast performance of algorithms across different lay.

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 Mulgaonkar Tabla Solo dataset (MTS) 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.

4.2.3 Tabla solo dataset

The **Mulgaonkar Tabla Solo dataset (MTS)** 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 Mulgaonkar²⁷. 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 into a machine readable format. The scores obtained from OCR were manually verified and corrected for errors, while 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

²⁷<http://musicbrainz.org/release/220c5efc-2350-43dd-95c6-4870dc6851f5>

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 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 repository²⁸. 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 built as a part of CompMusic: 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.

Mridangam stroke dataset

The Anantapadmanabhan Mridangam Strokes dataset (AMS)²⁹ 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

²⁸<http://compmusic.upf.edu/tabla-solo-dataset>

²⁹<http://compmusic.upf.edu/mridangam-stroke-dataset>

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

Table 4.16: The [Anantapadmanabhan Mridangam Strokes dataset \(AMS\)](#). 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 from Table 2.2

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 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 \(UMS\)](#) is a transcribed collection of two *tani-āvartanas* (percussion solo) played by the renowned mridangam maestro Padmavibhushan Dr. 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

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 [UKS Mridangam Solo dataset \(UMS\)](#) 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.

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. 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 as 16 bit WAV files.

The audio file of each [tani](#) has been segmented into short musically relevant phrases, and each phrase has been transcribed into its constituent strokes, represented as syllables. The segmentation of audio files and syllabic transcription of each phrase was done by professional Carnatic percussionists. The transcriptions also include pauses (denoted by ,) 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. The transcription is not time aligned, but only a sequence of the strokes

Dataset	Bangu	Daluo	Naobo	Xiaoluo	Total
Training	59	50	62	65	236
Test	1645	338	747	291	3021

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

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 purposes³⁰ 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**) is an annotated collection of Beijing opera percussion instruments, with audio and time aligned onset annotations (Tian et al., 2014). 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 audio in the dataset was recorded 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 **xiaoluo**. Unlike pitched instruments, most idiophones cannot be tuned. These percussion instruments are made from metal casting or wood 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

³⁰<http://compmusic.upf.edu/mridangam-tani-dataset>

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 Audacity³¹ into 30-second long tracks, with possibly simultaneous onsets to closely emulate the real world conditions. The examples in training and test dataset are mutually exclusive.

For the onset annotations, manual labeling of onset locations was 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 reuse³². The dataset can be used for training models for each percussion instrument class, and MIR tasks such as percussion instrument identification, source separation, and instrument-wise onset detection, as used by Tian et al. (2014).

4.2.6 Jingju percussion pattern dataset

The Jingju Percussion Pattern dataset (JPP) is a collection of audio examples and scores of percussion patterns played by the percussion ensemble in Beijing opera (Srinivasamurthy, Caro, et al., 2014). 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

³¹<http://audacity.sourceforge.net>

³²<http://compmusic.upf.edu/bo-perc-dataset>

ID	Pattern Class	# Instances	$\bar{T}_f (\sigma)$
1	daoban tou 【导板头】	66	8.70 (1.73)
2	man changchui 【漫长锤】	33	13.99 (4.47)
3	duotou 【夺头】	19	7.18 (1.49)
4	xiaoluo duotou 【小锣夺头】	11	8.16 (2.15)
5	shanchui 【闪锤】	8	10.31 (3.26)
Total		133	9.85 (3.69)

Table 4.19: The Jingju Percussion Pattern dataset (JPP dataset). The last column is the mean pattern length (\bar{T}_f) and standard deviation (σ) in seconds. Figure 2.5 shows the music scores for these patterns.

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 music pieces and audio 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.

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 class it belongs to, 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 a few 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)**³³ label files. There is also a single master label file provided with all the annotations for batch processing using **HTK**.

The annotations are publicly shared and available to all³⁴. The audio is from commercially available releases and can be easily accessed using the associated **MBIDs**. The dataset can be used for instrument-wise onset detection, and percussion pattern transcription and classification, as applied by Srinivasamurthy, Caro, et al. (2014).

4.2.7 Other evaluation datasets

As discussed in Section 2.3.3, there are several datasets available for beat tracking evaluation, used in **MIREX** or otherwise, e.g. SMC dataset (Holzapfel et al., 2012), Ballroom dataset (Gouyon et al., 2006), McKinney dataset (Moelants & McKinney, 2004), RWC database (Goto, 2006), Hainsworth dataset (Hainsworth & Macleod, 2003), and GTZAN-Rhythm dataset (Marchand et al., 2015). Of all these datasets, in addition to Indian art music rhythm datasets, we use the Ballroom dataset for evaluating algorithms and approaches presented in the thesis. There are other non-eurogenetic music (Turkish and Cretan music) datasets for testing automatic rhythm analysis approaches. We do not use them and present any evaluations on those datasets, but they are briefly described for completeness.

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)

³³HTK: <http://htk.eng.cam.ac.uk/>

³⁴<http://compmusic.upf.edu/bopp-dataset>

and has been widely used for several MIR tasks such as genre classification, tempo tracking, beat and downbeat tracking, e.g. by Gouyon et al. (2006); Krebs, Holzapfel, et al. (2015); Böck et al. (2014).

It consists of 697 thirty second 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.

Turkish rhythm dataset

The Turkish rhythm dataset was compiled and annotated by Dr. Andre Holzapfel (Holzapfel et al., 2014) and is an extended version of the annotated data used by Srinivasamurthy, Holzapfel, and Serra (2014). 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 makam 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 *Düyek*.

Cretan music dataset

The Cretan music dataset consists of 42 full length music pieces of Cretan leaping dances compiled and annotated by Dr. Andre Holzapfel (Holzapfel et al., 2014). 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 all the pieces of the dataset 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 dataset.

To summarize, the chapter presented a comprehensive discussion of the research corpora and test datasets useful for rhythm related MIR tasks, focusing on Indian art music. The corpora and datasets are easily accessible and hence are valuable resources for data-driven MIR. An illustrative analysis of the Indian art music rhythm datasets showed a good potential for data-driven computational musicology research. The Indian art music rhythm and percussion datasets, along with the Ballroom dataset will be extensively used for the experiments in meter analysis and percussion pattern discovery.

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 to formulate 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 for 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 available audio data and the related additional metadata. We continue and elaborate on the basis of the formulation presented in Section 3.3.1.

Meter inference

Given an audio music recording, meter inference aims to estimate the rhythm class (or meter type or *tāla*), possibly time-varying tempo, beats and downbeats. In the context of Carnatic music, the task of meter inference aims to recognize the *tāla*, estimate the time varying tempo (τ_o or τ_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*, estimate the time varying tempo (τ_b), 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 *tālas* having similar structure. In Carnatic music, it is quite possible that the same composition can be performed in two different *tālas*, which further can lead to confusion, e.g. the compositions in *rūpaka tāla* (12 *akṣaras* in a cycle) can be performed in *tiśra nade ādi tāla* (24 *akṣaras* in a cycle, see Figure 2.2). A common example is the composition Himadri Suthe¹ in *rāga Kalyāṇi*.

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 or *tāla*) 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

¹<https://musicbrainz.org/work/6155262b-601a-41ba-8dcf-6f5b15b744f6>

an audio music recording, given the **tāla**. For Hindustani music, given the **tāl**, the task aims to track the time varying τ_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 the most comprehensively addressed task in this thesis. We explore different approaches and evaluate them on the rhythm annotated datasets of Carnatic and Hindustani music. The proposed novel 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.

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 higher level information improves 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 is hypothesized 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 2.3.2 (one such algorithm for Carnatic music is also described later in this chapter in Section 5.2.1). Since the tempo of a piece can vary over time, a narrow range of tempo for 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-automatic meter tracking system, where a human listener can tap along to the first one (or few) downbeats of the piece and an automatic meter tracker would then 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 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 Plumley (2006) and 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 for 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, the approach 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 τ_o , the **akṣara** pulse locations, and the **sama**. For estimating these components, a set of rhythm descriptors are first computed from the audio that are indicative of 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 proposed by Grosche and Müller (2011b) is used to

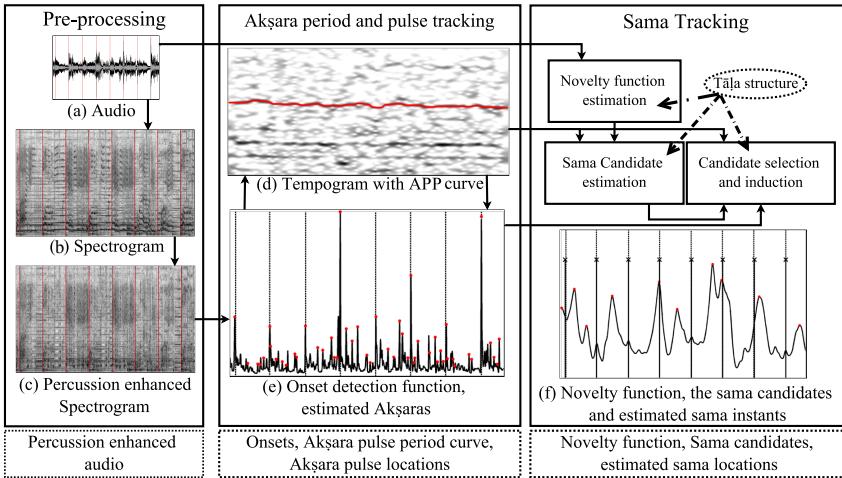


Figure 5.1: Block Diagram of the algorithm showing the signal flow and representative illustrations of different stages of the algorithm. The important outputs at each stage are also shown at the bottom. In each panel, the vertical lines that run through the panel indicate the **sama** ground truth instants. The estimated **sama/akṣara** candidates are shown with red dots and the estimated **sama** are shown with \times .

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 based on periodicity constraints, which leads to the final estimates. A block diagram of the approach is shown in Figure 5.1. 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 $f[n]$ obtained from a music piece z , as it has been shown to improve beat tracking performance in pieces with predominant vocals by Zapata and Gómez (2013). The predominant melody (F_0) 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

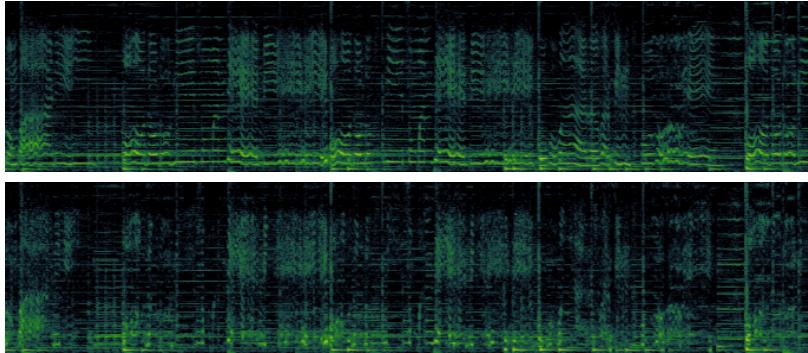


Figure 5.2: An illustration of percussion enhancement on a short audio excerpt of Carnatic music. The figure shows the spectrogram of the audio excerpt, before percussion enhancement (top panel) and after percussion enhancement by suppressing the lead melody (bottom panel). The lead melody is suppressed, while the **tambūra** (drone) is still present.

enhanced signal $f_p[n]$, with the harmonic component suppressed, is used for further processing (Figure 5.1(c)). An illustration of percussion enhancement for a short audio excerpt of Carnatic music is shown in Figure 5.2.

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. For each audio frame k ($k \leq K$), the first function ($y_f[k]$) uses the whole frequency range of the spectrogram and the other function computes the spectral flux only in the range of $0 - 120$ Hz ($y_l[k]$) and captures the low frequency onsets of the left (bass) drum head of the mridangam.

The function $y_f[k]$ is used to compute a Fourier-based Tempogram \mathbf{G} proposed by (Grosche & Müller, 2011b), computed every 0.25 second using a 8 second long window (Figure 5.1(d)). If the time indexes at which the tempogram is computed is denoted with i , ($1 \leq i \leq K_G$), the most predominant τ_o curve can be tracked by estimating the best path $\Gamma = \{\gamma_i : i = 1, 2, \dots, K_G\}$ through the tempogram matrix \mathbf{G} that provides a balance between tempogram amplitude at time index i , $\mathbf{G}_{\gamma_i, i}$, and the local continu-

ity of τ_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.

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

The function $\mathfrak{O}(\gamma_i/\gamma_{i+1})$ is an extra penalty term to penalize tempo doubling and halving between adjacent frames, and the parameters $\theta_1 (= 0.01)$ and $\theta_2 (= 10^6)$ provide different weights to the three terms. Based on observations from the CMR_f dataset, the search for the best path through the tempogram is restricted between the range of 120 to 600 APM (*akṣaras* per minute).

The above objective function is solved using a **Dynamic Programming (DP)** based approach to obtain a τ_o curve. Assuming the longest tracked τ_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 Γ^* (Figure 5.1(d), Γ^* is shown as a thick red line). Using the τ_o and the *tāla* information, we can obtain the time varying τ_s curve for the piece by multiplying the τ_o by the number of *akṣaras* in a cycle of the *tāla*. A further example of a tempogram and the estimated time varying τ_o curve for a piece of Carnatic music² from CMR_f dataset is shown in Figure 5.3. The figure shows the variations in tempo through a Carnatic music piece.

The *akṣara* pulse locations predominantly lie on strong mridangam onsets. The *akṣara* pulse candidates are estimated as the peaks of the function $y_f[k]$. Using these κ candidate peaks $\{o_i\}$, $i = 1, 2, 3, \dots, \kappa$, with locations t_i and peak amplitude ξ_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 *akṣara* period. The best set of candidates $\mathcal{O}_z = \{o_i^*\} \subset \{o_i\}$ are estimated using a **DP** approach (Figure 5.1(e)).

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

²Kamalamba, a *kṛti* in *rāga* Ānandabhairavi and *miśra chāpu tāla*, from the album Madrasil Margazhi 2005 by Aruna Sairam: <http://musicbrainz.org/recording/3baa722d-480e-4ae7-8559-a88dce41e1d4>

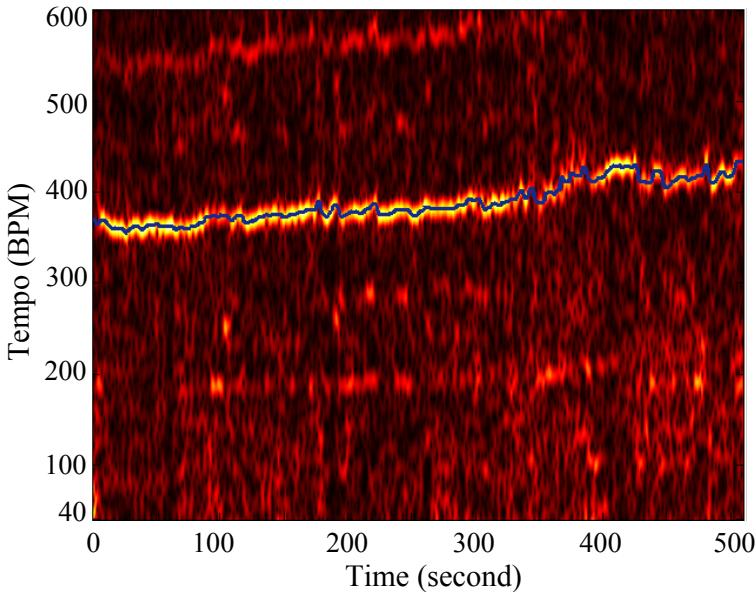


Figure 5.3: Estimated time varying tempo curve (shown as a bold blue line) plotted on top of a tempogram, for a Carnatic music piece. In the piece, apart the local tempo variations, we can see that the tempo increases with time. The tempogram shows high values in tempo octave related bands, with the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

The function $\Upsilon(t_i, t_{i+1}, \Gamma)$ is a function that returns an exponentially decaying weight based on the time difference between t_i and t_{i+1} in relation to the local **akṣara** period, γ_{t_k} . The parameter $\delta (= 3)$ provides a tradeoff between the two terms.

Sama tracking

The use of **MFCC** as features for timbral characteristics is explored. As a detection function for **sama** ($y_s[k]$), 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 5.1(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 **khaṇḍa chāpu** respectively so that the novelty function is computed over about an **āvartana**.

The peaks of the novelty function $y_s[k]$ 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 (~ 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 $y_s[k]$ and estimated τ_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 τ_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 τ_s is used with 10 randomly chosen **akṣara** locations from $\{o_i\}$ as **sama** candidates. RB-1 neither uses the τ_s , nor the candidate estimation using $y_s[k]$, while RB-2 uses the estimated τ_s but not the candidate estimation using $y_s[k]$.

Starting with the **sama** candidates obtained either from Method-A or Method-B, for each candidate, the **tāla** cycles are induced based on local τ_s period obtained from the τ_s curve. For each seed, the next and previous three estimated cycle periods are searched for onset peaks in $y_f[k]$ 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. For each candidate, an estimated **sama** sequence is thus obtained. Since all candidates are not necessarily **sama** locations, though the estimated τ_s is right, the sequences can have different offsets.

The final step of the algorithm is to shift, align and merge these

Measure	CML	AML
τ_o estimation	81.2	98.9
τ_o tracking	80.4	96.3

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

sequences obtained from each candidate. Starting with the longest **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 τ_s are merged into a single **sama** estimate aligned with the closest left stroke onset obtained from $y_l[k]$. This forms the final set of **sama** locations $\mathcal{S}_z = \{s_{t_i}\}$ estimated from the candidates and the onset detection function, as shown in Figure 5.1(f) with \times .

Results

The annotated **CMR_f** dataset has annotations only for beats and **samas** of the piece. From the **sama** locations, we can obtain the ground truth for τ_s curve, and hence the ground truth for τ_o curve. Since we do not have the ground truth for **akṣara** locations, we present the results only for tempo (τ_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 of the piece 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

Variant	p	r	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 p : Precision, r : Recall, f : f-measure, \mathfrak{I} : information gain, are shown. The values are mean performance over the whole **CMR_f** dataset. The last column shows the fraction (as a percentage) of the estimated *sama* candidates that are true samas.

estimation and *akṣara* period tracking and further, there is not a significant difference between their performances, indicating that the 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 τ_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 (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 (τ_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* candi-

date 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 τ_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 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 pro-

gramming is also ad hoc and loosely uses the tightly coupled information between the tempo, *akşaras* and the *sama*.

Considering these insights and limitations, we explore Bayesian models for meter inference, which provide an effective probabilistic 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 used by listeners and musicians.

As discussed earlier in Section 2.4.1, a **Dynamic Bayesian Network (DBN)** (Murphy, 2002) is well suited for meter analysis, since it relates variables over time through conditional (in)dependence relations. The bar pointer model is one such **DBN** 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 this chapter, 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 Car-

natic and Hindustani music test datasets presented in Chapter 4, with additional evaluations on the Ballroom dataset to test for generalization and to baseline performance. An extensive evaluation of the algorithms in the thesis is on the most relevant task of meter tracking, while meter inference and informed meter tracking tasks are 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). The following 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 (Srinivasamurthy et al., 2015).
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 (Srinivasamurthy et al., 2016).

Extensions and enhancements to inference schemes on the bar pointer model extensions are 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).

Finally, an evaluation of these algorithms is presented in Section 5.4, followed by a discussion and summary of the experiments and results.

5.3.1 The bar pointer model

The bar pointer model (or dynamic bar pointer model), referred to as BP-model in this chapter, is a generative model that has been successfully applied for meter analysis tasks. The model assumes a hypothetical time pointer within a bar that progresses at the speed of the tempo to traverse through the bar and then reinitializes at the end of the bar to track the next bar. The model also assumes that

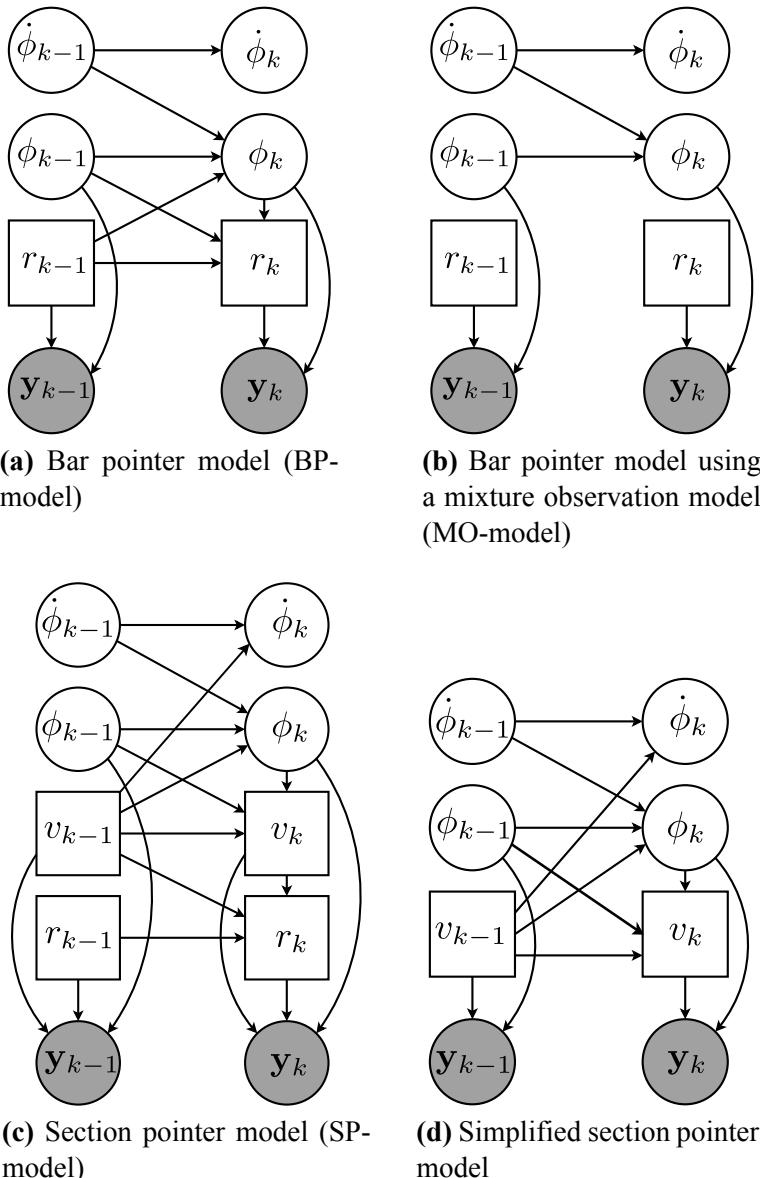


Figure 5.4: The meter analysis models used in the dissertation. In each of these DBNs, circles and squares denote continuous and discrete variables, respectively. Grey nodes and white nodes represent observed and latent variables, respectively.

specific bar length rhythm patterns are played in a bar depending on the rhythmic style, and uses these patterns to track the progres-

sion 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 and allowed in 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 (Whiteley et al., 2006). 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

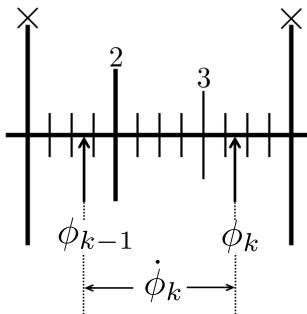


Figure 5.5: An illustration of the progression of bar position and instantaneous tempo variables over two consecutive audio frames in a cycle of *rūpaka tāla*. The effect of instantaneous tempo is greatly exaggerated for clarity in the illustration.

pattern of a rhythm class. Each pattern r has an associated length 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 analysis frame 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. A uniform initialization is used for meter inference and tracking, while a narrower informed initialization is done for informed meter tracking.

Transition model

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

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

The individual terms of the equation can be expanded as,

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

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,

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

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 μ and variance σ^2 . The value of $\sigma_{\dot{\phi}_k}$ depends on the value of tempo to allow for larger tempo variations at higher tempi. We set $\sigma_{\dot{\phi}_k} =$

$\sigma_n \cdot \dot{\phi}_{k-1}$, where σ_n is a user parameter that controls the amount of local tempo variations we allow in the music piece. The pattern transitions are governed by,

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

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 (e.g. the positions corresponding to downbeats, beats). 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: ≤ 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 64th 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 64th 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 (2×2), 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 of:

- downbeat (**sama**) instants: $\mathcal{S}_z = \{t_k^* \mid \phi_k^* = 0\}$
- beat instants: $\mathcal{B}_z = \{t_k^* \mid \phi_k^* = i \cdot M_r / B_r, i = 1, \dots, B_r\}$
- local instantaneous tempo: ϕ_k^*

- estimated rhythmic patterns: r^*

Two different inference schemes are now described, an inference using the Viterbi algorithm in a discretized state space, and an approximate inference using particle filters in the continuous space of ϕ 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 inference can be performed using the Viterbi algorithm to compute the most likely sequence of hidden states given the observed feature sequence.

We follow a discretization scheme that is identical to the method proposed by Krebs, Holzapfel, et al. (2015), by replacing the continuous variables ϕ 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 ϕ 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 Viterbi decoding inference algorithm in BP-model as described in the section will be denoted as **HMM₀** 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 $\mathcal{S} = M \cdot N \cdot R$ and needs a $\mathcal{S} \times \mathcal{S}$ 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 $\mathcal{S} = 96000$ states. The computational complexity of the Viterbi algorithm is $O(K \cdot |\mathcal{S}|^2)$. Even though the state transition matrix is sparse due to a lower 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 **SMC** 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 $\{w_K^{(i)}\}$, $i = 1, \dots, N_p$, $\mathbf{x}_{1:K}$ is the set of all state trajectories until

frame K , $\delta(x)$ is the Dirac delta function, and N_p is the number of particles.

With this particle system, 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 clearly 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 $P(\mathbf{x}_k | \mathbf{y}_{1:k})$ 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 ϕ . 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,

$$d(i, j) = \lambda_\phi \left[(\cos(\varphi^{(i)}) - \cos(\varphi^{(j)}))^2 + (\sin(\varphi^{(i)}) - \sin(\varphi^{(j)}))^2 \right] \\ + \lambda_{\dot{\phi}} \left(\dot{\phi}^{(i)} - \dot{\phi}^{(j)} \right)^2 + \lambda_r (r^{(i)} - r^{(j)})^2 \quad (5.16)$$

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 reclustering 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 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 BP-model as described in this section will be denoted as 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

Algorithm 1 An outline of the AMPF_0 algorithm (AMPF for inference in BP-model)

```

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  $\mathbf{x}_{1:K}^* = \mathbf{x}_{1:K}^{(i*)} | i^* = \arg \max_i w_K^{(i)}$  ▷ MAP sequence

```

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 BP-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 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 an 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 propose changes to the model structure of the BP-model, either by adding additional hidden variables or using different conditional independence relationships, and inference extensions that aim to improve inference in BP-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 as described in Section 5.3.1 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.

In meter tracking, since all the rhythmic patterns belong to the same type of meter, we can simplify BP-model to track only the ϕ and $\dot{\phi}$ 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 ϕ 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 $\phi_k = (\phi_{k-1} + \dot{\phi}_{k-1}) \text{ mod}(M)$ 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. Since MO-model assumes that the length of all rhythmic patterns are equal, it cannot be applied for the task of meter inference where many different **tālas** of different lengths are present, but can be applied for the task of meter tracking.

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 MO-model, denoted to as **HMM_m** in this dissertation. Similarly, the **AMPF** with the MO-model extension is outlined in Algorithm 2 and is denoted as **AMPF_m** in the rest of the chapter.

Algorithm 2 Outline of the **AMPF_m** algorithm (**AMPF** for inference in 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 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  $\dot{\phi}_k^{(i)} \sim P(\dot{\phi}_k^{(i)} | \phi_{k-1}^{(i)})$ 
15: Compute  $\beta_{1:K}^* = \beta_{1:K}^{(i*)} | i^* = \arg \max_i w_K^{(i)}$   $\triangleright$  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 the longer cycle length patterns, shorter patterns 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, the paper assumes that the beats form a regular isochronous sequence - an assumption that does not hold for many musics of the world, such as Indian, Turkish, Balkan, or Korean musics. Furthermore, the paper does not attempt to infer higher level metrical information, e.g. 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 BP-model: the restrictions to short cycles and isochronous beat sequences (Srinivasamurthy et al., 2016). The generalization of the BP-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 SP-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 BP-model, with the BP-model being a special case of the SP-model. Hence the SP-model can be applied to arbitrary music styles in a straight forward way, just like the BP-model.

Meter tracking in Indian art music is a suitable case for testing the SP-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 musically meaningful cycle divisions.

Hindustani music has *tāl* cycles that last over a minute (Clayton, 2000), which is a good test case for the SP-model. The large tempo range and the filler strokes in Hindustani music (especially *vilaṁbit* pieces) can provide a denser 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 *vilaṁbit* 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 SP-model, a hypothetical pointer traverses each section of a metrical cycle. Hence, in addition to the variables ϕ , $\dot{\phi}$, 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 as $\mathbf{x}_k = [\phi_k, \dot{\phi}_k, r_k, v_k]$, 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 and without loss of generality, 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 ϕ 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 . ϕ 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 ϕ 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,

$$\begin{aligned} P(\mathbf{x}_k \mid \mathbf{x}_{k-1}) &= P(\phi_k \mid \phi_{k-1}, \dot{\phi}_{k-1}, v_{k-1}) P(\dot{\phi}_k \mid \dot{\phi}_{k-1}, v_{k-1}) \\ &\quad P(v_k \mid v_{k-1}, \phi_k, \dot{\phi}_{k-1}) P(r_k \mid r_{k-1}, v_k, v_{k-1}) \end{aligned} \quad (5.20)$$

Each of the terms in Eq. 5.20 can be expanded as,

$$P(\phi_k \mid \phi_{k-1}, \dot{\phi}_{k-1}, v_{k-1}) = \mathbb{1}_\phi \quad (5.21)$$

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,

$$P(\dot{\phi}_k \mid \dot{\phi}_{k-1}, v_{k-1}) \propto \mathcal{N}(\dot{\phi}_{k-1}, \sigma_{\dot{\phi}_k}^2) \times \mathbb{1}_{\dot{\phi}} \quad (5.22)$$

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 μ and variance σ^2 . As before with the BP-model, the value of $\sigma_{\dot{\phi}_k}$ depends on the value of tempo to allow for larger tempo variations at higher tempi. In addition, $\sigma_{\dot{\phi}_k}$ also depends on the length of the section, providing higher flexibility of tempo in longer sections. We set $\sigma_{\dot{\phi}_k} = \sigma_n \cdot \dot{\phi}_{k-1} \cdot (M_{v_{k-1}}/M)$, where σ_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,

$$P(v_k | v_{k-1}, \phi_k, \phi_{k-1}) = \begin{cases} \mathbb{B}(v_{k-1}, v_k) & \text{if } \phi_k < \phi_{k-1} \\ \mathbb{1}_v & \text{else} \end{cases} \quad (5.23)$$

where, \mathbb{B} is the $V \times V$ time-homogeneous section 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,

$$P(r_k | r_{k-1}, \phi_k, \phi_{k-1}) = \begin{cases} \mathbb{A}(r_{k-1}, r_k) & \text{if } \phi_k < \phi_{k-1} \\ \mathbb{1}_r & \text{else} \end{cases} \quad (5.24)$$

where, \mathbb{A} is the $R \times R$ time-homogeneous pattern 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 also 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. The matrix \mathbb{A} closely follows \mathbb{B} and has non-zero probabilities only for allowed pattern transitions. For illustration, consider tracking **rū-pak tāl** (which has three **vibhāgs** $V = 3$) with the SP-model and two rhythmic patterns per section (hence, $R = 6$). The canonical forms of the section transition matrices \mathbb{B} and \mathbb{A} can then be illustrated as in Figure 5.6.

$$\mathbb{B} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \mathbb{A} = \begin{bmatrix} 0 & 0 & p_1 & 1-p_1 & 0 & 0 \\ 0 & 0 & p_2 & 1-p_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_3 & 1-p_3 \\ 0 & 0 & 0 & 0 & p_4 & 1-p_4 \\ p_5 & 1-p_5 & 0 & 0 & 0 & 0 \\ p_6 & 1-p_6 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 5.6: An illustration of the form of section (\mathbb{B}) and rhythmic pattern (\mathbb{A}) 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.

The observation model with the SP-model is similar to that of the BP-model, with an 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 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 64th note cells, all features within the cell are 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 (2×2), respectively. Hence, there is an observation GMM for each section, rhythmic pattern, and tied section position states.

It is straightforward to see that the BP-model is a special case of the SP-model, when the rhythmic patterns span the whole bar (cycle). By pooling in all the section length patterns from different *tālas* together, SP-model can also be applied for meter inference task. Further, 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 HMM_s . The AMPF inference in SP-model will be referred to in the rest of the chapter as AMPF_s and the algorithm is outlined in Algorithm 3.

5.3.3 Inference extensions

The proposed inference extensions aim for better approximate inference in the BP-model, either by making it faster, or by improving approximate inference.

End-of-bar pattern sampling

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 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 that is held fixed for the whole bar. 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. The strategy of sampling a rhythmic pattern at the beginning of the bar and fixing it for the whole bar is not intuitive and hence is suboptimal. An extension to AMPF_0 algorithm is proposed to address this limitation.

The extension, called end-of-bar pattern sampling extension to AMPF (called AMPF_e in short), defers the decision of sampling (and hence inferring) the pattern in the current bar to the end of the bar. In every bar being tracked, the algorithm accumulates weights for each of the patterns over the whole bar, and uses the final accumulated weight to choose the most likely pattern at the end of the bar.

Algorithm 3 Outline of the **AMPF_s** algorithm (**AMPF** for inference in SP-model)

```

1: for  $i = 1$  to  $N_p$  do
2:   Sample  $\mathbf{x}_0^{(i)} \sim P(\mathbf{x}_0)$  ▷  $\mathbf{x}_k = [\phi_k, \dot{\phi}_k, r_k, v_k]$ 
3:   Set  $w_0^{(i)} = 1/N_p$  ▷  $\boldsymbol{\alpha}_k = [\phi_k, \dot{\phi}_k, v_k]$ 
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  $\phi, r, v$ : Proposal and weights
7:     Sample  $\phi_k^{(i)} \sim P(\phi_k^{(i)} | \boldsymbol{\alpha}_{k-1}^{(i)})$ , Set  $c_k^{(i)} = c_{k-1}^{(i)}$ 
8:     if  $\phi_k^{(i)} < \phi_{k-1}^{(i)}$  then ▷ Section crossed
9:        $r_k^{(i)} \sim P(r_k^{(i)} | r_{k-1}^{(i)})$  ▷ Sample from  $\mathbb{A}$ 
10:       $v_k^{(i)} \sim P(v_k^{(i)} | v_{k-1}^{(i)})$  ▷ Sample from  $\mathbb{B}$ 
11:    else
12:       $r_k^{(i)} = r_{k-1}^{(i)}, v_k^{(i)} = v_{k-1}^{(i)}$ 
13:       $\tilde{w}_k^{(i)} = w_k^{(i)} \cdot P(\mathbf{y}_k | \phi_k^{(i)}, v_k^{(i)}, r_k^{(i)})$ 
14:    for  $i = 1$  to  $N_p$  do ▷ 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 ▷ Cluster, resample, reassign
17:      Cluster and resample  $\{\mathbf{x}_k^{(i)}, w_k^{(i)}, c_k^{(i)} | i = 1, 2, \dots, N_p\}$ 
      to obtain  $\{\hat{\mathbf{x}}_k^{(i)}, \hat{w}_k^{(i)} = 1/N_p, \hat{c}_k^{(i)}\}$ 
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  $\phi_k^{(i)} \sim P(\phi_k^{(i)} | \dot{\phi}_{k-1}^{(i)}, v_{k-1})$  ▷ Sample tempo
21: Compute  $\mathbf{x}_{1:K}^* = \mathbf{x}_{1:K}^{(i*)} | i^* = \arg \max_i w_K^{(i)}$  ▷ MAP sequence

```

The proposed enhancement can be formulated in a particle system using two different cluster groups. In addition to **AMPF** clustering based on metrical position and tempo (ignoring rhythmic patterns), an additional grouping is achieved with the rhythmic patterns. Within a single system of particles, we can then defer the inference of patterns till the end of a bar, as outlined in detail next.

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 metrical position and tempo $\beta = [\phi, \dot{\phi}]$, there are R different particles, one per rhythmic pattern. The weights can hence be organized in two dimensions (for the ease of understanding): one dimension denotes the subset of hidden variables β , and the other dimension stores the weights of all the R patterns for each β . 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,

$$w_k^{(i,j)} \propto w_{k-1}^{(i,j)} P(\mathbf{y}_k \mid \mathbf{x}_k^{(i,j)}) = w_{k-1}^{(i,j)} P(\mathbf{y}_k \mid \beta_k^{(i)}, r_k^{(i)} = j) \quad (5.28)$$

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, $\mathbf{w}_k^{(i,:)}$ is a vector of weights of all rhythmic patterns for each $\beta_k^{(i)}$, and $\Omega_k^{(i)}$ denotes the weight of $\beta_k^{(i)}$, computed as the marginal over all rhythmic patterns.

The **AMPF_e** algorithm is outlined in Algorithm 4. The algorithm can be interpreted to have two groups of particles in the particle system, one clustered based on β and the other is a group of R particles (for each β) representing rhythmic patterns. The weights $w^{(i,j)}$ accumulate the weight for every pattern j . For every $\beta^{(i)}$ that crosses

Algorithm 4 Outline of the **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:        $j^* = \arg \max_j (\mathbf{w}_k^{(i,:)})$ ; 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  $\{\beta_k^{(i)}, \Omega_k^{(i)}, c_k^{(i)} | i = 1, 2, \dots, N_p\}$ 
      to obtain  $\{\hat{\beta}_k^{(i)}, \hat{\Omega}_k^{(i)} = 1/N_p, \hat{c}_k^{(i)}\}$ 
21:    for  $i = 1$  to  $N_p$  do
22:      Set  $\beta_k^{(i)} = \hat{\beta}_k^{(i)}$ 
23:      for  $j = 1$  to  $R$  do  $\triangleright$  Weights redistributed
24:         $w_k^{(i,j)} = w_k^{(i,j)} \cdot \frac{\hat{\Omega}_k^{(i)}}{\Omega_k^{(i)}}$ 
25:      Sample  $\dot{\phi}_k^{(i)} \sim P(\dot{\phi}_k^{(i)} | \dot{\phi}_{k-1}^{(i)})$ 
26:       $\beta_{1:K}^* = \beta_{1:K}^{(i^*)} \mid i^* = \arg \max_i \Omega_K^{(i)}$   $\triangleright$  MAP sequence

```

In the algorithm, $\Omega_k^{(i)} = \sum_{j=1}^R w_k^{(i,j)}$

the end of a bar, the pattern j^* with maximum $w^{(i,j)}$ is assigned to the previous bar, thus deferring the decision of inference of rhythmic pattern to the end of the bar. Once the decision of previous bar is done, the weights in the vector $\mathbf{w}^{(i,:)}$ of the current frame are redistributed based on the transition probabilities of the patterns from the inferred pattern j^* of the previous bar. As with systematic resampling in the [AMPF](#) with BP-model, the resampling across $\beta^{(.)}$ is done at a fixed interval of T_s using the marginal summed up weights $\Omega^{(.)}$. Each of the two resampling/reweighting steps ensure that the new weights maintain a valid probability distribution over the particle system.

It is necessary that all the R rhythmic patterns associated with $\beta^{(i)}$ to be of the same length, and hence the [AMPF_e](#) algorithm can only be applied to 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 made faster by utilizing the time sparsity of onsets, 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.

Onset events are important cues to infer progression through metrical structures, and it is hypothesized that humans listen to these cues and use an inherent sense of time to track metrical structures accurately. We wish to analyze if automatic approaches can do a faster and accurate inference by just focusing on the onsets. Hop inference makes inference faster by skipping likelihood computation and sampling steps, and can speed up 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 (AMPF_p) : The peaks of the spectral flux feature sequence is an indicator of events such as onsets. Using a peak finding algorithm, the frames at which onset peaks occur are estimated. The progression of the particles by sampling from transition model and an update of their weights are done only at

these peak frames, skipping the non-peak frames. The transition model update equations 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 (AMPF_g) : Despite the advantage of a faster inference, peak hop inference can lead to sharp discontinuities in ϕ and tempo values due to large jumps in their values since they are sampled with large gaps of a significant number of frames. An improvement to peak hop while maintaining continuity is the onset gated weight update, where $\dot{\phi}$ and ϕ are sampled and updated every frame to maintain continuity, while weights of the particles (using the likelihoods from the observation model) are updated only at frames where there is a peak in the spectral flux feature, indicating a rhythmic event. The basic premise is to maintain the continuity in tracking the ϕ 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 computational advantage however is lower than that for peak hop inference.

The different meter tracking models were presented in this section are listed and summarized in Table 5.3. The table also shows the acronyms for the algorithms we use in the dissertation, along with the meter analysis tasks to which they can be applied. We now present the experiments 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 for different tasks and datasets with the models and algorithms 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

Acronym	Model	Inference algorithm	Meter Analysis	
			Inference	Tracking
$\S \text{HMM}_0$	BP-model [†]	Viterbi algorithm	✓	✓
$\S \text{AMPF}_0$	BP-model	AMPF	✓	✓
$^* \text{HMM}_m$	MO-model [†]	Viterbi algorithm	✗	✓
$^* \text{AMPF}_m$	MO-model	AMPF	✗	✓
$^* \text{HMM}_s$	SP-model [†]	Viterbi algorithm	✓	✓
$^* \text{AMPF}_s$	SP-model	AMPF	✓	✓
$^* \text{AMPF}_e$	BP-model	AMPF with end-of-bar pattern sampling	✗	✓
$^* \text{AMPF}_p$	BP-model	Peak hop inference in AMPF	✓	✓
$^* \text{AMPF}_g$	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 thesis. 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.

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 compare 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 Indian art music datasets (both Carnatic and Hindustani), 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 Experimental setup

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. This implicitly assumes that the music culture is known *a priori* in all experiments. Unless otherwise specified, the following global settings for the experiments is used.

The results on the Carnatic music dataset (**CMR**) and Hindustani music data subsets **HMR_s** and **HMR_I** are focused on. From the experiments, we see that the datasets **CMR** and **CMR_f** 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 (**vilābit**, 10-60 MPM), medium (**madhya**, 60-150 MPM), and fast (**dṛt**, > 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_I** (**vilābit** pieces) and **HMR_s** (**madhya** and **dṛt** pieces) datasets separately to assess performance individually on pieces with long and short cycle duration. The results are presented for each dataset as an average over the pieces in all the **tālas** (or meters), while specific comments on the performance on each **tāla** is discussed when needed. Performance on ballroom dataset is reported for meter inference and tracking tasks for comparison.

All results are reported as the mean performance over three runs in a two fold cross validation experiment. The train and the test data folds have equal number of pieces (with a maximum difference of one piece when there are odd number of pieces in a dataset). In meter inference experiments, the total set of **tālas** being tracked is known, along with their structure. The training data contains pieces pooled from all the **tālas** contained in the whole 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 **HMR_s** and **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, as discussed earlier, additional information is provided to the tracking algorithms on tempo and the first instance of downbeat in tempo-informed meter tracking and tempo-sama-informed meter tracking, respectively.

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_t (AML_{t,b}) and information gain (\mathfrak{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 dissertation, we used the evaluation toolkit developed by Matthew Davies³. To compute the f-measure in **CMR**, **HMR_s**, 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.

The computation of f-measure with **HMR_l** dataset is an exception, where a bigger margin window is allowed. Since cycles are of long duration in **HMR_l** dataset and current evaluation approaches were not designed with such long cycles in mind, an error tolerance window of 70 ms is very tight. To account for the length of the cycle in the error margin, a 6.25% median inter annotation interval is used as the tolerance window, as used in many other beat tracking evaluations (e.g. by Hockman et al. (2012)). This choice of a larger allowance window also corroborates well with the observation that in **vilābit** pieces of the **HMR_l** 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 **vilābit** pieces is also beyond the duration of what is called the perceptual present (Clarke, 1999). However, it is to be noted that this approach is a compromise and better evaluation measures that can handle these complexities are to be developed. The prob-

³We use the code available at <http://code.soundsoftware.ac.uk/projects/beat-evaluation/>

lem of evaluating the accuracy of meter tracking in *vilābit* pieces of Hindustani music (and other musics with long duration cycles) is itself a research problem that needs to be studied systematically, including musicians and listeners into the study.

For meter inference and tracking, we additionally report the results of median tempo estimation as computed from the output 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 that 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 (or meter) and hence the accuracy of *tāla* recognition is also reported for the task.

Most experiments are conducted for rhythmic patterns $R = 1$ and $R = 2$ (per rhythm class), but the results are presented only for $R = 1$ pattern per *tāla*. Experiments with $R = 2$ do not show any significant improvement/change. Hence they are not presented with all models but when necessary, performance with $R = 2$ is indicated and discussed. It is to be noted that with $R = 1$, model extension MO-model (with **AMPF_m**) and inference extension **AMPF_e** are equivalent to the baseline **AMPF₀**.

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 (equivalent to [35, 130] beats per minute in *ādi tāla*), that for Hindustani music is set as [10, 370] *mātrās* per minute, and for ballroom dataset as [60, 230] beats per minute. For meter tracking, the tempo range for each *tāla* is independently learnt. Further, with SP-model, the tempo ranges learned for a particular *tāla* are applied to track all the section length patterns of the *tāla*, assuming that the tempo ranges are properties of a *tāla* and not of its sections.

We use the number of bar positions, $M_r = 1600$ for the longest rhythmic pattern we encounter in the dataset and scale all other pat-

tern 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** (3 beats and 12 **akṣaras**) 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 are 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 64th note cells within which the observation probability is assumed to be constant.

For the **HMM** based Viterbi algorithm 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 set as $N_p = 1500 \times R$. We set the user parameter that controls tempo variance in Eq. 5.6 and Eq. 5.22 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 model and inference exten-

sions. Informed meter tracking is then discussed. A final summary of results over all the Indian art music datasets is also presented for a comparison of the performance of approaches.

5.4.2 Meter inference

The results of meter inference provide a baseline for meter analysis algorithms when the underlying metrical structure is unknown. It is the hardest and most uninformed task in meter analysis: estimating the *tāla*, the tempo, the beats and the *sama*. The results are presented for inference on the BP-model on **CMR**, **HMR_s**, **HMR_I** and Ballroom datasets for both **HMM₀** and **AMPF₀** algorithms in Table 5.4. The model training uses pooled data from all the rhythm classes within a particular dataset. The results are presented only for $R = 1$ per rhythm class, without much improvement seen 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_I** dataset is poor, showing the challenges in tracking long metrical cycle durations. The performance with **HMM₀** is marginal poorer than **AMPF₀** for Indian music datasets. Since metrical cycles in Indian music are longer in duration, it is necessary to have a finer discretization grid. The poorer performance is largely attributed to the coarse grained 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 **HMR_s** dataset with an accuracy between 60-70%. For long cycle duration pieces in **HMR_I** 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 difficult. Beat AML_t measure $AML_{t,b}$ is comparable to beat f-measure. It was reported by

	Algo.	f_b	$AML_{t,b}$	\mathfrak{I}_b Bits	f_s	Tempo		Tāla %
						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_1	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.

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 HMR_s and HMR_1 datasets, further emphasizing the difficulties of tracking long duration cycles. The tempo estimation at CML with HMR_1 dataset is as low at 12% with 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. $AMPF_0$ 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 HMM_0 . $\bar{\text{A}}\text{di t}\bar{\text{a}}\text{l}\bar{\text{a}}$ is the most popular $\text{t}\bar{\text{a}}\text{l}\bar{\text{a}}$ in Carnatic music and there is a huge variety of rhythmic patterns that are played in the $\text{t}\bar{\text{a}}\text{l}\bar{\text{a}}$. The large difference between beat tracking and sama tracking performance shows that though beats were estimated at the correct metrical level, sama estimation is difficult from the rhythmic patterns used here. This additionally means that capturing the wide variety of patterns of $\bar{\text{a}}\text{di t}\bar{\text{a}}\text{l}\bar{\text{a}}$ within a single rhythmic pattern used here is suboptimal and hence it is harder for the inference algorithms to get a cue of the metrical position from this pattern. In Hindustani music HMR_s dataset, the performance is best for $\text{d}\ddot{\text{r}}\text{t }\bar{\text{e}}\text{k}\bar{\text{t}}\bar{\text{a}}$ pieces that tend have high tempo and short duration cycles. Both these observations further emphasize that short duration cycles are better tracked by the inference algorithm than longer duration cycles.

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 $\text{t}\bar{\text{a}}\text{l}\bar{\text{a}}$ is known, and hence meter tracking is done for each $\text{t}\bar{\text{a}}\text{l}\bar{\text{a}}$ in the datasets separately. The training data also includes pieces from the specific $\text{t}\bar{\text{a}}\text{l}\bar{\text{a}}$ 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) with minor differences in formulation and experimental parameters.

Before presenting the results for model and inference extensions, we tabulate the performance of meter tracking with the bar pointer model on the Indian music datasets and Ballroom dataset with both HMM_0 and AMPF_0 algorithms in Table 5.5. This provides another baseline performance to compare with meter inference and all the extensions discussed in the thesis.

At a broad level, Table 5.5 shows an improvement in performance with meter tracking compared to meter inference (Table 5.4). In addition, we see a lower difference between beat and down-beat tracking f-measure values, indicating a larger improvement in downbeat estimation when the underlying metrical structure is known i.e. knowing the $\text{t}\bar{\text{a}}\text{l}\bar{\text{a}}$ improves the sama tracking performance. Similar to meter inference, the performance on short du-

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
	$AMPF_0$	0.827	0.840	1.97	0.671	0.955
HMR_s	HMM_0	0.835	0.796	1.39	0.733	0.663
	$AMPF_0$	0.884	0.858	1.64	0.772	0.844
HMR_l	HMM_0	0.353	0.305	0.86	0.429	0.294
	$AMPF_0$	0.374	0.513	1.40	0.396	0.390
Blrm.	HMM_0	0.929	0.921	2.78	0.821	0.987
	$AMPF_0$	0.909	0.895	2.56	0.735	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.

ration cycle datasets CMR, HMR_s , and Ballroom datasets is better than that on HMR_l dataset. As with meter inference, the poorer performance of HMM_0 compared to $AMPF_0$ is largely attributed to the coarse grained discretization of the state space.

The median tempo estimation performance with meter tracking is better than meter inference since a more narrow and accurate range of tempo is trained due to the presence only one *tāla* in the training dataset. The accuracy of tempo estimation is high for Carnatic music, and 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. Tempo estimation in Hindustani music datasets is however lower, with poor tempo estimation with HMR_l dataset. There is a significant scope for improvement in both CML and AML accuracy in Hindustani music. 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*, e.g. $f_s = 0.74$ for HMM_0 in *khaṇḍa chāpu tāla*, showing that the variety of rhythmic patterns

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 _s	0.886	0.864	1.66	0.783	0.837	0.942
HMR _I	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 (MO-model with 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 AMPF_0 algorithm.

in **ādi tāla** makes it harder to track. Compared to meter inference, **ādi tāla** shows an improvement in tracking, indicating that knowing the **tāla** and the underlying metrical structure helps to track longer cycles better.

With such a baseline of meter tracking using the bar pointer model, and given that AMPF_0 shows 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 also show that approximate inference methods such as particle filters can be effectively applied to meter analysis tasks. 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 among all the algorithms.

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 _s	0.924	0.890	1.88	0.850	0.855	0.971
HMR _I	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 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 ϕ to the whole *tāla* cycle, while the proposed SP-model applies ϕ 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 AMPF_s is shown in Table 5.7. It shows a significant improvement in **sama** tracking f-measure compared to AMPF_0 with bar pointer model. The beat tracking performance also improves, but to lesser extent than **sama** tracking, showing that using section length patterns help tracking the **sama** more accurately. There is no further improvement in tempo estimation in SP-model compared to BP-model. The utility of the SP-model is hence primarily in improving downbeat tracking performance.

The improvement with the long cycle duration pieces in HMR_I dataset is further encouraging to use shorter section length patterns to track longer cycles. A significant improvement is also observed in Carnatic music with *ādi tāla* with the sama tracking f-measure of $f_s = 0.46$ from $f_s = 0.39$ for AMPF_0 . Both these observations show that using shorter section length patterns have a potential ap-

	Dataset	f_b	$AML_{t,b}$	\mathfrak{I}_b Bits	f_s	Tempo	
						CML	AML
CMR	AMPF_e	0.826	0.842	1.97	0.668	0.958	0.997
	AMPF_p	0.519	0.561	0.67	0.213	0.927	0.969
	AMPF_g	0.756	0.756	1.51	0.580	0.938	0.98
HMR_s	AMPF_e	0.882	0.858	1.64	0.777	0.833	0.935
	AMPF_p	0.655	0.572	0.59	0.273	0.768	0.822
	AMPF_g	0.821	0.653	1.25	0.653	0.743	0.895
Blrm.	AMPF_e	0.908	0.895	2.56	0.734	0.98	0.98
	AMPF_p	0.631	0.694	1.49	0.322	0.922	0.923
	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.

plication in tracking long duration metrical cycles.

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 (AMPF_e), peak hop inference (AMPF_p), and onset gated weight update (AMPF_g). The goal of these experiments is to compare the performance of the inference extensions with AMPF_0 algorithm. The long cycle duration HMR_l dataset is excluded from evaluation of the inference extensions. Inference extensions are only evaluated on CMR and HMR_s datasets (Performance on Ballroom dataset also shown for reference) and compared with AMPF_0 . The results are shown in Table 5.8. The table shows results only for $R = 1$, which means that AMPF_e is equivalent to AMPF_0 . From the table, we see that AMPF_e has equivalent performance to AMPF_0 . It is important to note that AMPF_e does not show any significant improvement from $R = 1$ to $R = 2$.

For both AMPF_p and AMPF_g algorithms, a peak picking algorithm is used to select 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 highest valued peak is added into the peak sequence.

Though peak hop inference (AMPF_p) provides a significant boost in inference time (up to $10\times$ faster), we see from Table 5.8 that its performance is significantly 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 at the peaks of the spectral flux feature 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 (AMPF_g) overcomes this limitation by progressing the tempo and position variables of the bar pointer model every frame and hence maintains continuity. Though it also speeds up inference (to a lesser extent than peak hop), its performance is poorer since it fails to model the rhythmic events that happen between the two peaks as 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. Though these two ideas are promising to improve inference, they need further exploration to improve their performance. Further analysis and comparison of these extensions with 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 to improve performance. The results for informed meter tracking with BP-model (with AMPF_0) and SP-model (with 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 addi-

Dataset	Algo.	f_b	$AML_{t,b}$	\mathfrak{I}_b	f_s
CMR	$AMPF_0$	0.899	0.952	2.35	0.792
	$AMPF_s$	0.898	0.950	2.38	0.814
HMR_s	$AMPF_0$	0.939	0.941	1.99	0.882
	$AMPF_s$	0.943	0.943	2.00	0.918
HMR_1	$AMPF_0$	0.425	0.959	2.76	0.786
	$AMPF_s$	0.439	0.979	2.83	0.848

Table 5.9: Results of tempo informed meter tracking with $AMPF_0$ and $AMPF_s$ on Indian music datasets. The table shows beat and *sama* tracking performance with different measures.

tional 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 ϕ 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 ϕ within a tight range as before, and further use the first instance of *sama* to initialize the ϕ 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 with the bar and section pointer model (BP-model and SP-model, with $AMPF_0$ and $AMPF_s$, 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 $AMPF_s$ marginally

Dataset	Algo.	f_b	$AML_{t,b}$	\mathfrak{I}_b	f_s
CMR	$AMPF_0$	0.880	0.943	2.37	0.834
	$AMPF_s$	0.917	0.946	2.40	0.901
HMR_s	$AMPF_0$	0.959	0.920	2.02	0.911
	$AMPF_s$	0.958	0.915	2.01	0.933
HMR_1	$AMPF_0$	0.530	0.978	2.84	0.99
	$AMPF_s$	0.542	0.98	2.83	0.99

Table 5.10: Results of tempo-sama-informed meter tracking with $AMPF_0$ and $AMPF_s$ on Indian music datasets. The table shows beat and sama tracking performance with different measures.

performs better than $AMPF_0$ 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 a high f-measure for both sama and beat tracking is achieved in informed meter tracking. We also see that the beat tracking performance in the two informed tracking cases is similar, while sama tracking performance improves in tempo-sama-informed tracking. A high beat and downbeat tracking performance is achieved in CMR and HMR_s datasets. 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 limited 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 models and algorithms. 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 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 algorithms. Statistical significance tests are done for the meter inference, meter tracking (including model extensions MO-model and SP-model), and informed tracking methods by pooling the re-

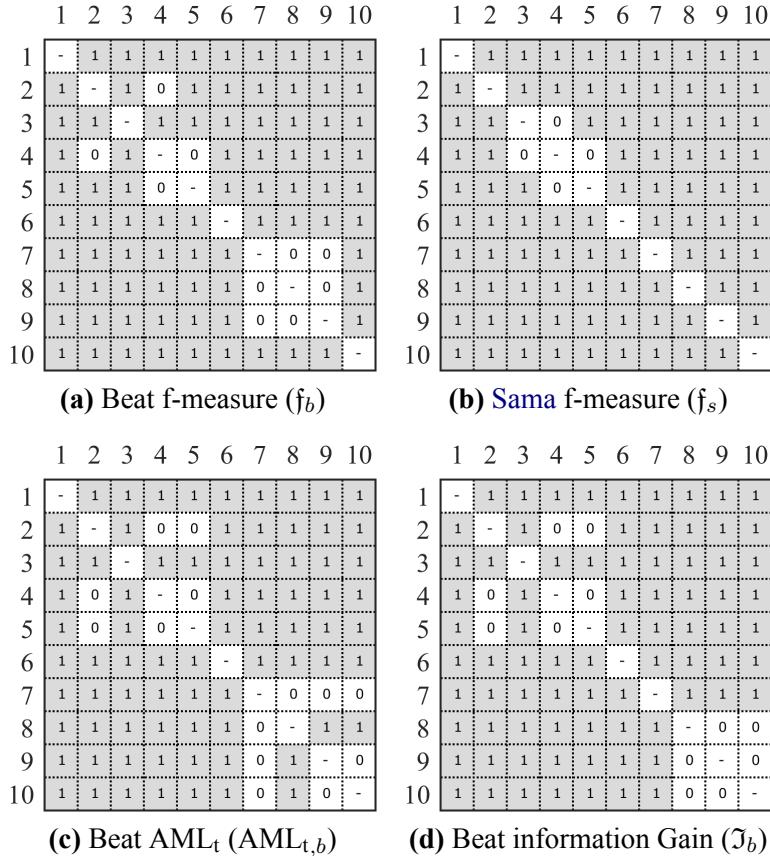


Figure 5.7: Results of statistical significance testing of meter analysis results on Indian art music datasets. The figure shows the results for the four different performance measures: Beat f-measure (f_b), **Sama** f-measure (f_s), Beat AML_t (AML_{t,b}) and Beat information gain (J_b) in panels (a), (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 column-3 of Table 5.11.

sults over all Indian music datasets (**CMR**, **HMR_s**, **HMR_I** datasets - 269 pieces in total). Statistical significance tests are done for BP-model inference extensions (**AMPF_e**, **AMPF_p**, **AMPF_g**) by pooling the results over **CMR** and **HMR_s** (210 pieces in total) to compare with **AMPF₀**.

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 AMPF_m is equivalent to 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 leads 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 AMPF_0 and 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 AMPF_0 algorithm in inference and tracking have statistically insignificant differences. In addition, AMPF_0 and AMPF_s in tempo-informed tracking have insignificant differences with AMPF_0 in tempo-sama-informed tracking. Sama f-measure (f_s) shown in panel (b) indicates statistically insignificant differences between HMM_0 and 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 significance results of beat $\text{AML}_{t,b}$ measure in panel (c) is comparable to that for beat f-measure. Informed tracking methods have several statistically insignificant differences among themselves with the beat $\text{AML}_{t,b}$ measure since the correct metrical level is already provided to the informed tracking algorithm, and hence leads to similar performance. An acceptable beat information gain ($\mathcal{I}_b > 1.5$ bits) is obtained in most cases, with several statistically insignificant differences 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 performance also improves, but to a lesser extent.

For an analysis and comparsion of inference extensions, we

Algo.	ID	f_b	$AML_{t,b}$	\mathfrak{I}_b Bits	f_s	Tempo	
						CML	AML
AMPF_0	1	0.852	0.848	1.83	0.715	0.90	0.97
AMPF_e	2	0.850	0.849	1.82	0.716	0.90	0.97
AMPF_p	3	0.579	0.566	0.63	0.240	0.85	0.93
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 **CML** and **HMR_s** 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.

Algo.	f_b	$AML_{t,b}$	\mathfrak{I}_b Bits	f_s	Tempo	
					CML	AML
AMPF_0	0.735	0.751	1.68	0.641	0.77	0.88
AMPF_m	0.749	0.773	1.75	0.660	0.78	0.89

Table 5.13: Comparing the meter tracking performance of AMPF_0 and AMPF_m algorithms on Indian art music datasets for $R = 2$ patterns. Numbers in bold for the beat and **sama** tracking measures indicate a statistically significant improvement.

pool the results on Indian music datasets **CML** and **HMR_s** and present it in Table 5.12. We compare the extensions AMPF_e , AMPF_p and AMPF_g with the baseline meter tracker AMPF_0 . In the table, since $R = 1$, note that AMPF_e is equivalent to AMPF_0 . Statistical tests indicate that for all measures, AMPF_0 and AMPF_e are equivalent and show no statistically significant difference in performance. In addition, for all measures, AMPF_p and AMPF_g both give significantly lower performance compared to AMPF_0 . The hop inference extensions need further improvement and do not match up to the performance of doing a full inference at every frame.

With that summary of results, we now focus on some more analysis with $R = 2$ comparing meter tracking performance of AMPF_0 with AMPF_m and AMPF_e . Table 5.13 shows a summary of results over all the Indian music datasets for meter tracking with AMPF_0 and AMPF_m and $R = 2$. An analysis showed that there is no statisti-

cally significant difference in results between $R = 1$ and $R = 2$ for either AMPF_0 or AMPF_m (for all measures). However, for $R = 2$, the beat tracking measures show an improvement for AMPF_m over AMPF_0 . For the case of AMPF_e compared with AMPF_0 however, there was no statistically significant improvement with more patterns, showing the need for further exploration of the end-of-bar sampling AMPF algorithm to improve its performance.

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. After 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 discussed 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 Bayesian models explicitly considered musically relevant information for meter analysis, leading to culture-aware 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 (*usul*) and to Arab-andalusian music (*mizān*). 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_s, and HMR_l datasets that consist of two 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. A future evaluation on a larger dataset with full length pieces, such as the rhythm annotated pieces in HMD_o and CMD_o collections will further boost such a claim.

One main limitation of the algorithms presented in the chapter was the assumption of a single *tāla* for the whole audio recording presented to the algorithm. While this is a fair and realistic assumption for Carnatic music, which is distributed as segmented recordings containing a single piece, Hindustani music recordings can have two or more pieces in different *tāl* and *lay*. A rhythm based segmentation might be necessary there before applying the meter analysis algorithms. Such segmentation could be performed using, e.g. Bayesian change point detection (Barber, Cemgil, & Chiappa, 2011), a problem that needs further exploration.

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 multiple 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.

Vilāmbit (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 **HMR₁** 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.

A comparison across different **tālas** showed that tracking longer length **tālas** that have quite a diversity of patterns played in them are more difficult to track, e.g. **ādi tāla** in Carnatic music. **Vilāmbit ēktāl** in Hindustani music is also difficult to track owing to its long duration cycles and equal length **vibhāgs**. The section length rhythmic patterns in **ēktāl** are equal in length and similar, which is confusing to a tracker that uses only spectral energy based features. In summary, longer **tālas** that have wider scope for improvisation are difficult to track, with longer duration cycles adding further to tracking complexity.

Apart from the percussion patterns played on mridangam and **tabla** that are indicative of the position in metrical cycle, melodic patterns also in many cases indicate the position in cycle. This is further true in compositional forms of both Hindustani and Carnatic music, 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 melody and lyrics based features into the observation model is hence hypothe-

sized to additionally help to improve meter analysis performance.

The presented Bayesian models can be further improved to incorporate other structures and priors that can be utilized for improving meter tracking - such as tighter bounds on tempo, tighter restriction on continuity, and allowance for errors such as 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. They can be better utilized with more rhythmic patterns modeling a *tāla*, which needs more diverse features. The use of better features and faster inference with better models is the focus in future research on the topic. Recent approaches that use deep learning to build observation models have also seen some success (Böck & Schedl, 2011).

The meter analysis algorithms discussed in the chapter were developed within the context of CompMusic project and hence are aligned with its goals to lead towards defining relevant rhythm similarity measures. Meter analysis is the first step towards that goal. Automatic meter analysis provides valuable content based metadata for a piece of music with several useful applications: 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)

Percussion plays an important role in Indian art music with a significant freedom to improvise, leading a wide variety of percussion patterns that help to create multiple layers of rhythm. An analysis of these percussion patterns hence is an important step towards developing rhythm similarity measures. We wish to discover percussion patterns from audio recordings in a data-driven way, while using a musically meaningful representation for percussion patterns.

We address the task of percussion pattern discovery in this chapter, taking the approach of transcription followed by a search for patterns. The work presented in the chapter is basic and exploratory, and only for demonstrating the utility of a syllabic percussion system in percussion pattern transcription and discovery. Most experi-

ments presented contain preliminary results, needing further work. The goals of the chapter are:

1. To discuss two broad approaches to percussion transcription. To focus on 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 jingju percussion patterns, as a simpler example test case of percussion pattern transcription and pattern classification. To extend the approach to Indian art music, and evaluate it on mridangam and tabla solo datasets.
4. To identify the advantages and shortcomings of such an approach, with possible future research directions to pursue.

We first start by describing 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 in two different broad ways. The two approaches differ mainly in the way they define 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 can then be described and represented using 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, each retaining their own musical identity. 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 2.3.10, event-based transcription algorithms (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 & Plumbley, 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 jingju percussion instrument ensembles, aiming at providing a baseline for further research in jingju 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 jingju 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 the *tabla* or the *mridangam*. In either case, the overall timbre represented by the syllable is sufficient to define a percussion pattern. In such a case, a percussion pattern can be represented solely based on these syllables and transcribed as such.

In this chapter, we explore only the second approach, using syllables to define overall timbres of percussion strokes. We then 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 a 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, Indian percussion has a great scope for improvisation within the framework of the *tāla*. This leads a large number of percussion patterns that can be played on the *mridangam* and *tabla*, which makes it further difficult to define relevant percussion patterns without ambiguity. We address this issue by making the discovery unsupervised and data-driven, by using transcribed ground truth music scores to define relevant patterns.

Given the complexity of the task in Indian art music due to ill-defined large number of patterns, we consider the case of percussion patterns in *jingju* as an initial test case for our hypothesis and methods. Percussion patterns in *jingju* are simpler in both these aspects when compared to Indian art music: *jingju* percussion patterns are well defined and limited in number. Once we demonstrate and validate our hypothesis with *jingju*, we extend the methodology to Indian art music, with a data-driven but extendable definition of

a percussion pattern. Since there are a limited number of jingju percussion patterns, the problem of pattern discovery in jingju is simplified into a pattern classification task, as is described further in the following section.

6.2 The case of Beijing opera (Jingju)

To recall from Section 2.2.5, percussion ensemble in jingju consists of five instruments played by four musicians, and can be grouped based on timbre into four instrument groups - **bangu** (clapper-drum), **xiaolu** (small gong), **daluo** (big gong) and **naobo** (cymbals). The waveform and spectrogram of an audio example with all four of these instrument classes is shown in Figure 6.1¹. We can see the amplitude dynamics and spectral shapes for each instrument. **Daluo** has a falling pitch, while **xiaolu** has a rising pitch profile. **Naobo** has a broadband spectrum with significant energy in higher frequencies, as is characteristic of cymbals. Onsets generated by **bangu** are sharp, have a much lower amplitude and shorter transient time, and happen in higher density than those generated by the other instruments. Hence, the **bangu** onsets are easily masked by the cymbals and gongs. We can also see how the **bangu** stroke is masked by an adjoining **xiaolu** 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 with an onomatopoeic oral mnemonic syllable. Since there are 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 jingju 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.

¹The audio example is available here: <http://www.freesound.org/people/ajaysm/sounds/205971/>

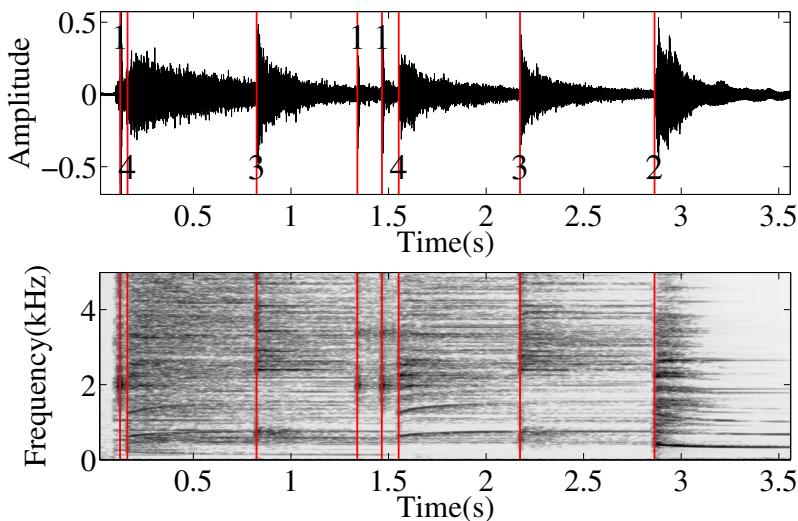


Figure 6.1: The waveform and spectrogram of an audio example containing all four instrument groups of jingju. The top panel shows the waveform and the bottom panel is the spectrogram, the x-axis for both panels is time (in seconds). The vertical lines (in red) mark the onsets of the instruments. The onsets are labeled to indicate the specific instruments: **bangu-1**, **daluo-2**, **naobo-3**, **xiaoluo-4**.

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

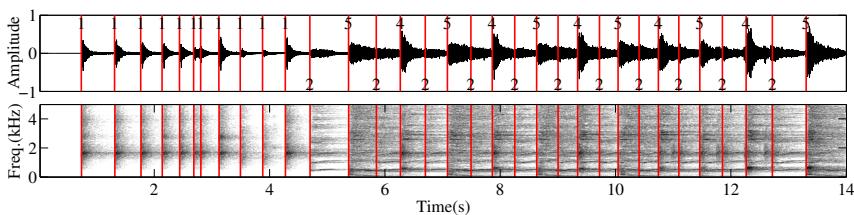


Figure 6.2: The waveform and spectrogram of an audio example of the pattern shanchui. The top panel shows the waveform and the bottom panel is the spectrogram. The vertical lines (in red) mark the onsets of the syllables. The onsets are labeled to indicate the specific syllable group: DA-1, TAI-2, QI-3, 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.

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 lack 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 *xiaolu*, but in the rest of the pattern is played by *xiaolu* and the *bangu* together. In fact, generally speaking, the strokes of the *bangu* 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 *bangu*), 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 *daluo*, or just the *daluo+naobo* combination.

In the cases where the percussion pattern is to accompany the movements of actors on stage, certain syllable subsequences 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 shanchui, along with time aligned markers to indicate the syllable onsets. The spectrogram also shows the timbral characteristics of the percussion 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 jingju 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 jingju - daoban tou, man changchui, duotou, xiaoluo duotou, and shanchui (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. daoban tou pattern is followed by an aria in *banshi* daoban.

We now present a formulation for transcription and recognition of syllable based audio percussion patterns, and evaluate it

on the [Jingju Percussion Pattern dataset \(JPP\)](#) (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. Their approach is different from ours in the sense that the onomatopoeic representations they used 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 that explored automatic 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 syllables $\mathcal{A} = \{A_1, A_2, \dots, A_{N_s}\}$, where N_s is the total number of syllables in the set. Hence, a percussion pattern is represented as $\mathbf{A}_i = [a_1, a_2, \dots, a_{L_i}]$ where $a_j \in \mathcal{A}$ and L_i is the length of \mathbf{A}_i . Given a test audio signal $f[n]$ containing a percussion pattern, 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](#) $\{\Lambda_j\}$, $1 \leq j \leq N_s (= 5)$, for each syllable A_j using features extracted from the training audio

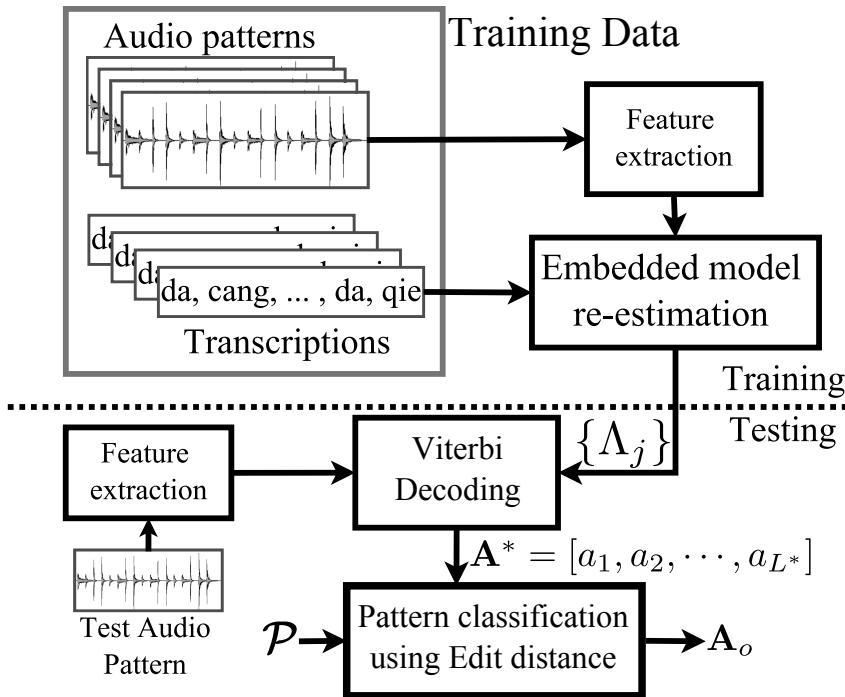


Figure 6.3: The block diagram jingju percussion pattern classification approach

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 zeroth 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 zeroth MFCC coefficient) in classification performance. Hence we have two sets of features, **MFCC_0_D_A**, the 39 dimensional feature including the zeroth, delta (velocity) and double-delta (acceleration) coefficients, and **MFCC_D_A**, the 36 dimensional vector without the zeroth 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 **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 **Hidden Markov model 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 pattern 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 $P(a_{k+1} = A_j \mid a_k = A_i) = 1/N_s$, $1 \leq i, j \leq N_s$, with k being the sequence index. This also forms a simple uninformed language model for forming the percussion patterns using syllables. Given the feature sequence extracted from test audio pattern, we use the **HMMs** $\{\Lambda_j\}$ 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 patterns 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 (Navarro, 2001).

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.

$$\mathbf{A}_o = \arg \min_{1 \leq j \leq N_a} d(\mathbf{A}^*, \mathbf{A}_j) \quad (6.1)$$

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 despite a low syllable transcription accuracy. We see that the feature **MFCC_0_D_A** leads to 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

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.

Pattern class	Total	ID	1	2	3	4	5
daoban tou	62	1	100				
man changchui	33	2		93.9			6.1
duotou	19	3	10.5		68.4		21.1
xiaoluo duotou	11	4			18.2	81.8	
shanchui	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).

provides several insights. 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 (zeroth) 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 not 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. However, 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 compute a confusion matrix (Table 6.2) with one of the well performing configurations: [MFCC_0_D_A](#) with d_1 distance. We see that duotou (ID = 3) has a low recall, and gets confused with shanchui (ID = 5) often. A close examination of the scores showed that a part of the pattern duotou (Figure 2.5c) is contained within shanchui (Figure 2.5e), which explains the source of confusion. From the scores, we also see that xiaolu duotou (ID = 4) and duotou (ID = 3) have similar structure, with the [daluo](#) and [naobo](#) strokes replaced by [xiaoluo](#), explaining the confusion between these two patterns. Such confusions can be handled with better language models for modeling percussion patterns, a topic of research that needs 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 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 jingju, with a larger variety of syllables and ill defined indefinite 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 this section, 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 pattern discovery in Indian art music percussion solo recordings, we follow a data driven transcription + search approach. The approach mainly has three sub-tasks:

Pattern library generation: Create 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 sub-tasks 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 task hence is 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 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 shorter patterns that are played very often, which are grouped in different combinations to create longer patterns. In a data-driven way, a library of such patterns can 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. In [tabla](#) solos, despite the differences across [gharānās](#), 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 the creation of a library of standard phrases or patterns across compositions of different [gharānās](#) present in the [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 musically relevant criteria for automatic discovery of patterns can also be used.

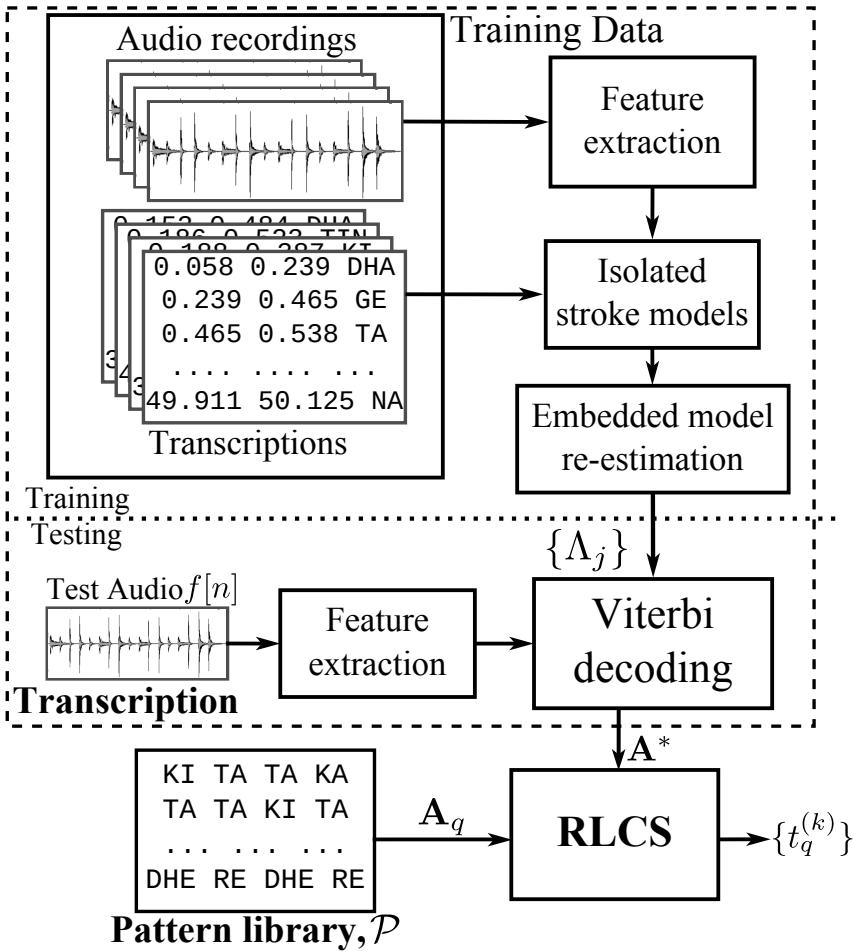


Figure 6.4: A block diagram of percussion pattern discovery approach in Indian art music. The figure considers the example of **tabla** solos for illustration.

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 $L = 4, 6, 8, 16$. These lengths were chosen based on the structure of **tālas** in the score collections (**ādi** and **rūpaka tāla** in

ID	Pattern	<i>L</i>	Count
A ₁	DHE, RE, DHE, RE, KI, TA, TA, KI, NA, TA, TA, KI, TA, TA, KI, NA	16	47
A ₂	TA, TA, KI, TA, TA, KI, TA, TA, KI, TA, TA, KI, TA, TA, KI, TA	16	10
A ₃	TA, KI, TA, TA, KI, TA, TA, KI	8	61
A ₄	TA, TA, KI, TA, TA, KI	6	214
A ₅	TA, TA, KI, TA	4	379
A ₆	KI, TA, TA, KI	4	450
A ₇	TA, TA, KI, NA	4	167
A ₈	DHA, GE, TA, TA	4	97

Table 6.3: Query [tabla](#) percussion patterns, their ID, length (*L*) and the number of instances in the [MTS](#) dataset (Total instances: 1425).

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 a bass component (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.

6.3.2 Automatic transcription

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

ID	Pattern	L	Count
A₁	DH3, TA, DH3, TA, TH, DH3, TH, TA	8	70
A₂	TA, DH3, TA, TH, DH3, TH, TA, TM	8	69
A₃	DH3, TA, DH3, TA, TH, DH3	6	89
A₄	DH3, TA, TH, DH3, TH, TA	6	70
A₅	TA, TH, DH3, TH, TA, TM	6	69
A₆	DH3, TA, TH, DH3	4	291
A₇	DH3, TA, DH3, TA	4	114
A₈	TH, DH3, TA, TH	4	102
A₉	TA, TH, DH3, TH	4	102

Table 6.4: Query mridangam percussion patterns, their ID, length (L) and the number of instances in the [UMS](#) dataset (Total instances: 976).

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

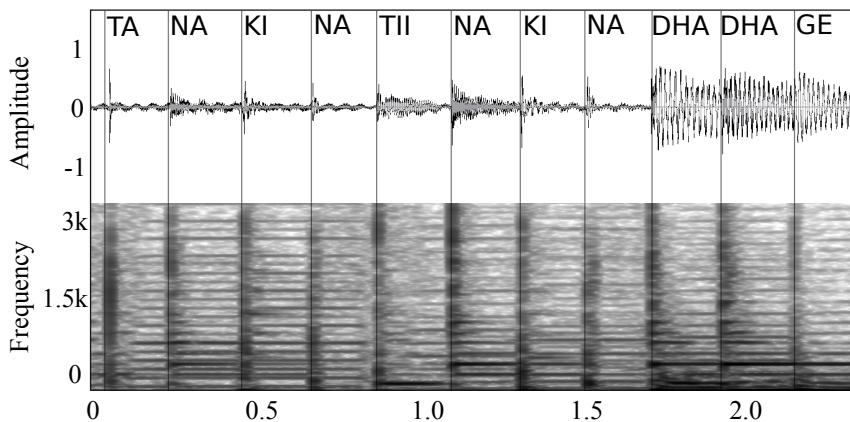


Figure 6.5: The waveform and spectrogram of an audio example of a tabla percussion pattern shown with the onsets and the mapped syllable names from Table 4.15.

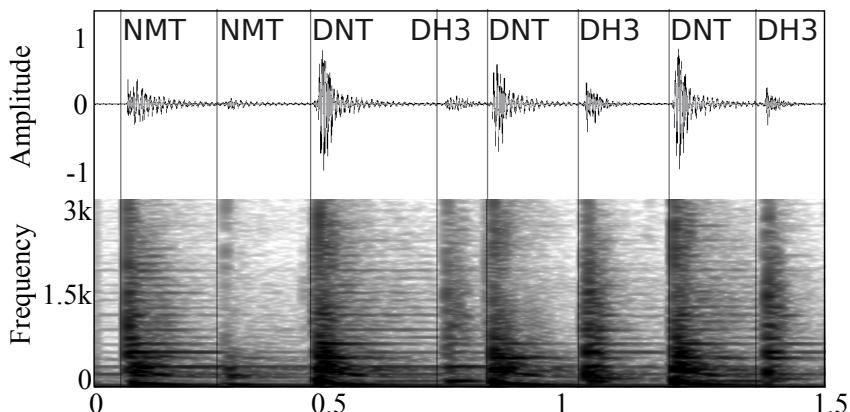


Figure 6.6: The waveform and spectrogram of an audio example of a mridangam percussion pattern shown with the onsets and the mapped syllable names from Table 4.17.

the use of energy (as measured by the zeroth MFCC coefficient) in transcription performance. Hence we have two sets of features, **MFCC_0_D_A**, the 39 dimensional feature including the zeroth, delta and double-delta coefficients, and **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 **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 transcriptions, 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/solkaṭus** tend to follow a **bōl/solkaṭu** 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 provides the most likely sequence of syllables and their onset timestamps, given as $\mathbf{A}^* = [(t_1, a_1), (t_2, a_2), \dots, (t_*, a_{L^*})]$, where t_i is the onset time of a_i and L^* is the length of the transcribed sequence. Similar to the experiments with Beijing opera, all the transcription experiments were done using **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 jingju 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. **Rough Longest Common Subsequence (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 se-

quence \mathbf{A}^* with i and the query syllable sequence \mathbf{A}_q with j in this section. 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. δ 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 substring containing the subsequence match ending at the i^{th} and the j^{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 σ 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^{th} and the j^{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 η 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 ρ 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 ξ . 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 robust framework for subsequence search. The parameters ρ , η , ξ and δ 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 $f(.)$ 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 (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** called the warped density **RLCS** (denoted as RLCS_s in this chapter). 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 κ lead to more deviation from an identity transformation. RLCS_0 is a limiting case of 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** that uses a timbral distance $d(a_i, a_j)$ as RLCS_d and experiment with different thresholds δ .

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 have already been segmented into short phrases with the query patterns being the same order of length as the test audio recording, while the recordings in [MTS](#) dataset are full length compositions spanning multiple [tāl](#) cycles, and hence much longer than the query patterns. We will also analyze the effect of this difference in datasets on the results of pattern search.

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 transcription 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 (\mathcal{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 learned 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 initialization of [HMMs](#) is shown in Table 6.5 and the results for [HMMs](#) initialized with isolated stroke examples are 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). In both tables, the best performing combination with highest test Accuracy is shown in bold. For test data performance, the values underlined in each column of the ta-

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	68.50	50.48	<u>65.33</u>	44.10
	MFCC_0_D_A	69.33	46.72	<u>64.49</u>	39.48

Table 6.5: Automatic transcription results on the MTS dataset (tabla) 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. All values are in percentage.

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	70.71	60.77	<u>66.23</u>	53.13

Table 6.6: Automatic transcription results on the MTS dataset (tabla) using HMMs initialized using isolated stroke examples for each syllable.

bles are statistically equivalent to the best result (in a paired-sample t-test at 5% significance levels).

Overall, we see a best test Accuracy of 53.13% for isolated stroke initialization with the **MFCC_0_D_A** feature and a bigram language model, which justifies the use of 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 transcrip-

Variant	Parameter	Precision (p)	Recall (r)	f-measure (f)
Baseline	-	0.479	0.254	0.332
RLCS_0	$\delta = 0$	0.384	0.395	0.389
RLCS_d	$\delta = 0.3$	0.139	0.466	0.214
RLCS_d	$\delta = 0.6$	0.084	0.558	0.145
RLCS_s	$\kappa = 1$	0.412	0.350	0.378
RLCS_s	$\kappa = 4$	0.473	0.268	0.342
RLCS_s	$\kappa = 7$	0.482	0.259	0.336
RLCS_s	$\kappa = 9$	0.481	0.258	0.335

Table 6.7: Performance of approximate pattern search on [MTS](#) dataset ([tabla](#)) using different [RLCS](#) variants using the best performing parameter settings for RLCS_0 ($\rho = 0.875$, $\eta = 0.76$ and $\xi = 0.6$).

tions. 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.

We use the output transcriptions from the best performing combination ([MFCC_0_D_A](#) and a bigram language model) to report the performance of pattern search with approximate string matching in Table 6.7, using different [RLCS](#) variants for the query patterns from 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 (p), recall (r) and their harmonic mean f-measure (f). 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 per-

formed a grid search over the values of η , ρ and ξ with RLCS_0 . The parameters η and ξ are varied in the range 0 to 1. To ensure that the minimum length of the pattern matched is at least 2, we varied ρ in the range, $1.1/\min(L_q) < \rho < 1$. The parameter η is the convex sum parameter for the contribution of the rough match length density of the reference and the query towards the final score. With increasing η , we give more weight to the reference length ratio, allowing more insertions. We observed a poor true positive rate with larger η , 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 RLCS_0 is reported in Table 6.7. We see that RLCS_0 improves the recall, but with a lower precision and an improved f-measure, showing that the flexibility in approximate matching provided by **RLCS** comes at the cost of additional false positives. 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 values of ρ , η and ξ that give the best f-measure with RLCS_0 are then fixed for all subsequent experiments to compare the performance of the proposed **RLCS** variants. The results with other variants of **RLCS** are also reported in Table 6.7. The results from RLCS_d show that the use of a timbral syllable distance measure with higher threshold δ 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 RLCS_s improves the precision with larger

LM	Feature	Training		Test	
		C	A	C	A
Flat	MFCC_D_A	76.66	59.43	<u>74.08</u>	55.64
	MFCC_0_D_A	76.63	63.79	<u>74.13</u>	60.23
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 (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.

values of κ . The penalties on matches with higher number of insertions and deletions is large and they are left out, leading to a good precision at the cost of a poorer recall. We observe that both the above mentioned variants improve either precision 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 flat start embedded HMM training. With the best performing combination, we then report results of pattern search using different RLCS variants using the query patterns from 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. For test data performance, the values underlined in each column of the table are statistically equivalent to the best result (in a paired-sample t-test at 5% signif-

Variant	Parameter	Precision (ρ)	Recall (τ)	f-measure (f)
Baseline	-	0.902	0.492	0.637
RLCS_0 -1	$\delta = 0$	0.902	0.492	0.637
RLCS_0 -2	$\delta = 0$	0.258	0.762	0.386

Table 6.9: Performance of approximate pattern search for baseline and RLCS_0 . RLCS_0 -1 shows the best f-measure in the experiments, obtained with parameter settings $\rho = 0.525$, $\eta = 0.51$ and $\xi = 0.95$, while RLCS_0 -2 shows the best recall achieved, obtained with parameter settings $\rho = 0.275$, $\eta = 0.11$ and $\xi = 0.45$.

icance levels). 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 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 from Table 6.4. Table 6.9 shows the average results of pattern search with the **UMS** dataset (mridangam) with the RLCS_0 algorithm, with an exact string search baseline also shown. We further establish the optimum parameter settings for **RLCS** us-

ing 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 RLCS_d and RLCS_s algorithms did not show any improvement in f-measure for **tabla** solos, only the results of RLCS_0 are reported for mridangam solos.

The results in Table 6.9 are significantly different compared to Table 6.7 and needs further explanation. We see a higher baseline using exact string match with a good precision with a poorer recall, indicating an improved transcription accuracy with mridangam. The poorest recall of 0.348 is achieved for pattern **A₃** in Table 6.4. Further, interestingly, the best performing f-measure with RLCS_0 (shown in the table as $\text{RLCS}_0\text{-}1$) is equivalent to the baseline, giving an identical performance. On closer inspection, we see that this is achieved at a high score threshold of $\xi = 0.95$. Such a high score threshold makes the **RLCS** algorithm to be equivalent to exact search, penalizing any approximate length scores and looking for exact matches. However, the best recall (of 0.762) with RLCS_0 (shown in the table as $\text{RLCS}_0\text{-}2$) is obtained for a lower score threshold of $\xi = 0.45$, but with a significantly lower f-measure of 0.386 as shown in the table. In the case of the mridangam dataset, **RLCS** does not show a significant advantage in improving f-measure. In addition, we see that the results on mridangam dataset are insensitive to a wide range of values of ρ and η .

Both these interesting observations can be explained from the nature of the **UMS** dataset. The dataset consists of audio files that contain short segmented phrases, with query patterns being in the same order of length as the test audio files. In such a case, the computed rough match lengths and densities are not ill defined, leading to the insensitivity of the mixing parameter (η) and the minimum fraction of match parameter (ρ). In addition, an algorithm that considers rough lengths but uses a binary distance measure (such as RLCS_0) would provide no significant advantage over an exact search. We can summarize that the present formulation of RLCS_0 algorithm is advantageous only when the query patterns are much shorter than the audio recordings in which these patterns are being queried. In addition, there is hence a need to explore improvements to pattern search in cases such as the mridangam dataset, where a query pattern is being searched over a large number of audio files that also contain short phrases. However, a more comprehensive

experimentation on larger datasets with such characteristics would be necessary to confirm this claim.

6.4 Conclusions

The chapter presented a detailed formulation of the task of percussion pattern discovery in music cultures syllabic percussion systems. The approaches utilized the overall timbres of percussion strokes (either from a single drum or from an ensemble) to define patterns. An evaluation on percussion datasets in jingju and Indian art music datasets formally showed the possibility of such an approach, along with its advantages and current limitations. The goal of evaluations on percussion datasets of jingju, *tabla* and mridangam was to present a methodology for transcription and discovery/classification of percussion patterns in syllabic percussion systems. The work presented was preliminary and not comprehensive, with a significant scope for improvement. However, the basic idea of using a musically meaningful representation system to define and describe patterns is valid and useful. Beijing opera provided a useful test case for percussion pattern classification, showing promising results.

We mainly 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**. Similar experiments on the **UMS** dataset showed a better transcription performance, while pointing out a limitation of the **RLCS** approach when querying patterns in segmented short audio files.

For future work, we aim to improve syllable boundaries output by transcription using onset detection. Inclusion of the rhythmic in-

formation 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, while addressing its limitations when searching short audio files.

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. It is followed by a summary of the work presented in the dissertation, along with some key results and conclusions. The thesis opens up a host of open problems: 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 resulted from the work in CompMusic so far, and those that have been planned within the efforts of the project. 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 it for defining rhythm similarity measures between (and within) music pieces in large corpora of music. While this has not been addressed formally 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.

Enhanced music listening is a primary application of the meter analysis methods presented in the thesis. The additional rhythm related information such as the *tāla*, time varying tempo, beats and the *sama* can all enhance the music listening experience. It finds audience both in 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 syllabic vocalization. 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. Such a query by vocals system can be further used to build a machine improvisation system.

During Hindustani music concerts, it is common to have a call-response improvisatory passages between the 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-Jawaab¹ 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 onomatopoeic nature of the *tabla bōls*. 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 and well curated with useful metadata, and 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 otherwise 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 applications is provided for a better understanding, and then we emphasize on how the rhythm analysis methods developed in the thesis apply and integrate into these applications.

¹Further details and a demo available at http://labrosa.ee.columbia.edu/hamr_ismir2015/proceedings/doku.php?id=sawaal-jawaab or <http://compmusic.upf.edu/ismir-15-hacks>

7.1.1 Dunya

As described earlier, Dunya² 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, helping us to evaluate the research results from a user perspective. Dunya is aimed at the music community 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 selected sources such as music archives, Wikipedia and MusicBrainz, and makes it available to users for an enriching and engaging listening experience with music. Audio, automatically and manually extracted features, and curated metadata can also be accessed through Dunya. Dunya has a front end 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. Essentia³ is an audio analysis and audio based MIR toolkit (Bogdanov et al., 2013a, 2013b). The Dunya backend uses Essentia to extract features. Hence,

²<https://dunya.compmusic.upf.edu>

³<http://essentia.upf.edu/>

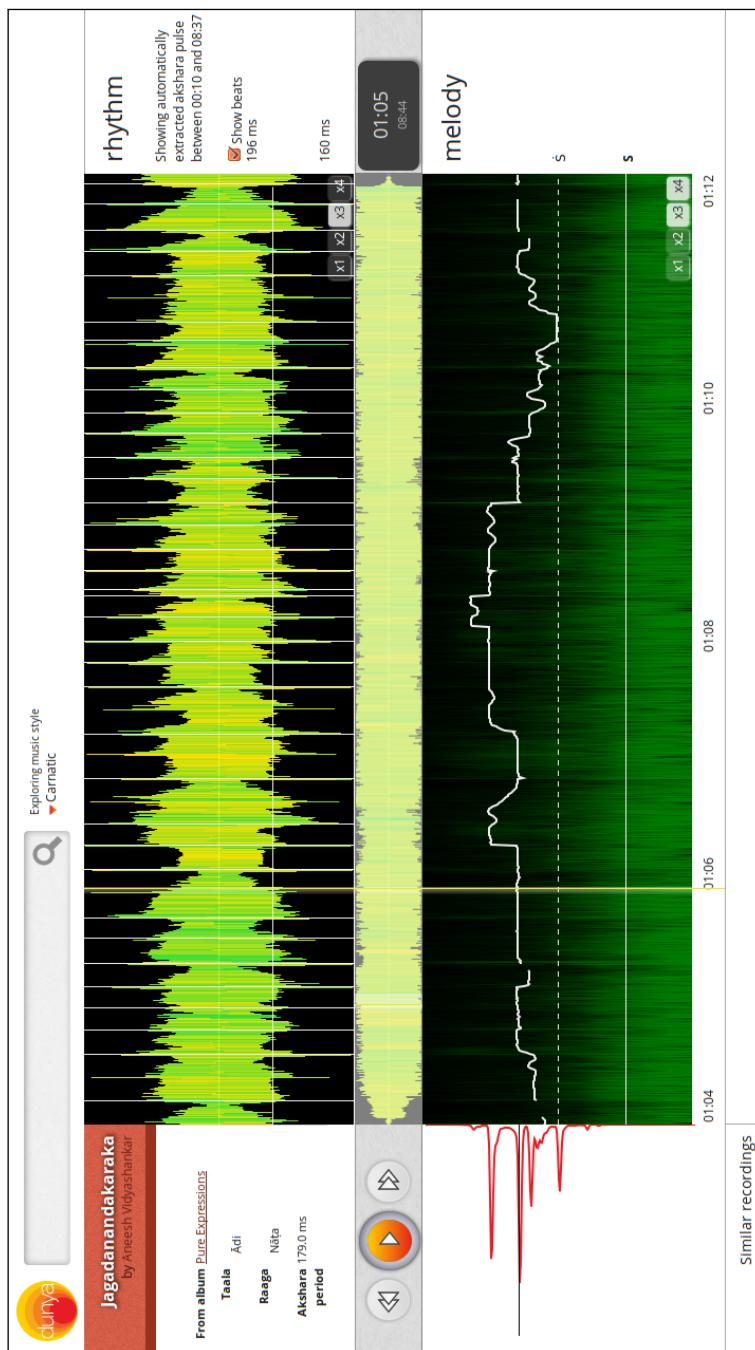


Figure 7.1: A screenshot of the recording page of Dunya showing rhythm related metadata in the top panel superimposed on top of the waveform.

specific rhythm extractors from the developed algorithms are also be added to Essentia.

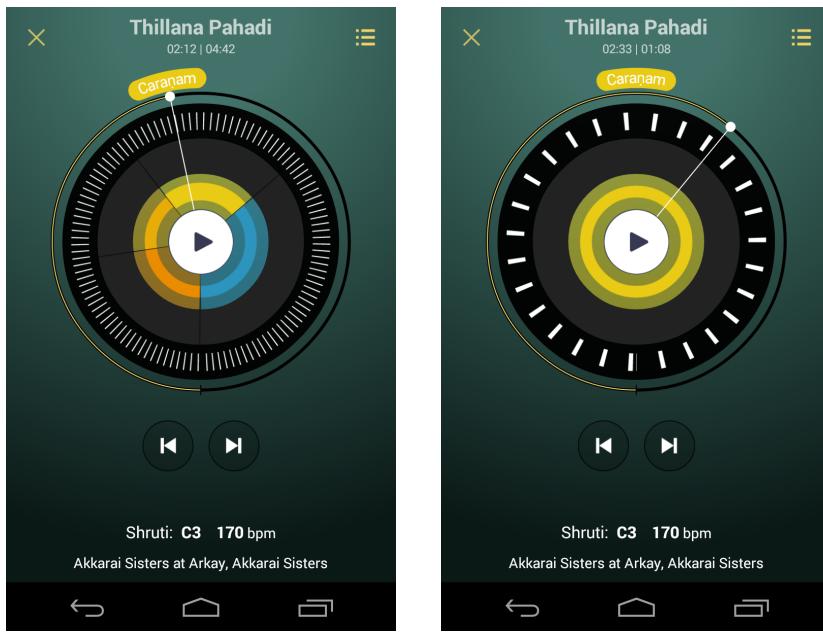
Drawing information from various data sources and relating them with ontologies, Dunya is the best platform to showcase the tools and algorithms developed as a part of the thesis. A screenshot of the Dunya recording page interface for a Carnatic music recording⁴ is shown in Figure 7.1. The recording page shows important rhythm related metadata related to the recording such as *tāla* editorial metadata and automatically extracted median *akṣara* pulse period (τ_o). In addition, the waveform panel on the top shows the time varying τ_o curve along with *akṣara* pulse 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 commercial applications, aiming to build technologies to foster learning and teaching of Indian music forms. Sarāga⁵ 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 provide a user with better understand-

⁴The 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

⁵Application 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 visualizing a music 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.

ing and appreciation of these music traditions. It additionally features unique compound filters over *rāgas*, *tālas*, instruments and artists for finding songs.

A screenshot of the application in Figure 7.2 shows the rich and novel visualization of a music recording⁶ including several different associated metadata. Figure 7.2a shows all the sections of the piece while Figure 7.2b 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

⁶The 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>)

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 semi-automatically extracted from audio using AMPF_0 algorithm with 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 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 author are listed below. The links to access all these datasets are provided in Appendix B.

- CompMusic Carnatic Music Rhythm (CMR_f) dataset: *Tāla*, beat and *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 with pieces spanning four popular *tālas*. A representative subset of the dataset with 118 pieces (CMR dataset) was also built (Section 4.2.1).
- CompMusic Hindustani Music Rhythm (HMR_f) dataset: *Tāl*, *mātrā* and *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 full dataset has about 5 hours of audio with pieces spanning four popular *tāls* and three different *lay* (tempo classes). Two subsets of the dataset grouped based on *lay*: HMR_l and HMR_s datasets were also built (Section 4.2.2).
- CompMusic Carnatic open music (CMD_o) collection: The *sama* annotations for the music pieces of the CMD_o collection in collaboration with Vignesh Ishwar. The collection presently contains

over 41 hours of music with 197 pieces and 16880 **sama** annotations (Section 4.1.4).

- CompMusic Hindustani open music (**HMD_o**) collection: The **sam** and section annotations for the music pieces of the **HMD_o** collection in collaboration with Kaustuv Kanti Ganguli. The collection presently contains over 43 hours of music with over 108 tracks, 11260 **sam** and 215 section annotations (Section 4.1.4).
- CompMusic Mulgaonkar **tabla** solo dataset (**MTS**) dataset: A **tabla** solo dataset comprising audio recordings, scores and time aligned syllabic transcriptions of 38 **tabla** solo compositions from different **gharānās** with time-aligned syllabic transcription was built with Swapnil Gupta (Section 4.2.3).
- CompMusic Jingju percussion pattern (**JPP**) dataset: The **JPP** dataset was built with Rafael Caro Repetto, and consists of 133 audio percussion patterns spanning five different pattern classes, with about 22 minutes of audio and over 2200 percussion syllables (Section 4.2.6).
- Jingju percussion instrument (**JPI**) 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 jingju (Section 4.2.5).

Technical and scientific contributions

- Identification of challenges, opportunities and applications of automatic rhythm analysis of Indian art music (Section 3.1).
- 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 culture-aware methods that can incorporate higher level music information (Chapter 3).
- Engineering formulation of meter analysis (meter inference, meter tracking and informed meter tracking) and percussion pattern discovery (transcription and search) in Indian art music (Section 3.3).

- An illustrative evaluation of the Carnatic and Hindustani music research corpora based on the methodology by Serra (2014) (Section 4.1).
- 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 (Section 4.2.1-4.2.2).
- 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 (Chapter 5).
- 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 (Chapter 6, joint work with Swapnil Gupta and IIT Madras CompMusic team).
- 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 (Section 6.2).

7.3 Conclusions and Summary

We present a summary, the conclusions and the 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 by 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 for **MIR** tasks extracting 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 limited to rhythm analysis in audio collections of Indian art music using Bayesian approaches, emphasizing on data and models. The thesis also addressed the question of the need for culture-aware data-driven approaches to rhythm analysis and its applications.

An introduction to rhythm in Indian art music was presented 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. Jingju (Beijing opera) percussion is a suitable case to study percussion patterns and hence a basic introduction to jingju 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 automatic analysis and to 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 to 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**, addressed in Chapter 4 shows that significant efforts are needed to build relevant datasets for research. It is possible to

build relevant high quality 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. To promote the ideas of open data and reproducible research, all the metadata and some of the audio from the research corpora are openly available through the Dunya API, while the copyrighted commercial audio is easily accessible.

Meter analysis was one of the main problems addressed in the thesis. Chapter 5 presented a comprehensive analysis of meter inference, 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 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 for the tasks 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. Contrary to intuition, more number of rhythmic patterns in the observation model did not improve meter analysis performance, which can be attributed to the simplistic spectral flux based audio feature. Further, the MO-model and the inference extensions did not show much improvement in performance, but are promising ideas to make inference better and faster with these Bayesian models. The models and inference extensions are capable of generalizing to other music cultures, as was

demonstrated with the evaluation on the 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 jingju percussion patterns. The approach was 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 along with a simple language model. A set of query patterns were automatically derived from symbolic syllabic scores. To make the approach robust to errors in automatic transcription, an approximate search algorithm (such as [RLCS](#)) was used to search for these query patterns in transcribed recordings. 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. The [UMS](#) dataset consists of short audio files segmented by phrases and hence [RLCS](#) does not show much improvement in overall f-measure 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 a potential to be explored in the future.

The goal of automatic rhythm analysis is to define musically 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 explored as a part of future work.

The sizeable curated research corpora and datasets provide an immense opportunity to 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 evolves over time and to sustainably improve the research corpora and build additional datasets for rhythm analysis tasks is an important task for future. 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 shorter excerpts, to make it more practi-

cal. 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 part of 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 better audio features that can perhaps also include information from other dimensions such as melody and lyrics 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 small datasets. While the framework using syllabic representation for percussion patterns is promising, an extensive evaluation on larger datasets spanning different instrument timbres, playing styles, schools (*gharānās*), and variability in patterns is to be done in the future. Better transcription can be achieved in real world scenarios with the availability of 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.

The tasks of meter analysis and percussion transcription and discovery were addressed as independent tasks in the thesis, while there is also significant interplay between the tasks. Meter analysis could benefit from percussion instrument timbre based features to track metrical structures (the strokes of *tabla* played in a *ṭhēkā* are indicative of position in the *tāl* cycle). Percussion pattern discovery can benefit significantly if the *sama* and beat information is available, since percussion patterns are often aligned with beats (and sometimes *sama*) of the *tāla*. New approaches that can address the two tasks together and build joint analysis algorithms are promising.

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 to improve the algorithms through user feedback, while aiming to provide users with engaging 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 further research in the field, and this is perhaps where a new PhD thesis can begin!

List of Publications

The following is a list of publications by the author in the context of the thesis and CompMusic project. The text in parentheses outlines the specific contribution of the author in each publication.

Peer-reviewed journals

- 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. (Contributed to identifying problems, challenges, opportunities and state of the art, building the Indian music datasets, conducting experiments on Indian music datasets, analysis and interpretation of results and writing of the manuscript. Companion webpage for the paper: <http://compmusic.upf.edu/jnmr-2014-rhythm>)

Full articles in peer-reviewed conferences

- 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 (ICASSP2016)* (pp. 76–80). Shanghai, China. (Formulated the section pointer model, built the Hindustani rhythm dataset, conducted the experiments and wrote the manuscript. Companion webpage for the paper: <http://compmusic.upf.edu/icassp-2016-spm>)

- **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. (Formulated the mixture observation model, built the Carnatic rhythm data subset, conducted the experiments and wrote the manuscript. Companion webpage for the paper: <http://compmusic.upf.edu/ismir-2015-pf>)
- 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. (Contributed to formulating the problem of pattern discovery in tabla solos, building the dataset, transcription experiments, analysis of the results and writing the manuscript. Companion webpage for the paper: <http://compmusic.upf.edu/ismir-2015-tabla>)
- **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. (Contributed to formulating the problem of percussion pattern transcription in jingju, helped in building the percussion pattern dataset, conducted all the experiments and wrote the manuscript. Companion webpage for the paper: <http://compmusic.upf.edu/ismir-2014-bo>)
- 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)* (pp. 425–430). Taipei, Taiwan. (Contributed to building the Carnatic dataset, analysis of results and writing the manuscript. Companion webpage for the paper: <http://compmusic.upf.edu/ismir-2014-odd>)
- **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. Athens, Greece. (Contributed to collection and analysis of corpora data, writing the manuscript. Companion webpage for the paper: <http://compmusic.upf.edu/smc-2014-corpora>)

- 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. (Formulated the tempo, *akṣara* and *sama* tracking problem, built the Carnatic rhythm dataset, conducted the experiments and wrote the manuscript. Companion webpage for the paper: <http://compmusic.upf.edu/icassp-2014-talaTrack>)
- 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. (Formulated the problem of instrument-wise onset detection in jingju, designed and conducted the experiments, and wrote the manuscript. Companion webpage for the paper: <http://compmusic.upf.edu/icassp-2014-onsetbo>)
- 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. (Formulated the problem of tracking sub-beat structure and cycle length from audio recordings, compiled the dataset, conducted the experiments and wrote the manuscript.)
- Dzhambazov, G., Srinivasamurthy, A., Senturk, S., & Serra, X. (2016). On the use of Note Onsets for Improved Lyrics-to-audio alignment in Turkish Makam Music. To appear in the *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, New York, USA. (Contributed to the formulation of a variable-time HMM for lyrics-to-audio alignment - content not a part of the dissertation.)

Other contributions to conferences

- Krebs, F., Holzapfel, A., & **Srinivasamurthy, A.** (2014). MIREX 2014 Audio Downbeat Tracking Evaluation: KHS1. 10th Music Information Retrieval Evaluation eXchange (MIREX), extended abstract. Taipei, Taiwan. (Algorithm from Holzapfel et al. (2014) submitted for evaluation.)
- Gulati, S., Ganguli, K. K., Gupta, S., **Srinivasamurthy, A.**, & Serra, X. (2015). RAGAWISE: A Lightweight Real-time Raga Recognition System for Indian Art Music. In Late-Breaking Demo Session of the 16th International Society for Music Information Retrieval Conference. Malaga, Spain. (Contributed to conceptualization, feedback, debugging and testing.)
- Caro R., **Srinivasamurthy, A.**, Gulati, S., & Serra, X. (2014). Jingju music: concepts and computational tools for its analysis. A Tutorial in the 15th International Society for Music Information Retrieval Conference, Taipei, Taiwan. (Presented the rhythm part of the tutorial, discussing automatic rhythm analysis problems and methods in jingju music.)

Resources

This appendix is a compendium of links to resources and additional material related to the work presented in the thesis. An up-to-date set of links and material can also be obtained from the companion webpage <http://compmusic.upf.edu/phd-thesis-ajay> or its mirror on www.ajaysrinivasamurthy.in/phd-thesis. Latest updates on the CompMusic project can be obtained from <http://compmusic.upf.edu/>.

Some of the results not reported in the dissertation and audio examples showcasing the results are presented on the companion webpage. The companion webpage will also be updated with any additional resources and material that will be built in the future.

Music concepts and audio examples

A resource page for Carnatic *tālas*, with additional explanation of the structure of many different *tālas*, with audio examples of music pieces in popular Carnatic *tālas*

<http://compmusic.upf.edu/examples-taala-carnatic>

A resource page for Hindustani *tāls*, with additional explanation of the structure of many different *tāls*, with audio examples of music pieces in popular Hindustani *tāls*

<http://compmusic.upf.edu/examples-taal-hindustani>

Audio examples for the different percussion instruments used in Beijing opera

<http://compmusic.upf.edu/examples-percussion-bo>

A resource page for percussion patterns in Beijing opera, including scores and audio examples of popular percussion patterns

<http://compmusic.upf.edu/bo-perc-patterns>

A resource page for *usul*, the cyclic rhythmic framework in Turkish makam music, with audio examples and scores

<http://compmusic.upf.edu/examples-usul-mmt>

Corpora and datasets

Access to the corpora and datasets will be through the Dunya API, providing access to audio recordings, metadata and features. Standalone archives of datasets are also distributed in some cases outside of the Dunya API. All the research corpora and datasets related to the thesis are also listed at the links below.

Research corpora - <http://compmusic.upf.edu/corpora>

Test datasets - <http://compmusic.upf.edu/datasets>

The Dunya Carnatic collection on MusicBrainz that forms the Comp-Music Carnatic music research corpus

[http://musicbrainz.org/collection/
f96e7215-b2bd-4962-b8c9-2b40c17a1ec6](http://musicbrainz.org/collection/f96e7215-b2bd-4962-b8c9-2b40c17a1ec6)

The Dunya Hindustani collection on MusicBrainz that forms the CompMusic Hindustani music research corpus

[http://musicbrainz.org/collection/
213347a9-e786-4297-8551-d61788c85c80](http://musicbrainz.org/collection/213347a9-e786-4297-8551-d61788c85c80)

The Carnatic Creative Commons music collection (CMD_o) on MusicBrainz with openly accessible music

[http://musicbrainz.org/collection/
a163c8f2-b75f-4655-86be-1504ea2944c2](http://musicbrainz.org/collection/a163c8f2-b75f-4655-86be-1504ea2944c2)

The Hindustani Creative Commons music collection (HMD_o) on MusicBrainz with openly accessible music

[http://musicbrainz.org/collection/
6adc54c6-6605-4e57-8230-b85f1de5be2b](http://musicbrainz.org/collection/6adc54c6-6605-4e57-8230-b85f1de5be2b)

The Carnatic Music Rhythm dataset (CMR_f) containing rhythm annotated pieces of Carnatic music, from which a subset CMR dataset is also available

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

The Hindustani Music Rhythm dataset (HMR_f) containing rhythm annotated pieces of Hindustani music, from which two subsets HMR_s and HMR_l datasets are also available

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

The Anantapadmanabhan Mridangam Strokes dataset (AMS) containing audio examples of individual strokes of the mridangam in various tonics

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

The UKS Mridangam Solo dataset (UMS) containing a transcribed collection of two $tani$ - $\bar{a}vartana$ played by the renowned mridangam maestro Padmavibhushan Umayalpuram K. Sivaraman

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

The Mulgaonkar Tabla Solo dataset (MTS) containing a transcribed collection of tabla solo audio recordings spanning compositions from six different $gharānās$ of tabla, compiled from the album *Shades of Tabla* by Pandit Arvind Mulgaonkar

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

The Jingju Percussion Instrument dataset (JPI) containing isolated strokes spanning the four percussion instrument classes used in Beijing opera

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

The [Jingju Percussion Pattern dataset \(JPP\)](#) containing a collection of audio percussion patterns covering five pattern classes in Beijing opera

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

Results

An extended set of results, along with a few audio examples analyzed with the models and algorithms presented in the dissertation are available on the companion page.

Audio examples

<http://compmusic.upf.edu/phd-thesis-ajay#examples>

Extended results

<http://compmusic.upf.edu/phd-thesis-ajay#results>

Tools and code

The links to tools and code related to the thesis are listed. Up-to-date links to code (including future releases) will be available on:
<http://compmusic.upf.edu/phd-thesis-ajay#code>

Essentia audio analysis library

<http://essentia.upf.edu/>

Dunya API

<https://github.com/MTG/pycompmusic>

Dunya front end

<http://dunya.compmusic.upf.edu/>

Dunya server and back end

<https://github.com/MTG/dunya>

A MATLAB package for meter analysis (Florian Krebs)

<https://github.com/flokadillo/bayesbeat>

A MATLAB package for beat tracking evaluation (Matthew Davies)

<https://code.soundsoftware.ac.uk/projects/beat-evaluation/>

Rhythm analysis tools for jingju, from the tutorial in ISMIR 2014

<http://compmusic.upf.edu/jingju-tutorial>

Sawaal-Jawaab Code and Demo

<http://compmusic.upf.edu/ismir-15-hacks>

Sonic Visualizer, for visualization and annotation of audio

<http://www.sonicvisualiser.org/>

BeatStation, an interface to record beat tapping

<https://github.com/ajaysmurthy/beatStation>

Glossary

C.1 Carnatic music

- akṣara** The lowest metrical pulse (subdivision)
- 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
- ghaṭam** A percussion instrument used in Carnatic music (specially made clay pot with a narrow mouth)
- khañjira** A tambourine like percussion instrument used in Carnatic music
- konnakōl** The art form of reciting percussion syllables
- kṛti** A common compositional form in Carnatic music
- mōrsiṅg** The Indian jaw (jew's) harp
- mṛdaṅgam** The primary percussion accompaniment in Carnatic music (common spelling mridangam)
- 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 beginning of an āvartana (equivalent to a downbeat)

- śyāmā śāstri** A prominent Carnatic music composer
- solkatṭu** The onomatopoeic oral percussion syllables
- 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
 - miśra 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 performances
- tyāgarāja** A prominent Carnatic music composer
- vīṇā** A stringed instrument used in Carnatic music

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 music
- bōl** The onomatopoeic oral percussion syllables of the tabla
- dhrupad** A music style in Hindustani music
- gharānā** The stylistic schools of Hindustani music
- khālī** A hand wave in the tāl cycle (unaccented)
- khyāl** A music style in Hindustani music
- lay** The tempo class
 - dr̥t** Fast tempo class
 - madhya** Medium tempo class
 - vilāmbit** Slow tempo class
- mātrā** The lowest defined metrical pulse in Hindustani music (equivalent to a beat)
- 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
- santūr** A trapezoid-shaped hammered dulcimer or string instrument

- sāraṅgi** A bowed music instrument used in Hindustani music
- sarōd** A fretless plucked string instrument used in Hindustani music
- sitār** A plucked string 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
- tabla** The primary percussion accompaniment in Hindustani music
- bāyān** The left drum
 - dāyān** The right drum
 - diggā** Alternative name for the left drum
- tānpura** The drone music instrument used in Hindustani music
- thālī** A hand clap in the tāl cycle (accented)
- gaṭ** A compositional form in tabla
- kāyadā** A compositional form in tabla
- palaṭā** A compositional form in tabla
- pēśkār** A compositional form in tabla
- rēlā** A compositional form in tabla
- ṭhēkā** The basic bōl pattern associated with a tāl
- vibhāg** The sections of a tāl cycle

C.3 Beijing opera (Jingju)

- bangu** Clapper-drum
- banshi** Rhythmic modes of Beijing opera
- daluo** Big gong
- naobo** Cymbals
- luogu jing** Percussion pattern
- xiaoluo** Small gong

C.4 Acronyms

- AMS** Anantapadmanabhan Mridangam Strokes dataset
- JPI** Jingju Percussion Instrument dataset
- JPP** Jingju Percussion Pattern dataset

- CMR_f** Carnatic Music Rhythm dataset
- CMD_o** Carnatic Creative Commons music collection
- CMR** Carnatic Music Rhythm dataset (subset of CMR_f)
- HMR_f** Hindustani Music Rhythm dataset
- HMR_l** Hindustani Music Rhythm dataset (subset of HMR_f with **vi-lambit** and **madhya lay** pieces)
- HMD_o** Hindustani Creative Commons music collection
- HMR_s** Hindustani Music Rhythm dataset (subset of HMR_f with **drt lay** pieces)
- HMM_m** HMM (Viterbi) inference algorithm with discretized mixture observation model (Srinivasamurthy et al., 2015)
- HMM₀** HMM (Viterbi) inference algorithm with discretized bar pointer model
- HMM_s** HMM (Viterbi) inference algorithm with discretized section pointer model
- AMPF_e** AMPF inference algorithm with end-of-bar sampling
- AMPF_m** AMPF inference algorithm with mixture observation model (Srinivasamurthy et al., 2015)
- AMPF_g** AMPF inference algorithm with onset gated weight update
- AMPF_p** Peak hop inference with AMPF
- AMPF₀** AMPF inference algorithm with bar pointer model
- AMPF_s** AMPF inference algorithm with the section pointer model (Srinivasamurthy et al., 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
- AML** Allowed Metrical Levels
- AMPF** Auxiliary Mixture Particle Filter
- APF** Auxiliary Particle Filter
- CML** Correct Metrical Level
- DBN** Dynamic Bayesian Network

- DP** Dynamic Programming
GMM Gaussian Mixture Model
HMM Hidden Markov model
HPSS Harmonic-Percussive Source Separation
HTK Hidden Markov model Toolkit
ICT Information and Communication Technologies
IOI Inter-Onset Interval
LCS Longest Common Subsequence
LSTM Long Short-Term Memory
MAP maximum *a posteriori*
MBID MusicBrainz IDentifier
MFCC Mel-Frequency Cepstral Co-efficients
MFCC_0_D_A MFCC features with energy, velocity and acceleration coefficients
MFCC_D_A MFCC features without energy but with velocity and acceleration coefficients
MIR Music Information Research
MIREX Music Information Retrieval EXchange
MMA Madras Music Academy
MPF Mixture Particle Filter
NMF Non-negative Matrix Factorization
RLCS Rough Longest Common Subsequence
RNN Recurrent Neural Network
SIS Sequential Importance Sampling
SISR Sequential Importance Sampling/Resampling
SMC Sequential Monte Carlo
STFT Short-Time Fourier Transform
WAQ Width Across Query
WAR Width Across Reference

Bibliography

The numbers in brackets at the end of each bibliographic entry indicate the pages in which it is cited.

- 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. [57, 245]
- 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*. London, UK. [86]
- 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)* (pp. 181–185). Vancouver, Canada. [86, 168]
- 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. [118]

- Aucouturier, J. J., & Pachet, F. (2002, October). 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. [267]
- Barber, D., Cemgil, A. T., & Chiappa, S. (2011). *Bayesian time series models*. Cambridge University Press. [239]
- 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. [43, 45, 168, 182, 184, 264]
- Bello, J. P., Rowe, R., Guedes, C., & Toussaint, G. (2015). Five Perspectives on Musical Rhythm. *Journal of New Music Research*, 44(1), 1–2. [2]
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 34–43. [88]
- Beronja, S. (2008). *The Art of the Indian tabla*. Rupa and Co. New Delhi. [32, 33, 34]
- Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011, October). The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)* (pp. 591–596). Miami, USA. [116]
- Bhatkhande, V. N. (1990). *Hindustani Sangeet Paddhati: Kramik Pustak Maalika Vol. I-VI*. Sangeet Karyalaya. [67, 129]
- Böck, S., Krebs, F., & Schedl, M. (2012, October). Evaluating the Online Capabilities of Onset Detection Methods. In *Proceedings of the 13th International Society for Music Information Retrieval Conference* (pp. 49–54). Porto, Portugal. [45, 140]
- Böck, S., Krebs, F., & Widmer, G. (2014, October). 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)* (pp. 602–607). Taipei, Taiwan. [52, 175, 191, 208]
- Böck, S., & Schedl, M. (2011, September). Enhanced Beat Tracking with Context-aware Neural Networks. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)* (pp. 135–139). Paris, France. [49, 50, 241]
- Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., & 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. [186, 280]
- Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., & 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. [280]
- Brachman, R., & Levesque, H. (2004). *Knowledge Representation and Reasoning (The Morgan Kaufmann Series in Artificial Intelligence)*. Morgan Kaufmann. [88]
- 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. [101, 134, 172]
- 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)* (pp. 313–318). Taipei, Taiwan. [118]
- Chandola, A. (1988). *Music as Speech: An Ethnomusicological Study of India*. Navrang. [261]
- 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 (ISMIR 2012)* (pp. 445–450). Porto, Portugal. [56]
- Chordia, P. (2005a). *Automatic Transcription of Solo Tabla Music* (Doctoral dissertation). Stanford University. [85]
- Chordia, P. (2005b, September). Segmentation and recognition of tabla strokes. In *Proceedings of 6th International Society for Music Information Retrieval Conference (ISMIR 2005)* (pp. 107–114). London, UK. [85]
- Chordia, P. (2006, August). Automatic transcription and representation of solo tabla music. *Computing in Musicology*, 14, 123–138. [67]

- Chordia, P., Sastry, A., & Albin, A. (2010, October). Evaluating Multiple Viewpoints Models of Tabla Sequences. In *Proceedings of ACM Multimedia Workshop on Music and Machine Learning* (pp. 21–24). Florence, Italy. [85, 91]
- 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. [85]
- Chordia, P., Sastry, A., Mallikarjuna, T., & Albin, A. (2010, August). Multiple viewpoints modeling of tabla sequences. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)* (pp. 381–386). Utrecht, Netherlands. [85, 91]
- Clarke, E. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *The Psychology of Music* (2nd ed., pp. 473–500). Academic Press, San Diego. [66, 208, 221]
- 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. [37]
- Clayton, M. (2000). *Time in Indian Music : Rhythm, Metre and Form in North Indian Rag Performance*. Oxford University Press. [21, 29, 30, 33, 66, 76, 80, 81, 177, 208, 209]
- Cohen, K. B., Ogren, P. V., Fox, L., & Hunter, L. (2005, October). Empirical data on corpus design and usage in biomedical natural language processing. In *AMIA Annual Symposium Proceedings 2005* (pp. 156–160). Washington, DC, USA. [116]
- Cooper, G., & Meyer, L. B. (1960). *The rhythmic structure of music*. University of Chicago Press. [20]
- D'Ambrosio, B. (1999). Inference in Bayesian networks. *AI magazine*, 20(2), 21–36. [59]
- Damerau, F. J. (1964, March). A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, 7(3), 171–176. [124]
- 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*. [52]
- 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. [51, 109, 110, 181]
- 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. [47, 48, 77, 102, 306]
- 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. [45, 46]
- Dixon, S. (2007). Evaluation of The Audio Beat Tracking System Beatroot. *Journal of New Music Research*, 36(1), 39–50. [47, 48]
- Dixon, S., Guoyon, F., & Widmer, G. (2004, October). Towards Characterisation of Music via Rhythmic Patterns. In *Proceedings of the 5th International Conference on Music Information Retrieval*. Barcelona. [50]
- Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Non-linear Filtering*. [60, 201, 202]
- Durand, S., Bello, J. P., & David, B. (2016, March). Feature adapted convolutional neural networks for downbeat tracking. In *Proceedings of the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016)* (pp. 296–300). Shanghai, China. [52]
- Durand, S., Bello, J. P., David, B., & Richard, G. (2015, May). Downbeat tracking with multiple features and deep neural networks. In *Proceedings of the 40th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*. Brisbane, Australia. [52]
- Dutta, A. E. (1995). *Tabla: Lessons and Practice*. Ali Akbar College. [33]
- 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)* (pp. 1–6). Kanpur, India. [264, 265]
- Ellis, D. P. W. (2007). Beat Tracking by Dynamic Programming. *Journal of New Music Research*, 36(1), 51–60. [47, 48, 77, 107]

- Fitzgerald, D. (2010, September). Harmonic/percussive separation using median filtering. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*. Graz, Austria. [46]
- 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. [57]
- Fletcher, N. H., & Rossing, T. D. (1998). *The Physics of Musical Instruments*. Springer. [57]
- Foote, J. (2000, August). Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo 2000* (Vol. 1, pp. 452 – 455). New York, USA. [47, 56, 186]
- 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. [50]
- Fouloulis, A., Papadelis, G., Pastiadis, K., & Papanikolaou, G. (2010, March). 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. [55]
- 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. [50]
- Gillet, O., & Richard, G. (2004a, October). Automatic Labelling of Tabla Signals. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*. Barcelona, Spain. [85]
- 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. [57, 245]
- Gillet, O., & Richard, G. (2007, September). Supervised and unsupervised sequence modeling for drum transcription. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)* (pp. 219–224). Vienna, Aus-

- tria. [85]
- 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. [57, 245]
- 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. [88]
- Goto, M. (2006, October). AIST Annotation for the RWC Music Database. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)* (p. 359-360). Victoria, Canada. [50, 174]
- 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. [57, 245]
- Gottlieb, R. S. (1993). *Solo Tabla Drumming of North India: Its Repertoire, Styles, and Performance Practices*. Motilal BanarsiDass Publishers. [32, 34, 258]
- Gouyon, F. (2005). *A Computational Approach to Rhythm Description* (Doctoral dissertation). Universitat Pompeu Fabra, Barcelona, Spain. [43, 44]
- Gouyon, F., & Dixon, S. (2005). A Review of Automatic Rhythm Description Systems. *Computer Music Journal*, 29(1), 34–54. [43]
- Gouyon, F., Herrera, P., & Cano, P. (2002, June). Pulse-dependent Analyses of Percussive Music. In *Audio Engineering Society: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Espoo, Finland. [57, 245]
- 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. [50, 174, 175]
- Grosche, P., & Müller, M. (2011a). Extracting Predominant Local Pulse Information from Music Recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6),

- 1688–1701. [47]
- Grosche, P., & Müller, M. (2011b, October). Tempogram Toolbox: MATLAB tempo and pulse analysis of music recordings. In *12th International Conference on Music Information Retrieval (ISMIR 2011), late-breaking contribution*). Miami, USA. [47, 56, 182, 184]
- 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. [76]
- 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. [76, 101, 306]
- Gupta, S. (2015). *Discovery of Percussion Patterns from Tabla Solo Recordings* (Master's Thesis). Universitat Pompeu Fabra, Barcelona, Spain. [168, 258]
- 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. [15, 16, 168, 258]
- Hainsworth, S., & Macleod, M. (2003, October). Beat tracking with particle filtering algorithms. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 91–94). New Paltz, New York. [49, 50, 174]
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. [49]
- 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) (pp. 169–174). Porto, Portugal. [51, 110, 111, 181, 221, 306]
- Holzapfel, A., Davies, M., Zapata, J. R., Oliveira, J. L., & Gouyon, F. (2012, November). Selective Sampling for Beat Tracking Evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9), 2539–2548. [48, 49, 50, 66, 100, 174, 225]
- 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. [54, 55]
- Holzapfel, A., Krebs, F., & Srinivasamurthy, A. (2014, October). 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)* (pp. 425–430). Taipei, Taiwan. [15, 141, 175, 191, 205, 208, 226, 296]
- Holzapfel, A., & Stylianou, Y. (2009, October). Rhythmic Similarity in Traditional Turkish Music. In *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR 2009)* (pp. 99–104). Kobe, Japan. [55]
- Holzapfel, A., & Stylianou, Y. (2011). Scale transform in rhythmic similarity of music. *IEEE Transactions on Speech and Audio Processing*, 19(1), 176–185. [55, 102, 306]
- Huang, X., & Deng, L. (2010, February). An Overview of Modern Speech Recognition. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed., pp. 339–366). Chapman and Hall/CRC. [60, 246]
- Hughes, A., & Gerson-Kiwi, E. (2001). Solmization. In *Grove music online. oxford music online*. Oxford University Press. Retrieved from <http://www.oxfordmusiconline.com/subscriber/article/grove/music/26154> [84]
- Hughes, D. (2000). No nonsense: the logic and power of acoustic-iconic mnemonic systems. *British Journal of Ethnomusicology*, 9(2), 93–120. [84]
- Huron, D. (2002). Music information processing using the hum-drum toolkit: Concepts, examples, and lessons. *Computer Music Journal*, 26(2), 11–26. [85]
- ISO/TC. (2001). *ISO/TC 15919:2001 Information and documenta-*

- tion – Transliteration of Devanagari and related Indic scripts into Latin characters.* Geneva, Switzerland: International Organization for Standardization. [22]
- Jehan, T. (2005). *Creating music by listening* (Doctoral dissertation). Massachusetts Institute of Technology. [51]
- Jha, R. (2001). *Abhinav Geetanjali Vol. I-V.* Sangeet Sadan Prakashan. [67, 129]
- Johansen, A., & Doucet, A. (2008). A note on auxiliary particle filters. *Statistics and Probability Letters*, 78(12), 1498–1504. [202]
- 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. [57]
- 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. [243]
- 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. [46]
- Klapuri, A., & Davy, M. (2006). *Signal Processing Methods for Music Transcription*. Springer. [56]
- 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. [47, 48, 52, 100, 102, 107, 306]
- 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. [88]
- 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. [91]
- 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. [73]
- Koduri, G. K., & Serra, X. (2013, October). A knowledge-based approach to computational analysis of melody in Indian art music. In *Proceedings of the International Workshop on Semantic Music and Media, International Semantic Web Conference* (pp. 1–10). Sydney, Australia. [88]
- Kolinski, M. (1973). A cross-cultural approach to metro-rhythmic patterns. *Ethnomusicology*, 17(3), 494–506. [20]
- Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press. [59]
- 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)* (pp. 227–232). Curitiba, Brazil. [15, 52, 140, 174, 191, 197, 223]
- 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)* (pp. 72–78). Malaga, Spain. [191, 199]
- 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. [15, 52, 175, 191, 195, 199, 201, 202, 203, 223]
- Kuriakose, J., Kumar, J. C., Sarala, P., Murthy, H. A., & Sivaraman, U. K. (2015, February). Akshara Transcription of Mrudangam Strokes in Carnatic Music. In *Proceedings of the 21st National Conference on Communication (NCC)*. Mumbai, India. [86, 169, 171]
- Lang, D. (2004). *Fast Methods for Inference in Graphical Models and Beat Tracking the Graphical Model Way* (Master's Thesis). The University of British Columbia, Vancouver, Canada. [49]
- Lang, D., & Freitas, N. D. (2005). Beat Tracking the Graphical Model Way. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17* (pp. 745–752). MIT Press. [49]

- Lee, Y.-Y., & Shen, S.-Y. (1999). *Chinese Musical Instruments (Chinese Music Monograph Series)*. Chinese Music Society of North America Press. [39]
- Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. MIT Press Cambridge. [20, 66]
- Liberman, M., & Cieri, C. (1998, May). The Creation, Distribution and Use of Linguistic Data: The Case of the Linguistic Data Consortium. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Granada, Spain. [116]
- Ligeti, G. (2007). Brief an Kai Jakobs. In M. Lichtenfeld (Ed.), *Gesammelte Schriften*. Paul Sacher Stiftung. [20]
- Lin, H., Wu, H., & Wang, C. (2011). Music Matching Based on Rough Longest Common Subsequence. *Journal Information Science and Engineering*, 27(1), 95–110. [61, 264, 265]
- London, J. (2001). Metre. In L. Macy (Ed.), *Grove music online. oxford music online*. Oxford University Press. Retrieved from <http://www.oxfordmusiconline.com/subscriber/article/grove/music/18519> [20]
- London, J. (2004). *Hearing in time: Psychological aspects of musical meter*. Oxford: Oxford University Press. [20]
- 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. [254]
- Marchand, U., Fresnel, Q., & Peeters, G. (2015, October). GTZAN-Rhythm: extending the GTZAN test-set with beat, downbeat and swing annotations. In *Extended abstracts for the Late-Breaking Demo Session of the 16th International Society for Music Information Retrieval Conference*. Malaga, Spain. [50, 174]
- Mauch, M., & Dixon, S. (2012, October). A Corpus-based Study of Rhythm Patterns. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR 2012)* (pp. 163–168). Porto, Portugal. [57, 84]
- 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. [52, 189]

- Mehta, R. C. (2008). *Indian Classical Music & Gharana Tradition*. Readworthy Publications, India. [23]
- Miron, M. (2011). *Automatic Detection of Hindustani Talas* (Master's Thesis). Universitat Pompeu Fabra, Barcelona, Spain. [33, 77, 85]
- Moelants, D., & McKinney, M. F. (2004, August). 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)* (pp. 558–562). Evanston, IL, USA. [50, 174]
- Mu(穆文义), W. (2007). *Jingju dajiyue jiqiao yu lianxi: 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. [39, 40, 41, 42]
- Müller, M., Ellis, D. P. W., Klapuri, A., Richard, G., & Sagayama, S. (2011). Introduction to the Special Issue on Music Signal Processing. *IEEE Journal of Selected Topics in Signal Processing*, 5(6), 1085–1087. [20]
- Murphy, K. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning* (Doctoral dissertation). UC Berkeley, Computer Science Division. [56, 59, 191]
- Naimpalli, S. (2005). *Theory and practice of Tabla*. Popular Prakashan. [33]
- 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)* (pp. 550–553). [57, 251]
- Navarro, G. (2001, March). A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, 33(1), 31–88. [60, 253]
- 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)* (pp. 1–4). Lausanne, Switzerland. [46]
- 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. [267]

- Pan, S., & Weng, W. (2002, July). Designing a speech corpus for instance-based spoken language generation. In *Proceedings of the 2nd International Conference on Natural Language Generation* (pp. 49–56). New York, USA. [116]
- Parncutt, R. (1994). A Perceptual Model of Pulse Salience and Metrical Accent in Musical Rhythms. *Music Perception*, 11(4), 409–464. [20]
- Parry, M., & Essa, I. (2003, October). Rhythmic Similarity through Elaboration. In *Proceedings of 4th International Conference on Music Information Retrieval (ISMIR 2003)*. Baltimore, USA. [55]
- Paulus, J., & Klapuri, A. (2009). Drum Sound Detection in Polyphonic Music with Hidden Markov Models. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009. [58]
- Paulus, J., & Virtanen, T. (2005, September). Drum transcription with non-negative spectrogram factorisation. In *Proceedings of the 13th European Signal Processing Conference (EUSIPCO 2005)* (pp. 4–8). Antalya, Turkey. [57, 245]
- 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. [97, 116]
- 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. [49, 52]
- 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. [50, 101, 102, 105, 306]
- 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. [54, 55, 102, 306]
- Porter, A., Bogdanov, D., Kaye, R., Tsukanov, R., & Serra, X. (2015, October). AcousticBrainz: a community platform for

- gathering music information obtained from audio. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (pp. 786–792). Malaga, Spain. [116]
- 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)* (pp. 101–106). Curitiba, Brazil. [4, 91, 131]
- 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). [59, 60, 200]
- Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of Speech Recognition*. PTR Prentice Hall. [60]
- Rae, A., & Chordia, P. (2010, January). Tabla Gyan: An Artificial Tabla Improviser. In *Proceedings of the First International Conference on Computational Creativity (ICCC X)* (pp. 155–164). Lisbon, Portugal. [85]
- Raimond, Y. (2008). *A Distributed Music Information System* (Doctoral dissertation). University of London. [88]
- Raman, C. V. (1934). The Indian Musical Drums. *Journal of Mathematical Sciences*, 1(3), 179–188. [85, 86]
- Raman, C. V., & Kumar, S. (1920). Musical Drums with Harmonic Overtones. *Nature*, 104. [85]
- Ranjani, H., & Sreenivas, T. (2013, November). Grouping carnatic music notes using a multi-gram language model. In *Proceedings of Acoustics 2013*. New Delhi, India. [91]
- Ranjani, H., & Sreenivas, T. (2015, April). Multi-instrument detection in polyphonic music using Gaussian Mixture based factorial HMM. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 191–195). Brisbane, Australia. [86]
- Ravikiran, C. N. (2008). *Perfecting Carnatic Music, Vol. I-II*. The International Foundation for Carnatic Music, www.ravikiranmusic.com. [67]
- Sachs, C. (1953). *Rhythm and tempo*. W. W. Norton & Co. [19]
- 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. [47, 183]
- Sambamoorthy, P. (1998). *South Indian Music Vol. I-VI*. The Indian Music Publishing House. [24, 26]
- 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 Conference (ISMIR 2013)* (pp. 487–492). Curitiba, Brazil. [87, 95]
- 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. [85]
- Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition* (Doctoral dissertation). Stanford University. [47]
- 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. [183]
- 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. [3, 118]
- 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. [5, 14, 15, 97, 116, 118, 286]
- Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jordà, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., & Widmer, G. (2013). *Roadmap for Music Information ReSearch*. Retrieved from http://www.mires.cc/sites/default/files/MIRES_Roadmap_ver_1.0.0.pdf [5]
- Shankar, V. (1999). *The art and science of carnatic music*. Parampara. [19]
- 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, pp. 494–499). Springer Berlin Heidelberg. [57, 245]
- 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*. [57]
- Smith, R. (2007, September). An Overview of the Tesseract OCR Engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 2, pp. 629–633). Washington, DC, USA. [167]
- 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. [15, 16, 172, 174, 251]
- 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)* (pp. 344–356). London, UK. [91]
- 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. [67, 129]
- 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. [15, 191, 192, 226, 239, 306]
- 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)* (pp. 76–80). Shanghai, China. [15, 147, 191, 192, 208, 213, 226, 306]
- 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. [3, 7, 14, 79, 99, 175]
- 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*. Athens, Greece. [15, 118]
- 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. [xxvi, 15, 82, 133, 182, 282]
- 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. [77, 102, 306]
- Swartz, A. (2002, January). MusicBrainz: a semantic Web service. *IEEE Intelligent Systems*, 17(1), 76–77. [88]
- T. K. Govinda Rao. (2003a). *Compositions of Muddusvāmi Dīk-shitar*. Ganamandir Publications. [122]
- T. K. Govinda Rao. (2003b). *Compositions of Śyāmā Śāstri, Subbarāya Śāstri and Aññasvāmi Śāstri*. Ganamandir Publications. [122]
- T. K. Govinda Rao. (2009). *Compositions of Tyāgarāja*. Ganamandir Publications. [122]
- Thompson, L., Dixon, S., & Mauch, M. (2014, October). Drum Transcription via Classification of Bar-Level Rhythmic Patterns. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (pp. 187–192). Taipei, Taiwan. [58]
- Thoshkahna, B., & Ramakrishnan, K. R. (2011, November). 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)* (pp. 251–256). Miami, USA. [46]

- 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. [15, 40, 82, 171, 172, 245]
- Toussaint, G. T. (2004, October). A comparison of rhythmic similarity measures. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*. Barcelona, Spain. [54]
- 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. [60]
- Tzanetakis, G., & Cook, P. (2002, July). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 293–302. [50]
- 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. [50]
- 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)* (pp. 136–140). Brisbane, Australia. [87]
- 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. [202]
- 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. [87]
- 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 Communications*. Guwahati, India. [87]
- Viswanathan, T., & Allen, M. H. (2004). *Music in South India*.

- Oxford University Press. [23]
- Vos, J., & Rasch, R. (1981). The perceptual onset of musical tones. *Perception & Psychophysics*, 29(4), 323–335. [45]
- 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 2006)* (pp. 29–34). Victoria, Canada. [191, 194]
- 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. [191]
- Wichmann, E. (1991). *Listening to Theatre: The Aural Dimension of Beijing Opera*. University of Hawaii Press, Honolulu. [39]
- Widdess, R. (1994). Involving the Performers in Transcription and Analysis: A Collaborative Approach to Dhrupad. *Ethnomusicology*, 38(1), 59–79. [37]
- Wilpon, J. G., Rabiner, L. R., Lee, C.-H., & Goldman, E. R. (1990). Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11), 1870–1878. [97]
- 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. [48, 185]
- Wynne, M. (Ed.). (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. [116]
- Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. C. (2006). *The HTK book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department. [253, 263]
- Zapata, J. R., & 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 Process-*

- ing (*ICASSP 2013*) (pp. 51–55). Vancouver, Canada. [50, 183]
- 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). [50, 189]

Index

- Additive meter, 21, 66
Bar pointer model, 192
Bayesian model, 10, 58, 191
Beat tracking, 48, 80
Beijing opera, 38, 171
Carnatic music, 22, 120
CompMusic, 3
Downbeat tracking, 51
Eurogenetic music, 3
Fundamental frequency, 46, 71
Graphical model, 58
Heterophony, 71
Hindustani music, 22, 128
Idiophone, 39, 57
Isochronicity, 20, 24, 66
Language model, 253, 263
Membranophone, 57
Meter, 20
Meter analysis, 79, 177
Meter inference, 79, 178
Meter tracking, 52, 79, 179
Novelty function, 45, 47, 183
Onomatopoeia, 28, 32, 37, 57
Onset, 43
Onset detection, 43, 82, 184
Onset patterns, 54
Ontology, 88
Oral mnemonic syllable, 28, 32, 39, 68, 84, 247
Particle filter, 60, 200
Perceptual present, 66, 208
Research corpus, 115
Rhythm similarity, 7, 54, 88
Section pointer model, 208, 229
Segmentation, 86
Similarity matrix, 77, 186
Solmization, 84
Spectral flux, 45, 140, 184, 197
Speech recognition, 60
Syllabic percussion, 12, 38, 69, 96, 245, 257
Tactus, 20

- Tatum, 20
Tempo tracking, 47, 80, 184
Tempogram, 47, 56, 182
Test dataset, 115
Tonic note, 27, 168, 170
Tāla recognition, 75, 222
Viterbi decoding, 200, 251, 263
Vocal percussion, 57