A Corpus-based Approach to the Computational Modeling of Melody in Raga Music

A thesis submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

by

Kaustuv Kanti Ganguli
(Roll No. 134076013)

Under the guidance of
Prof. Preeti Rao

Department of Electrical Engineering
Indian Institute of Technology Bombay, Mumbai, India.

May 2019

To my grandparents ...

Thesis Approval

The thesis entitled

A Corpus-based Approach to the Computational Modeling of Melody in Raga Music

by

Kaustuv Kanti Ganguli
(Roll No. 134076013)

is approved for the degree of

Doctor of Philosophy

Examiner

Examiner

Guide

Chairman

Date: 20·12·2019

Place: Mumbai

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY, INDIA

CERTIFICATE OF COURSE WORK

This is to certify that **Kaustuv Kanti Ganguli** (Roll No. 134076013) was admitted to the candidacy of Ph.D. degree on July 2013, after successfully completing all the courses required for the Ph.D. programme. The details of the course work done are given below.

S.No	Course Code	Course Name	Credits
1	EE 635	Applied Linear Algebra	6
2	EE 603	Digital Signal Processing and its Applications	6
3	EE 679	Speech Processing	6
4	EE 601	Statistical Signal Analysis	6
5	EE 610	Image Processing	6
6	EE 779	Advanced Topics in Signal Processing	6
7	EE 712	Embedded System Design	6
8	EE 678	Wavelets	6
9	EE 692	R & D Project	6
10	EES 801	Seminar	4
11	HS 699	Communication and Presentation Skills	PP
12	CS 725	Foundations of Machine Learning	AU
		Total Credits	**58**

IIT Bombay

Date: Dy. Registrar (Academic)

Abstract

Indian art music is predominantly an oral tradition with pedagogy involving the oral transmission of raga lessons. There exist text resources for musicology, wherein the lexicon finds its place in a rather prescriptive manner. Raga performance allows for considerable flexibility in interpretation of the raga grammar in order to incorporate elements of creativity via improvisation. It is therefore of interest to understand how the musical concepts are manifested in performance, and how the artiste improvises i.e. uses the stock musicological knowledge in "new" ways, while carefully maintaining the distinctiveness of each raga in the ears of trained listeners. Alongside addressing the issue of subjectivity, scalability, and reproducibility in musicological research, this work proposes novel methods for relevant music information retrieval (MIR) applications like rich transcription, melody segmentation, motif spotting, raga recognition. While a general engineering approach is to optimize certain evaluation metrics, we aimed to ensure that our findings are informed by musicological knowledge and human judgment. To achieve this, our approach is two-fold: computational modeling of the melody on a sizable, representative corpus; then validating the models through behavioral experiments towards understanding the learned schema by trained musicians.

We propose computational representations that robustly capture the particular melodic features of the raga under study while being sensitive enough to the differences between ragas tested within a sizable, representative music corpus. To make a good foundation for tuning of hyper-parameters, we exploit the notion of "allied ragas" that use the same tonal material but differ in their order, hierarchy, and phraseology. Our results show that computational representations of distributional and structural information in the melody, combined with suitable distance measures give insights about how the aspect of raga distinctiveness is manifested in practice over different time scales by creative performers. Finally, motivated by the parallels between musical structure and prosodic structure in speech, we present listening experiments that explore musicians' perception of ecologically valid synthesized variants of a raga-characteristic phrase. Our findings suggest that trained musicians clearly demonstrate elements of categorical perception in the context of the technical boundary of a raga. The last leg of the thesis discusses the application areas, namely mainstream MIR tasks as well as computational musicology paradigm. In sum, we aim to impart music knowledge into a data-driven computational model towards modeling human-judgment of melodic similarity. This cognitively-based model can be useful in music pedagogy, as a compositional aid, or building retrieval tools for music exploration and recommendation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The broad perspective

The melodic form in Indian art music (IAM) is governed by the system of ragas. A raga can be viewed as falling somewhere between a scale and a tune in terms of its defining grammar which specifies the tonal material, tonal hierarchy, and characteristic melodic phrases [141, 153]. Rao et al. [152] remarks that raga, although referred to as a concept, really escapes such categories as concept, type, model, pattern etc. A raga is brought out through certain phrases that are linked to each other and in which the svaras have their proper relative duration. This does not mean that the duration, the recurrence and the order of svaras are fixed in a raga; they are fixed only within a context [191]. The svaras form a scale, which may be different in ascending and descending phrases, while every svara has a limited possibility of duration depending on the phrase in which it occurs. Furthermore, the local order in which the svaras are used is rather fixed. The totality of these musical characteristics can best be laid down in a set of signature phrases which is a gestalt that is immediately recognizable to the expert. The rules, which constitute prototypical stock knowledge also used in pedagogy, are said to contribute to the specific aesthetic personality of the raga. In the improvisational tradition, performance practice is marked by flexibility and creativity that coexist with the strict adherence to the rules of the chosen raga. The empirical analyses of raga performances by eminent artists can lead to insights on how these apparently divergent requirements are met in practice.

Music information retrieval (MIR) is an interdisciplinary science of retrieving information from music. MIR is a small but growing field of research with many real-world applications, and also mentions that MIR researchers should have a background in paradigms such as musicology,

psychology, academic music study, signal processing, machine learning, or some combination of these. Few of the potential applications are: music recommendation, structural segmentation, instrument recognition, rich transcription, query-by-humming, automatic music categorization, music generation, and so on. The scope of audio signal processing, which is our domain of expertise, comes into play in the 'feature extraction and representation' stage, though statistics and machine learning are integral components of this domain. Another interesting application is the human-computer interaction and music interfaces. But music being a consumable entity, it is an important issue to address what 'music information' indeed interests the community. One of the popular practice of music research is in the domain of musicology, which by definition, is a scholarly analysis of music as a part of digital humanities. The broad topics that interest ethnomusicologists across cultures and repertoires are: music theory, music performance practice and biographies (artistic research), history of music, music education and pedagogy, philosophy (particularly aesthetics and semiotics). Another key segment of researchers are in the domain of acoustics and psychoacoustics that looks into the topics like music acoustics, the science and technology of acoustical musical instruments, and the musical implications of physiology (music therapy), psychology; generally referred to as systematic musicology.

Having seen the applied fields of musicology, let us motivate few application areas that are interesting but beyond the scope of traditional musicology research. This lays the foundation for the significance of technological research: (i) audio mining: exploring music archives, libraries, and digital collections for a better interactive navigation; (ii) audio fingerprinting: intellectual property (IP) rights, copyright issues, identification and traceability; (iii) sociology and economy of music: distribution, consumption chain, user profiling, validation, user needs and expectations, evaluation of MIR systems, building test collections, experimental design and metrics etc. We choose to explore the acoustic research paradigm, this can be reframed as a study of computational musicology. The thesis centres around the theme of 'music similarity'. To briefly introduce the idea, music similarity is one of the widely-addressed problems in MIR research. This covers a combinatorial study of several different musical dimensions at varying time-scales. These musical dimensions include timbral texture, melody, rhythmicity, lyrics etc., whereas the similarity can be computed at song-level, phrase-level, or even note-level. In the Western music repertoire, much work have been carried out in 'music similarity' research involving lyric, instrumentation, score, rhythm and so on. In the current work, we focus on 'melodic' similarity on the time-scale of musical 'phrases'. Our broad goal is to develop

computational models that take two music signals as input, to obtain a quantitative measure of (dis)similarity between the signals that ideally correlates with human judgment. The pedagogy in IAM involves acquiring a knowledge of raga motifs and the ability to recall and render the phrases in the context of a performance. Further, musical meaning is associated with each raga that may be considered to depend on the listener's acquired knowledge of the musical idiom [12, 208]. The perception of similarity in music, on the other hand, is central to MIR, where pieces of music must be categorized using different musical features in order to facilitate search. Research on musical features and the associated (dis)similarity measures has therefore remained pertinent lately. The findings from studies on Western folk-song tune families closely match the notion of characteristic motifs of ragas in Indian art music [141, 202].

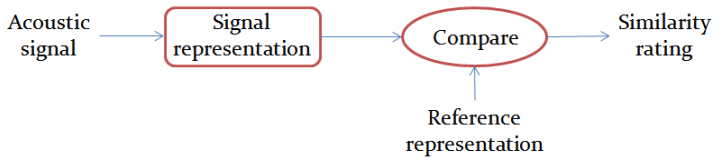

Figure 1.1: Block diagram of the computational model for melodic similarity.

Figure 1.1 shows the block diagram of our approach where the two main computational blocks are: (i) representation derived from an acoustic signal, and (ii) distance measure of the two given signals (one input vs. one reference signal). The final output of the system is the similarity rating in terms of a score. Ideally both these aspects (representation-cum-similarity) must be informed by human perception knowledge.

There are two traditional, distinctive approaches to MIR research: (i) data-driven approach, and (ii) knowledge-driven approach. While the former stresses on big-data and statistical modeling, the latter deals with knowledge representation and retrieval strategies informed by musicological rules and possibly human judgments. Both approaches are quite well-accepted in their own realm, both have their own pros and cons; ideally a combination (where the two complement each other) might be something worth exploring. However the trade-off often is very clear, in fact the thesis portrays a case where we would motivate how big-data analysis could help. In contrast, another subproblem needs in-depth analysis of a small representative corpus to establish a hypothesis. Having portrayed these diverse perspectives, now the crucial question to ask is, whether the standard approaches in MIR have any correlation with human judgments. But what does 'human judgment' refer to, being a loaded term by itself? We may assume that the term refers to epistemology, i.e. how humans 'think'. The branch that studies such

phenomenon in music, also referred to as cognitive musicology, deals with music perception, cognition, affect, and emotions. Cognitive musicology is a branch of cognitive science concerned with computationally modeling musical knowledge with the goal of understanding both music and cognition [2]. This is a branch of music psychology that emphasises on computer modeling to study music-related knowledge representation with roots in artificial intelligence (AI). This field seeks to model how musical knowledge is represented, stored, perceived, performed, and generated. Thus we raise the question: what is the form of human memory for the musical attributes? We shall briefly address this issue, but before that it is worth reviewing a comparison of the man vs. machine understanding.

The principle philosophy of machine learning or AI is that the computational models should be designed such that it is consistent with human perception and cognition. But one major difference is that feature modeling is a bottom-up procedure, whereas human perception is top-down[1]. Thus, to closely model human cognition, an interdisciplinary research is a must, but this often is beyond the scope of a single-handed research endeavor. Through this thesis we would motivate the fact that an informed design strategy in a specific realm, with certain reasonable assumptions, could make it plausible to propose a cognitively-based computational model. At the same time, it is a difficult problem, by itself, to estimate the human judgment, i.e. to obtain an objective measure out of a subjective experiment. Researchers have proposed interesting solutions through psychoacoustic/cognitive experiments and behavioral studies, though it is extremely difficult to generalize [2]. Precise experimental methods in a very controlled framework may lead to a small yet significant insight.

We propose computational models towards understanding how the musical concepts are manifested in performance, and how the artiste improvises. An important function of this model would be to capture the notion of grammaticality in performance. Such an exercise could eventually lead to computational tools for assessing raga performance accuracy in pedagogy together with the complementary aspect of creative skill. A popular notion of grammaticality in performance is preserving a raga's essential distinctiveness in terms of the knowledge-

[1]It would as well be interesting to interpret the Gestalt principles in auditory perception, i.e. to know what is the manifestation of symmetry, proximity, closure, continuity in human audition. If music perception is truly at a 'global' level, what is the time-scale thereof? This might lead to judging the role of a 'pause' in music.

[2]There also exists is a 'semantic gap' between human abstract perceptual attributes and bottom-up machine accessible close-to-signal features / representations. This is now well acknowledged in most machine learning tasks especially touching on realizing human-perceptual tasks, e.g. multi-media indexing retrieval.

able listener's perception [14, 37, 145]. Thus, a performance with possibly many creative elements is considered not to transgress the raga grammar as long as it does not "tread on another raga" [102, 145, 199]. The technical boundary of a raga should therefore ideally be specified in terms of limits on the defining attributes where it is expected that the limit depends on the proximity of other ragas with respect to the selected attribute. We consider deriving a computational representation of distributional and structural information based on maximizing the discrimination of "close" ragas. The notion of "allied ragas", i.e. ragas with identical scales but differing in attributes such as the tonal hierarchy and characteristic phrases [113], is intersting in terms of designing such a case study to investigate discriminating features between "close" ragas. Within the improvisational framework there exists a highly structured lexicon. The performer's mandate is to elaborate upon the raga's essential features without compromising the cues to its identity. Raga grammar in music theory texts comprises only the essential specification of distributional and structural components in terms of tonal hierarchies and typical phrases respectively. A goal of the thesis is to develop a computational representation for distributional and structural information that can eventually be applied in the empirical analyses of audio recordings of performances.

One of the keywords in the thesis title is 'modeling', we need to understand how a mathematical model fits a melodic movement. We therefore should consider only those melodic segments that are relevant to human cognition. This opens up the necessity for exploring in what form a melody is encoded in human memory, and which are the main anchors/cues in a melodic segment that humans (musicians and listeners)[3] assign more weightage to. In the light of experiments on speech production and perception (also prosody), where the motor theory states that human production system is dependent on perception and vice-versa; some related questions thus become very pertinent. Short-term vs. long-term memory (and also working memory) in music training and performance, time-scale of psychoacoustic relevance of musical events etc. are few examples of them. We shall briefly address these aspects in course of discussion.

Last but not the least, we mention few of the potential application-fields of our work. The dissemination of musical material in the pedagogical scenario of IAM involves oral and aural

[3]Do listeners and musicians perceive the same melody in a similar way? What is it that makes musicians a 'musician'? Is it only that musicians are capable of vocally producing music? Or is their perception, in a way, advanced? These are some open-ended questions worthy of thinking.

mode of transmission. Though musicological texts describe musical ideas (e.g. thaat, raga, tala, pakad and so on, there is a lack of documentation of the nuances in an actual raga performance. We carry out melodic analyses on concert audios sung by eminent artists. This would facilitate music listeners to efficiently navigate through a concert recording in an interactive manner, with a visually enhanced experience of appreciating the pitch movements, percussion beats etc. Also, a music learner can be evaluated against an artiste's (of his/her choice) rendition of a melodic phrase, with visual feedback of the (dis)similarity in the melodic shapes. Another key area is content-based retrieval systems like robust music 'search enablers' and recommendation systems. While existing top-ranked search engines employ tag-based recommendation system, we would like to exploit melodic features for rank-ordering 'similarity' across ragas. This might lead to finding certain hidden relationships between ragas (e.g. murchhana or key-transposed melody) which is otherwise not so explicit. A bigger aim would be to achieve an 'audio summarization' system wherein raga-characteristic phrases from an hour-long concert could be discovered and stitched together in order to auto-generate a 'trailer' of the concert. In sum, we aim to impart music knowledge into a data-driven computational model towards modeling human-judgment of melodic similarity. This cognitively-based model can be useful in music pedagogy, as a compositional aid, or building retrieval tools for music exploration and recommendation.

1.2 Organization of the thesis

Chapters 2 through 4 lay the foundation of the thesis via relevant literature survey of the different disciplines, namely the musical and computational backgrounds. We choose melody contour modeling as the major task for the thesis. Raga recognition has been the most widely addressed MIR problem in IAM, majority of the approaches used first-order pitch distributions as the discriminating feature. We consider the same representation (and a few variations thereof), limiting ourselves to a slightly constrained problem with allied raga-pairs. With unavailability of datasets that represent (un)grammatical performances, we chose concerts by eminent musicians as a proxy for the grammatical performance for the labeled raga, and the same as an ungrammatical performance for its allied raga. Chapters 5 and 6 discuss the distributional and structural attributes of melodic representation-cum-distance measures towards modeling ungrammaticality in raga performances via the task of allied raga-pair discrimination.

Finally, we want to ensure our computational models conform with human perception and cognition. If improvisation is, to certain extent, predictable by musicians and knowledgeable listeners, it is interesting how they use: (i) schematic expectations based on long-term memory of the raga, and (ii) dynamic expectations based on short-term memory of the local context. One way to address this problem is to understand: in what form in musicians' long-term memory the prototypical melodic shapes are stored, and to what surface resolution musicians attend small deviations from the prototype. In Chapter 7, we present a novel approach to stylize the raw melodic contour and incorporate controlled variations to generate synthetic stimuli to be used in a behavioral experiment. Chapters 8 and 9 are compilations of a few of the retrieval tasks attempted – in a query-by-humming task, a proof-of-concept framework to explore the structure of a raga performance, and also a deliverable interface for real-time raga recognition running on a web browser that corroborates the ideas presented in the thesis. Chapter 10 summarises the contributions, with a focus on the achieved insights. Following is a set of appendices that describe certain demographic/methodological details which is necessary for completeness of the thesis, but would create digression in the main discourse.

Chapter 2

Literature on melodic similarity in MIR and computational musicology

2.1 Motivation

Computational models for melodic similarity at the level of a musical phrase require the definition of a representation and a corresponding distance measure. In the case of Western music, the representation draws upon the well-established written score where pitch intervals and note timing are unambiguously captured. The distance measure is then typically cast as a string matching problem where musically informed costs are associated with the string edit distance formulation [123, 194]. Research on melodic similarity measures, of course, also extends to more cognitive modeling based approaches using higher-level features extracted from melodic contours rather than simply surface features [142]. From the papers of Mullensiefen, we learn: (i) importance of using tested musically trained subjects, (ii) designing of stimuli, (iii) specifying the task and rating scale, and (iv) drawing conclusions about the predicting power of various representation-cum-similarity measures. The author has reported his works on a similar philosophy in couple of other papers. These are: "Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments" [121], and "Modelling experts' notions of melodic similarity [124]. Another summary paper "High-level feature descriptors and corpus-based musicology: Techniques for modelling music cognition" [126] discusses about application of the aforementioned philosophy on a large dataset of Western folk- and pop-songs. The author compares a number of similarity algorithms and judges the retrieval performances on different melody representations. It is remarked that the ones which correlates with human

perception are better in MIR applications as well. Russo et al. [195] designied an empirically derived measure (EDM) based on multiple regression with five predictor features, from a behavioral experiment of human judgment. Finally the authors compare their proposed similarity measure with the state-of-the-art string edit-distance Mongeau-Sankoff measure (MSM) and show the superiority of the proposed approach. The authors point to two main applications of melodic similarity based algorithms: content-based music search (cover song detection, music plagiarism detection etc.), and music composition tool that takes user-preference (in terms of genre, rhythm etc.).

Non-western and folk musics do not lend themselves well to the Western score based transcription [29, 198, 202]. Having originated and evolved as oral traditions, they typically do not have a well-developed symbolic representation thus posing challenges for pedagogy and also for computational tasks such as music retrieval. Relevant past work on the computational analyses of non-Western music includes the representation of flamenco singing with its highly melismatic form with smooth transitions between notes where onsets are not clearly specified [29]. Volk et al. [202] express their concern that MIR work on similarity typically faces the problem of ground-truth i.e. what musical pieces can be considered to be similar. Authors used musicological experts to annotate their pieces. A symbolic transcription based on sequences of several short notes was fitted to the continuous time-varying pitch contour using dynamic programming based optimization using probability functions for pitch, energy, and duration. In the case of Indian art music (IAM) forms, melodic phrases are identified not only by the stable notes at precise pitch intervals but also by the shapes of the continuous transient pitch segments connecting them. The continuous pitch curves have certain dynamics that are not accurately represented by a succession of short notes. Some musicological literature proposes a set of 15 basic melodic shapes (alankar) [14]. However there is no general acceptance of this in contemporary Hindustani classical music nor has there been any computational research that exploits this.

Several qualitative musicological works bring out new musical insights but are prone to criticism of not having supported their findings using a sizable corpus. Contrary to that, quantitative computational studies manage to scale to sizable data sets, but fall short of discovering novel musical insights. In the majority of cases, computational studies attempt to automate a task that is well known and is fairly easy for a musician to perform. There have been some studies that try to combine these two types of methodologies of working and corroborate several

concepts in musical theories using computational approaches. In Chinese opera music, [156] performed a comparison of the singing styles of two Jingju schools where the author exploits the potential of MIR techniques for supporting and enhancing musicological descriptions. Autrim[1] (Automated Transcription for Indian Music) has used MIR tools for visualization of Hindustani vocal concerts that created a great impact on music appreciation and pedagogy in IAM. There are also several studies that perform computational modeling of different melodic aspects in IAM. These studies address computational research tasks such as raga recognition [36, 81], melodic similarity [59, 83, 151], discovery and search of melodic patterns [68, 82], segmentation of a musical piece [197] and identification of specific landmarks in melodies [80]. These approaches typically employ signal processing and machine learning methodologies to computationally model the relevant melodic aspect. These studies can provide a ground for developing tools and technologies needed to navigate and organize sizable audio collections of music, perform raga-based search and retrieval from large audio archives and in several other pedagogical applications.

In the next sections, we shall discuss the relevant literature on both Western and non-Western repertoires with respect to the distributional and structural attributes of melodic similarity. Towards the end of this chapter, we briefly introduce the cognitive perspective and some methodological review of related literature. The detailed discussion of experimental paradigm-specific works, however, are described in the corresponsing chapter.

2.2 The distributional view

2.2.1 Introduction

In Western music, the psychological perception of "key" has been linked to distributional and structural cues present in the music [90]. Krumhansl [99] derived key profiles from tonality perceived by listeners in key-establishing musical contexts thereby demonstrating that listeners are sensitive to distributional information in music [173]. The distribution of pitch classes in terms of either duration or frequency of occurrence in scores of Western music compositions has been found to correspond with the tonal hierarchies in different keys [146, 173, 181]. Further, the task of automatic key detection from audio has been achieved by matching pitch chroma (the

[1] https://autrimncpa.wordpress.com/

octave-independent relative strengths of the 12 pitch classes) computed from the audio with template key profiles [54, 75, 139]. As in other tonal music systems, pitch distributions have been used in characterizing raga melodies. In contrast to the discrete 12-tone pitch intervals of Western music, raga music is marked by pitch varying continuously over a range as demonstrated via a recent raga music transcription visualization [13]. Recognizing the importance of distributional information, we find that the pitch-continuous nature of Indian classical traditions gives rise to several distinct possibilities in the choice of the computational parameters of a histogram representation, such as bin width. While the raga grammar as specified in music theory texts is not precise enough to resolve the choices, suitable questions posed around observed performance practice can possibly facilitate a data-driven solution.

2.2.2 The theory of tonal hierarchy in music

Tonal hierarchy, as discussed by Krumhansl [100] refers to both a fundamental theoretical concept in describing musical structure and a well-studied empirical phenomenon. As a theoretical concept, the essential idea is that a musical context establishes a hierarchy of tones. Author [99] proposed a key-finding algorithm that is based on a set of "key-profiles" first proposed by her earlier work [101], representing the stability or compatibility of each pitch-class relative to each key. The key-profiles are based on experiments in which participants were played a key-establishing musical context such as a cadence or scale, followed by a probe-tone, and were asked to judge how well the probe-tone "fits" given the context (on a scale of 1 to 7, with higher ratings representing better fitness). Given these key-profiles, the algorithm judges the key of a music piece by generating a 12-element input vector corresponding to the total duration of each pitch-class in the piece. The correlation is then calculated between each key-profile vector and the input vector; the key whose profile yields the highest correlation value is the preferred key. In other words, the listener's sense of the fit between a pitch-class and a key is assumed to be highly correlated with the frequency and duration of that pitch-class in pieces in that key [100].

Other studies also found that the tonal hierarchy measured in empirical research mirrors the emphasis given to the tones in compositions by way of frequency of occurrence and duration [146, 174, 182]. This relationship between subjective and objective properties of music provides a strong musical foundation for the psychological construct of the tonal hierarchy. Castellano et al. [30] used contexts from 10 North Indian ragas. As reviewed earlier, raga music theory describes a hierarchy of the tones in terms of importance. In a probe tone study, ratings of

the 12 tones of the Indian scale largely confirmed these predictions suggesting that distributions of tones can convey the tonal hierarchy to listeners unfamiliar with the style. Both American and Indian groups of listeners gave the highest ratings to the tonic and the fifth degree of the scale. These tones are considered by Indian music theorists to be structurally significant, as they are immovable tones around which the scale system is constructed, and they are sounded continuously in the drone. Relatively high ratings were also given to the *vadi* svara, which is designated for each raga and is considered the dominant note. The ratings of both groups of listeners generally reflected the pattern of tone durations in the musical contexts. This result suggests that the distribution of tones in music is a psychologically effective means of conveying the tonal hierarchy to listeners whether or not they are familiar with the musical tradition. Beyond this, only the Indian listeners were sensitive to the scales (lit. *thaat*) underlying the ragas. For Indian listeners, multidimensional scaling of the correlations between the rating profiles recovered the theoretical representation of scale (in terms of the tonal material) described by theorists of Indian music. Thus, the empirically measured tonal hierarchy induced by the raga contexts generates structure at the level of the underlying scales, but its internalization apparently requires more extensive experience with music based on that scale system than that provided by the experimental context [30].

In a similar study, Raman et al. [146] used four Carnatic (south Indian classical music repertoire) ragas to come up with a baseline template and then employed two techniques of modulation, namely *grahabedham* (changing the tonic while keeping the tones unchanged) and *ragamalika* (shifting to any distinctly different raga). By tracking responses of both Western and Indian listeners to the modulations in Carnatic music, they identified the various cues that listeners utilized in order to discern the modulation. The authors reported that the Westerners' responses matched those of the Indians on ragas with structures similar to Western scales, but differed when ragas were less familiar, and surprisingly, they registered the shifts more strongly than Indian musicians. These findings, according to the authors, converged with previous research in identifying three types of cues: (i) culture-specific cues – schematic and veridical knowledge – employed by Indians, (ii) tone-distribution cues – duration and frequency of note occurrence – employed by both Indians and Westerners, and (iii) transference of schematic knowledge of Western music by Western participants.

There has been no notable work on analyzing tonal hierarchy from symbolic notation in raga music. This could be due to the availability of scores, if at all, only for very well-known

compositions. Performances are dominated by improvisation and rarely transcribed. A recent work [159] analyzed the scores of 3000 compositions across 144 ragas to find that the pitch class distribution served well to cluster ragas with similar scales together.

2.2.3 First-order pitch distributions in audio MIR

The methods mentioned above for key estimation worked with the symbolic representation of music. A large number of the approaches for key estimation in audio recordings of Western music are essentially motivated by the same methodology. However, the task of key estimation from audio recordings becomes much more challenging due to the difficulty of extraction of a reliable melody representation, such as a music score, from polyphonic music recordings [75]. Based on pitch chroma features computed from the audio signal spectrum, a 12-element pitch-class profile (PCP) vector is estimated [75, 139]. Next the correlation of this estimated tonal profile is implemented with all possible theoretical key-profiles derived in [101]. The key-profile that results in the maximum correlation is marked as the key of the music piece.

Another music tradition where the concept of tonality has been studied with reference to pitch distributions is Turkish Makam music. Due to the presence of the micro-tonalities and continuous melodic movements between the notes, a fine grained pitch distribution is considered as a feature for modeling tonality. Bozkurt et al. [25, 70] use a pitch distribution with bin-width of $\frac{1}{3}$ Holdrian comma (approximately 7.5 cents). This results in a 159-dimensional pitch-class distribution (PCD) vector that performs significantly better in a Makam recognition task compared to a 12, 24, or 36-dimensional PCP vector often used for tonal analysis of Western music.

The computational modeling of the distinctive attributes of a raga has been the subject of previous research motivated by the task of raga recognition from audio [18, 35, 36, 43, 96]. The tonal material has been represented by a variety of first order pitch distributions as depicted in Figure 2.1. Experimental outcomes based on recognition performance have been used to comment on the relative superiority of a given representation as a feature vector in either a template-based or trained model-based classification context. Motivated by the pitch-continuous nature of the melody, histograms of different bin widths computed from octave-folded instantaneous pitch values have been used as templates in raga recognition tasks. The taxonomy, presented in Figure 2.1, summarizes the distinct first-order pitch distribution representations proposed in the raga recognition literature. The top-level classification is based on whether the continuous

melodic contour is used as such, or segmented and quantized prior to histogram computation.

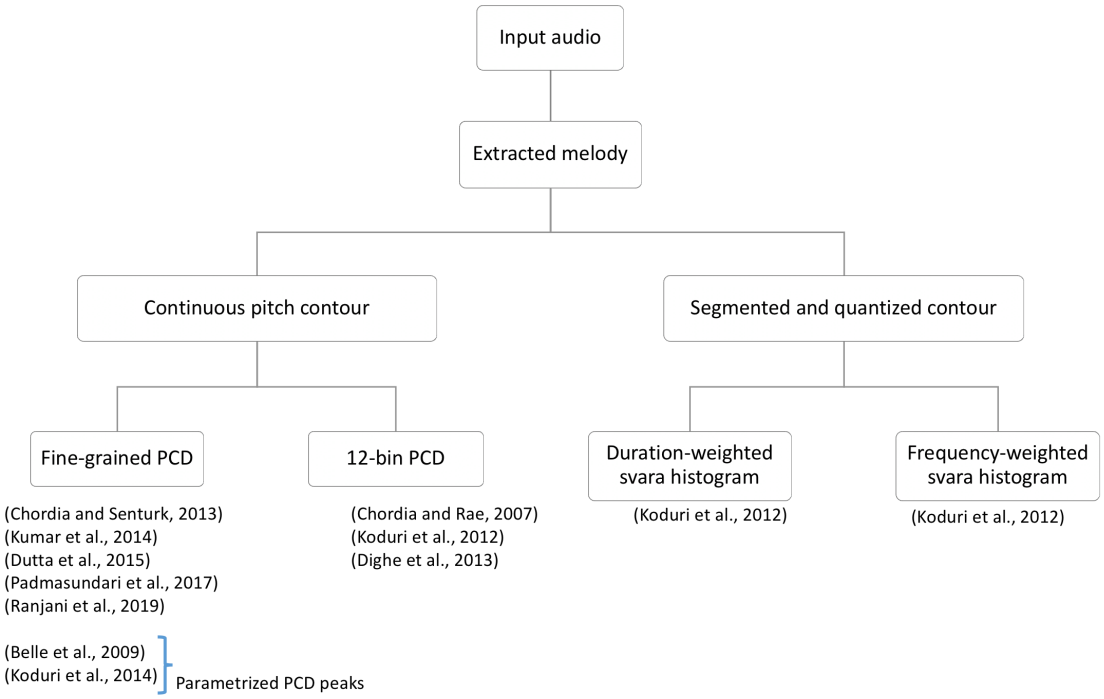

Figure 2.1: Taxonomy of the previous endeavors in raga recognition from first order pitch distributions.

As we have seen earlier, along with the set of svaras, raga grammar defines the functional roles of these svaras in terms of their saliences. A formal definition of the salience of a svara in a melody does not exist, and therefore, several methods have been proposed to quantify it. Chordia [35] and Dighe [43] represent svara saliences using a 12-bin PCD computed as a histogram of the tonic-normalized and octave-folded pitch sequence. The pitch is detected at uniform intervals (audio frames) across the recording to obtain a time series representing the melodic contour. The salience of a bin in the histogram is therefore related to the total duration of the (octave-folded) pitch in the melody. This global feature is robust to pitch octave errors and is shown to perform well on a sizable dataset. A simple extension to the 12-bin PCD feature mentioned above is to compute the pitch distribution using fine grained bins, e.g. at 1 cent resolution. Such a fine grained PCD is used widely [18, 36, 96, 104]. These studies report a superior performance in recognizing ragas by using the high resolution PCD as compared to a 12-bin PCD. Belle [18] and Koduri [97] proposed a parametrized version of the PCD, wherein the parametrization is performed for the pitch distribution shape across each svara region. Both works exploit the distinctions between ragas in the intonation of shared svaras via peak position, amplitude and shape in the high-resolution PCD.

The above reviewed distributions used the continuous pitch contour (time-series) extracted from the audio. An alternate approach is that of Koduri [95] who computed the distribution from only the stable-pitch regions of the contour. Based on certain heuristic considerations, such regions were segmented out and used to construct the PCD corresponding to the 12 svaras. Two variants of svara salience estimation were implemented. One of their proposed approaches treats the total duration of a svara as its salience similar to the previously mentioned approaches with the continuous pitch contour [e.g. 35]. The other approach considers the frequency of occurrence (i.e. the count of instances) of the svara (irrespective of the duration of any specific instance) as its salience. Thus, three types of first order pitch-class distributions were tested: (i) $P\_continuous$: unconstrained pitch-class histogram with a choice of bin resolutions (1, 2, $\cdots$, 100 cents), (ii) $P\_duration$: constrained to stable notes only, weighted by the durations thereof (12-bin), and (iii) $P\_instance$: constrained to stable notes only, weighted by the count of instances (12-bin). The only distance measure tested was the symmetric KL divergence. Overall, $P\_duration$ performed the best. For $P\_continuous$, there was no noticeable difference across different choices of bin width below 50 cents.

2.3 The structural view

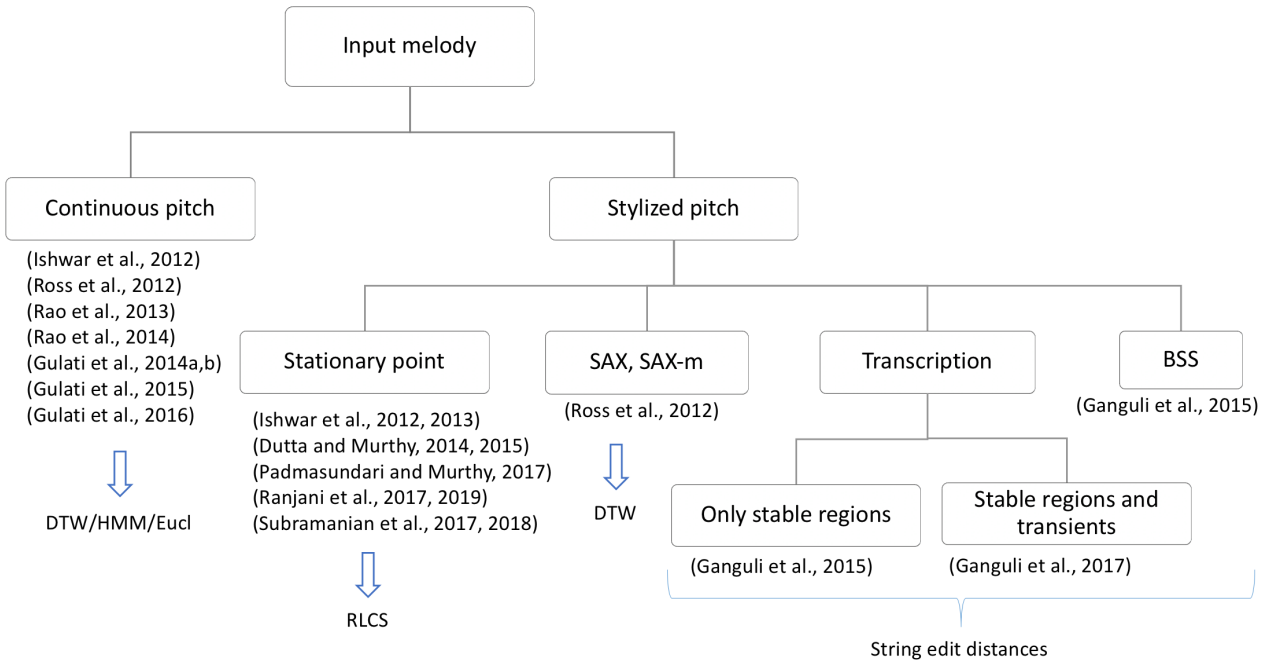

Figure 2.2: Taxonomy of the previous endeavors of phrase level melodic similarity.

Analysis of melodic patterns is a well studied research task in MIR and computational musicology. In Indian art music, however, this task has gained attention only recently. We see that there are two closely related but different pattern processing tasks that these approaches address: (i) Pattern recognition: given a query melodic pattern the objective is to retrieve its other occurrences in the test audio recordings, and (ii) Pattern discovery: given a collection of music recordings the objective is to discover melodic patterns in the absence of any ground truth annotations of the melodic patterns. Figure 2.2, summarizes the distinct strucctural time-series representations proposed in the raga recognition literature.

A majority of the existing methods follow a supervised approach and focus on the task of pattern recognition. Based on the description of these approaches there are two main processing units involved in this task: melody representation and (dis)similarity computation. There is often an interplay between the choices made within these units. In the context of MIR in Flamenco music style, Gomez [75] remarks that pitch and timing variations across occurrences of the melodic patterns can be either handled in the melody representation or during the computation of the melodic similarity. In the former case, a generic or a musically agnostic distance measure might be sufficient for the computation of melodic similarity, whereas, in the latter, a music specific distance measure that can incorporate domain knowledge and can handle timing and pitch variations is required.

2.3.1 Melodic representation

With only a couple of exceptions of our previous work [57, 68, 161], many approaches work with a fine grained continuous melody representation. This is attributed to the characteristics of the melodies in IAM, due to which the extraction of a reliable symbolic or discrete melody representation becomes a challenging task [206]. Moreover, the transitory melodic regions between the svaras in a melody are found to be important in the computation of melodic similarity [39–41, 85], which are lost in a simple transcription of melodies. Our previous works [57, 68, 161] examine the effect of abstracting the melody representation by using techniques such as symbolic aggregate approximation (SAX) [111] and behavioral symbol sequence (BSS) [179]. Another previous work [68] proposed a heuristic-based pseudo melody transcription approach to obtain a discrete melody representation. A more recent work [57] additionally proposed a vector quantization technique to create a codebook of melodic transient shapes and thereby use them in a similar framework as before. It was found that these abstracted representations reduce the

computational cost by a significant factor. However, the accuracy remains inferior compared to a continuous melody representation (considering the best performing distance measure for both the types of melody representations) [68, 161]. Ishwar [19, 91] and Dutta [45, 46] used an abstracted melodic representation that exploits specific melodic characteristics of Carnatic music. Authors consider only the stationary points (where the slope of the melody line becomes zero) of a continuous melody representation. However, such a representation is too coarse to compute a reliable melodic similarity. It is primarily used to prune the search space in order to reduce the computational cost, however, the final computation is done by using a continuous melody representation. Viraraghavan et al. [200, 201] attempts to describe both Hindustani and Carnatic music using CPNs, and STAs. Padmasundari [134] attempts to perform raga identification for 182 Carnatic ragas where motif based locality sensitive hashing approach is used. She clearly shows that pallavi lines in CM, early portions of alapana enable raga identification easily. Dutta's work [47] is relevant for both Hindustani and Carnatic music where attempts are made to find common motifs in ragas by analysis of various compositions/getting typical motifs from musicians. Ranjani's work attempts to transcribe Carnatic and Hindustani music using critical points [147, 148]. However, we notice that a continuous melody representation, which places minimal assumptions on the melodic style, appears to be the most versatile representation.

While there are well studied models for melody segmentation, per se, in symbolic representation of music [26], to the best of our knowledge, segmentation models for IAM in audio recordings are almost inexistent. Some attempts in Carnatic music repertoire include [163, 169]. As a result, the existing approaches either tend to use a brute force segmentation strategy or use a local alignment-based distance measures that do not require an explicit melody segmentation. Ross et al. [160, 161] detect specific rhythmic and melodic landmarks (sam locations and nyas svara onsets) in the audio recordings to determine the location of the potential melodic pattern candidates. However, detecting these landmarks in itself is a challenging task [80, 177]. Thus, these approaches might not generalize and scale to other types of melodic patterns and to large music collections.

2.3.2 Melodic (dis)similarity

A majority of the approaches use a dynamic programming based similarity measure. Our previous works [150, 151, 160, 161] use a similarity measure based on different variants of dynamic time warping (DTW). Ishwar [91] and Dutta [45, 46] use a rough longest common subsequence

(RLCS)-based similarity measure, and more recently our works [57, 68] employ a modified version of the Smith-Waterman algorithm to compute melodic similarity [175]. The dominance of dynamic programming based similarity measures can be attributed to the fact that the melodic patterns in IAM undergo a large degree of non-linear timing variations [58], which can further be attributed to the improvisatory nature of this music tradition. Computing sequence similarity without any temporal alignment, such as in the Euclidean distance, falls short of measuring a meaningful melodic similarity in IAM [161]. Although, a thorough comparison of the Euclidean distance with the dynamic programming based similarity measures for the same melody representation is lacking in the literature. Some of the existing studies also propose enhancements to the well-known distance measures. Dutta [46] also propose to modify the intermediate steps involved in the computation of the RLCS distance to make it more suitable to melodic sequences. These modifications are reported to result in an improvement in the precision of the system, while maintaining the same recall. Rao et al. [151] propose to learn an optimal shape of the global path constrained applied in the DTW based distance measure. However, as reported by the authors, the learned global constraint degraded the performance of the method. Moreover, since the constraint learning is performed for a particular pattern category, such a technique is not applicable to an unseen data, which is the case in the task of pattern discovery. Sridharan's work [163] represents the pallavi using Cent Filterbank Energy slope features to perform matching. Pitch quantisation was not performed so that errors due to pitch are avoided. He has also analysed khayals (wherever repeating patterns are present) of Hindustani music and found the length of the composition; though this model had a dependency on the starting point of the song.

2.4 Literaure on perceptual experiments

As discussed in the introductory chapter, we aim to build a cognitively-based computation model. Survey on perceptual methods are hence presented next. The domain specific literature would be discussed in Chapter 7 for better integrity with the stimuli design aspects and experiments. Here we discuss some of the methodological aspects that are common to any perceptual experiment. An important aspect that we found in [202] is the subtle distinction between judgment of melodic similarity and categorization into tune families. This work is relevant to us in terms of designing the right scenario and rating-task. Authors also argue that there are two

distinct facets of a music cognition research – melodic similarity and categorization into tune-families – to demonstrate a dissociation between categorization and similarity, showing that "categorization can be done in two ways, by similarity and by rule". Thus we speculate that Western musicians would use similarity, and not rule, when presented Hindustani raga phrases, leading to a non-categorical perception which is a common phenomenon in foreign language perception.

There are few distinct approaches in the methodology adopted by different researchers. These are nicely consolidated by Marsden in his recent article [116], the summary is as follows.

- Many studies ask experimental subjects, often experts, to judge the similarity between pairs of melodies or extracts of melodies on a rating scale [48, 49, 121, 124, 165]. This has the advantage of directly generating measures of difference which will almost certainly have the first three of the four properties (non-negativity, self-identity, symmetry, and triangle inequality) of a similarity metric. One objection to experimental procedures like this is that they are not realistic: musicians are rarely (if ever) in a situation when they have to match the similarity between melodies to a number.

- Such direct rating was avoided in another study which also used expert judgment but subjects were asked to rank a set of melodies by their similarity to a reference melody rather than to simply compare pairs of melodies [187, 188]. A measure of difference can be derived from the relative positions of melodies in the rankings, but this measure can only be relative, unlike the potentially absolute measure derived from direct rating of similarity.

- Another paradigm which avoids an artificial direct rating of similarity is to present subjects with three melodies and ask them to indicate the pair which are most alike and the pair which are least alike [10, 132]. This approach is the one which places the least burden on experimental subjects, and it appears to have been successful for non-expert subjects, unlike the paradigms mentioned above. On the other hand, deriving measurements from these observations requires a method such as multi-dimensional scaling, and a large quantity of observations.

- Other studies have avoided direct judgment of similarity, whether by experts or naive listeners. Some have depended on categorization of melodies either from existing musicological studies [124, 203] or on the basis of geographical origin [94].

- Yet other studies have attempted to judge similarity on the basis of some real musical activity. Studies aimed at producing measurements for use in query-by-humming systems have been based on asking subjects to sing a known melody [88, 136].

- Subjects can also be asked to deliberately vary a melody [20], and once again the variations are assumed to be more similar to the original than to other melodies.

We studied methodological constraints and design aspects of psychoacoustic experiments [11, 106, 117, 138]. We also reviewed the approach of modeling melodic improvisations as 'networks of elaborations' [115] and 'cognitive demand' in listeners [114] where the author proposes a scheme of cognitive representation of music. The first work [115] aims to find an underlying model that is 'common' amongst the improvised patterns (ornamented variations) of a 'template' melodic phrase even when the surface features of the music differ. This brings out the concept of 'deep' vs. 'surface' features of musical memory that [32, 202] also point to. In the second work [114], the author claims that the cognitive demand might derive from a combination of truly cognitive and simpler perceptual processing that are the two ways of listening to a recurrent musical material. Therefore the cognitive mode taken by the subject is important for the correct interpretation of the human ratings, e.g. it is important to note whether a musician is using the cognitive mode of a trained musician's or merely a listener's.

Neuro-musicologists take a few principled approaches to study the effect of incongruity in music stimuli via EEG experiments. The aim is typically to capture whether the incongruity is semantic, syntactic, or just as a mismatch of expectancy. We review three recent studies [22, 69, 183, 184] that provide neuro-physiological evidence concerning the mechanism underlying music perception. The authors of these studies also comment on the dependence on familiarity of the stimuli, similarities and differences between musical and linguistic processing for musicians vs. non-musicians. The difference in brain response depending on which cognitive mode a subject engages to, is also addressed. A detailed discussion on the features are presented in Appendix.

Chapter 3

Hindustani music concepts

3.1 Motivation

Indian art music (IAM) is quintessentially an improvisatory music form in which the line between 'fixed' and 'free' is extremely subtle [191]. In a raga performance, the melody is loosely constrained by the chosen composition but otherwise improvised in accordance with the raga grammar. In a typical Hindustani music concert, an artist executes variations of the raga characteristic phrases that represent the raga identity in the course of the process related to improvisation. The characteristic phrases of a raga (pakad) are typically referred to in terms of notation but are fully described via the surface acoustic realization. The artist or performer uses his knowledge of the raga grammar to interpret the notation when it appears in a written composition in the specified raga. The shape of a recurring melodic motif within and across performances of the raga shows variability in terms of one or more of the following aspects: pitch interval, relative note duration and shape of alankar (ornaments), if any, within the phrase [150]. The rules of the raga grammar are manifested at different time scales, at different levels of abstraction, and demand different degrees of conformity. While some of the elements of raga grammar are explicit, others are implicit and can take years of musical training to master. A number of textbooks and scholarly studies exist that describe different improvisatory aspects of melodies in IAM [14, 15, 33, 37, 55, 93, 152, 191]. These works also attempt to uncover some of the implicit aspects of raga grammar. A majority of these are musicological in nature which typically involve either a thorough qualitative analysis of a handful of chosen musical excerpts, or a compilation of expert domain knowledge. In this chapter, we discuss the relevant concepts of Hindustani music repertoire in the light of "improvisation at the time-scale of melodic phrases"

that would facilitate discussion of the interpretation of the computational findings with respect to musicology. We also briefly address "musical expectancy" for framing interesting musical questions that could be approached via computational methods.

3.2 Raga grammar and performance

The melodic form in Indian art music is governed by the system of ragas. A raga can be viewed as falling somewhere between a scale and a tune in terms of its defining grammar which specifies the tonal material, tonal hierarchy, and characteristic melodic phrases [141, 153]. The rules, which constitute prototypical stock knowledge also used in pedagogy, are said to contribute to the specific aesthetic personality of the raga. In the improvisational tradition, performance practice is marked by flexibility and creativity that coexist with the strict adherence to the rules of the chosen raga. The empirical analyses of raga performances by eminent artists can lead to insights on how these apparently divergent requirements are met in practice. Among the relatively few such studies, Widdess [207] presents a detailed analysis of a sitar performance recording from the point of view of a listener to obtain an understanding of how raga characteristics are manifested in a well structured but partly improvised performance. In this work, we consider a computational approach towards using recordings of raga performance to investigate how the tonal hierarchy of a raga, as a prominent aspect of music theory, influences performance practice. A computational model would need to incorporate the essential characteristics of the genre and be sufficiently descriptive to distinguish performances of different ragas.

van der Meer [191] comments that technically a raga is a musical entity in which the intonation of svaras, as well as their relative duration and order, is defined. Many authors [14, 37, 145, 152, 191] refer to the importance of certain svaras in a raga. From the phrase outline we may filter certain svaras which can be used as rest, sonant or predominant; yet the individual function and importance of the svaras should not be stressed [191]. A nyas svara is defined as the resting svara, also referred to as 'pleasant pause' [42] or 'landmark' [80] in a melody. Vadi and samvadi are best understood in relation with melodic elaboration or vistar ('barhat'). Over the course of a barhat, artists make a particular svara 'shine' or 'sound'. There is often a corresponding svara which sustains the main svara and has a perfect fifth relation with it. The subtle inner quality of a raga certainly lies in the duration of each svara in the context of the phraseology of the raga. A vadi, therefore, is a tone that comes to shine, i.e., it becomes attrac-

tive, conspicuous and bright [191]. Another tone in the same raga may become outstanding that provides an answer to the former tone. This second tone is the samvadi and should have a fifth relationship with the first tone. This relationship is of great importance.

There is a prescribed way in which a *khayal* performance develops. The least mixed variety of khayal is that where an alap is sung, followed by a full sthayi (first stanza of the composition) in madhya (medium) or drut (fast) laya (tempo), then layakari (rhythmic improvisation) and finally tan (fast vocal improvisation). When the elaboration reaches the higher (octave) Sa, the antara (second stanza of the composition) is sung. If the composition is not preceded by an alap, the full development of the raga is done through barhat. The composition is based on the general lines of the raga, the development of the raga is again based on the model of the composition [191]. In Table 3.1, we observe both distributional and structural attributes: the tonal material specifies the scale while the *vadi* and *samvadi* refer to the tonal hierarchy, indicating the two most dominant or salient svaras in order. The remaining notes of the scale are termed *anuvadi* (allowed) svara; the omitted notes are termed *vivadi* or *varjit* svara. The typical note sequences or melodic motifs are captured in the comma separated phrases in the *aroha-avaroha* (ascent-descent) and the Characteristic Phrases row of Table 3.1. Finally, the *shruti* indicates the intonation of the specified scale degree in the given raga in terms of whether it is higher or lower than its nominal place (i.e. in Just Intonation tuning). The only other inscription we see is the parentheses, e.g. (R) that indicate that the note R is "weak" (*alpatva*) in Deshkar [152].

Other descriptors that are often available are the type of emotion evoked by the raga, tempo or pace at which it is performed, and the pitch register of major activity. Finally, most texts discussing a raga's features explicitly mention the corresponding allied raga, if any, which has the same scale but supposedly different "treatment" of notes [152]. In music training too, the allied raga pairs are taught concurrently so that learners can infer the technical boundaries of each of the ragas more easily.

3.2.1 Allied ragas

The notion of "allied ragas" is helpful in delineating the concepts for two "close" ragas. These are ragas with identical scales but differing in attributes such as the tonal hierarchy and characteristic phrases [113], as a result of which they may be associated with different aesthetics. For example, the pentatonic ragas Deshkar and Bhupali have the same set of notes (scale degrees or *svaras*): $\mathrm{S}, \mathrm{R}, \mathrm{G}, \mathrm{P}, \mathrm{D}$, corresponding to 0, 200, 400, 700, and 900 cents respectively from

the tonic (see solfege in Figure 3.1). Learners are typically introduced to the two ragas together and warned against confusing them [102, 152, 191]. Thus, a deviation from grammaticality can be induced by any attribute of a raga performance appearing suggestive of an allied, or more generally, different raga [113, 145]. The overall perception of the allied raga itself is conveyed, of course, only when all attributes are rendered according to its own grammar.

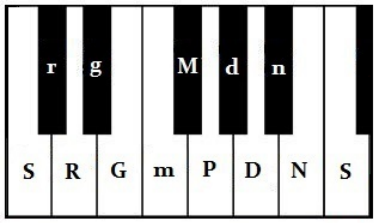

Figure 3.1: The solfege of Hindustani music shown with an arbitrarily chosen tonic (S) location.

Feature	Deshkar	Bhupali	Puriya	Marwa	Multani	Todi
Tonal material	SRGPD	SRGPD	SrGMDN	SrGMDN	SrgMPdN	SrgMPdN
Aroha	SGPD $\overline{\text{S}}$PD$\overline{\text{S}}$	SRG PD$\overline{\text{S}}$	$\underline{\text{N}}$(r)G MN(D)$\overline{\text{S}}$	$\underline{\text{DN}}$r GMD N$\bar{\text{r}}\overline{\text{S}}$	$\underline{\text{N}}$SgMP PN$\overline{\text{S}}$	SrgMP MdN$\overline{\text{S}}$
Avaroha	$\overline{\text{S}}$ PDGP DPG(R)S	$\overline{\text{S}}$DP GDP GRS	NM DG MG(r)S	$\bar{\text{r}}$ND DMGr $\underline{\text{D}}$rS	$\overline{\text{S}}$N(d)P gMg (r)$\underline{\text{N}}$S	$\overline{\text{S}}$NdP MrgrS
Vadi, Samvadi	D, G	G, D	G, N	r, D	S, P	d, g
Char. phrases	SGPD, P(D)$\overline{\text{S}}$P	R$\underline{\text{D}}$S, RPG, PD$\overline{\text{S}}$	$\underline{\text{N}}$(r)S, $\underline{\text{N}}$(r)G, MDG	$\underline{\text{DN}}$r, rGMD, D$\bar{\text{r}}\overline{\text{S}}$	r$\underline{\text{N}}$S, gMP, PN$\overline{\text{S}}$	Srg, PMd, dN$\overline{\text{S}}$
	DGP, DPG(R)S	$\overline{\text{S}}$DP, GDP, GRS	DNM, DG, MG(r)S	$\bar{\text{r}}$ND, DMGr, $\underline{\text{ND}}$rS	NdP, MgM, g(r)S	NdP, Md, rgrS
Shruti	Higher R, G, D	Natural R, G, D	Lower r	Higher r, D	Higher N	Lower r, g

Table 3.1: Specification of raga grammar for three allied raga-pairs [13, 15, 92, 102, 113, 133, 152, 178]. Gaps in *Aroha-Avaroha* indicate temporal breaks (held svara or pause), under/overline are lower/higher octave markers.

Table 3.1 presents a comparison of the melodic attributes corresponding to the grammars of the allied ragas as compiled from musicology texts [13, 92, 102, 113, 133, 152, 178]. It may be noted that the presented information is common across the cited sources. Apart from Deshkar-Bhupali, a well-known allied raga-pair is Puriya-Marwa. In the description of raga Puriya, Rao et al. [152] mention that if either r or D were emphasized in Puriya, it would immediately create an impression of Marwa; therefore strict care should be taken to highlight only G and N. The complementary warning finds its place in the description of raga Marwa. The third allied raga-pair that we include for the study is Multani-Todi where r and d are weak in the former and omitted from ascending phrases.

In performance, a raga is typically presented in the form of a known composition (*bandish*) set in any appropriate metrical cycle (*tala*) within which framework a significant extent of elab-

oration or improvisation takes place [152, 206]. The performer is free to choose the tonic that is played on the drone in the background throughout the concert. A concert begins with slow elaboration of the chosen raga's characteristics in an unmetered section called the *alap*. This is followed by the *vistar* which comprises the chosen composition and elements of improvisation. The progress of the concert is marked by the gradual increase of tempo with the duration of the rhythmic cycles decreasing and slightly more fast paced and ornamented rendering of the melodic phrases [191, 208]. While the raga grammar specifies a characteristic melodic phrase by a sequence of svara, as in Table 3.1, the pattern serves only as a mnemonic to the trained musician to recall the associated continuous melodic shape or motif. Overall, we note that the raga grammar specification, while skeletal, comprises clear distributional (tonal hierarchy) and structural (melodic phrases) attributes. In this work, we focus on a computational representation of the former in a manner suited to the empirical analysis of raga performances. We observe that there are variations in phrase shape across instances and concerts in a given raga. We wish to examine whether: (i) there are context based explanations for these variations; (ii) whether the distinctiveness of the raga is maintained (in terms of discriminating it from the allied raga) in spite of the variations. The changing contexts to be studied are suggested by this description of the nature of improvisation in Indian music.

3.3 Datasets

Serra [171] stressed on the fundamental differences between a research corpora and a test corpus in terms of the ability of the former to capture the essence of a particular music culture. The author advocated the relevance of five criteria that were taken care of during compilation of the CompMusic [1] collection, namely purpose, coverage, completeness, quality, and reusability. The types of data collection that a research corpora demands are audio recordings and editorial metadata. In addition to the information available on the album cover-art, culture-specific elements (e.g. terminology for a given concept in a repertoire), in consultation with domain experts, add value to the metadata. Serra also mentions the usefulness of sharing the corpora and associated metadata on open platforms, so researchers of diverse background can use them for subtasks. This facilitates discovery of similarity and differences across music cultures; or to address the bigger question of cross-cultural universality of music concepts.

[1] https://compmusic.upf.edu/

We have carefully chosen our datasets ensuring the representativeness and diversity of the Hindustani repertoire. The scope of the thesis has been limited to vocal music concerts only, as this paradigm is referred to as the most spontaneous expression of the raga nuances – even instrumentalists are taught ragas by vocalizing them. There are a few overlapping/disjoint datasets depending on the task at hand. While some datasets are used as a training corpus for a mainstream retrieval task in a 'bigdata' sense, there are other datasets which are manually curated for a focused ethnomusicological study. Few of these selected material is then minutely hand-curated for realizing artificial stimuli for conducting controlled behavioral experiments. We describe the two main datasets, namely the allied-ragas dataset and phrase dataset that constitute the distributional and structural views respectively.

The music collection used in this study was compiled as a part of the CompMusic project [170]. The audio recordings used in this study are drawn from the Hindustani music corpus from "Dunya" [2] as a representative set of the vocal performances in the *khayal* genre [171]. The editorial metadata for each audio recording is publicly available on the metadata repository MusicBrainz[3]. We augmented the dataset derived from the Dunya corpus with concerts from personal collections to obtain the overall test dataset presented in Table 3.4. Additionally, the metadata associated with each concert in our dataset is presented in Table 3.3. The artists are selected as the stalwarts of Hindustani vocal music spanning the past 7 or 8 decades, hailing from different *gharana*s (lit. school of thought).

3.3.1 Allied-ragas dataset

The selected music material in our collection is diverse in terms of the number of artists (26) and recordings. The chosen allied raga-pairs belong to the set of 7 most popularly performed allied pairs [113, 145]. Our collection includes a total of 55 concerts from the 3 widely performed allied raga-pairs, as presented in Table 3.2, that are diverse both in terms of the number of svaras and their pitch-classes. Note that our dataset comprises diversity in terms of tonal material (pentatonic, hexatonic, and heptatonic scales). All of the allied raga pairs happen to share the phenomenon of a duration constraint on the second scale degree (r/R svara) occurring in one of the ragas of the pair. All the concerts belong to either *vilambit* (slow) or *madhya* (medium) *laya* (tempo). Some concerts consist of a *drut bandish* towards the end for a short

[2] https://dunya.compmusic.upf.edu/Hindustani

[3] https://musicbrainz.org/

Raga	# Concerts	Duration (hours)	# Artists
Deshkar	6	2:16:50	5
Bhupali	11	5:12:23	9
Puriya	7	2:30:26	7
Marwa	12	4:19:12	12
Multani	7	2:30:57	7
Todi	12	4:39:55	12
# Total	55	21:29:43	52 (26 unique)

Table 3.2: Description of the dataset in terms of number of artists, concerts, and their durations.

duration. Standalone *drut* (fast tempo) concerts are not included due to their typically short durations (less than 12 min) severely limiting both the composition and improvisation sections available for analyses. The accompanying instruments are *tanpura* (drone), *tabla* (percussion), and harmonium or *sarangi* as melodic accompaniment. The pitch range of any performance spans approximately two octaves (octave of the tonic and about half of each of the lower and upper octaves). All of the concerts comprise elaborations based on a selected *bandish*. The number of concerts is unequally distributed across ragas due the greater availability of concerts in the slow and medium tempo ranges in some ragas.

3.3.2 Phrase dataset

The second set of annotations is the occurrences of the GRS (/PDS) phrases. Shyamrao [102] remarks that, notably the key-phrases of a raga do not consist of any ornamentation but unmistakably consist only of notes. The author takes examples of the ragas Deshkar and Bhupali to state that delineating the fundamental phrases becomes quite complex when the ragas in question comprise the same notes. The GRS (/PDS) phrase is a common phrase between the two ragas and it happens to mark the end of the descending (/ascending) line of the scales. The annotator is instructed to mark the phrase from the onset of the note G (/P) till the offset of the note S in Praat [4] software, however we do not stress on the precision of the boundary markings. The GRS (/PDS) phrases are distributed across three octaves (upper, middle, and lower octaves), although lower (/upper) octave instances are fewer. A count of the phrases used in this

[4] http://www.fon.hum.uva.nl/praat/

Raga	Audio ID	Artist	Tonic in Hz	*Tala* (*Laya*)	Duration in min:sec
Deshkar	DK_AC-1	Ajoy Chakrabarty	146.8	Ikwai (Madh)	21:25
	DK_AC-2	Ajoy Chakrabarty	138.5	Teental (Madh)	16:22
	DK_KA	Kishori Amonkar	233.0	Teental (Madh)	29:55
	DK_RK	Rashid Khan	155.5	Teental (Madh)	21:54
	DK_VK	Venkatesh Kumar	155.5	Jhaptal (Madh)	20:45
	DK_UK	Ulhas Kashalkar	155.5	Teental (Madh)	16:29
Bhupali	BP_AC-1	Ajoy Chakrabarty	130.7	Ektal (Vil)	33:46
	BP_AC-2	Ajoy Chakrabarty	146.8	Ektal (Vil)	37:57
	BP_RK	Rashid Khan	155.5	Teental (Madh)	22:21
	BP_OD	Omkar Dadarkar	155.5	Ektal (Vil)	55:46
	BP_DVP	D V Paluskar	158.7	Ektal (Vil)	28:11
	BP_BGAK	B Ghulam Ali Khan	140.7	Ektal (Vil)	23:59
	BP_ARK	Abdul Rashid Khan	137.3	Teental (Madh)	17:30
	BP_KA-1	Kishori Amonkar	236.9	Teental (Vil)	29:40
	BP_KA-2	Kishori Amonkar	235.9	Teental (Vil)	21:34
	BP_KG	Kumar Gandharv	192.2	Teental (Madh)	16:37
	BP_SK	Sameehan Kashalkar	146.8	Tilwada (Vil)	25:02
Puriya	PR_RSM	Rajan Sajan Mishra	155.5	Ektal (Vil)	20:03
	PR_J	Jasraj	138.5	Ektal (Vil)	15:44
	PR_BJ	Bhimsen Joshi	155.5	Teental (Madh)	18:17
	PR_DVP	D V Paluskar	155.5	Ektal (Vil)	28:40
	PR_AK	Amir Khan	146.8	Ektal (Vil)	17:30
	PR_RMS	Ratan Mohan Sharma	138.5	Ektal (Vil)	19:23
	PR_BM	Brajeswar Mukherjee	146.8	Ektal (Vil)	30:49
Marwa	MR_VD	Vasantrao Deshpande	155.5	Ektal (Vil)	18:57
	MR_BGAK	B Ghulam Ali Khan	138.5	Ektal (Vil)	16:32
	MR_AC	Ajoy Chakrabarty	146.8	Teental (Madh)	21:31
	MR_RK-1	Rashid Khan	155.5	Ektal (Vil)	25:52
	MR_RK-2	Rashid Khan	155.8	Ektal (Vil)	24:06
	MR_OD	Omkar Dadarkar	155.5	Ektal (Vil)	32:17
	MR_AK	Amir Khan	138.5	Jhumra (Vil)	18:29
	MR_CRV	C R Vyas	157.9	Jhaptal (Madh)	14:58
	MR_HB	Hirabai Barodekar	210.6	Teental (Madh)	13:32
	MR_J	Jasraj	139.7	Ektal (Vil)	27:44
	MR_SJ	Srinivas Joshi	155.2	Ektal (Madh)	17:02
	MR_SK	Sameehan Kashalkar	146.8	Ektal (Vil)	33:59
Multani	ML_RSM	Rajan Sajan Mishra	155.5	Ektal (Vil)	29:56
	ML_AC	Ajoy Chakrabarty	146.8	Ektal (Vil)	21:05
	ML_KC	Kaushiki Chakrabarty	233.0	Teental (Madh)	25:04
	ML_HB	Hirabai Barodekar	207.0	Ektal (Vil)	13:01
	ML_BJ	Bhimsen Joshi	146.8	Ektal (Vil)	19:07
	ML_OT	Omkarnath Thakur	146.8	Ektal (Vil)	14:42
	ML_OD	Omkar Dadarkar	155.5	Teental (Vil)	27:52
Todi	TD_OT	Omkarnath Thakur	155.5	Ektal (Vil)	16:50
	TD_HB	Hirabai Barodekar	207.0	Ektal (Vil)	13:36
	TD_VS	Veena Sahasrabuddhe	220.0	Teental (Madh)	19:48
	TD_KA	Kishori Amonkar	207.0	Ektal (Vil)	29:23
	TD_BJ	Bhimsen Joshi	146.8	Ektal (Vil)	19:06
	TD_J	Jasraj	138.5	Ektal (Vil)	20:05
	TD_AK	Amir Khan	139.4	Jhumra (Vil)	27:34
	TD_KG	Kumar Gandharv	186.3	Jhumra (Vil)	32:01
	TD_KKG	Kaivalya K Gurav	174.4	Ektal (Vil)	29:03
	TD_OD	Omkar Dadarkar	153.1	Teental (Vil)	17:36
	TD_PA	Prabha Atre	237.8	Ektal (Vil)	26:14
	TD_RRM	Ritesh Rajnish Mishra	156.0	Rupak (Madh)	28:39

Table 3.3: Metadata associated with each concert in our dataset. We list the artist, tonic, *tala*, *laya*, and duration of each concert. Vil ≡ *vilambit*; Madh ≡ *madhyalaya*.

study is presented per concert in Table 3.5 and 3.6 for GRS and PDS phrases respectively.

Raga	Audio ID	Artist	Tonic in Hz	Tala (Laya)	Bandish	Duration in min
Deshkar	DK_AC-1	Ajoy Chakrabarty	146.8	Ikwai (Madh)	Aai Ri Badariya	21:25
	DK_AC-2	Ajoy Chakrabarty	138.5	Teental (Madh)	Aai Ri Badariya	16:22
	DK_KA	Kishori Amonkar	233.0	Teental (Vil)	Piya Jaag	29:55
	DK_RK	Rashid Khan	155.5	Teental (Madh)	Hoon To Tore	21:54
	DK_VK	Venkatesh Kumar	155.5	Jhaptal (Madh)	Chidiya Chutava	20:45
	DK_UK	Ulhas Kashalkar	155.5	Teental (Madh)	Ab Na Sahe	16:29
Bhupali	BP_AC-1	Ajoy Chakrabarty	130.7	Ektal (Vil)	Prabhu Rang	33:46
	BP_AC-2	Ajoy Chakrabarty	146.8	Ektal (Vil)	Prabhu Rang	37:57
	BP_RK	Rashid Khan	155.5	Teental (Madh)	Karoge Tum	22:21
	BP_OD	Omkar Dadarkar	155.5	Ektal (Vil)	Jab Hi Sab	55:46
	BP_DVP	D V Paluskar	158.7	Ektal (Vil)	Jab Hi Sab	28:11
	BP_BGAK	B Ghulam Ali Khan	140.7	Ektal (Vil)	Prabhu Rang	23:59

Table 3.4: Metadata associated with each concert in our dataset. We list the artist, tonic, tala, laya (Vil: vilambit or slow tempo, Madh: madhyalaya or medium tempo), and duration of each concert. Ragas Deshkar and Bhupali comprise two concerts by the same artist (Ajoy Chakrabarty) performing the same bandish.

3.4 Musicological viewpoint of improvisation

Improvisation in Indian music can be understood in terms of a small number of fundamental processes of development. These include the five (mentioned by Widdess [208]), of which 2 relevant to our work on raga-specific phrase variations are: (i) rhythmic intensification: gradual increase in tempo melodic motif is reduced in length in successive repetitions; (ii) development of individual pitches: with a single pitch treated as the focus of attention for a while ("subject of discussion"). The music cues to this are the introduction of the note (has not yet appeared) in the gradually widening melodic range that slowly includes successively higher and/or lower pitches and the perceived emphasis. Therefore, the contexts we should consider are: changes of tempo, approach of sam, local focal note.

van der Meer [192] describes improvisation as a range of different processes. Even a phrase may be unique since durations and intonations are not restricted to quavers, semi-quavers, etc. or naturals, sharps and flats. Such differences can be meaningful – great musicians

Audio ID	# Tala cycles	Cycle len. range: sec	#GRS phr. (in alap)	Octave: H, M, L	Mukhda phr.: Yes, No	Mukhda appr.: Yes, No
DK_AC-1	30	24 – 19	18 (1)	2, 16, 0	0, 17	11, 6
DK_AC-2	24	31 – 27	14 (1)	1, 13, 0	0, 13	9, 4
DK_KA	45	23 – 18	20 (1)	2, 18, 0	0, 19	9, 10
DK_RK	38	14 – 8	22 (2)	3, 19, 0	0, 20	12, 8
DK_VK	28	9 – 7	18 (3)	2, 16, 0	0, 15	11, 4
DK_UK	44	10 – 6	18 (1)	3, 15, 0	0, 17	13, 4
BP_AC-1	26	59 – 53	25 (2)	6, 18, 1	16, 7	4, 19
BP_AC-2	31	45 – 39	28 (1)	4, 22, 2	16, 11	5, 22
BP_RK	45	11 – 9	54 (4)	2, 52, 0	18, 32	8, 42
BP_OD	31	56 – 52	33 (2)	4, 29, 0	24, 7	5, 26
BP_DVP	21	47 – 42	25 (1)	3, 22, 0	15, 9	6, 18
BP_BGAK	25	39 – 36	23 (2)	3, 18, 2	15, 6	4, 17

Table 3.5: Manual annotation of the GRS phrase in both ragas and related context-based measurements. Rhythm based: no. of cycles and range of cycle lengths, concert-location based: distribution in alap and vistar, pitch-range based: distribution of phrases in the three octaves, and mukhda based: the phrase being the mukhda or the penultimate phrase before the mukhda.

Audio ID	# Tala cycles	Cycle len. range: sec	#PDS phr. (in alap)	Octave: H, M, L	Mukhda phr.: Yes, No	Mukhda appr.: Yes, No
DK_AC-1	30	24 – 19	16 (1)	0, 13, 3	10, 5	3, 12
DK_AC-2	24	31 – 27	13 (1)	0, 11, 2	9, 3	2, 10
DK_KA	45	23 – 18	11 (1)	0, 9, 2	0, 10	5, 5
DK_RK	38	14 – 8	19 (3)	0, 13, 6	12, 4	5, 11
DK_VK	28	9 – 7	12 (1)	0, 10, 2	0, 11	6, 5
DK_UK	44	10 – 6	15 (2)	0, 12, 3	10, 3	6, 7
BP_AC-1	26	59 – 53	19 (2)	0, 14, 5	0, 17	9, 8
BP_AC-2	31	45 – 39	15 (1)	0, 12, 3	0, 14	7, 7
BP_RK	45	11 – 9	25 (3)	0, 20, 5	0, 22	14, 8
BP_OD	31	56 – 52	24 (2)	0, 18, 6	0, 22	16, 6
BP_DVP	21	47 – 42	18 (1)	0, 15, 3	0, 17	13, 4
BP_BGAK	25	39 – 36	14 (1)	0, 9, 5	0, 13	6, 7

Table 3.6: Manual annotation of the PDS phrase in both ragas and related context-based measurements. Concert-location based: distribution in alap and vistar, pitch-range based: distribution of phrases in the three octaves, and mukhda based: the phrase being the mukhda or the penultimate phrase before the mukhda.

give the feeling that every step was inevitable. So, at each instance there are many possibilities and the musician chooses one with a good reason. Author further questions whether this spontaneous process is improvisation or interpretation of bandish; or whether it is actually different extents of memorization. Nettl [129] remarks that every performance must embed some of the 'points of reference' of the raga, else it will be considered "ignorance" (rather than creativity). That is, the chosen raga provides the building blocks in terms of the tonal material and characteristic melodic motifs. In his scholarly review of the notion of improvisation by Nettl [129], one view is that it applies to music for which there is basically a notation system from which the improviser departs. In Carnatic music tradition, different styles of improvisation have names. E.g. in niraval, it is related to the composition; in alapana and tanam, it has only the characteristics of the raga (i.e. different degrees of freedom). Different cultures draw different lines between composition and improvisation. Both draw from building blocks at various time scales.

Many authors have attempted to formalize a performance model to be able to come up with any universal structure of musical improvisation. Nettl [129] suggested that all music consists of a sequence of fixed structural points. Author remarks that if these are close together, the music is pre-composed; but if they are further apart, the performer must navigate from one to the next by interpolating blocks of material, whether spontaneously created or selected from a memory bank. Widdess [208] assumes this model of improvisation to fit Indian music well in many respects: the metrical cycle of tala provides fixed temporal points for action, especially the first beat of the cycle, called sam; a pre-composed block, or "composition" (bandish), alternates with improvised episodes; the melodic grammar of the raga provides pathways from note to note and motivic material for filling space within the tala cycle. But this oversimplified form seems a rather one-dimensional concept of performance [208], as if the performer needs only to think about one thing at a time and needs only to make decisions about how to get to the next "structural point" at the beginning of each "block". A number of studies (e.g. [191]) have suggested that improvisation in IAM is less spontaneous than it appears to be, and is almost always reliant on extensive memorized materials and procedures. But whether a singer truly improvises or not seems to vary between individuals; some musicians are clearly adept at on-the-spot composition [206]. Widdess [208] further remarks that perhaps the only way to tell that everything has not been worked out in advance, is where an improvised passage threatens not to turn out as intended, and we can hear the performer adjusting his materials in real time, in order to hit a structural point or to compensate for not hitting it – "it is the ongoing negotiation of

conflicting requirements and structures that makes performance of this kind particularly exciting for the listener."

3.4.1 Musical expectancy

Huron [90] has drawn some basic distinctions among the kinds of musical expectation that listeners to any music might have. Widdess [208] remark that listening is in the mode of "schematically induced dynamic expectations", i.e. expectations based on predicting what is coming next based on familiar melodic, rhythmic, and formal patterns; and "dynamic expectations" based on the perception of patterns emerging as the music unfolds. The other type of prediction is veridical where the listener has heard the same music before and knows exactly what to expect. So, we have schematic expectations based on long-term memory of the raga, alap etc. and dynamic expectations based on short-term memory of the performance itself. Apart from the above unconscious expectations, there exist certain conscious expectations based on explicit, accessible knowledge acquired through training or experience. They might be triggered by, e.g. by the performer's identity as a member of a particular stylistic lineage (lit. gharana), implying to the knowledgeable listener that (s)he will display particular techniques of sound production or other stylistic preferences.

These categories undoubtedly overlap, and all depend on repetition and memory [208]. In the case of IAM, where "improvisation" is emphasized more than the repetition of pre-composed pieces, schematic and dynamic expectations may be more significant than veridical. In particular, Widdess [208] suggests that expectations are generated by cognitive schema, whether these are learned through cumulative listening or generated dynamically within a single performance. This model of improvisation would help to account for the kind of unpremeditated performance practice. But at the same time it should not be assumed that those performers who rely more on memorized material are any less reliant on schema. Author argues that these schema are equally important for listeners to form an understanding of what is heard, whether by recognizing schema that they already know, or by perceiving dynamically the schematic structure of music as it unfolds. Widdess [208] suggests that empirical and analytical study of the schematic structure of orally composed and transmitted music, and of processes of schema combination, would illuminate the understanding of phenomena that is, in general, imprecisely characterized as "improvisation".

In other works, Widdess [130, 208] proposed that in Indian music, improvisation can

be partly understood in terms of a small number of fundamental processes of development. The major one relevant to our discussion is the melodic expansion (vistar) which describes gradual widening of melodic range to include successively higher, and/or lower, pitches. This process can be seen both in the development of an individual phrase, and in the structuring of each large section of a performance. The reverse process, a gradual contraction of melodic range, can also occur, e.g. in the final descent of an alap or the descending phase of a tan. The next process, rhythmic intensification, defines on the large scale, existence of a gradual increase in tempo or rhythmic density, with different technical procedures becoming available at each new tempo. On the small scale, an individual motif can be progressively reduced in length at successive repetitions. Another important process as proposed by Widdess [130] is the development of individual pitches. According to this process a single pitch may, for a time, be treated as the focus of attention or "subject of discussion"; such a pitch may be repeated, emphasized, prolonged, and/or taken as the concluding note of successive phrases. Typically, the next higher pitch is hinted at during the development of the previous pitch, before becoming the focus of attention in its turn.

Now the interesting question is, since improvisation is so culture-specific, it would be good to try to answer some of the questions as raised by Nettl [129] for IAM. E.g. how much do performers discriminate between alap and vistar in terms of improvised content, and especially phrase shapes. Do they consider that they perform the same raga similarly or differently each time? We would like to build a theoretical model that provides a way to express the relationship between machine analysis of recordings and the musicological concepts that govern how scholars understand their structures. van der Meer [191] provides also an anecdote about time-scaling the melodic shape of a phrase: "The only concern is to return to the sama. For continuity and to avoid the feel of undecidedness, he (the performer) has a number of stereotyped reductions and enlargements of the phrase." We will see whether we can predict which stereotype is used in a given context.

3.5 Scope of the thesis

Based on the review of literature on both computational and musicological aspects, we summarise the scope of this thesis.

- In the distributional front, we consider a computational approach towards using record-

ings of raga performance to investigate how the tonal hierarchy of a raga, as a prominent aspect of music theory, influences performance practice. A computational model would need to incorporate the essential characteristics of the genre and be sufficiently descriptive to distinguish performances of different ragas.

- In the structural front, we focus on a computational representation of the former in a manner suited to the empirical analysis of raga performances. We observe that there are variations in phrase shape across instances and concerts in a given raga. We wish to examine whether: (i) there are context based explanations for these variations; (ii) whether the distinctiveness of the raga is maintained (in terms of discriminating it from the allied raga) in spite of the variations. The changing contexts to be studied are suggested by this description of the nature of improvisation in Indian music. We aim to build a theoretical model that provides a way to express the relationship between machine analysis of recordings and the musicological concepts that govern how scholars understand their structures.

- The chief objective of the perceptual experiments is to investigate whether a characteristic melodic shape behaves like a prototypical motif. We first determine all the distinct independent dimensions of actual physical variability by observing actual instances from concerts. Next, we verify whether the existence of a "prototype" only applies to raga-characteristic phrases or it extends to any melodic pattern.

- The final agenda of the thesis is to show how all of the above computational models can effectively be used in mainstream MIR applications as well as focused ethnomusicological studies via empirical analysis.

Chapter 4

Audio processing for melodic analysis

4.1 Motivation

This chapter discusses the signal processing methods that constitute the backbone of empirical research on computational musicology in audio MIR, i.e. detection and labeling of the musicological entities in the acoustic signal domain. The low-level acoustic features that characterize a melody are pitch, loudness, and timbre. The acoustic correlates thereof are fundamental frequency (F0), intensity, and spectral shape. We limit the scope of our work to vocal melody only. Also, as the timbral features capture more of voice quality and phonetic content than the melodic content, timbre is not discussed in this thesis. However, there is a notion that pitch and loudness co-vary due to voice production mechanism [158]. We compute the intensity, or vocal energy, from the harmonic sum of the predominant F0. For any voice synthesis purpose (to be discussed in Chapter 7), we make use of the intensity information. The majority of the analyses in this thesis deal with F0 only, as the primary melodic feature.

As discussed in the introductory chapter, we characterize both distributional and structural representations of melody which involve different sets of features to be captured at different time-scales. While the distributional representation is devoid of any timing information, the structural representation is a time-series data (fundamental frequency versus time) that we shall refer to as "pitch contour" or "melodic contour". Henceforth, any mention of "melody" in an acoustic signal context will refer to the melodic contour. Figure 4.1 shows the tonic normalized melodic contour (pitch in cents) from a phrase taken from raga Tilak Kamod performance by Ashwini Bhide. We see that there is a certain melodic shape which can be further broken down into melodic events. This chapter aims to define the relevant melodic events and propose novel

computational representations of the melodic events and suitable (dis)similarity measures to compare them.

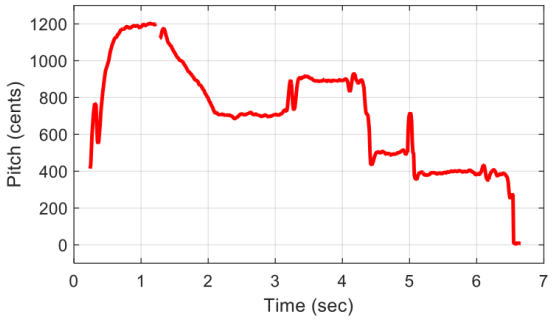

Figure 4.1: Sample melodic contour from a phrase taken from raga Tilak Kamod performance by Ashwini Bhide, normalized with respect to the concert tonic. The svara transcription (note sequence) would be SPDmGS.

4.2 Background

Our goal is to propose computational representations that robustly capture the particular melodic features of the raga in a performance. First order pitch distributions have been used as a key feature to the classical problem of raga recognition with varied degree of success. This distributional feature, though devoid of any timing information, has shown to be effective in such a classification scenario [18, 35, 36, 43, 96]. The power of this representation lies in the fact that it captures the long-term tonal hierarchy of a raga performance. Hence it was appropriate for us to include this feature in our analyses. However, there are several configurations and tunable parameters that need to be selected in a principled manner.

We present the block diagram to obtain the two categories of distributions, viz. continuous pitch based and segmented svara based, as depicted in Figure 4.2. In the block diagram itself, we introduce the structural representation of melodic shape, in the framework of time-series analysis. Structural representation can also be in both continuous (as in Figure 4.1) or discrete domain. The discrete form includes segmenting the melodic shape into further sub-units or melodic events, but how micro- or macro-level the time-scale should be is a parameter to be chosen. We go by the top-down musical definition of melodic events to have a quantitative model of melodic structure. The continuous time-series or the discrete symbol sequence have been successfully used in various tasks, e.g. motif discovery, query by humming, raga recognition. We attempted few of them as a proof of concept (to be discussed in Chapter 8) to advocate

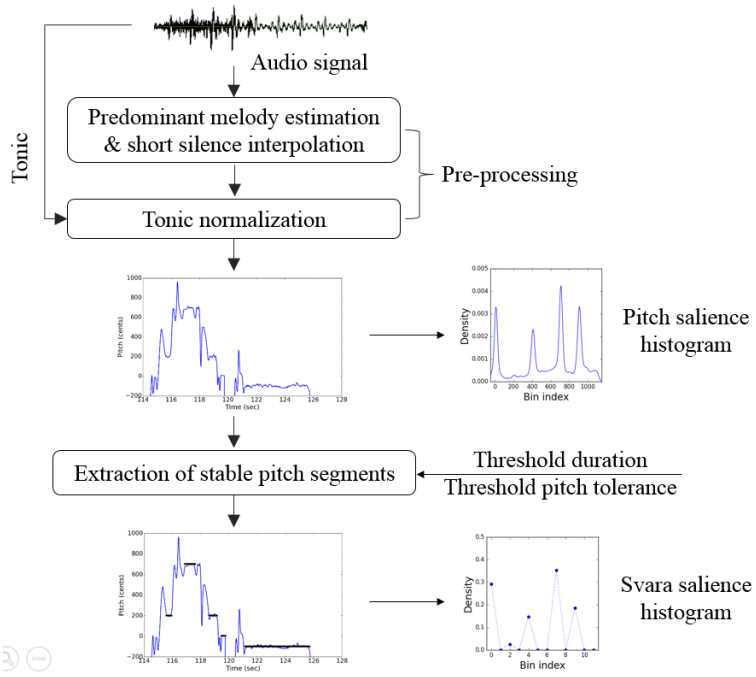

Figure 4.2: Block diagram of the signal processing chain from audio signal to the distributional and structural representations.

our proposed representation; and then further addressed the task of allied raga discrimination to show the power of these melodic features.

4.2.1 Objective modeling for musical entities

As stated earlier, we bring in musical knowledge to come up with the segmentation schema in the structural representation of melodic shapes. The available musicological text resources use the musical notation (using solfege only, rarely with certain added symbols) to describe the melodic structure. The symbols are interesting, but has to be interpreted with the Guru's (instructor) guidance and is left to subjectivity. The question we ask is whether we can objectively interpret the symbols used in musicological texts to understand the way it is manifested in performance. Some examples follow.

- G P D ∼, P D S \P, D \G – '\' indicates glide (meend), '∼' indicates a nyas [13].
- D P^P G ^{G}R S – superscripts are indicative of touch notes [152].
- D (P)G P, D P G (R)S – commas (',') are resting points that delimit phrases, braces denote less emphasis/short duration [102].

Now that the sequence of svaras are used to characterize a melodic shape, we observe from the melodic contour that these map to the relatively stable horizontal regions. Hence we chose to mimic the textbook notation by segmenting the pseudo-steady svaras. Given that we have a method to segment the stable svaras, the residue are the transient melodic shapes joining two consecutive stable svara segments or starting/ending at a pause (silent region). The stable svara transcription algorithm and event segmentation would be discussed in detail in the subsequent sections. For the transient segments too, we propose a dictionary modeling technique using vector quantization. Finally, we have to make a choice of the numeric representation of the segmented melodic events. E.g. the stable svaras can be represented by the mean/median/mode of the segmented region for characterizing intonation (lit. *shruti* or microtonal variation) of a certain svara in context of a given raga. The pitch glides can be characterized by assigning a "broad" shape (convex/concave) apart from the curve fitting parameters. Also, a *gamak* (pitch oscillation between two svaras) can be characterized by a "flag" and counting the number of oscillations.

4.3 Methodology

4.3.1 Pitch time-series extraction from audio

Predominant-F0 detection is implemented by an algorithm from [154] that uses harmonicity-based grouping of sinusoidal partials detected in the polyphonic audio to estimate one or more F0 candidates in each frame. Next, the spectral and temporal properties of the singing voice are exploited to discriminate its partials from those of the melodic accompaniment. Further temporal smoothness constraints minimize pitch octave errors. The melodic pitch is detected at 10 ms intervals throughout the audio recording with no F0 values assigned in the detected purely instrumental regions. The algorithm also returns the intensity (energy of the vocal harmonics) at each time step. The predominant-F0 detection algorithm is based on a small set of tunable parameters, which are selected from preset values, for pitch tracking with high accuracy using available knowledge such as singer gender and whether the expected pitch variation is rapid or slow.

As our dataset comprises performances by a number of artists, male and female, the detected vocal melody must be normalized with respect to the tonic pitch. The fundamental fre-

quency (F0) values in Hz are converted to the cents scale by normalizing with respect to the concert tonic determined using a classifier based multi-pitch approach to tonic detection [77]. With an accuracy of over 90%, any gross errors are easily corrected based on raga (or rather, allied raga group) information. As all concerts in the dataset belongs to ragas Deshkar and Bhupali that constitute the same tonal material, any gross error in the tonic (e.g. transposed to any other scale degree with respect to the actual tonic) is rectifiable.

The final preprocessing step is to interpolate short silence regions below a threshold (250 ms as proposed by [56]) indicating musically irrelevant breath pauses or unvoiced consonants, by cubic spline interpolation, to ensure the integrity of the melodic contour. The value 250 ms has been empirically chosen from the distribution of "pauses" (contiguous silent segments). The distribution is bimodal in nature, with a valley around 250 ms indicating a threshold between unvoiced frames and musician-intended breath pauses. Median filtering with a 50 ms window is performed to get rid of irrelevant local pitch fluctuations. Eventually, we obtain a continuous time series of pitch values representing the melody line throughout the vocal regions of the concert.

4.3.2 Svara segmentation and labeling

The stylization of the continuous pitch contour has been of interest in both music and speech. In Western music, piece-wise flat segments are used to model the melody line corresponding to the note values and durations in the underlying score. Speech signals, on the other hand, have smoothly varying pitch which can be stylized, for example, with polynomial fitting [71]. Researchers have tried stylizing the pitch contour with several innovative strategies, such as polynomial fit [71], SAX [112] and its variations [68], melody transcription [50], melodic shape assignment [8, 89, 90].

We observe smoothly varying melodic contours but peaky overall pitch distributions coinciding with the discrete svara intervals. The pitch contour of a melodic phrase can thus be viewed as a concatenation of events of two categories: (i) a pseudo-steady segment closely aligned with a raga svara, and (ii) a transitory segment which connects two such consecutive steady segments. The latter is often referred to as an *alankar* or an ornament comprising figures such as *meend* (glide), *andolan* (oscillation), *kan* (touch note), etc. A reduced representation corresponding to a sequence of svara can be achieved by detecting the "stable" segment boundaries and discarding the time segments connecting these.

The underlying scale interval locations or svara are estimated from the prominent peaks of the long-term tonic-normalized pitch histogram across the concert. The allowed pitch deviation about the detected svara location, T_{tol}, is empirically chosen to be ± 35 cents. This is based on previous work where a closer value (± 25 cents) was found to optimize the recognition of the svara sequence corresponding to a phrase based on the time series representing the melodic shape across many different instances of the same phrase extracted from audio recordings [151]. The authors adjusted the local cost (difference of corresponding pitches of reference and test templates) so that perceptually unimportant pitch differences do not affect the DTW optimal path estimate apart from the global path constraint. For our task, we started with this value as T_{tol} which gave reasonable results. But for certain boundary cases (e.g. where the peak shape is asymmetric with respect to the detected peak due to raga-specific intonation), ± 25 cents was not enough to capture all onset/offsets. Thereafter, we empirically chose ± 35 cents which was able to guard this issue consistently. We also tried with a more moderate quarter-tone (50 cents) tolerance, but the same did not sound acceptable upon hearing the segmented notes [1].

The above steps provide segments of the pitch time-series that approximate the scale notes while omitting the pitch transition regions. Next, a lower threshold duration of T_{dur} is applied to the fragments to discard fragments that are considered too short to be perceptually meaningful as held svaras [150]. T_{dur} is empirically set to 250 ms, supported by previous subjective listening experiments [198], where authors reported that 250 ms was found to be optimal duration for a "khada svara" (stable note) for the task of classifying between Hindustani and Carnatic vocal styles. This leaves a string of fragments each labeled by the corresponding svara. Fragments with the same note (svara) value that are separated by gaps less than 100 ms are merged. The svara sequence information (i.e. scale degree and absolute duration of each steady pitch segment) across the concert recording is stored. The algorithm for the proposed melody transcription is presented in Algorithm 1.

[1] This was also tested against the stylization procedure where we take both stable notes and modeled contour of the transients. The tuning of ± 35 cents had performed better segmentation of stable notes so the joining of transient and stable note segments were smooth and sounded acceptable upon resynthesis. We shall see the resynthesis procedure in Section 4.4.2.2.

Algorithm 1 Stabla svara transcription algorithm

procedure HEURISTICTRANSCRIPTION
    tDur ← Threshold *duration*$(\#frames)$
    tTol ← Threshold *pitch tolerance*
    i ← Current frame *pointer*
    top:
    $diff \leftarrow$ *pitch*(i) − *nearestSvara*
    if $diff >$ *tTol* **then return** false
    loop1:
    $j \leftarrow i$
    $i \leftarrow i + 1.$
    $noteOnset \leftarrow j.$
    $noteName \leftarrow nearestSvara.$
    if $(i - j) > tDur$ **then**
        $noteOnset \leftarrow i.$
        goto *loop1*.
        close;
    loop2:
    if $noteName(i) == noteName(j)$ **then**
        $noteOffset(j) \leftarrow noteOffset(i).$
    goto *top*.

4.3.2.1 Hysteresis thresholding

The following is another independent approach to svara (and thereby) transient segmentation. Note that this is less sophisticated an algorithm that was developed prior to the above, we report the same for completeness. This method was used in characterizing melodic shapes [58], the shortcomings of this motivated to develop a more robust transcription methodology which gave rise to the above. In subsequent chapters, the term "transcription" shall refer to the Algorithm 1.

The most simplistic approach to segment the stable svaras by their onset/offset is to threshold around the defined svara positions. However, given the fact that the vocal jitter or intended pitch oscillation (*andolan*) can give rise to multiple unwanted segments, we implemented a double thresholding to get rid of these. This method is known as the hysteresis thresholding method. Figure 4.3 shows the DnDP phrase in raga Alhaiya Bilawal by Ashwini Bhide and its segments by indicating the onsets/offsets of the melodic events.

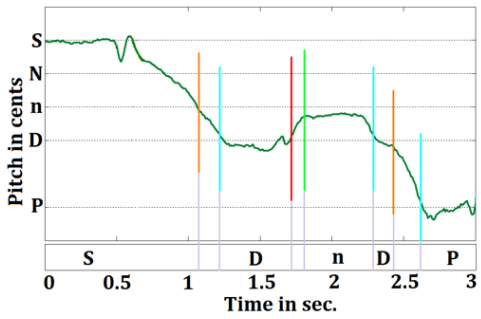

Figure 4.3: Onset and offset detection of melodic events using hysteresis thresholding method. Different colors indicate onset/offset from/to upper/lower pitch intervals. The test phrase in the plot is a DnDP phrase in raga Alhaiya Bilawal by Ashwini Bhide.

The two thresholds are set as 50 and 20 cents. If a contour segment enters within ± 50 cents and then eventually within ± 20 cents of a semitone pitch interval, an onset is marked at the 20-cent location. In the case of offset marking, an offset is marked at the 20-cent location if the contour subsequently exits the ± 50 cents region. This is reasonable in a way that the vocal jitter does not create spurious onsets/offsets. However, touch notes (*kan* svaras) can create extra offset and onset around the same svara, which can be taken care of by appropriate merging strategies in the post-processing stage. The onsets/offsets shown in Figure 4.3 are color-coded: orange (red) indicates offset of a svara exiting towards a lower (higher) svara; cyan (green) indicates onset of a svara entering from a higher (lower) svara.

4.4 Melodic representations

As discussed in the introductory chapter, the thesis revolves around the theme of melodic representation-cum-distance measure. The two broad aspects of the representation block are the distributional and structural features. While the former captures the distribution of tonal material at a larger time-scale, giving rise to a tonal hierarchy; the latter models the local pitch shapes as a sequence of melodic events. Both of the two representations are measured in continuous as well as discrete domain. We discuss, in detail, the different variants of distributional and structural representations proposed (or modified from literature) in this thesis and the tunable hyperparameters thereof.

4.4.1 Distributional representations

Given the pitch-continuous nature of raga music, we are faced with multiple competing options in the definition of a tonal representation. Closest to the tonal hierarchy vector of Krumhansl [99] is the 12-bin histogram of the total duration of each of the svara segments detected from the melodic contour as described in Section 4.3.2. Considering the importance of the transitions connecting stable notes as well as micro-tonal differences in intonation between the same svara in different ragas, a higher dimensional histogram derived from all the pitch values in the continuous melodic contour would seem more suitable. The bin width for such a pitch continuous distribution is also a design choice we must make.

4.4.1.1 Pitch salience histogram

The input to the system is the tonic normalized pitch contour (cents versus time). The pitch values are octave-folded (0 – 1200 cents) and quantized into p bins of equal width (i.e. the bin resolution is $\frac{1200}{p}$). The bin centre is the arithmetic mean of the adjacent bin edges. The salience of each bin is proportional to the accumulated duration of the pitches within that bin. A probability distribution function is constructed where the area under the histogram sums to unity. This representation is equivalent to $P\_continuous$ as proposed by Koduri [96]. Given the number of bins, the histogram is computed as:

$$H_k = \sum_{n=1}^{N} \mathbb{1}_{[c_k \leq F(n) < c_{k+1}]} \tag{4.1}$$

where H_k is the salience of the k^{th} bin, F is the array of pitch values $F(n)$ of dimension N, (c_k, c_{k+1}) are the bounds of the k^{th} bin and $\mathbb{1}$ is an indicator random variable. Figure 4.4 shows the pitch salience histogram for $p = 1200$ (1 cent bin resolution) where different colors indicate different concerts in the corresponding raga. Comparing the Deshkar and Bhupali distributions, differences in the heights of the R peak (around bin 200) and in the precise location (intonation) of the G peak (around bin 400) are observed. For a bin resolution of 100 cents, the representation is equivalent to the PCD [35].

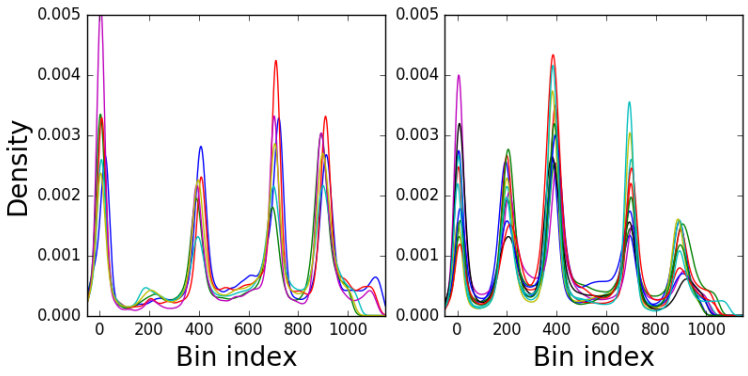

Figure 4.4: Pitch salience histograms (octave folded, 1 cent bin resolution) of 6 concerts in raga Deshkar (left) and and 11 concerts in raga Bhupali (right).

4.4.1.2 Svara salience histogram

The svara salience histogram is not equivalent to the PCD. The input to the system is the string of segmented stable svaras extracted from the melodic contour as described in Section 4.3.2 (and similar to the $P\_duration$ proposed by Koduri [96]). The svara salience histogram is obtained as:

$$H_k = \sum_{n=1}^{N} \mathbb{1}_{[F(n) \in S_k, k \in (1,2,\cdots,12)]} \tag{4.2}$$

where H_k is the salience of the k^{th} bin, F is the array of pitch values $F(n)$ of dimension N, and S_k is the k^{th} svara of the octave. H_k is always a 12-element vector. Figure 4.5 shows the tonal hierarchy in the form of svara salience histogram where different colors indicate different concerts in the corresponding raga. One major difference between pitch salience histogram and svara salience histogram is that the precise intonation information is lost in the latter.

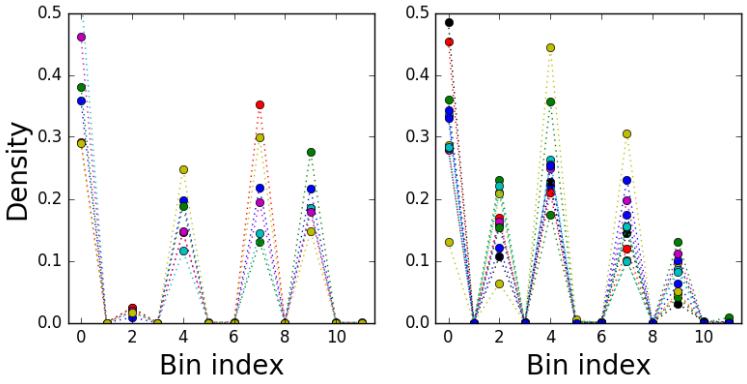

Figure 4.5: Svara salience histograms (octave folded) of 6 concerts in raga Deshkar (left) and and 11 concerts in raga Bhupali (right).

4.4.1.3 Svara count histogram

The frequency of occurrence of the notes was reported by Smith et al. [174] to strongly correlate with the hierarchy of tones; hence we decide to investigate the same as a potential measure of svara salience. This is equivalent to the $P\_instance$ proposed by Koduri [96], where salience is proportional to the frequency of occurrence of each svara. The svara count histogram is obtained as:

$$H_k = \sum_{j=1}^{J} \mathbb{1}_{[S(j)=S_k, k\in(1,2,\cdots,12)]} \tag{4.3}$$

where H_k is the salience of the k^{th} bin, S is the array of segmented svaras $S(j)$ of dimension J, and S_k is the k^{th} svara of the octave. H_k is always a 12-element vector. Figure 4.6 shows the tonal hierarchy in the form of a svara count histogram where different colors indicate different concerts in the corresponding raga. We observe a high visual similarity between the svara salience and count histograms.

4.4.1.4 Extension of svara salience histogram to different time-scales

The svara salience histogram can be computed at different time-scales, e.g. at a breath phrase (BP) level [2] or a tala cycle level [3]. Figure 4.7 shows the svara salience histogram for each BP of the concert of raga Todi sung by Ajoy Chakrabarty. The 12th bin along the y-axis corresponds to the tonic svara Sa (0 cents), 24th for its octave (1200 cents). The svara salience peaks are indicated in the color scale (dark indicates a strong peak). We also compute the svara

[2] Refer to Chapter 8

[3] Refer to Chapter 5

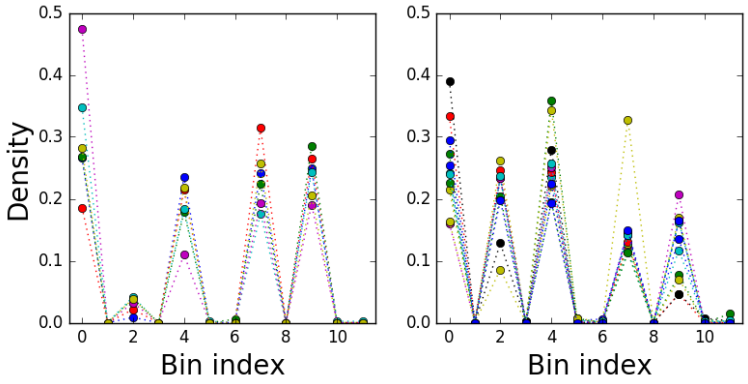

Figure 4.6: Svara count histograms (octave folded) of 6 concerts in raga Deshkar (left) and 11 concerts in raga Bhupali (right).

salience histogram corresponding to each *tala* cycle of a concert. Figure 4.8 shows the histogram versus cycle index for one concert each in raga Deshkar and Bhupali by the same artist Ajoy Chakrabarty.

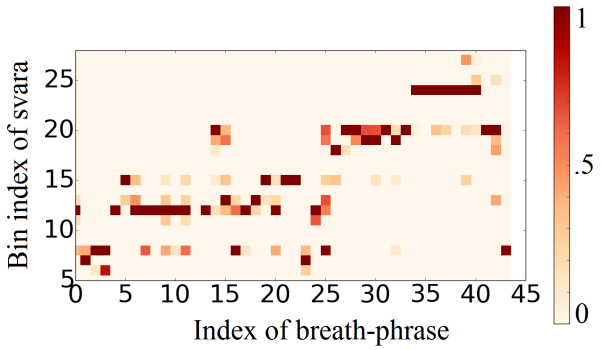

Figure 4.7: Pitch histogram of svaras for each breath-phrase in a case-study concert of raga Todi sung by Ajoy Chakrabarty.

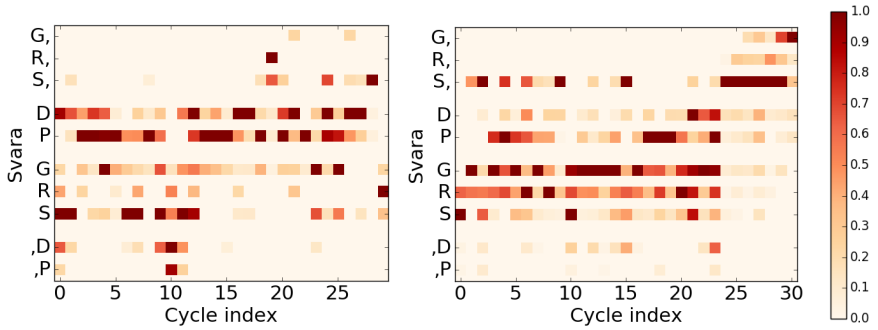

Figure 4.8: Svara salience histograms across tala cycles for DK_AC-1 (left) and BP_AC-2 (right). The concert metadata is presented in Chapter 3.

4.4.2 Structural representation

Similar to that of the distributional attribute, we have certain design choices for the representation of melodic shapes. The two major distinct approaches are continuous versus discrete representations. In the continuous domain, we segment melodic events by their corresponding onsets/offsets, but preserve the original pitch samples to characterize the event. In the discrete domain, on the other hand, the events are mapped to discrete symbols. The most fundamental such parameter is the pitch interval (the solfege, to mimic the textbook notation representation) for stable note events. Further, we can encode the duration (either absolute, or relative to certain tempo), intonation (e.g. median cents value), slope, shape and so on.

4.4.2.1 Behavioral sequence modeling

With a goal to preserve the characteristic shape of the melodic shape including the pitch transitions in the mapping to the symbol sequence, we consider the approach of Tanaka [180] who proposed "behavioral symbols" to capture distinct types of local temporal variation in a human motion capture system. A melodic phrase can be viewed as a sequence of musical gestures by the performer, with a behavioral symbol then potentially corresponding to a single (arbitrary movement) in pitch space. A sequence of symbols would serve as a sketch of the melodic motif. In Tanaka's system, the symbols are purely data-dependent and evolve from the analysis itself [179, 180]. We bring in musical context constraints as presented in the algorithm description next.

The pitch time-series is segmented into fixed duration windows centered at uniformly spaced intervals so that the windows are highly overlapping as illustrated in Figure 4.9. The pitch contour within each window is replaced by a piecewise flat contour where each piece represents a fixed fraction of the window. While Tanaka recommends normalization of the pitch values within the window to [0,1] range in order to eliminate vertical shifts and scaling between otherwise similar shapes, we omit this step given that we are not looking for transposition or scaling invariance in the mukhda detection task. The piece-wise flat sub-segments are obtained by the median of the pitch values in the corresponding subsegment. We choose median as opposed to mean [179] as it is less sensitive to the occasional outliers in the pitch contour. We bring in further domain constraints by using the discrete scale intervals for the quantization of the piecewise sub-segments that describe a specific behavioral symbol (BS). We obtain a

sequence of BS, one for each window position. Due to the high overlap between windows, repetitions are likely in consecutive symbols. These are replaced by a single BS which step brings in the needed time elasticity. Figure 4.9 illustrates the steps of construction of the BS sequence (BSS) and its repetition removed version (the modified BSS) from a simulated pitch time-series.

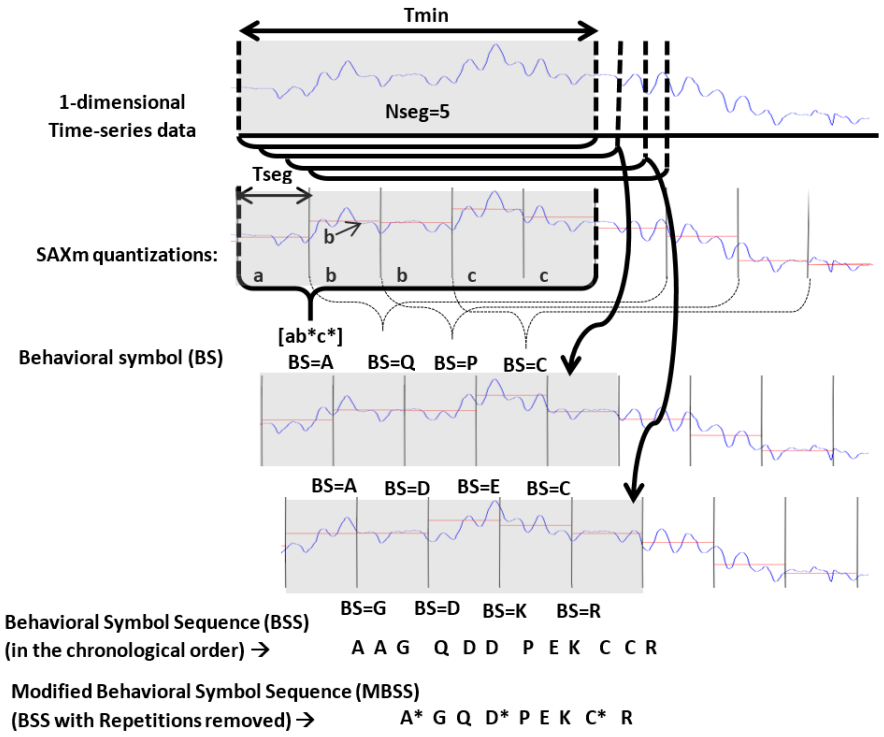

Figure 4.9: Construction from a pitch time series of the BS sequence (BSS) and the modified BSS.

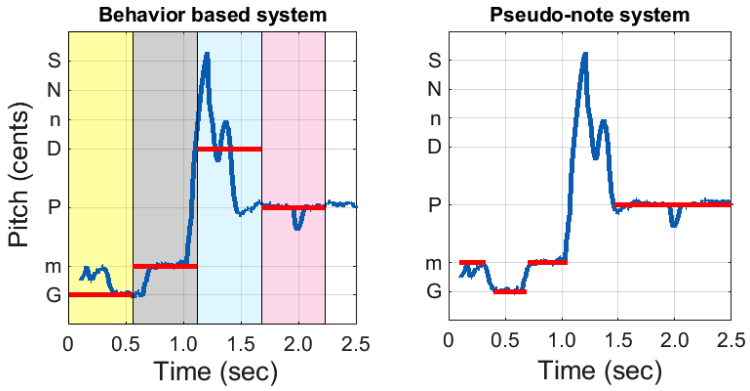

Figure 4.10: Comparison of BSS method with proposed stable svara transcription. The test phrase in the plot is a mnDP phrase in raga Alhaiya Bilawal by Ashwini Bhide.

4.4.2.2 Transcription inclusive of transient segments

Different approaches for representing the pitch contour transients with several innovative strategies are found in the literature, e.g. polynomial fit [71], melodic shape assignment [8, 90]. While some of these approaches addressed the task of melodic representation from a purely retrieval viewpoint, others proposed different approaches from a musicology and pedagogy perspective. Datta et al. [40, 41] had used 2^{nd} degree polynomial to automatically extract (and hence classify) 'meend' from the performances in Hindustani vocal music. Gupta et al. [85] had reported superiority of a 3^{rd} degree polynomial over a second-degree (i.e. parabolic) contour for the task of objective assessment of ornamentation in Indian classical singing.

The stable svara transcription representation discarded all melodic transients and only preserved the sequence of stable note segments. We propose a way to consider the transient regions in the modeled contour, with an additional step of quantizing them to a set of codebook vectors. We first normalize each transient segment to lie within 0 to 1 range. A 3^{rd} degree polynomial is fitted and we generate a candidate shape by constructing a unit length (100 samples) contour from the polynomial coefficients. The K-means clustering algorithm with Euclidean distance measure is used to generate a codebook of distinct representative transient shapes (refer to Figure 4.11). The quantization of a test transient segment is achieved through a nearest neighbor classifier (on the fitted and normalized 3^{rd} degree polynomial) with the same Euclidean distance measure as used during training. If the achieved representation corresponds to some invariant skeleton of the melodic shape of the phrase via a low-degree polynomial, we would anticipate obtaining better matches across variations of the melodic phrase. We address the question of how to bring domain knowledge into this transformation. We will use this method in a retrieval task to be discussed in Chapter 8.

The training corpus comprises 30 songs from 30 different ragas (22,092 transient segments) from the 'Raga Dataset' of 300 songs [79] from the CompMusic [171] collection. From each raga, one song is randomly chosen to constitute the training corpus.

Stepwise reconstruction

This section is not directly related, but presents a stepwise procedure for the above representation. The steps involved to reconstruct the stylized contour is to denormalize the codebook vector to the original time-scale and pitch range. We take the exact reverse steps of the encod-

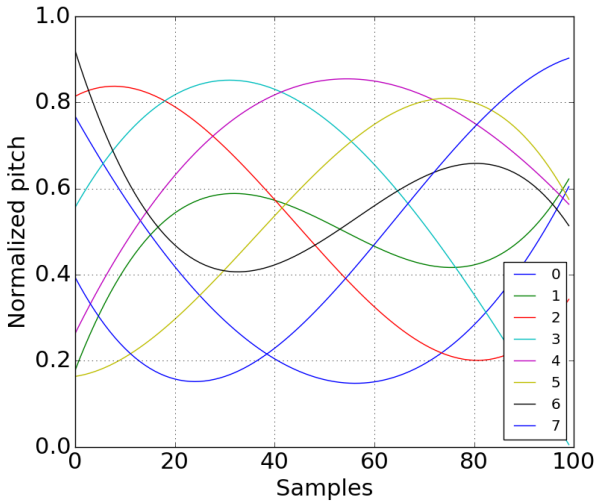

Figure 4.11: 8 centroids obtained corresponding to each cluster index from the codebook. Each vector is normalized between [0,1] and contain 100 samples.

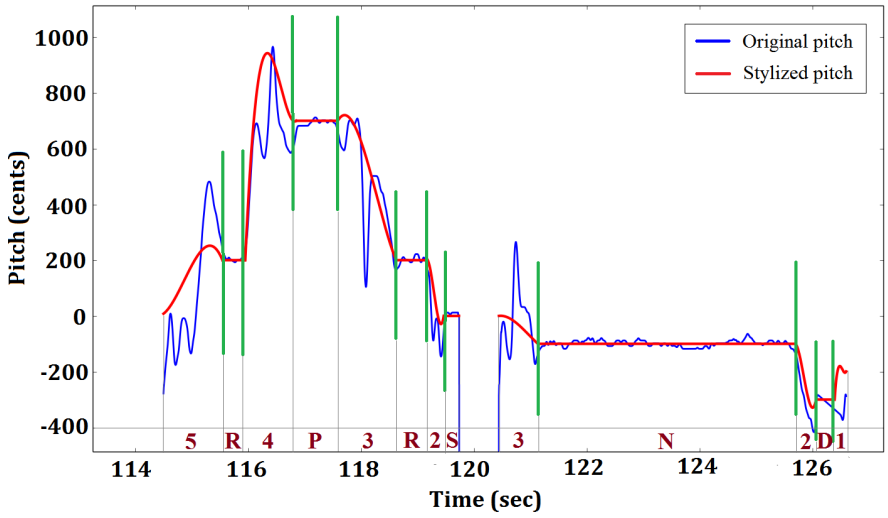

Figure 4.12: Reconstruction of stylized contour from stable svaras and modeled codebook shapes for transients. The test phrase in the plot is from a concert in raga Sudh Sarang by Ajoy Chakrabarty.

ing procedure to decode the transient segments (refer to Figure 4.13). Given a time-series pitch sequence p, where $a = min(p)$ and $b = max(p)$, normalized (within range [0,1]) contour q is obtained as $q = \frac{p-a}{b-a}$. For denormalizing $q = [q[1], q[2], ..., q[L]]$ ensuring end-point matches between two pitch points c and d, we arrive at the equation:

$$q' = \frac{q - \min(q[1], q[L])}{\max(q[1], q[L]) - \min(q[1], q[L])} * [\max(c, d) - \min(c, d)] \quad + \quad \min(c, d) \qquad (4.4)$$

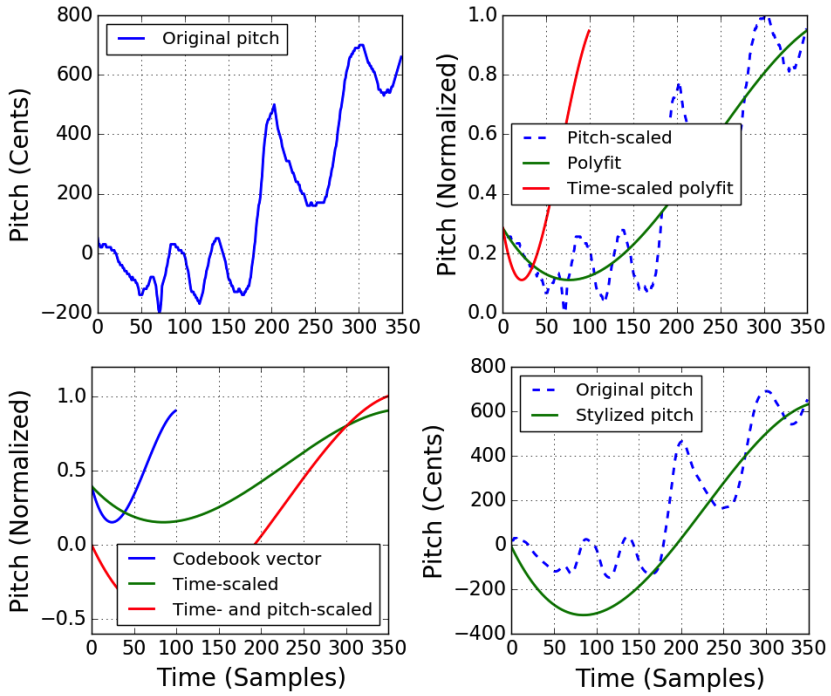

Figure 4.13: Steps of (de)normalization of a transient segment to the corresponding stylized contour. The corresponding codebook vector (bottom left) is of index 7. Bottom right shows the (superimposed) original and the stylized contours. The time unit is shown in samples on purpose.

We observe from Figure 4.13 that the pitch oscillations are neglected and a lowpass trend is obtained as a stylized transient segment. This phenomenon is favorable for retrieval applications, because the oscillations (lit. gamak) is not an essential, but occasional, part of a melodic phrase. Musicians choose to either apply or omit a gamak on based on local context. Hence we are at a better chance of retrieving a phrase independent of the presence of gamak. We have not added smoothness constraints at end-points, a derivative-based approach to ensure smooth transition between stable notes and transients could be worth investigating.

4.4.3 Event segmentation for melodic motifs

A motif is a sequence of svaras whose melodic realization includes specific intonations and transitions to/from neighboring svaras [102]. While computational models for measuring melodic similarity between phrases have employed distance measures between time-series of pitch values of the phrase segments, we might expect that a more discriminative representation is possible by explicitly incorporating features that contrast between ragas in question. This is reminiscent of the 'discriminative DTW' approach of Rabiner [143, Sec. 5.6]. In his 'discriminative methods in speech recognition' especially in the DTW framework – towards identifying 'discriminative' parts of phones/letters/words from their temporal representation and learning the discriminative weighting of their constituent parts by a Fisher discriminant like approach. We take a similar approach to identify discriminative features of raga characteristic phrases. Figure 4.14 shows a representative GRS phrase from each of the ragas. Distinctive features suggested by the comparison are: (i) durations of each of the stable svara regions, (ii) the durations of the glides connecting the svaras, and (iii) the pitch interval of the svara G. The implementation of these features would involve decisions on segmentation of stable svaras, and determining the pitch interval value from the pitch continuum in the region. Further, it is important to figure out the kind of normalization that is needed to reduce possible variability due to the tempo of the performance.

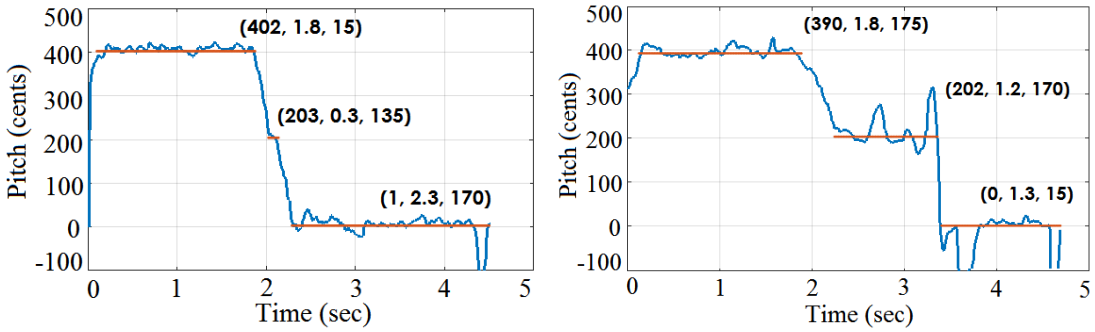

Figure 4.14: Melodic contours (blue: pitch versus time) arbitrarily chosen GRS phrases from ragas Deshkar (left) and Bhupali (right). The horizontal lines (orange) indicate the segmented svaras, the tuple corresponding to each svara denotes the extracted features ($Intonation$, $Duration$, $Slope$) for the corresponding events. The test phrases in the plots are taken from concerts DK_AC-1 and BP_AC-1.

We describe a phrase as a sequence of melodic 'events' that can each be described by the chosen features. For the GRS phrase in question, we consider the following five events, i.e. svaras G, R, S, and the G − R and R − S transitions. The selected features are: (i) $Start\_time$:

onset of an event, (ii) End_time : offset of an event, (iii) $Duration$: difference of the two, (iv) $Intonation$: precise pitch interval location of a stable svara in the octave obtained as the median pitch value over the duration of the svara, and (v) $Slope$: gradient between the mean of last 20% and the first 20% pitch samples of a stable svara segment.

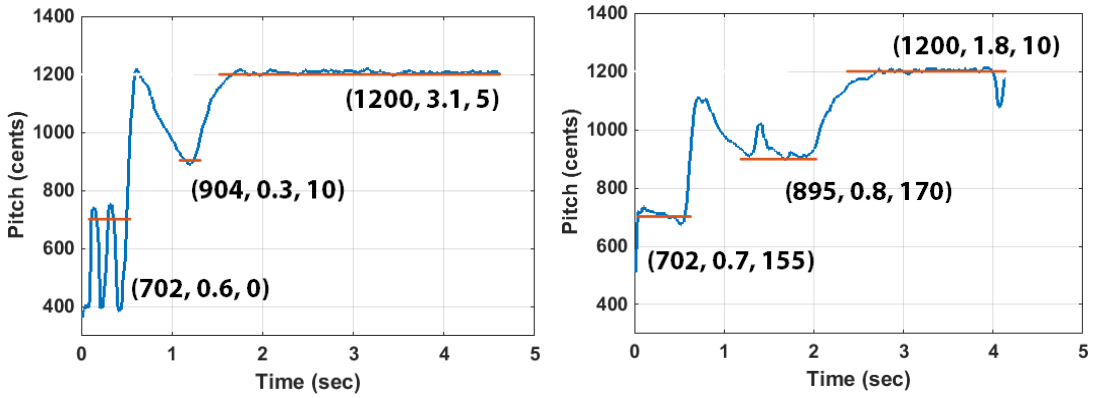

Figure 4.15: Melodic contours (blue: pitch versus time) arbitrarily chosen PDS phrases from ragas Deshkar (left) and Bhupali (right). The horizontal lines (orange) indicate the segmented svaras, the tuple corresponding to each svara denotes the extracted features ($Intonation$, $Duration$, $Slope$) for the corresponding events. The test phrases in the plots are taken from concerts DK_AC-1 and BP_AC-1.

One may suspect from Figure 4.15 that the PDS phrases in raga Deshkar have a lot of alankars. The observation is not generalizable though. It is true that Deshkar PDS phrases have consistent gamak/kan (also seen from the hierarchy in [65]) whereas Bhupali does not show the same. The common shape of the PD glide is that it always has an overshoot towards upper S. For the current study we did not consider the glide shape but only duration. Figure 4.16 presents 6 PDS instances from raga Deshkar concert by Ajoy Chakrabarty (DK_AC-1). We observe that the ornamentation on the P note is not consistently present (only present in subplots 4, 5). However, the overshoot to S before the D note is common to all of them. Note that the subplot 3 presents PDS phrase in lower octave.

Note that we have given 3 different svara segmentation (transcription to obtain onset/offsets and labeling) methods. These are, in principle, different implementations of the same philosophy i.e. segmenting pseudo-steady notes which are within a pitch tolerance around a given pitch interval for a contiguous threshold duration. For the experiments to be discussed in the subsequent chapters, the event segmentation method will use the transcription as described in Algorithm 1. To summarise, we break down the event segmentation steps as follows.

- Find the continuous pitch distribution (10 ms intervals, 12.5 cents binning) across the

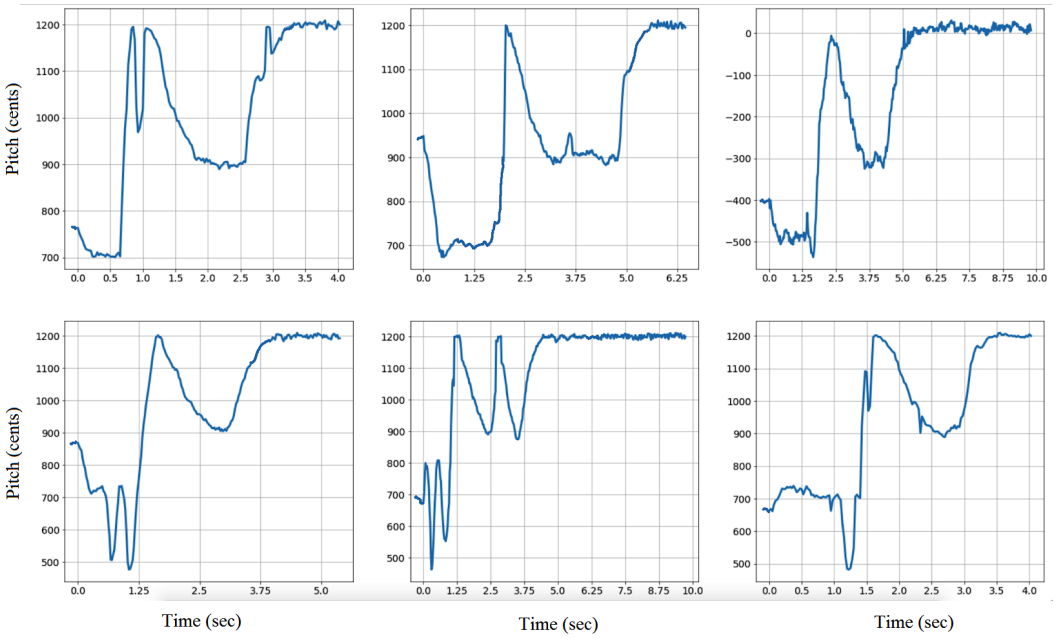

Figure 4.16: Instances of PDS phrase from raga Deshkar by Ajoy Chakrabarty (DK_AC-1).

vocal regions of the entire concert.

- Find the peaks in the above distribution corresponding to the notes of interest e.g. G, R, S and P, D, S.

- For each pre-segmented phrase, consider the continuous pitch vs time (10 ms) is obtained.

- The pre-segmented phrase was coarsely marked by a musician and therefore contains regions slightly beyond the desired phrase boundaries. In this contour, we mark bands of width ± 35 cents around the note pitch as determined from the picked histogram peaks.

- Mark all fragments of duration > 250 ms as notes. Next merge all fragments separated by < 100 ms.

- Now we have only 3 note segments e.g. P, D, S. Each of the inter-note regions is a "transient between the corresponding notes". We measure the durations of each of the 3 note regions and 2 transient regions. We also measure the median pitch of each of the note regions.

4.5 Distance measures between pitch distributions

An end-to-end system for melodic similarity consists of a distance computation block. The distance measures are often coupled with the representation; e.g. comparing two continuous melodic shapes would demand a vector measure like dynamic time warping (DTW) distance, whereas we need a edit distance-based measure like Smith-Waterman distance to compare a discrete sequence of strings. Again, certain distance measures (e.g. KL divergence, Bhattacharyya distance) are probabilistic measures and hence can be applied only to distributions (pitch histograms). Certain other distance measures like the Correlation distance can universally be applied to both distributional and structural representations, provided that the two metrics are of the same length.

Finally, we need a distance measure computable between the histogram representations that correlates well with closeness of the compared performances in terms of raga identity. There exist several distinct distance measures between probability distributions with different physical interpretations [31]. In the case of first-order pitch distributions, we are looking for a similarity in the tonal hierarchy captured by the distribution. The psychological model of Krumhansl [99] is the most influential one and presents one of the most frequently applied distance measures in previous studies, the Correlation distance [70]. This measure is often used in cognitive studies. This measure does not require the compared entities to be probability distributions but rather any two patterns of same dimension. The correlation distance is given by:

$$d_{\text{corr}}(P,Q) = \frac{\sum_{i=1}^{N}(p_i - \overline{p})(q_i - \overline{q})}{\sqrt{\sum_{i=1}^{N}(p_i - \overline{p})^2(q_i - \overline{q})^2}} \tag{4.5}$$

where P, Q refer to the two distributions under test, p_i, q_i are the masses of i^{th} bins of the distributions, and $\overline{p}$, $\overline{q}$ are the means.

We consider the Euclidean, d_{eucl} (L-2 norm), and City-block, d_{ctbl} (L-1 norm) distances as they have been successfully used for pitch histogram similarity in previous studies. [70] advocate the City-block (Manhattan) distance for its superior performance in the shift-and-compare method for automatic tonic detection and they used the same for Makam recognition. The Euclidean distance between two n-dimensional vectors P and Q in Euclidean n-space is the length of the line segment connecting them, given by:

$$d_{\mathrm{eucl}}(P,Q) = \sqrt{\sum_{i=1}^{N}(p_i - q_i)^2} \tag{4.6}$$

The Cityblock (also known as Manhattan) distance between two vectors P, Q in an n-dimensional real vector space $\mathbb{R}$, is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes, given by:

$$d_{\mathrm{ctbl}}(P,Q) = \sum_{i=1}^{N}|p_i - q_i| \tag{4.7}$$

We also consider the Bhattacharyya distance as a suitable measure for comparing distributions. It is reported to outperform other distance measures with a PCD-based, as well as with higher-order distribution based, features in the raga recognition task [36, 79]. For two probability distributions P and Q over the same domain, the Bhattacharyya distance measures the similarity between the distributions, and is given by:

$$d_{\mathrm{bhat}}(P,Q) = -\log\left(\sum_{i=1}^{N}\sqrt{p_i \cdot q_i}\right). \tag{4.8}$$

Another good distance measure for comparing distributions should reflect the extent of similarity between their shape. We choose the Kullback–Leibler (KL) divergence measure [103] as a suitable measure for comparing distributions. Symmetry is incorporated into this measure by summing the two values which was first introduced to PCD-based MIR studies by [18]. The KL distance between two probability distributions is a measure of how one distribution diverges from the second, given by:

$$d_{\mathrm{KL}}(P||Q) = \sum_{i=1}^{N} p_i \cdot \log\left(\frac{p_i}{q_i}\right) \tag{4.9}$$

$$D_{\mathrm{KL}}(P,Q) = d_{\mathrm{KL}}(P||Q) + d_{\mathrm{KL}}(Q||P) \tag{4.10}$$

where i refers to the bin index of the distribution, and P and Q refer to the distributions of two concerts under test.

4.5.1 A note on dynamic programming approaches

The most commonly used distance metric, implemented by dynamic programming, is the dynamic time warping (DTW) distance. The pitch sequences are warped in a non-linear fashion

and hence DTW has been extensively used to find the alignment between the reference and candidate sequence. The Euclidean distance between the aligned time-series samples is referred to as DTW distance. Previous work [151] has shown the relevance and capacity of DTW distance to classify raga-characteristic melodic motifs. In this chapter, we shall look at the required modifications for a partial matching of time-series (or subsequence matching) and relevant configurations of the different parameters involved.

The objective of DTW is to compare two (time-dependent) sequences $X := (x_1, x_2, ..., x_n)$ of length $n \in \mathbb{N}$ and $Y := (y_1, y_2, ..., y_m)$ of length $m \in \mathbb{N}$. These sequences may be discrete signals (time-series) or, more generally, feature sequences sampled at equidistant points in time. Typically, the cost function $c(x, y)$ is small (low cost) if x and y are similar to each other, and otherwise $c(x, y)$ is large (high cost). Then the goal is to find an alignment between X and Y having minimal overall cost. An (n, m)-warping path $p = (p_1, p_2, ..., p_l)$ defines an alignment between two sequences X and Y by assigning the element x_{n_l} of X to the element y_{m_l} of Y. The boundary condition enforces that the first elements of X and Y as well as the last elements of X and Y are aligned to each other. In other words, the alignment refers to the entire sequences X and Y . The monotonicity condition reflects the requirement of faithful timing: if an element in X precedes a second one this should also hold for the corresponding elements in Y, and vice versa. Finally, the step size condition expresses a kind of continuity condition: no element in X and Y can be omitted and there are no replications in the alignment (in the sense that all index pairs contained in a warping path p are pairwise distinct). Furthermore, the DTW distance generally does not satisfy the triangle inequality even in case c is a metric. For $1 < i \leq n$ and $1 < j \leq m$:

$$D(i, j) = min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\} + c(x_n, y_m) \tag{4.11}$$

The Smith-Waterman algorithm does local sequence alignment of two sequences. Given two strings, it compares segments of all possible lengths and gives a score based on a similarity measure. The algorithm doesn't penalize the gaps occurring before the first symbol and after the last symbol of any of the strings. The parameters are: (i) similarity measure: this measure assigns a positive for each pair of symbols that are aligned, (ii) gap penalty: this measure assigns a penalty for not aligning a sequence of k successive symbols of one string with any symbol in the other string. It assigns a score for each position p of the song string corresponding to an alignment with the query string where the last character of the query aligns with position p in

the song. There could be many alignments of the query string which end at the same position p, but the assigned score corresponds to the alignment giving maximum score. Let the length of the song string be n and the length of the query string be m. In the dynamic programming implementation of this algorithm, we maintain $(m+1)*(n+1)$ matrix H of scores where the rows are indexed from 0 to m and columns are indexed from 1 to n. The i^{th} row of the matrix gives the scores for each position in the song index for a query string corresponding to the first i characters of the original query string.

The 0^{th} row and column of this matrix are initialized to 0 and the remaining entries are calculated by the following procedure

$$H(i,0) = 0; 0 \leq i \leq m$$
$$H(0,j) = 0; 0 \leq j \leq n$$
$$H(i,j) = max \left\{ \begin{array}{cc} 0 & \\ H(i-1,j-1) + s(a_i, b_j) & Match/Mismatch \\ max_{k \geq 1} H(i-k,j) + W_k & Deletion \\ max_{l \geq 1} H(i,j-l) + W_l & Insertion \end{array} \right\}; 1 \leq i \leq m, 1 \leq j \leq n \tag{4.12}$$

where W_k corresponds to gap length of k.

4.5.1.1 Comparing the time complexities

DTW method looks for all windows of length $\frac{m}{2}$ to $2*m$ in the song and computes the DTW distance between the query and all such windows. Hence, for each time instant we consider $(2*m - \frac{m}{2}) = \frac{3*m}{2}$ windows. Thus, a total of $\frac{3*mn}{2} \sim O(mn)$ windows are considered over the whole song. Since the length of each window is $O(m)$ and the length of query is n, we incur $O(m^2)$ time in calculating DTW distance for each of $O(nm)$ windows. Hence the total time complexity of this algorithm is $O(nm^3)$.

In the case of string edit distances, calculating the $(i,j)^{th}$ entry of the matrix needs a total of $O(i+j)$ comparisons. Hence, the running time of the algorithm is

$$\sum_{i=1}^{m} \sum_{j=1}^{n} O(i+j) = \sum_{i=1}^{m} \sum_{j=1}^{n} O(m+n) = O(mn^2 + m^2 n) = O(n^2), \because m \ll n \tag{4.13}$$

In the above procedure, we are allowing for all possible gap lengths. However, practically an alignment having a continuous gap of size equal to the query length might be the optimum

amongst all alignments ending at the same position but it is very likely to not be a match good match for the query. Since we are interested in finding a very good match t the query string, we can limit the maximum gap size to the length of the query. This essentially makes our algorithm linear in terms of size of song string. The running time of the modified algorithm is

$$\sum_{i=1}^{m}\sum_{j=1}^{n} O(i+j) = \sum_{i=1}^{m}\sum_{j=1}^{n} O(m) = O(m^2 n) = O(n), \because m \ll n \tag{4.14}$$

Thus we are able to achieve a linear-time query search algorithm, over a modified brute force search which is $O(n^3)$. This is a reasonably good improvement, and might facilitate fast (if not real time) search through large databases if the symbolic representation is precomputed and indexed.

4.6 Statistical model fit: regression

One of the important statistical model that we shall use is the regression model. In Chapter 6, we will see the use of both linear and multiple regression models for objectively understanding the prediction power of different melodic and contextual features, and their possible coupled effect. In Chapter 7, we will use the logistic regression model for modeling the categorical nature of music perception by trained musicians.

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables, i.e. the average value of the dependent variable when the independent variables are fixed.

One of the inputs to the logistic regression model is the raw outcome (x) and the second input to the model is a categorical array of the stimulus continuum. We provide a categorization of 5 class A + 5 class B candidates for an example problem to understand the model. The model is fitted in a logged odds space. The log odds transformation converts proportions in the range 0 to 1 into logits in the range $-\infty$ to $+\infty$. Logit values from the fitted model can be

converted to probabilities so that fitted curves in the log odds space become sigmoidal curves in the probability space. The model included a bias coefficient, and a duration-tuned coefficient, as given by Equation 7.4.

$$ln(p/(1-p)) = c + m * x => p = \frac{1}{e^{-(c+m*x)} + 1} \tag{4.15}$$

where p is the expected value in a category, c is the intercept i.e. displacement along x-axis, and m is the regression coefficient [120] i.e. the growth rate.

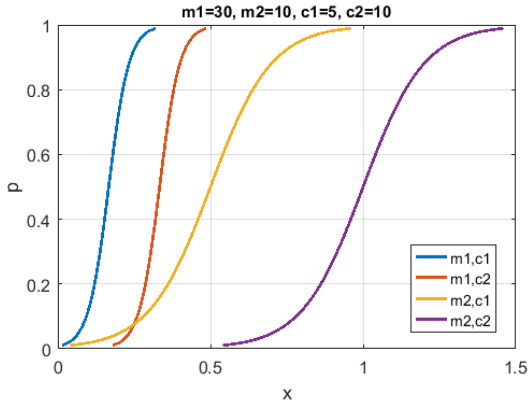

Figure 4.17: Effect of m and c in logistic regression model. We observe that m governs the tilt while c is the horizontal shift.

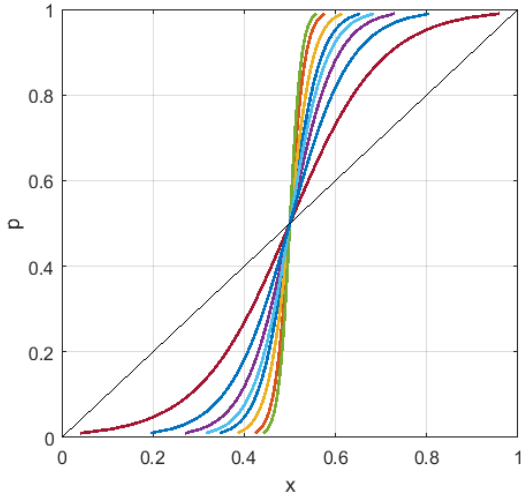

Figure 4.18: Coupled effect of m and c in logistic regression model. The relation $m = -2 * c$ satisfies the coordinate (0.5,0.5) and the slope/intercept covaries for different sets of (m, c).

Chapter 5

The distributional view

5.1 Motivation

Raga performance allows for considerable flexibility in interpretation of the raga grammar in order to incorporate elements of creativity via improvisation. It is therefore of much interest in pedagogy to understand what ungrammaticality might mean in the context of a given raga, and possibly develop means to detect this in an audio recording of the raga performance. One prominent notion is that ungrammaticality is considered to occur only when the performer "treads" on another, possibly allied, raga in a listener's perception. With this view, we consider modeling the technical boundary of a raga as that which separates it from another raga that is closest to it in its distinctive features. We wish to find computational models that can indicate ungrammaticality using a data-driven estimation of the model parameters; i.e. the raga performances of great artists are used to obtain representations that discriminate most between same and different raga performances. We choose three well-known pairs of allied ragas in Indian art music for an empirical study of computational representations and coupled distance measures for the distinctive attributes of tonal hierarchy.

[0]This chapter is largely drawn from the following papers:

- K. K. Ganguli and P. Rao. "On the distributional representation of raga and melody," Transactions of International Society for Music Information Retrieval (TISMIR), 1(1): 79–95, 2018. [65]
- K. K. Ganguli and P. Rao. "Towards computational modeling of the ungrammatical in a raga performance," in Proc. of the 18th International Society for Music Information Retrieval Conference (ISMIR), October 2017, Suzhou, China. [63]

5.2 Background

An important function of our proposed computational model would be to capture the notion of grammaticality in performance. Such an exercise could eventually lead to computational tools for assessing raga performance accuracy in pedagogy together with the complementary aspect of creative skill. A popular notion of grammaticality in performance is preserving a raga's essential distinctiveness in terms of the knowledgeable listener's perception [14, 37, 145]. Thus, a performance with possibly many creative elements is considered not to transgress the raga grammar as long as it does not "tread on another raga" [102, 145, 199]. The technical boundary of a raga should therefore ideally be specified in terms of limits on the defining attributes where it is expected that the limit depends on the proximity of other ragas with respect to the selected attribute. We therefore consider deriving a computational representation of distributional information based on maximizing the discrimination of "close" ragas.

A goal of this study is to develop a distributional representation for raga music that links the mandatory raga grammar with observed performance practice. The parameters of the computational model are derived by maximizing the recognition of ungrammaticality in terms of the distributional information in a performance of a stated raga [63]. We consider the structural information in the form of phrase characteristics in Chapter 6. We use audio recordings of performances of eminent Hindustani vocalists as very creatively rendered, but grammatically accurate, references of the stated raga. Performances in the allied raga serve as proxies for the corresponding ungrammatical performances. The obtained model will further be applied to obtain insights on consistent practices, if any, observed in raga performances apart from what is specified by the grammar.

As shown in the review in Chapter 2, several methods based on pitch distribution have been applied to the raga recognition task. Although the outcomes are expected to depend on the design of the dataset, this aspect has received hardly any careful consideration. The test datasets used in previous work typically comprised a number of performance audio recordings arbitrarily chosen for an equally arbitrarily chosen set of ragas. In the face of this diversity of datasets, it is difficult to justify the conclusions or predict how the results generalize to other datasets of performances and ragas. The ragas in the test sets often correspond to different scales; given this distinction in the set of notes, the precise implementation of the first-order distribution is probably not relevant. We propose to develop and tune the parameters of a computational

representation for the distribution using a dataset and evaluation methods that are sensitive to changes in the parameters within the reasonable space of parameter choices. This is achieved with the use of allied ragas and a more musicologically meaningful criterion related to the technical boundary of the raga in the distributional feature space.

There has also been a lack of attention, in the literature, to the choice of probability distribution distance measures. This, of course, partly owes itself to the emphasis on classifier-based approaches with input melodic features such as first-order distributions. Here the focus has been on gross performance of the raga recognition system in terms of classification accuracy rather than on obtaining insights into the computational equivalents of musicological concepts. [95–97] proposed a handful of representations (in terms of first-order pitch distributions) but always applied the KL divergence as a distance measure. In contrast, [70] used a bunch of distance measures but did not report their sensitivity to the different possible pitch distribution representations. While [39] have commented on many possible configurations of the bin centres/edges and their precise locations, we do not expect the precise locations of bin centres/edges to affect the performance of a high bin-resolution representation in the raga recognition task. In this chapter, we systematically investigate the choice of bin width and distance measure for continuous pitch distributions computed from performance recordings. We also consider discrete-pitch representations derived from svara-segmented note regions. Given that melody transcription for raga music itself is a challenging (or rather, ill-defined) task [206], it is necessary to rely on heuristic methods for svara segmentation and transcription. We present a section to advocate our parameter tuning choices through the framework of current experiments. The dataset used in these experiments are discussed in Chapter 3. We next discuss the evaluation criteria and experiments for the distributional attribute of raga melodies.

5.3 Evaluation criteria and experiments

With a view to identifying the choices in bin-width, type of histogram and distance measure between histograms that best serve in the representation of tonal hierarchy for raga music, we present experiments on our dataset of allied raga performances. The evaluation criteria relate to the achieved separation of performances belonging to different ragas in an allied pair. More specifically, we evaluate the performance of unsupervised clustering with k-means (k=2) with its implicit Euclidean distance measure on each set of allied raga performances. The performance

in unsupervised clustering can be quantified by the cluster validation measure *cluster purity* (hereafter Clu_Pur) which is obtained by assigning each obtained cluster to the underlying class that is most frequent in thae given cluster, and then computing the classification accuracy.

$$Clu_Pur = \frac{1}{N}\sum_{i=1}^{k} max_j \left|c_i \cap t_j\right| \tag{5.1}$$

where N is the number of data points, k is the number of clusters, c_i are the clusters and t_j are the underlying classes. In our context (k=2), a Clu_Pur value of 1 indicates perfect clustering, whereas 0.5 implies random clustering.

We also evaluate the effectiveness of the distance between representations in estimating the "grammaticality" of a given concert with reference to a selected raga represented by a histogram *template*. For example, every performance in the Deshkar-Bhupali dataset serves once as the grammar template of the corresponding raga. This template is paired with every performance in the 2-raga data subset to obtain grammatical and ungrammatical instances of pairings with reference to the chosen template. Likewise for the other allied raga-pair subsets. We term the inverse of the distance computed between the concert histogram and the template histogram in a pair as the *grammaticality coefficient* of the corresponding concert. A low value of the coefficient would indicate an ungrammatical performance with reference to the given raga template. The receiver operating characteristic (ROC) can serve to evaluate the efficacy of this measure across the entire dataset of performances.

An ROC curve [52] provides a visualization of the trade-off between the true positives and false positives in a detection context. We consider our context to be the detection of ungrammatical instances (pairs) from the complete set of pairs. An ROC curve is obtained by varying the threshold applied to the obtained array of grammaticality coefficients, and computing the true positives and false positives. Given a raga template histogram, the detection of an ungrammatical instance is considered a true positive (TP) if the performance under test belongs to the allied raga. It is considered a false positive (FP) if it belongs to the same raga as the template histogram. To evaluate the different tuning parameters of the tonal hierarchy representation, we use the area under curve (AUC) measure (higher values are better) and the Equal Error Rate (EER), where the false positive rate equals the false negative rate (lower values are better).

In summary, there are two main features of the tonal hierarchy model under investigation: (i) the histogram representation, (ii) the between-histograms distance measure. Both continuous-pitch (various bin-widths) and segmented-svara options are tested in combination

with a number of distance measures between histograms.

5.3.1 Experiment 1: unsupervised clustering

Our base representation is the octave folded pitch salience histogram, normalized so that it is interpreted as a probability distribution. We test with different uniform bin widths, ranging from 1 to 100 cents, with centres coinciding with the tonic and semitone locations and their integer sub-multiples.

Figure 5.1 shows the cluster purity values at the considered bin widths (namely, 1, 12.5, 20, 25, 27, 30, 35, 40, 50 and 100 cents). We note that no degradation in purity is observed for 1 through 30 cent bin resolution. Each value on the curve is obtained by an average of 5 runs of the clustering algorithm using different initializations. For the cases of svara salience and count histograms, the average cluster purity values obtained are 0.96 and 0.84 respectively, indicating the slight superiority of the higher dimensional continuous-pitch distributions. Note that the svara histogram performances are considerably higher than the performance of the pitch salience distribution of the same dimensions (100-cent bin width). That the observed clustering in all configurations actually captures raga characteristics was confirmed by noting that each discovered cluster was heterogeneous in performer and performance metadata.

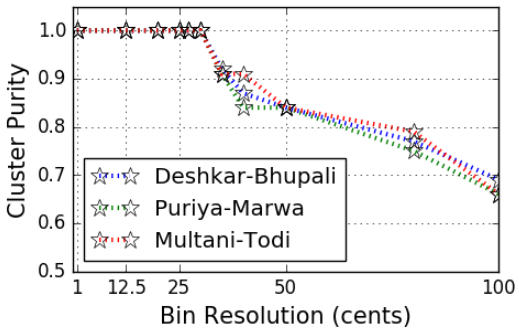

Figure 5.1: Cluster purity (CP) values obtained for different values of bin resolution for the pitch salience histograms in each of the 3 raga-pairs.

5.3.2 Experiment 2: detecting ungrammaticality

For the dataset of concerts corresponding to a given allied raga pair, we create pairs of each concert with every other concert in the set. For example, we obtain 17x17 (=289) pairs out of the 17 concerts in the Deshkar-Bhupali allied raga dataset. Of these, 2x6x11 (=132) pairs

contain mismatched ragas and hence serve as ungrammatical instances. With this logic, we obtain a total of 1011 (= 17x17 + 19x19 + 19x19) distance values across the dataset of the 3 allied raga pairs with 468 instances corresponding to the ungrammatical pairings (i.e. a raga and its corresponding allied raga).

The ROCs of Figure 5.2 show the true positive rate versus the false positive rate achieved in the detection of ungrammatical instances for each of 3 histogram representations and all the 4 considered distance measures. The 3 histogram representations are the pitch salience with 25 cent bin width (p=48), given that our previous experiment indicate undegraded clustering up to this bin width, and the 2 svara based histograms. The ROCs for the individual allied raga pairs are provided in the Appendix (Figures I, VI, and IX). The ROC evaluation measures, AUC and EER, for each allied raga-pair subset are presented in Table 5.1. The similarity in ROC shapes across raga-pairs indicates that the computed distance measure is independent of raga, and can serve as a robust measure of ungrammaticality that does not need recalibration with change of raga. Accordingly, a common threshold was applied to each set of 1011 distances computed over the full dataset to obtain the corresponding ROC in Figure 5.2.

5.4 Discussion

5.4.1 Performance across allied raga-pairs

Salient aspects of the experimental outcomes presented in Figure 5.2 and in Table 5.1 are summarized next.

- The histogram representation obtained from the continuous melodic contour at the finest bin resolution (p=96) with correlation distance is either as good as or, sometimes, better than any of the svara-based histograms. This indicates that capturing melodic movements such as glides and ornaments in the distributional representation is of value over relying on the stable segments only.

- The pitch salience histogram with p=48 comes close to the performance of p=96 with the correlation distance measure but is overall slightly worse with the other distance measures. As bin width is increased further to obtain the p=24 and p=12 pitch salience histograms, we note a sharp degradation, irrespective of the distance measure, in both the AUC and EER values.

Allied raga-pair	Distance measure	Evaluation metric	Pitch salience p=96	p=48	p=24	p=12	Svara salience	Svara count
Deshkar-Bhupali	Correlation	AUC	**.98**	.97	.95	.89	.85	.90
		EER	.04	.06	.08	.19	.19	.15
	Euclidean	AUC	.97	.95	.92	.86	.84	.88
		EER	.08	.09	.15	.20	.18	.18
	City-block	AUC	**.98**	**.98**	.94	**.90**	.85	.88
		EER	.05	.05	.11	.18	.20	.18
	Bhattacharyya	AUC	**.98**	**.98**	**.96**	.81	**.93**	**.95**
		EER	.03	.05	.10	.25	.09	.10
Puriya-Marwa	Correlation	AUC	**.99**	**.95**	**.90**	**.84**	.93	.94
		EER	.02	.08	.18	.23	.15	.13
	Euclidean	AUC	.92	.90	.82	.80	.94	**.95**
		EER	.13	.17	.19	.28	.15	.13
	City-block	AUC	.94	.93	.82	.77	.95	**.95**
		EER	.09	.12	.21	.29	.13	.11
	Bhattacharyya	AUC	.95	.87	.79	.72	**.98**	.94
		EER	.07	.19	.23	.42	.03	.11
Multani-Todi	Correlation	AUC	**.98**	**.98**	**.93**	**.94**	.97	**.99**
		EER	.04	.05	.12	.15	.06	.02
	Euclidean	AUC	.95	.96	.87	.89	.93	.97
		EER	.10	.08	.18	.16	.12	.06
	City-block	AUC	.96	.97	.92	.92	.96	.98
		EER	.08	.05	.12	.11	.08	.05
	Bhattacharyya	AUC	.90	.93	.90	.89	**.99**	.98
		EER	.17	.17	.11	.16	.03	.03

Table 5.1: Summary of results: evaluation measures AUC and EER for all combinations of representations and distance measures for each of the three allied raga-pairs. In bold font is the highest AUC across distance measures for a given representation and raga-pair.

- The svara-based histograms show clearly superior performances relative to pitch salience histograms of comparable dimension (p=12). The Bhattacharyya distance works best for svara based histograms and this performance comes close to that of the p=96 pitch salience histogram.

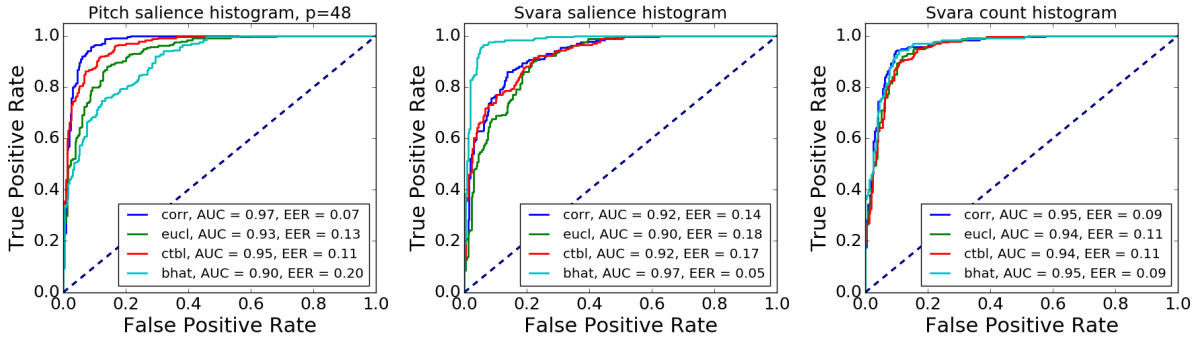

Figure 5.2: Combination of all three raga-pairs (full concerts): ROCs obtained for four different distance measures from pitch salience (left), svara salience (middle), and svara count (right) histograms from the combined distance vectors for all three raga-pairs.

5.4.2 Bin resolution: coarse or fine?

Given that the minimum interval between svaras in raga music is a semitone, one may argue that a 12-bin pitch-class distribution should be sufficient to discriminate ragas with different distributional information. However for ragas which share a scale, as with the allied ragas, a finer bin resolution may bring in further value by capturing differences in the precise intonation of the svaras. For example, this is the case with the Deshkar-Bhupali pair in Table 3.1 where at least 3 svaras ($\mathrm{R}, \mathrm{G}, \mathrm{D}$) have a difference in intonation (*shruti*) for the same scale degree. The question arises about how fine a bin resolution is needed to capture the intonation differences. [39] reported 27 cents as the optimal bin resolution ($p = 44$) for visually locating the precise shruti positions of different svaras. Our findings, in terms of cluster purity measure in Figure 5.1 agree with the observation in that no degradation is observed for 1 through 30 cent bin resolution. However the ROC based evaluation of the grammaticality coefficient showed a slightly improved performance, in terms of AUC measure, for a finer ($p = 96$) bin resolution over a coarser one ($p = 48$). While this has a theoretical justification in terms of the fine intonation differences (e.g. raga Deshkar uses a higher *shruti* of G, by an order of 10 cents), we note that the svara histograms (where such intonation information is lost) perform nearly as well. This

indicates that the relative saliences of the svaras (both in terms of duration and frequency of occurrence), as implemented here, are adequate features as well.

5.4.3 Validation of svara transcription parameters

It is interesting also to consider the computed histograms in terms of the distributional information provided by music theory. We consider here the information captured by the distribution in terms of musicological interpretations. From the pitch salience histograms (p=1200), we observe certain phenomena which are musicologically interesting. In Figure 4.4, there is a small peak observed for N svara (1100 cents $\approx 1100^{th}$ bin) for raga Deshkar, which is theoretically a *varjit* (disallowed) svara, but its salience is comparable to that of the allowed R svara. In the svara salience (and count) histograms in Figures 4.5 and 4.6, the peak corresponding to the N svara (11^{th} bin) is insignificant in comparison to that of the R svara. This indicates that the usage of the N svara is different. We confirmed, by interviewing musicians (including a couple of artists from our dataset), that the N svara is used as a *kan* svara (touch note) in raga Deshkar. This contributed to the pitch salience histogram, but not the svara histograms computed via the stable svara segmentation step. We see that our empirically chosen segmentation parameters (T_{tol} = 35 cents, T_{dur} = 250 ms) provide a representation that is consistent with the theory.

Further, the chosen svara segmentation parameters ensure that the correlation between the svara salience histogram and the svara count histogram is high. If T_{dur} is set less than 250 ms, the *varjit* (disallowed) svaras would appear in the svara count histograms (the svara salience histogram would not be similarly affected because of the short durations). Additionally, the slow glides would get segmented into svaras and add to the count in the svara count histogram. In contrast, if T_{dur} is set higher, the svaras with *alpatva* (shorter duration, e.g. R svara in raga Deshkar) usage would go undetected and hence vanish from both svara salience and count histograms. This would lead to an inaccurate representation of the raga grammar.

5.4.4 New insights on time scales

Given that the proposed histogram representations capture the distributional information in the concert, it is of interest to investigate the time scale at which the estimated tonal hierarchy can be considered to be stable and therefore representative of the raga. We carry out the previous allied raga discrimination experiments on segmented concerts. We divide each concert uniformly

into n segments (n = 1, 2, $\cdots$, 5) and construct the array of grammaticality coefficients across all the pairings associated with the set of $\left(\frac{1}{n}\right)^{th}$ duration segments. The goal is to determine the smallest fraction of the full concert that is necessary to robustly discriminate between the matched and mismatched raga pairs. Here, as before, one segment acts as a reference template for its raga grammar; the other member of the pair is a segment drawn from either the same raga (giving us a grammatical instance) or from the allied raga (giving us an ungrammatical instance). In every case, the distance measure chosen is the one with the best performance for the given histogram representation as indicated by Figure 5.2.

ROCs computed for segmentations with $n \geq 4$ were seen to lead to AUC < 0.5, which indicates that this time scale is too small to create a stable tonal hierarchy. We therefore consider only the cases of half and one-third segmentation further, giving us two datasets of concert segments of sizes 110 and 165 segments respectively from the original set of 55 full concerts. Figure 5.3 shows a comparison of ROCs across the full (n=1) and partial (n=2,3) concerts for the various representations where, as in the case of full concerts, the considered pairs for distance computation are constrained to be within allied-raga data subsets.

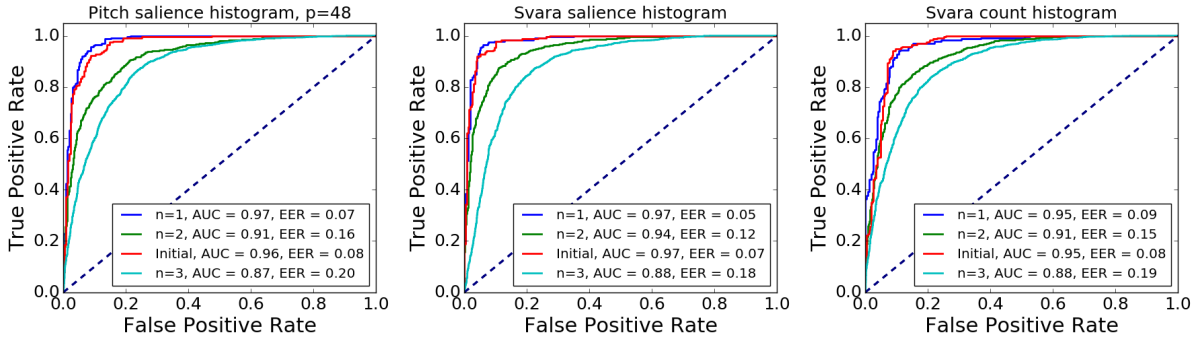

Figure 5.3: Combination of all three raga-pairs (partial and initial concerts): Comparison of ROCs obtained with best distance measures (Correlation distance for pitch salience and Bhattacharyya distance for svara histograms) at different time-scales (n=1,2,3) and with the initial portion (*alap+vistar*).

5.4.4.1 Raga delineation in initial portion

A good performer is continually engaged in exposing the raga through the improvisation interspersed throughout the concert. This is particularly important in the initial phase of the concert where establishing the raga identity in the mind of the listener is the primary goal. The ROCs (from Figure 5.3) indicate that at lower than the one-half portions, the duration is too small to constitute a stable tonal hierarchy, based on averaging a number of segments drawn from dif-

ferent regions of the concert. On the other hand, the initial phase of the concert comprising the *alap* and the initial part of the *vistar* (e.g. *sthayi* or the chorus line of the composition) is considered by musicians and listeners to fully embody the raga's melodic features. The histogram representation of the initial segment would be expected to be more stable across concerts in a given raga. We consider segments corresponding to the initial slow elaboration (from start of the concert till the end of *vistar* of the first bandish) as annotated by a trained Hindustani musician. The initial portion, so annotated, typically constitutes half the duration of the concert or less than this. We note in Figure 5.3 that the ROCs of the concert-initial segments are indeed as good as those of the full concerts and superior to those obtained by considering all segments of similar duration drawn from different locations.

5.4.4.2 Distribution at cycle level

One of the smallest recognizable time-scales in the concert is that of the rhythmic cycle (*tala*). Each cycle, which can range in duration from 5 sec to 90 sec (*madhya* to *vilambit laya*), contains one or several melodic phrases. A performer typically has a plan for the overall evolution of the melodic content, based on individual and stylistic influences [191]. We explore the application of the histogram representation at the cycle level, to the visualization of local melodic features. This could be interesting in view of the fact that the *vistar* (lit. *barhat*, meaning expansion) of a raga performance refers to the gradual "unfolding" of a raga by focusing on the different svaras in succession on a broad time-scale [14, 102, 207, 208]. The precise duration spent on each svara in the course of this progression is not discussed in the musicology literature.

We select the 6 concerts in our dataset corresponding to Deshkar raga and further choose 6 concerts corresponding to Bhupali raga based on artist diversity while also retaining two same-artist concerts for the comparison of features in this context. The rhythmic cycle boundaries are marked based on the detection of the main accent (*sam*) location [161, 177]. We compute the svara salience histogram corresponding to each *tala* cycle of a concert. Figure 5.4 shows the histogram versus cycle index for three concerts in raga Deshkar. The svara salience peaks are indicated in the color scale (dark indicates a strong peak). Figure 5.5 shows the same representation for three concerts in raga Bhupali. We choose to show two concerts by the same artist (Ajoy Chakrabarty) in both the ragas. As the concert progresses in time, we observe a clear shifting of the location of the most salient svara in a cycle as well as variation in the melodic range covered in the cycle. The salient svara is seen to move from the lower to higher pitches

accompanied by an overall increase in melodic range from the beginning of the concert to the end. The nature of the above variation is similar across the three concerts of each given raga. While in raga Deshkar, the two concerts by the same artist Ajoy Chakrabarty (DK_AC-1 and DK_AC-2, as presented in Table I in the Appendix) show nearly identical melodic progression with respect to the *tala* cycles, the two Bhupali concerts by the same artist differ more. This indicates that the grammar of raga Deshkar, being less flexible, causes the artist to be somewhat more constrained during the improvisation.

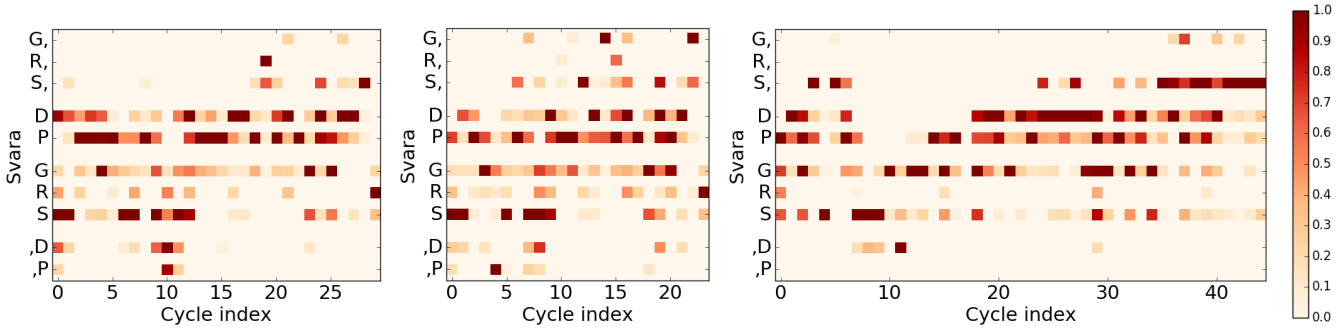

Figure 5.4: Svara salience histograms across tala cycles for DK_AC-1 (left), DK_AC-2 (middle), and DK_KA (right).

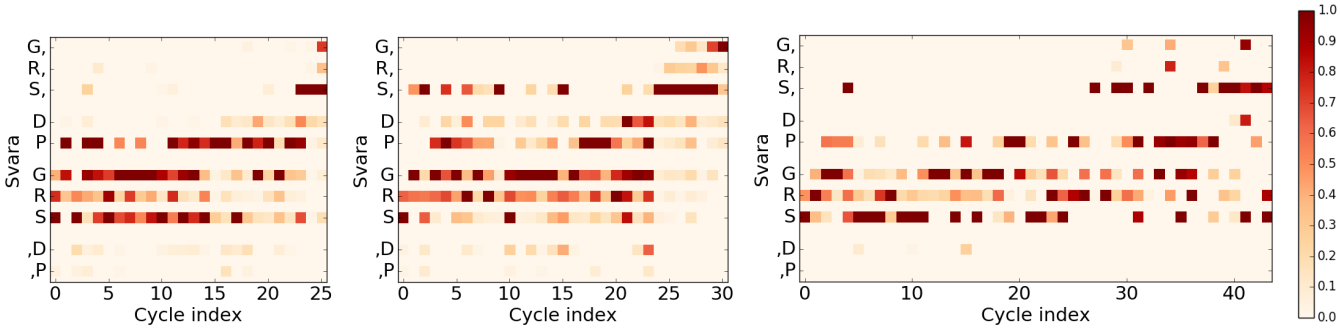

Figure 5.5: Svara salience histograms across tala cycles for BP_AC-1 (left), BP_AC-2 (middle), and BP_RK (right).

The common practice of gradually unfolding a raga over the course of the *vistar* is brought out by the time-normalized summaries of the 6 concerts of each raga in Figure 5.6. A plot of the most salient svara in each cycle versus the cycle index is computed for each concert. This contour is smoothed using a 7-point median filter. Next the individual concert summaries are each time-normalized and superposed in Figure 5.6. We clearly observe the upward shifting with time of the "focal" note through the allowed svaras. The relative duration spent on each svara is concert dependent. We note the omission of the R svara in Deshkar curves as expected from its grammar. The step-wise movement of the salient svara bears a high resemblance to the

time evolution of melody over the course of the *vistar*, as shown by [207, Figure 11], including the sharp fall in pitch of the salient svara at the end of many concerts. The latter depicts the descent to a lower svara marking the end of the *vistar*.

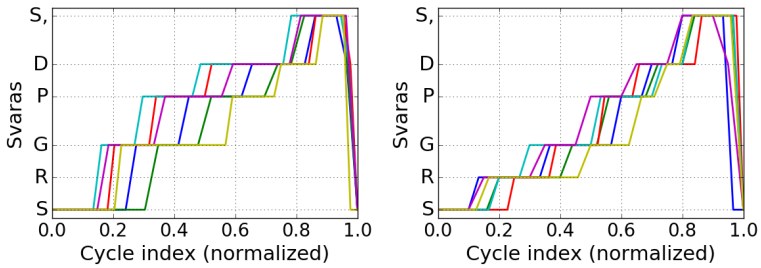

Figure 5.6: Time-normalized and smoothed cycle-level salient svara curves over the chosen 6 concerts in raga Deshkar (left) and raga Bhupali (right).

Similar observations were reported by [56], where the authors computed the svara salience histograms at the time scale of the breath phrase [1] across 75 concerts drawn from a variety of ragas and performers. By aggregation over several breath phrases, an evolution of the salient svara similar to that in Figure 5.6 was observed. The visual similarity among Figure 5.6 in the current work, Figure 11 in [207], and Figure 4 in [56] indicates that performers stick to a broad schema of progressing in the melody from a lower to a higher svara (and swiftly returning to a lower svara to mark the intended ending). This is an important result, we will see manifestation of a similar idea in Chapter 9.

5.5 Conclusion

Indian art music is a highly structured improvisational tradition. Based on the melodic framework of raga, performances display much creativity coexisting with adherence to the raga grammar. The performer's mandate is to elaborate upon the raga's essential features without compromising the cues to its identity. Raga grammar in music theory texts comprises only the essential specification of distributional and structural components of the grammar in terms of tonal hierarchies and typical phrases respectively. In the oral tradition, it is expected that there is much to be learnt from analysis of performances of great practitioners. A goal of the present work was to develop a computational representation for distributional information that can eventually be applied in the empirical analyses of audio recordings of performances. This would enable greater

[1] The authors defined "breath phrase" as a continuous segment of the melodic contour delimited by musician-intended breath pauses.

insights into the practical realization of a raga's distinctiveness in performance with potential applications to pedagogy.

The tonal hierarchy can be estimated from the detected melody line of a performance. In Western music, pitch-class profiles extracted from music pieces, both written scores and audio, have served well in validating the link between theoretical key profiles and their practical realization. In the pitch-continuous tradition of Indian art music where melodic shapes in terms of the connecting ornaments are at least as important in characterizing a raga as the notes of the scale, it becomes relevant to consider the dimensionality of the first order pitch distribution used to represent distributional information. Music theory however is not precise enough to help resolve the choices of bin width and distance measure between pitch distributions. We use a novel musicological viewpoint, namely a well-accepted notion of grammaticality in performance, to obtain the parameters of the computational representation based on audio performance data, we maximized the discrimination of allied raga performances using well-known evaluation metrics. Pitch salience histograms, as well as the stable segment based svara salience and count histograms, were considered as distinct representations of tonal hierarchy. We considered a variety of distance measures in order to derive a combination of histogram parameters and distance metrics that best separated same-raga pairs from allied-raga pairs.

It was found that svara salience histograms were as good as the high-resolution pitch salience histograms at the time-scale of full concerts, and superior to the svara count histograms. This observation continues to hold for the partial segments of concerts with a few exceptions. For the Deshkar-Bhupali dataset, we found that the continuous pitch salience histograms with the fine bin resolution of 25 cents served to capture raga grammar better for the segmented shorter portions of concerts (see the Appendix: Figure II). That is, the pitch distributions between the main peaks contributed usefully to the discrimination, indicating the importance of continuous melodic phrase shapes in this pentatonic raga-pair where the two most prominent svaras (*Vadi-Samvadi* in Table 3.1) are shared. Overall the best performing distance measures were correlation distance for the continuous pitch histograms and Bhattacharyya distance for discrete svara histograms. The proposed grammaticality coefficient served well to quantify the distributional difference across a pair of performances from same/allied raga independent of the raga.

Insights into the practical realization of the musicological concepts of raga delineation and melodic progression at the concert time scale were obtained. This points to the future possi-

bility of developing the proposed methods for large-scale comparative studies in musicology. Although not the main focus of this work, the obtained outcomes can also be applied to the general raga recognition task, given the performance demonstrated on the relatively challenging sub-problem of discriminating allied ragas.

In this work, we used the allied raga performance as the ungrammatical realization of a given raga. A more direct, but considerably more challenging, validation of the proposed computational model would involve relating the predicted ungrammaticality of a performance to the ungrammaticality actually perceived by expert listeners. Future work must also address the modeling of the structural aspect of raga grammar, corresponding to the phrases, since this is a more easily accessed cue to raga identity for listeners [63]. Finally, it would be of interest to investigate the relative weighting of the different raga attributes for an overall rating of grammaticality, possibly at the different time scales, based on observed expert judgments. This is, in a way, similar to the discourse of Rabiner [143] as discussed in Section 4.4.3.

5.6 Discussions on individual raga-pairs

5.6.1 Raga-pair Deshkar-Bhupali

The correspondence of the figures for the Deshkar-Bhupali allied raga-pair for distance measures is Figure 5.2 ≡ Figure 5.7 and time-scale is Figure 5.3 ≡ Figure 5.8. The trend of the shape of the ROCs (and relative differences of the AUC values) for the corresponding distance measures is observed to be similar.

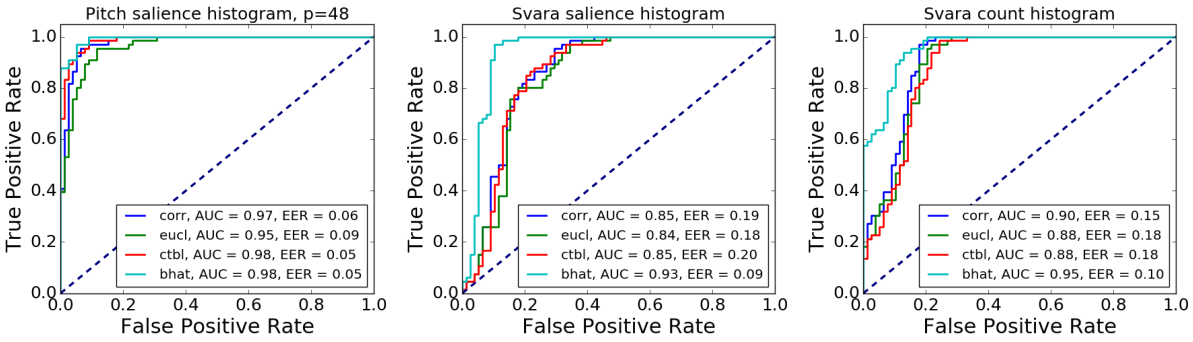

Figure 5.7: Raga-pair Deshkar-Bhupali (full concerts, octave-folded): ROCs obtained for four different distance measures from pitch salience (left), svara salience (middle), and svara count (right) histograms.

We obtain 17x17 (=289) pairs out of the 17 concerts in the Deshkar-Bhupali allied raga

dataset. Of these, 2x6x11 (=132) pairs contain mismatched ragas and hence serve as ungrammatical instances. The ROCs for the detection of ungrammatical instances were obtained from pairing concerts drawn from the Deshkar-Bhupali raga dataset. We observe that the svara histograms show distinctly superior performance with the Bhattacharyya distance measure.

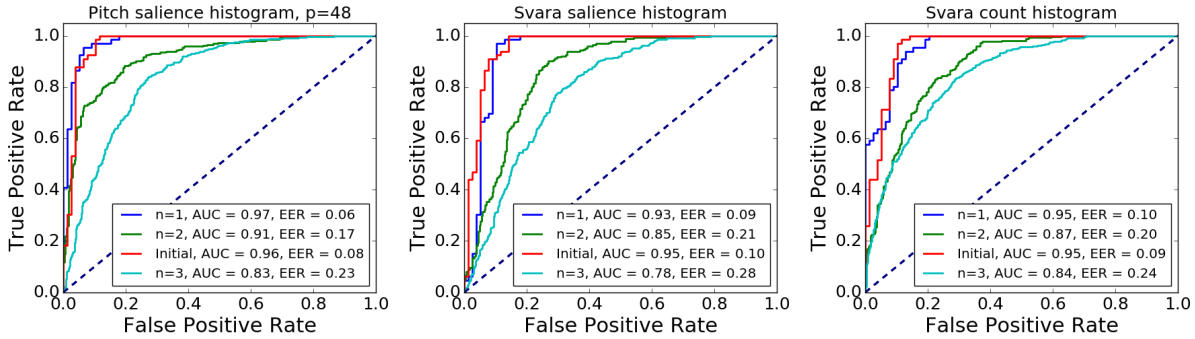

Figure 5.8: Raga-pair Deshkar-Bhupali (partial and initial concerts): Comparison of ROCs obtained with best distance measures (Correlation distance for pitch salience and Bhattacharyya distance for svara histograms) at different time-scales (n=1,2,3) and annotated initial portion (*alap+vistar*).

5.6.2 Raga-pair Puriya-Marwa

We carry out the same set of experiments for representation, distance measures for the Puriya-Marwa allied raga-pair. The correspondence of the figures that describe the tonal hierarchy representation is as follows: Figure 4.4 $\equiv$ Figure 5.9, Figure 4.5 $\equiv$ Figure 5.10, Figure 4.6 $\equiv$ Figure 5.11.

From the pitch salience histogram (Figure 5.9) of raga Puriya (left), we observe the svaras G and N as most prominent, which corresponds to the *vadi* and *samvadi* svaras. However, the peak height for the *vadi* G being less than that for the *samvadi* N is a result of octave-folding. N has more salience (accumulated duration) because of its presence in both lower and middle octaves, whereas G had most of its occurrences in the middle octave only. The r and D svaras have very low salience, which reinforces the grammaticality in the eminent artists' performance. The r peak is almost merged with the S peak, this is indicative of the fact that r is taken only within the glide (e.g. $\underline{\mathrm{N}}(\mathrm{r})\mathrm{S}, \underline{\mathrm{N}}(\mathrm{r})\mathrm{G}, \mathrm{G}(\mathrm{r})\mathrm{S}$) and is never sustained as an individual svara. In contrast, the r svara has the highest salience for raga Marwa (right), which also happens to be its *vadi* svara. The second most salient peak corresponds to the *samvadi* D svara.

The interpretations from the svara salience histograms (Figure 5.10) are similar. There is a high visual similarity between the pitch salience and svara salience histograms for the

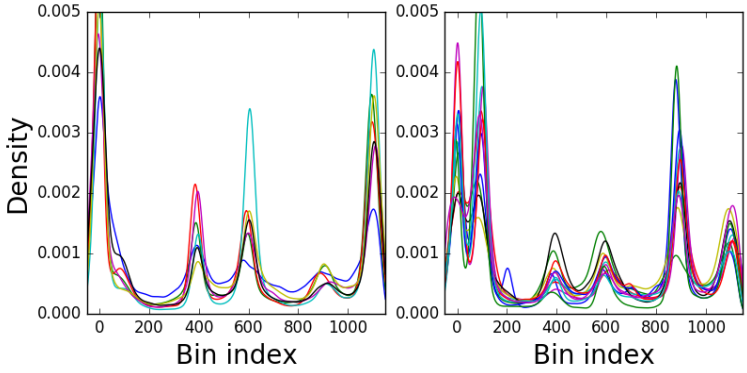

Figure 5.9: Pitch salience histograms (octave folded, 1 cent bin resolution) of 7 concerts in raga Puriya (left) and 12 concerts in raga Marwa (right).

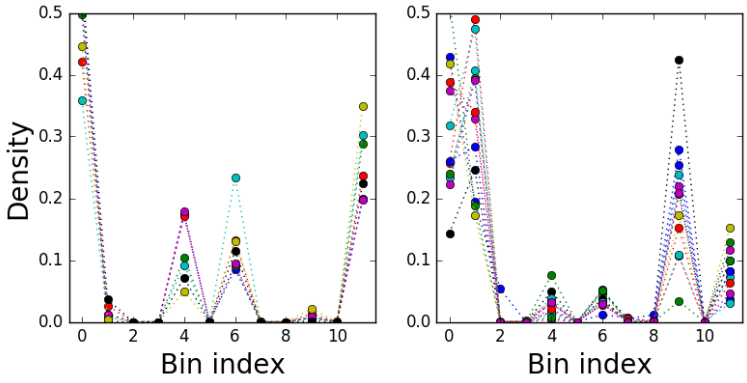

Figure 5.10: Svara salience histograms (octave folded) of 7 concerts in raga Puriya (left) and 12 concerts in raga Marwa (right).

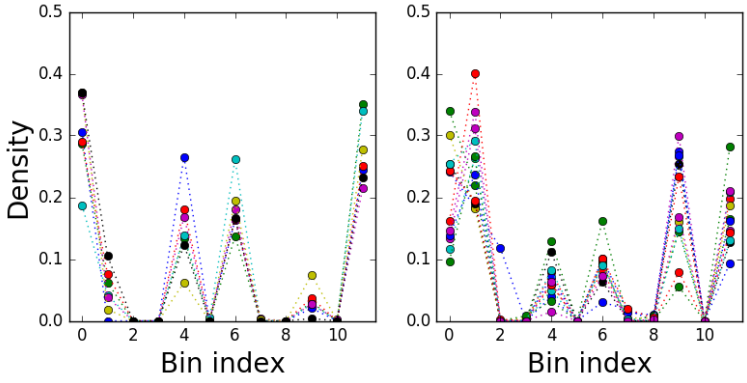

Figure 5.11: Svara count histograms (octave folded) of 7 concerts in raga Puriya (left) and 12 concerts in raga Marwa (right).

corresponding ragas (ignoring the bin mapping due to difference in resolution), this supports the musicological knowledge that major portion of a phrase in course of the raga development is covered by stable svaras. However, in comparison, we find an interesting observation on the relative salience of the r svara in the svara count histograms (Figure 5.11). In raga Puriya (left), the svara count is relatively high though the same has a very low salience in the svara salience histograms (Figure 5.10). This indicates a large number of detected r svaras that were of short durations, thereby accumulated to a low peak in the svara salience histogram. In contrast, the average duration of the r svaras in raga Marwa (right) are high, thereby accumulating to a high peak in the svara salience histogram for a moderate peak in the svara count.

The correspondence of the figures for the Puriya-Marwa allied raga-pair for distance measures is Figure 5.2 ≡ Figure 5.12 and time-scale is Figure 5.3 ≡ Figure 5.13. The trend of the shape of the ROCs (and relative differences of the AUC values) for the corresponding distance measures is observed to be similar.

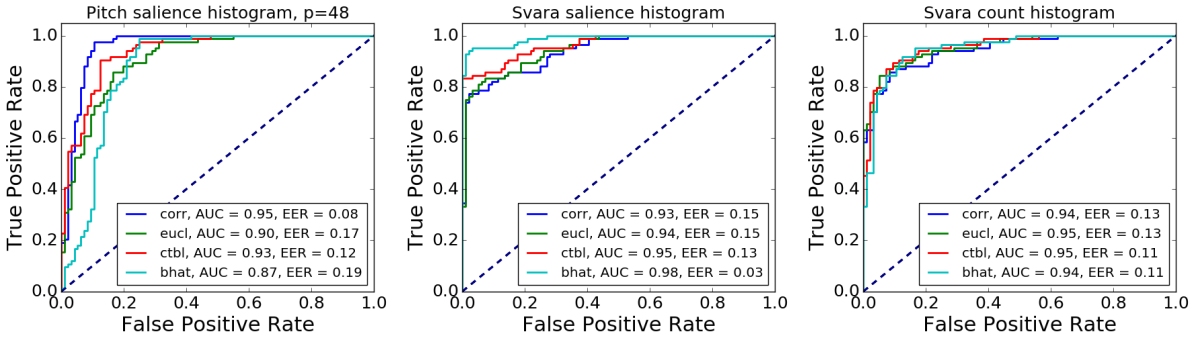

Figure 5.12: Raga-pair Puriya-Marwa (full concerts, octave-folded): ROCs obtained for four different distance measures from pitch salience (left), svara salience (middle), and svara count (right) histograms.

5.6.3 Raga pair Multani-Todi

Similarly, we perform the same experiments for the Multani-Todi allied raga-pair. The correspondence of the figures that describe the tonal hierarchy representation is as follows: Figure 4.4 ≡ Figure 5.14, Figure 4.5 ≡ Figure 5.15, Figure 4.6 ≡ Figure 5.16.

The pitch salience histogram (Figure 5.14) for raga Multani (left) has the salience of r and d svaras so feeble (and the peaks are merged with that of S and P respectively) that the tonal hierarchy almost resembles a pentatonic scale. This is also indicative of the fact that r and d svaras are taken only within the glide (e.g. $\underline{\mathrm{N}}(\mathrm{r})\mathrm{S}, \mathrm{N}(\mathrm{d})\mathrm{P}, \mathrm{Mg}(\mathrm{r})\mathrm{S}$) and is never sustained as

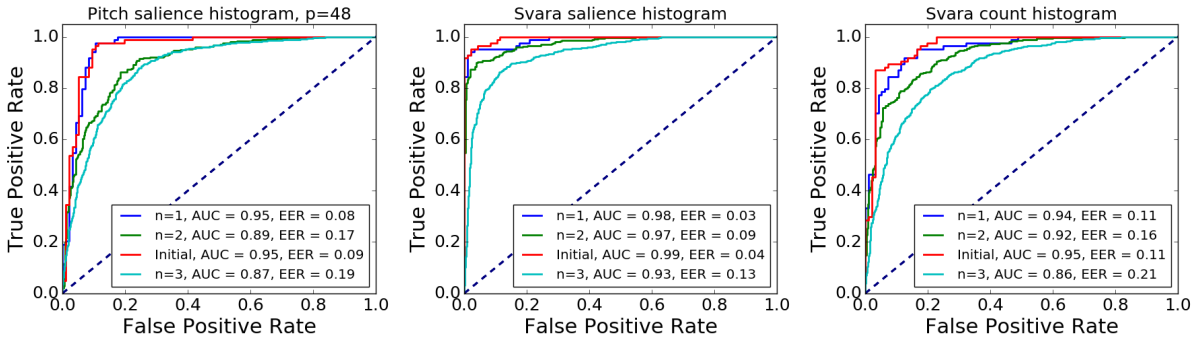

Figure 5.13: Raga-pair Puriya-Marwa (partial and initial concerts): Comparison of ROCs obtained with best distance measures (Correlation distance for pitch salience and Bhattacharyya distance for svara histograms) at different time-scales (n=1,2,3) and annotated initial portion (*alap+vistar*).

an individual svara. The *vadi* svara S and *samvadi* P, as hypothesized, are placed at the top two ranks in the tonal hierarchy. In contrast, the svaras in raga Todi (right) are distinctly visible in the tonal hierarchy. While the *vadi* svara d has a relatively high peak, *samvadi* g is lower down the order. The salience of r, again, is contributed by its presence in the middle and higher octaves.

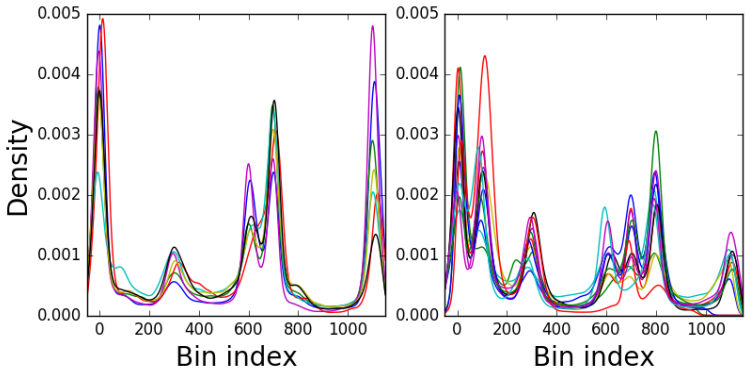

Figure 5.14: Pitch salience histograms (octave folded, 1 cent bin resolution) of 7 concerts in raga Multani (left) and 12 concerts in raga Todi (right).

The svara salience histogram (Figure 5.15) for raga Multani (left) also preserves the salience order of the S and P svaras. r and *samvadi* d svaras, in a similar way, have negligible peak heights. For raga Todi (right) also, visually, correlation (ignoring bin-mapping) between the pitch salience and svara salience histograms are high. The svara count histograms (Figure 5.16) are highly correlated with the svara salience histograms. For raga Multani (left), the count of the g svara is relatively high as one might expect from its salience in the svara salience histogram (Figure 5.15). This indicates presence of larger no. of detected g svaras,

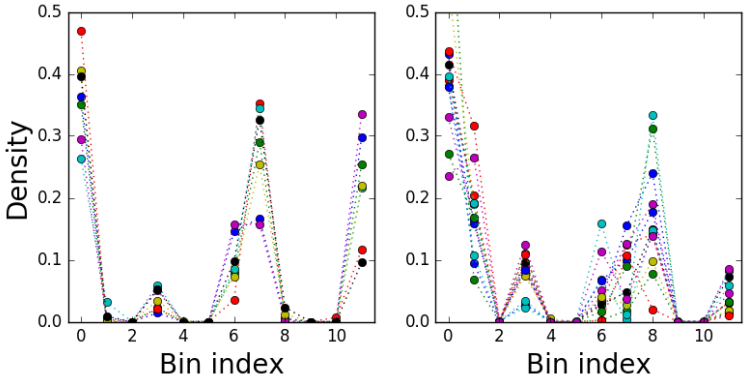

Figure 5.15: Svara salience histograms (octave folded) of 7 concerts in raga Multani (left) and 12 concerts in raga Todi (right).

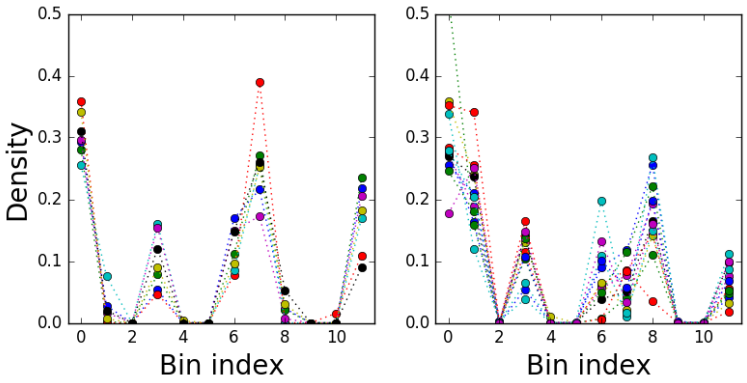

Figure 5.16: Svara count histograms (octave folded) of 7 concerts in raga Multani (left) and 12 concerts in raga Todi (right).

each of short duration.

The figure correspondence for the Multani-Todi allied raga-pair for distance measures is Figure 5.2 ≡ Figure 5.17 and time-scale is Figure 5.3 ≡ Figure 5.18. The trend of the shape of the ROCs (and relative differences of the AUC values) for the corresponding distance measures is observed to be similar.

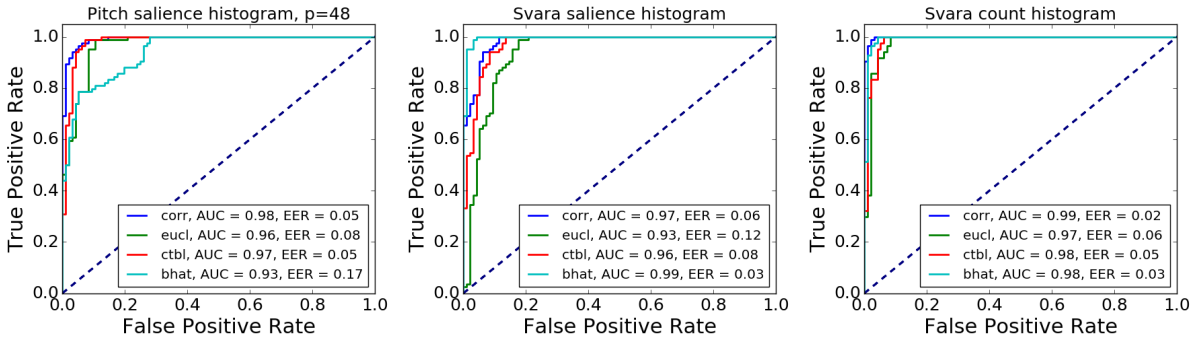

Figure 5.17: Raga-pair Multani-Todi (full concerts, octave-folded): ROCs obtained for four different distance measures from pitch salience (left), svara salience (middle), and svara count (right) histograms.

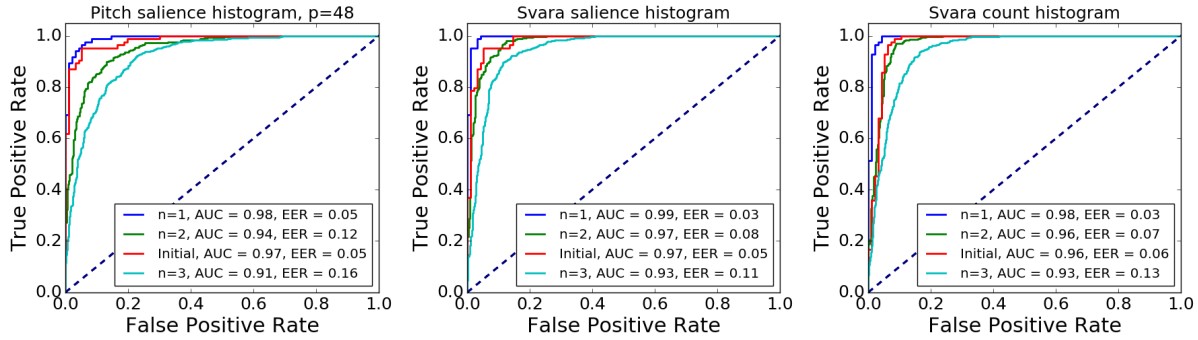

Figure 5.18: Raga-pair Multani-Todi (partial and initial concerts): Comparison of ROCs obtained with best distance measures (Correlation distance for pitch salience and Bhattacharyya distance for svara histograms) at different time-scales (n=1,2,3) and annotated initial portion (*alap+vistar*).

Chapter 6

The structural view

6.1 Motivation

Indian art music is an oral tradition with classical music training imparted principally through demonstration. Even so, there exist text resources that provide explicit accounts of musicological knowledge related to the basic underlying elements of raga and tala. The texts are written forms of what a musician might verbalize while teaching a student. In this work, we consider the computational analyses of recordings of raga performances by eminent Hindustani vocal artistes in terms of common available musicological knowledge about the raga lexicon/grammar. Our

[0]This chapter is largely drawn from the following papers:

- K. K. Ganguli and P. Rao. "A parametric approach to the structural representation of melody in Hindustani raga music," in preparation for the Journal of New Music Research (JNMR). [67]
- K. K. Ganguli and P. Rao. "Towards computational modeling of the ungrammatical in a raga performance," in Proc. of the 18th International Society for Music Information Retrieval Conference (ISMIR), October 2017, Suzhou, China. [63]
- P. Rao and K. K. Ganguli. "Linking prototypical, stock knowledge with the creative musicianship displayed in raga performance," Invited talk at the Frontiers of Research on Speech and Music (FRSM), December 2017, Rourkela, India. [149]
- K. K. Ganguli and P. Rao. "Validating stock musicological knowledge via audio analyses of contemporary raga performance," Invited talk at the 20th Quinquennial Congress of the International Musicological Society (IMS): Digital Musicology Study Session, March 2017, Tokyo, Japan. [64]
- K. K. Ganguli and P. Rao. "Tempo Dependence of Melodic Shapes in Hindustani Classical Music," in Proc. of the Frontiers of Research on Speech and Music (FRSM), March 2014, Mysore, India. [58]

analyses of performance audio use computational models to achieve an understanding of: (i) how the musicological knowledge is manifested in performance, and (ii) how the artiste improvises i.e. uses the stock musicological knowledge of the raga lexicon/grammar in "new" ways. We extract features that closely model: (i) raga notes, their relative saliences and intonation, (ii) occurrences of expected motifs, and (iii) melodic shape of raga-characteristic phrases.

6.2 Background

In a typical Hindustani music concert, an artiste executes improvisations of the raga characteristic phrases that represent the raga identity. The characteristic phrases of a raga (pakad) are typically referred to in terms of notation but are fully described via acoustic realization. The artist or performer uses his knowledge of the raga grammar to interpret the notation when it appears in a written composition in the specified raga. The shape of a recurring melodic motif, in terms of continuous pitch vs. time, within and across performances of the raga shows variability in terms of one or more of the following aspects: pitch interval, relative note duration and shape of alankars (ornaments), if any, within the phrase [150]. The rules of the raga grammar are manifested at different time scales, at different levels of abstraction, and demand different degrees of conformity. A majority of these are musicological in nature which typically involve either a thorough qualitative analysis of a handful of chosen musical excerpts, or a compilation of expert domain knowledge. Though these studies often present interesting musical insights, there are several potential caveats. Some of these caveats are: (i) small repertoire used in the studies challenge the generalizability of the proposed musical models, (ii) bias introduced due to the subjectivity in the analysis of musical excerpts, (iii) absence of concrete quantitative evidences supporting the arguments, (iv) the kind of analysis that can be done (manually) is limited by human capabilities, limited memory (both short- and long-term), and (v) difficulty in reproducibility of the results.

Several qualitative musicological works bring out new musical insights but are prone to criticism of not having supported their findings using a sizable corpus. Contrary to that, quantitative computational studies manage to scale to sizable data sets, but fall short of discovering novel musical insights. In the majority of cases, computational studies attempt to automate a task that is well known and is fairly easy for a musician to perform. There have been some studies that try to combine these two types of methodologies of working and corroborate several

concepts in musical theories using computational approaches. In Chinese opera music, [156] has performed a comparison of the singing styles of two Jingju schools where the author exploit the potential of MIR techniques for supporting and enhancing musicological descriptions. Autrim[1] (Automated Transcription for Indian Music) has used MIR tools for visualization of Hindustani vocal concerts that created a great impact on music appreciation and pedagogy in IAM. The main highlights of the work are as follows.

- We carry out acoustic measurements on two dimensions (pitch and duration) on a set of annotated raga characteristic phrases in order to find any possible trend(s) of independent variations (or co-variations) among the two dimensions from "production data".

- We validate by (dis)proving certain musicological assertions with support of statistical data which shows that it is reasonable to draw generalizations about these phrase variations (and possibly about other phrases that might fall in their "class").

- The broader aim of this study is to come up with the space of melodic variations and be able to comment on the novelty (or (in)correctness) of an unseen phrase. This has a wide scope in music pedagogy, especially for IAM where an objective assessment is seldom the case.

We consider the computational modeling of phrase shape based on maximizing the discrimination of close ragas with respect to the given attribute. The notion of "allied ragas" is helpful here where we consider ragas that share the same grammar in major attributes while differing in a few. In the present work, we consider the pair of allied ragas, Deshkar and Bhupali, and use the performances of eminent Hindustani vocalists as proxy for creatively expressed, but grammatically accurate, examples of the stated raga. The performances in the allied raga are likewise considered to be examples of the corresponding ungrammatical renderings. We evaluate known, as well as some new, representations in terms of the achieved discrimination on a dataset of performances across the two ragas. Although the scope of the experiments is restricted to the given pair of ragas and chosen attributes, we expect our outcomes to be generalizable.

The broad goal of this work is to model the variability of a melodic shape corresponding to a raga characteristic phrase within and across concerts. Thus it is important to discuss the

[1] https://autrimncpa.wordpress.com/

relevance of our choice of the phrase for a case study. We consider the GRS and PDS phrase (in two allied ragas Deshkar and Bhupali) shapes to carry out reasonable acoustic measurements in order to model the variability in different degrees of freedom. Given that we have a statistically reasonable dataset, we aim to discover the artist-dependence, tempo-dependence (tala-dependence if it has any), and context-dependence of the melodic shape variability. It is not only interesting to know which factors the variability can be explained by, but also the factors it does not depend upon. The amount of variability (quantifiable) explained (predicted) by a certain axis would enable us model a schema for the systematization of the melodic shape variation.

The word 'improvisation' is used by musicologists in different perspectives and it is important for us to disclaim that we carry out only tonal measurements to model the variability across realizations of the same melodic phrase. Improvisation, per se, might indicate (as discussed in literature) occurrence of an unforeseen variation which is unpredictable. We rather find that there is a strong dependence of the melodic and rhythmic events on the observed shape of a melodic phrase. The model would be applicable in both music retrieval, and in pedagogy in terms of correctness of a sung phrase and its degree of belongingness to a raga. We consider alap (and subsequent vistar of the vilambit/madhyalaya bandish) performance of vocal concerts by eminent artists for the task. In the alap (meaning introduction) section, artists introduce the raga by rendering the raga-characteristic phrases in their 'canonical' (exemplar) form. This also acts as a cue to the listeners for the raga identifiability. We exclude the *tan* section because there are no representative raga phrases in this section, the raga being already established. We build computational models that can address novel musicological questions about the underlying "fixed" structure within the so-called "free" improvisation. These pertain to the possible systematization of the behavior of a melodic phrase depending on its role in the bandish, e.g. whether a mukhda phrase of a bandish is expected to behave differently from the same melodic shape in the non-mukhda locations. The question is whether we can verify the invariance as well the systematic variability, if any, of the shape of a given characteristic melodic phrase over the course of a performance.

- What is the influence of tempo on the melodic shapes of the same phrase instances across the concert? Does the same depend on the location of the phrase within a cycle?

- Is there a systematization if the phrase is a mukhda or non-mukhda phrase?

- Which is the focal note in the region where the phrase is realized? Does the focal note affect the phrase shape?

6.3 Methodology and experiments

One issue for using the Dunya corpora in entirety, in terms of our specific task, is that the corpus also includes concerts of *drut laya* (fast tempo) whereas the raga phrases in their prescribed form manifest only in alap and slow development of the bandish [102]. Hence we filter the collection and augment a few more recordings that fulfill our criteria to propose a balanced (in terms of # concerts) dataset; the consolidated dataset is described in Chapter 3, as the phrase dataset (refer to Table 3.2). However, we observe a noticeable difference in the average duration in the concerts of the two ragas. The durations are of the usable segment (corresponding to the alap and vistar of the bandish) and the time-stamps are marked by a trained Hindustani musician which is later validated by a second musician. This truncation is to make sure that we include no concert segment where our musicological assumptions do not hold true. The following measurements are carried out, described in terms of the GRS phrase – the parallel concepts for the PDS phrase hold true.

1. Segmental measurements

   (a) Duration based measurements

   - Overall duration of the phrase: we define this as the duration from G onset till the S onset. This includes the two svaras G and R, and the transients G-R and R-S.
   - Duration of each constituent segments (svaras and transients).
   - Relative duration of the events within a phrase.

   (b) Intonation based measurements

   - Intonation of the stable svaras (slope, mean/median/mode, standard deviation, presence of kan (as a flag)) [2].

[2]The median intonation of the segmented stable svara is used in the experiments as the median was found to be least affected by the presence of alankars (causing pitch inflections). This was also tested by synthesizing the stable svaras by replacing the original pitch samples with the median intonation value. In a pilot listening experiment, musicians rated the synthesized phrase to be most natural when median was used out of mean/median.mode.

- Octave information (lower, middle, or higher).
- Tonic chosen by the artist.

2. Contextual measurements

   (a) Temporal context

   - Whether belonging to the unmetered (alap) or metered composition (vistar).
   - Tala (#beats) and laya (cycle length).
   - Location of cycle within concert: cycle index (absolute and normalized with respect to cycle count).
   - Location within cycle: proximity of the approaching sam (both in terms of #beats and absolute time).

   (b) Metadata-based measurements

   - Whether the phrase in question is a mukhda or a non-mukhda instance.
   - The "focal note" of the local region in the concert.

6.3.1 Musicological hypotheses testing

6.3.1.1 H1: frequency of occurrence of the characteristic melodic motifs

The motivation of this hypothesis is to observe how frequently the characteristic motif is invoked in the concerts. The most natural way to compute this is to calculate the total # phrases / total concert duration (measure A: phrases/sec). Or in the Western music sense where tempo would change phrase duration, we would be interested in calculating the total # phrases / total # tala cycles (measure B: phrases/cycle). The third measure is to compute fraction of cycles that carry at least a part of a phrase in terms of # cycles containing the phrase or part phrase / total # tala cycles (measure C: fraction of cycles for a partial phrase). This measure is particularly interesting, because certain non-mukhda phrases position themselves across two consecutive tala cycles and the reason is further explained in Section 6.3.1.3. The frequency of occurrence is shown in terms of the above three measures in Table 6.1. The measure A considers all phrases, but measures B and C exclude the alap phrases.

Measure of frequency of occurrence (mean values over the dataset)	Deshkar		Bhupali	
	GRS	PDS	GRS	PDS
A: phrases/sec	0.014	0.011	0.015	0.009
B: phrases/cycle	0.48	0.36	0.98	0.58
C: fraction of cycles containing phrase	0.39	0.51	0.99	0.62

Table 6.1: Frequency of occurrence of the motifs in terms of proportion of the phrase count with respect to the (A) concert duration, (B) absolute tala cycle count, (C) effective tala cycle count. For the actual count of phrases and cycles, refer to Tables 3.5 and 3.6

.

6.3.1.2 H2: variability of event duration/intonation across instances and concerts

Given the *Duration* values of each event in the GRS phrases (110 instances in raga Deshkar and 188 instances in raga Bhupali), we present the distributions of the event *Duration*s, in Figure 6.1 (left), in the form of boxplots of the raw measurements in seconds. We observe distinctions between the two ragas in nearly all the duration parameters, most notably in the R *Duration*s. That the R duration is small and constrained in raga Deshkar is supported by the raga grammar specification (as provided in Table 3.1) which indicates R in parantheses, suggesting a "weak note that is never sustained" [152]. Overall, the dispersion in the parameters is smaller in the phrase in raga Deshkar compared with Bhupali, consistent with the fact that it is a grammatically more constrained raga [13, 92, 152]. An exception is the realization of the S svara with several outlying values of duration due to its location at phrase end, where a number of contextual considerations influence the note offset.

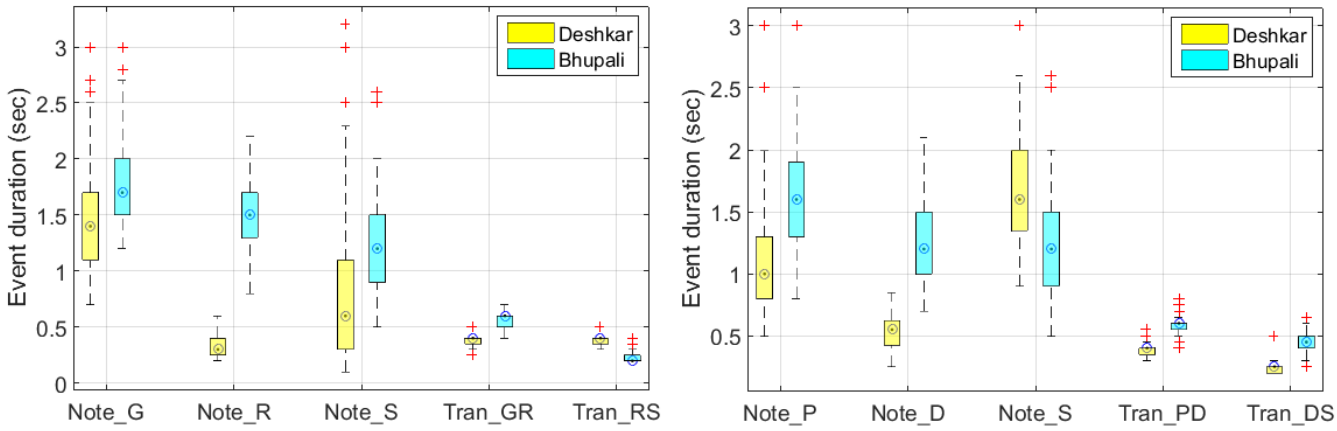

Figure 6.1: Distributions of event *Duration*s across the GRS (left) and PDS (right) phrase instances in the two ragas.

Similarly for the set of PDS phrases (86 instances in raga Deshkar and 115 instances in raga Bhupali), the distribution of the event $Durations$ are shown in Figure 6.1 (right). The most distinctive feature observed is the DS-transient $Durations$, while D $Durations$ also bear a small overlap in the two ragas. Table 6.2 lists the precise values of event duration statistics. The distribution of D duration for raga Deshkar PDS phrases are observed to be narrow. This can be explained by the role of the D svara as a *graha, amsa,* or *nyas* [191] within a phrase. While D is the *vadi* svara of raga Deshkar and would be realized as long nyas svaras for phrases ending on a D nyas, e.g. SGPD (refer to Table 3.2), the PDS phrase has less emphasis on D and hence the narrow distribution. The P(D)SP phrase (Table 3.2) would have even smaller D duration, if at all detected as a stable svara.

Raga	Phrase	Event duration statistics (mean, median, std dev) in sec				
		1: G/P	2: R/D	3: S	4: GR/PD	5: RS/DS
Deshkar	GRS	1.47, 1.4, 0.48	0.33, 0.3, 0.11	0.85, 0.6, 0.75	0.38, 0.4, 0.05	0.39, 0.4, 0.05
	PDS	1.13, 1.0, 0.48	0.55, 0.55, 0.14	1.72, 1.6, 0.54	0.4, 0.4, 0.07	0.24, 0.25, 0.05
Bhupali	GRS	1.75, 1.7, 0.31	1.55, 1.5, 0.32	1.25, 1.2, 0.46	0.58, 0.6, 0.07	0.23, 0.2, 0.06
	PDS	1.61, 1.6, 0.43	1.22, 1.2, 0.31	1.21, 1.2, 0.45	0.59, 0.6, 0.08	0.45, 0.45, 0.09

Table 6.2: Summary table of the different event duration statistics for both GRS and PDS phrases in ragas Deshkar and Bhupali.

The variation of intonation of stable svaras are presented in terms of the distribution of the median values. Figure 6.2 shows the dispersion for G svara in the GRS phrase (left) and D svara in the PDS phrase (right). The G intonations are observed to have a difference of ≈ 10 cents of the medians, with no overlap in the distributions between ragas Deshkar and Bhupali. In the case of PDS phrase, though the difference of medians is comparable (i.e. ≈ 10 cents), but there exists an overlap between the distributions corresponding to the two ragas. This is relatable to the comment on *shruti* in the raga grammar (Table 3.1) discussion which states that Deshkar has a higher intonation for R, G, D.

It is difficult to comment on the R intonation, because the duration spent on R (specially in raga Deshkar) is very low. As both G and R are a part of descending movement, a negative (downward) slope is observed for both the notes. In contrast, the slope of the S segments are observed to be flat, this might be because S is the last note of the phrase. Sometimes (in raga Bhupali) the S segment has a positive slope due to presence of a touch note (kan)

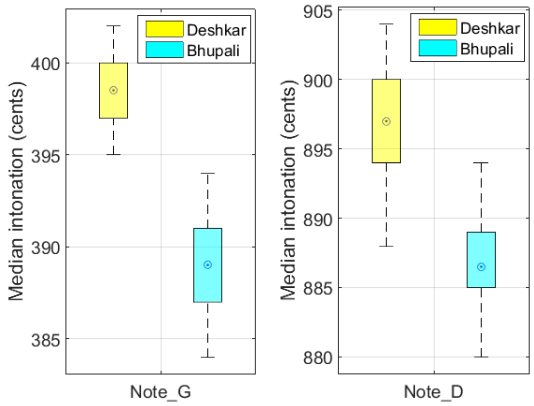

Figure 6.2: Distribution of median intonation of the G svara in GRS phrase (left) and D svara in PDS phrase (right) in the two ragas.

of lower D. No consistent trend is observed for presence of embellishment on a svara. The GRS phrase in raga Deshkar is observed to have no alankars, whereas raga Bhupali is often observed to have superimposed kan svaras on the G or R svara. For the PDS phrase, both ragas show occasional oscillations on the P svara but not on the D and S svaras. It is obvious to expect more embellishments on the PDS phrase, being a part of the higher tetrachord, as the raga already gets established before introduction (and repetitions) of this phrase. However, the distinct phrase shape of the D-S glide (and hence the overall shape of the latter half of the phrase) is observed to be devoid of much embellishments. This is to be noted that the segmented phrases may belong to xxGRS or xxPDS phrases where the preceding xx segment often guides the amount of embellishment realized on the actual GRS or PDS segment.

6.3.1.3 H3: dependence of phrase duration on cycle location, location in cycle

The hypothesis bases on the assumption that the tempo of a concert gradually increases, i.e. the cycle duration decreases with a certain rate. To verify the assumption, and to measure the rate of change of cycle duration, we fit a linear regression model and record the 'slope' and 'goodness of fit' (in terms of $\mathcal{R}^2$). Figure 6.3 shows a linear model to the tempo profile (inverse of cycle duration, adjusted with a scaling factor proportional to beat-count in the specified tala), corresponding to each concert in our dataset. We show this for a single concert (and validate others individually [3]) as there is no means of combining the data points of different

[3]The range of slopes obtained for the madhyalaya concerts are [0.69,0.86] with high $\mathcal{R}^2$ values indicating a good model fit. The same for vilambit concerts are observed to be [0.54,0.72] with lower $\mathcal{R}^2$ values because of a steplike increase in tempo as opposed to gradual increase in madhyalaya.

concerts varying in tempo-ranges. However, in the subsequent plots we time-warp the cycle indices between [0,1] by normalizing with respect to the cycle-count of each concert. The different event duration statistics for both GRS and PDS phrases in ragas Deshkar and Bhupali is presented in Table 6.2.

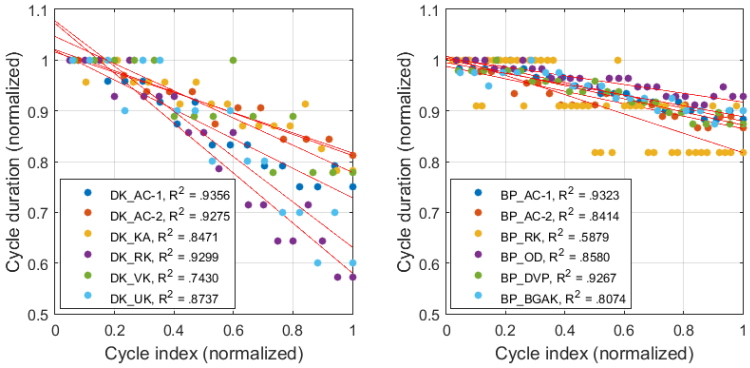

Figure 6.3: Scatter plot of the normalized cycle duration (with respect to maximum duration for each concert) versus normalized cycle index (with respect to maximum count for each concert). The $\mathcal{R}^2$ value indicates the goodness of linear regression model fit.

We observe a systematic variation in the form of an inverse relation between the cycle index and cycle duration. Figure 6.3 (the actual cycle length range is provided in Table 3.5) shows that there is a gradual increase of tempo, realized through decrease in cycle length at certain cycle boundaries. We hypothesize that the duration of the phrases would be proportional to the cycle length, and hence would decrease in course of the time-evolution of the concert. Figure 6.4 shows a scatter plot of phrase duration versus normalized cycle duration for each concert with different markers for mukhda (diamond) and non-mukhda (circle) instances. A systematic reduction in the phrase durations is observed as supported by the positive slope values of the model fit (red line), in both phrases in both ragas. However the $\mathcal{R}^2$ goodness of fit values are not high. This indicates, there might be other predictors that can explain the variability in the phrase duration.

One immediate question could be whether the variance (around the regression line) is contributed differently by the mukhda versus the non-mukhda instances. Hence we plot the distribution of the phrase durations where the same phrase occur both as mukhda and a non-mukhda, i.e. GRS phrase in all Bhupali concerts and PDS phrase in 4 (out of 6) Deshkar concerts. The proxy feature capturing the concert progression in time (as follows from Figure 6.3) is to consider the normalized cycle count for each concert. We show the distribution of phrase durations in these two cases in Figure 6.5 where the non-mukhda instances are observed to have higher

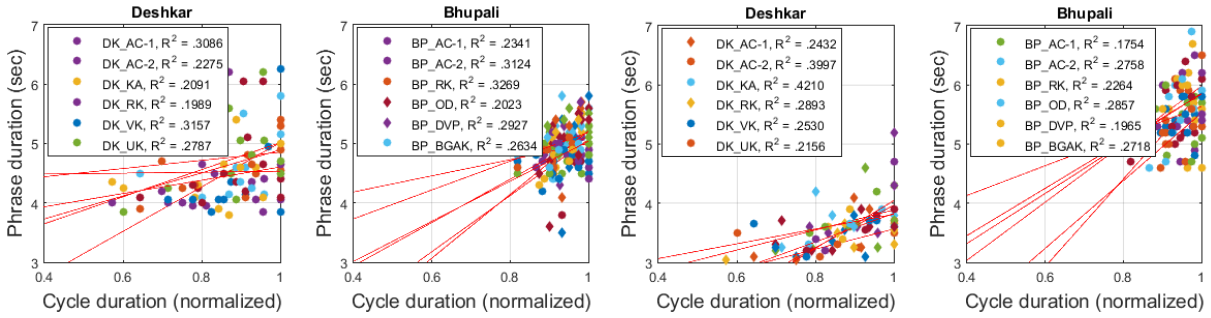

Figure 6.4: Scatter plot of phrase duration (sec) versus normalized cycle duration (with respect to maximum duration for each concert) with different markers for mukhda and non-mukhda instances: GRS (left) and PDS (right) phrases in raga Deshkar and Bhupali.

variance. The boxplots for the associated concerts are separately shown in Figure 6.6.

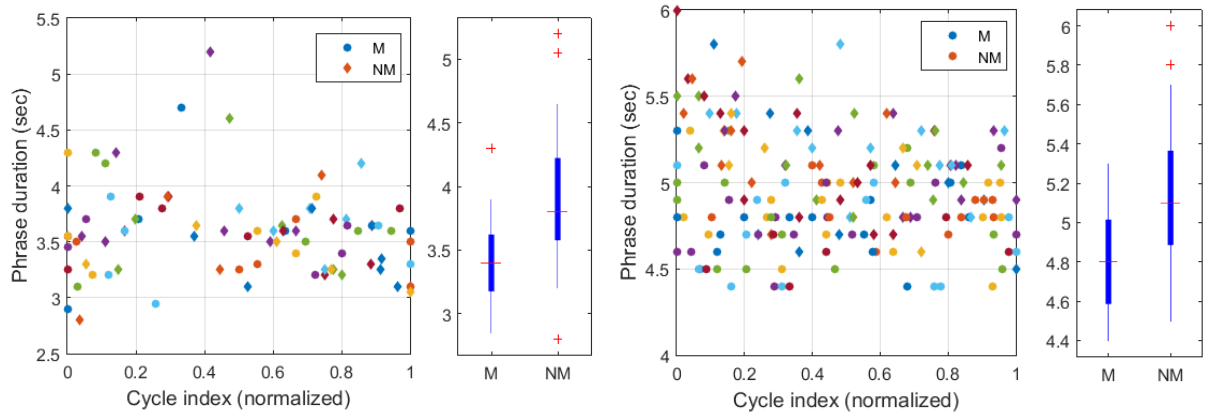

Figure 6.5: Scatter plot of phrase duration (sec) versus normalized cycle index (with respect to maximum count for each concert) with different markers for mukhda and non-mukhda instances: PDS phrase in raga Deshkar (left) GRS in raga Bhupali (right).

Considering that variations in phrase duration for the mukhda instances are trivial. Hence we pull out from all applicable concerts the non-mukhda instances and pool them together for an across-concert design to understand if the variability can be predicted by the concert progression in time. The correlation between these two metrics are shown in Figure 6.7.

We observe a weak reduction in the GRS phrase durations from the alap till the end of vistar in the Deshkar concerts. However, there is a second peak [4] (where the duration increases for a few cycles) observed for raga Bhupali [5] concerts. Close examination reveals that these phrases were rendered in the upper octave in course of the antara (second stanza of the com-

[4]This is visible more clearly in individual concerts with actual cycle indices. The plots in Figure 6.7 are with normalized cycle index, this effect is seen around the x-axis range (0.65,0.85).

[5]The isolated jumps in Deshkar concerts are explained through another tempral context-based measurement to be discussed next.

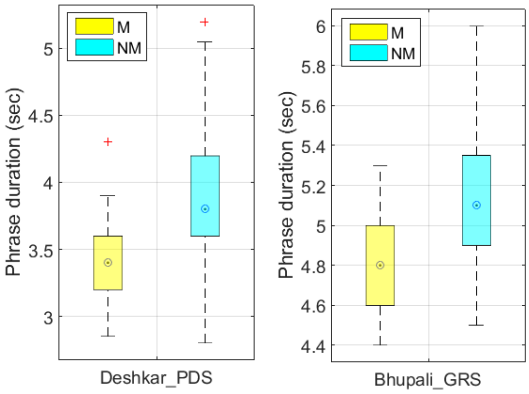

Figure 6.6: Boxplot of phrase duration (sec) for mukhda (M) and non-mukhda (NM) instances: PDS phrase in raga Deshkar (mean = 3.39 (M), 3.88 (NM); SD = 0.29 (M), 0.52 (NM)) and GRS phrase in raga Bhupali (mean = 4.78 (M), 5.14 (NM); SD = 0.22 (M), 0.36 (NM)).

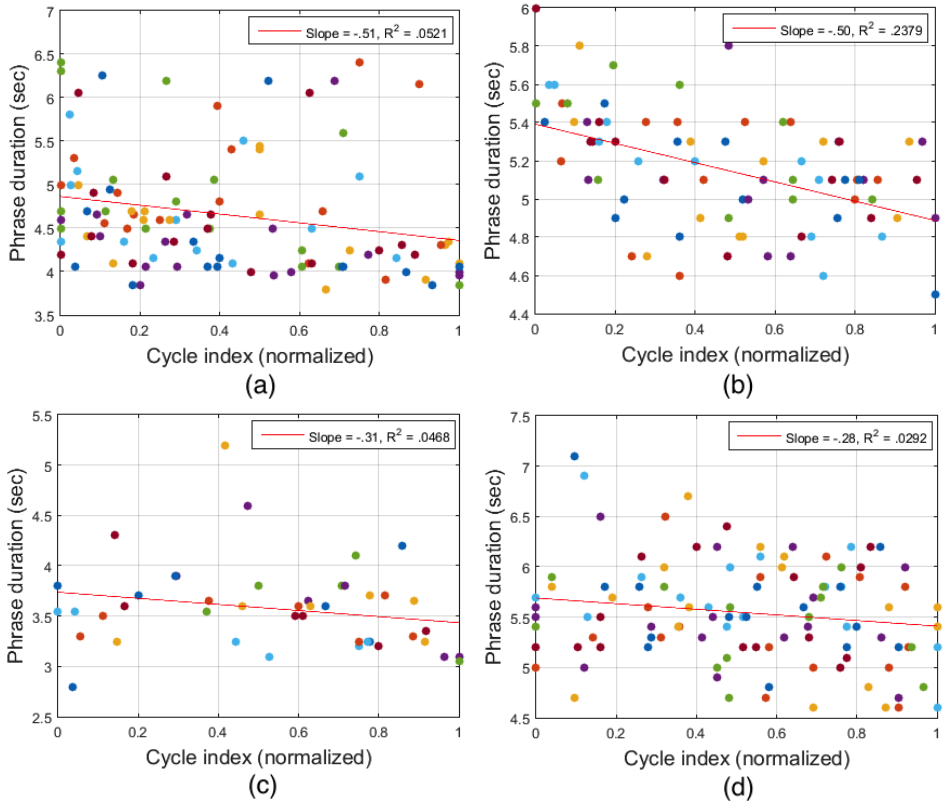

Figure 6.7: Scatter plot (different colors indicating six concerts in each raga) of phrase duration versus normalized cycle index for non-mukhda instances: GRS phrase in (a) raga Deshkar and (b) raga Bhupali. The same for PDS phrase in (c) raga Deshkar and (d) raga Bhupali. The red line is the model fit, with slope and 'goodness of fit' ($\mathcal{R}^2$) in legend.

position) vistar. Additionally, we observe that there are many cycles in raga Deshkar concerts where the GRS phrase is absent, while such a sparsity is not observed for raga Bhupali concerts. The reason is that the GRS phrase is a part of the mukhda (refrain of the song) phrases in all three Bhupali concerts, but not in Deshkar concerts.

Effect of proximity to the approaching sam

The higher variance in the non-mukhda instances, we speculate, can be explained through another temporal predictor: location in cycle, measured through 'proximity of the approaching sam'. We therefore plot a scatter of the phrase duration versus the proximity, shown in Figure 6.8. We observe a specific pattern. If the proximity feature lies approximately between 5 to 10 sec (irrespective of the cycle length, vis-a-vis, no. of beats), the phrase durations are less than the mean phrase duration. If the proximity is either less than 5 sec (region to the left) or more than 10 sec (region to the right), the phrase durations are more than the mean value. The applicable phrases are GRS in raga Deshkar and PDS phrase in raga Bhupali.

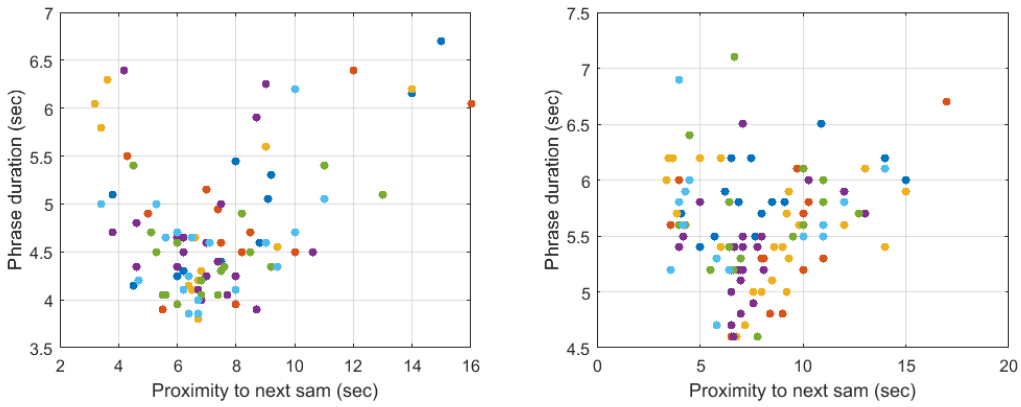

Figure 6.8: Effect of proximity of the approaching sam on phrase duration: GRS in raga Deshkar (left) and PDS phrase in raga Bhupali (right).

In general the the phrase duration is proportional to the time-difference (absolute, not in beats) of the phrase onset to the immediate next sam. This means if the onset of the phrase is away from the next sam location, the phrase is stretched freely, ensuring enough room for the approaching mukhda at the intended location. But if the sam is in proximity, the phrase duration is adjusted so that the mukhda phrase does not have to be time-compressed. This indicates that maintaining the mukhda duration is more important that it can override the non-mukhda phrase duration. On a further extreme case, when the phrase onset is too close to the approaching sam that both the phrase (in question) and the mukhda phrase cannot be fit within the available time-limit, musicians tend to stretch the phrase quite long to enter the next rhythmic cycle,

thereby aborting the mukhda phrase. This phenomenon is common in madhyalaya bandishes, but are rarely found in vilambit compositions. The cycle lengths of vilambit bandishes are long, hence musicians generally do not miss any mukhda (unless intended) due to such adjustments. A related point to note is that we observe many cycles in raga Deshkar concerts where the concerned (GRS/PDS) phrase is absent, while such a sparsity is not observed for raga Bhupali concerts being slow in tempo.

Now that we have two temporal predictors of the phrase duration, we fit a multiple regression model using these two. Figure 6.9 shows a 3-dimensional visualization where the phrase duration (in color scale) is plotted as a function of two predictors: (i) normalized cycle index, and (ii) proximity of the approaching sam. We observe that the dark blue dots (indicating phrase duration less than the mean) lie along a band of proximity values of 5 to 10 sec, whereas the light blue (or yellow) dots (indicating phrase duration higher than mean) lie either below or above this band. The regression of the phrase duration with the temporal predictors in terms of coefficient(s), intercept, and goodness of fit is shown in Table 6.3. It is to be noted that the goodness of fit values are low. The theoretical explanation and visual interpretation supports the musicological hypothesis, but the statistical model could be further investigated.

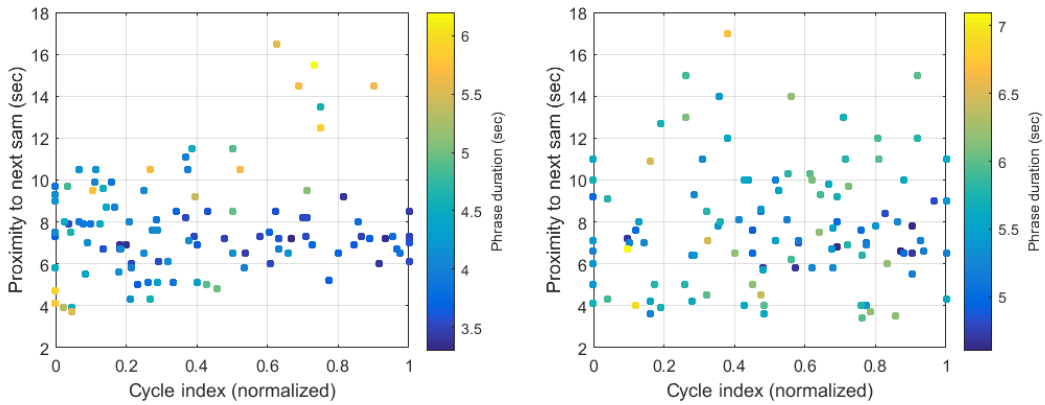

Figure 6.9: Coupled effect of normalized cycle index and proximity of the approaching sam on phrase duration: GRS phrase in Deshkar (left) and PDS phrase in Bhupali (right).

6.3.1.4 H4: relation of phrase duration with constituent svara durations

As discussed before in Chapter 3 that raga grammar constraints certain svaras to be compressed in time (e.g. alpatva usage of the R svara in raga Deshkar as listed in Table 3.1), it is of interest therefore to investigate how the relative duration of the constituent segments vary when a phrase is time-warped. Figure 6.10 shows the correlation the stable svara (separately) and

Raga	Phrase	Regression coefficients			$\mathcal{R}^2$
		Proximity	Norm cyc ind	Intercept	
Deshkar	GRS	–	-0.505	4.86	0.0521
	PDS	–	-0.169	3.66	0.0210
Bhupali	GRS	–	-0.215	5.05	0.0462
	PDS	–	-0.279	5.69	0.0292
Deshkar	GRS*	0.117	-0.632	3.51	0.2070
Bhupali	PDS*	0.026	-0.288	5.49	0.0528

Table 6.3: Summary table of the linear and multiple (marked with *) regression of the phrase duration with the temporal predictors in terms of coefficient(s), intercept, and goodness of fit.

transient (accumulated) durations with the overall duration for the GRS and PDS phrases in the two ragas.

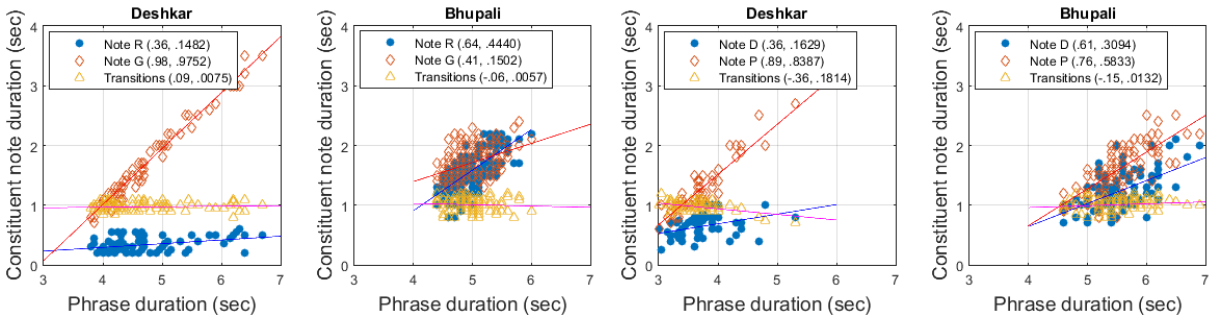

Figure 6.10: Correlations between R,G (or P,D) svaras and GR (or PD) transition durations versus GRS (or PDS) phrase duration for Deshkar (left) and Bhupali (right) phrases.

Raga	Phrase	Events (Correlation, $\mathcal{R}^2$)		
		G/P	R/D	Transients
Deshkar	GRS	0.98, 0.9752	0.36, 0.1482	0.09, 0.0075
	PDS	0.89, 0.8387	0.36, 0.1629	-0.36, 0.1814
Bhupali	GRS	0.41, 0.1502	0.64, 0.4440	-0.06, 0.0057
	PDS	0.76, 0.5833	0.61, 0.3694	-0.15, 0.0132

Table 6.4: Summary table of correlation of phrase duration with constituent event duration.

Table 6.3.1.4 presents the correlation values as observed from the legends in Figure 6.10. We observe that as prescribed in raga grammar for raga Deshkar, the R svara has almost zero

correlation with the phrase duration. We see that the correlation value and the $\mathcal{R}^2$ value as a goodness of fit of a linear regression model on the data points. We see that variation of the overall phrase duration can be modeled as a function of the durations of only the constituent stable svaras. This is because the fact that the variation in transient duration is negligible and it only contributes marginally to the overall variance.

6.3.2 Realization of emphasis

A focal note, as defined by Widdess [207] is defined as a note that the artist is intending to establish, hence reinforces repeatedly. The common pattern is a gradual unfolding of a focal note from low to high notes in an octave. The forms of emphasis, as summarised by Widdess [207] and van der Meer [191] and complemented by our observation, are often realized in: (i) duration of a particular note being increased – manifested as a focal svara in the local concert region, (ii) accent coinciding with a rhythmic stroke at (sub)matra level – often onset of a syllable of the lyric, and (iii) intensity co-varying with pitch, loosely coupled with accent – relative (harmonic) energy of the stable pitch regions are proportional to their pitch intervals.

Our previous work [65] (also discussed in Chapter 5) shows on the current dataset that the common trend of gradually unfolding a raga in course of *vistar* is maintained is brought out very nicely by the time-normalized summaries over the 6 concerts of each raga in Figure 5.6. A summary is obtained by applying smoothing by way of a 7-point median filter to the cycle-level salient svara curve of each concert, and then superposing the time-normalized curves in a single plot. This phenomenon bears a high resemblance to the time evolution of melody in course of the *vistar*, as shown in Figure 11 in [207].

6.3.2.1 Effect of "focal note" on the phrase shape

There is an observed pattern in the constituent svara durations within a phrase along the progression of focal note as observed in Figure 5.6. The duration of each svara depends on the local context. E.g. in raga Bhupali, the initial part of the concert where R is the focal note (local nyas svara) of the concert, the R duration of the GRS phrase is observed to be higher. The same phenomenon holds true for the G note. Specifically for raga Deshkar where R is a feeble note (only applied as an alpa svara), the R duration is observed to vary minimally. This example can also be explained by the same hypothesis, because there is no segment observed

in Deshkar performance with R as a focal note. The constituent svara durations (R & G in raga Bhupali as a test case) vs. the normalized tala cycle index is shown in Figure 6.11. We see that the focal-context svara durations are more than that of the non-focal contexts. The distribution of R duration while R is a focal note has a higher mean than the non-focal regions. The non-focal regions show a large dispersion of R durations, much of which is contributed by the last few cycles where the phrases belong to the antara vistar in the higher octave. While the R and G focal regions are clear, the scatter distribution in the start and end can be explained by the intensity plots that are presented next.

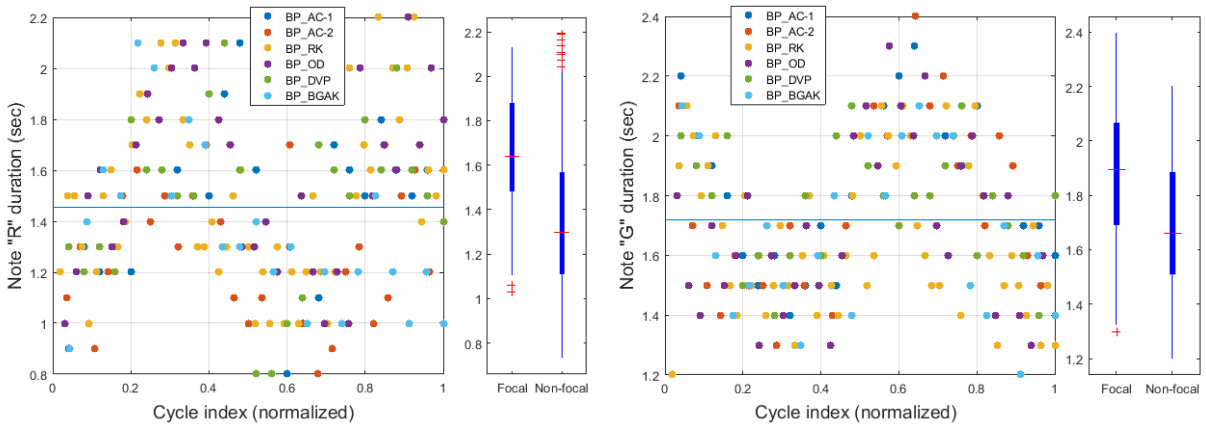

Figure 6.11: Variation of constituent svara durations in focal vs. non-focal context: R svara (left) and G svara (right) vs. the normalized tala cycle index in focal and non-focal contexts.

6.3.2.2 Observation: correlation of pitch and intensity are universal

The intensity contour is obtained from the multi-pitch F0 detection algorithm [154] where each 10 ms time frame is assigned with its F0 (Hz) and sum of harmonic energy (dB) with an assignment of 0 Hz for silent frames. Pitch (cents) and intensity (dB) often covary, which is possibly a physiological phenomenon. At concert level, the correlation coefficient is high (Mean=.72, SD=.4). The same phenomenon is observed to translate at the time-scale of phrase level realization. However, there are certain challenges involved: (i) the mukda phrases (or bol vistar regions) have a drop of harmonic energy at the consonant onset locations, affecting the local correlation [6], and (ii) the jitter for intensity contour is observed to be more than its corresponding pitch contour, hence we apply a 50 ms (equivalent to 5 samples) median filter smoothing for better comparison.

[6]The pitch is also affected by the same, but not so much.

6.3.2.3 Salient observations

- Pitch stability vs. intensity decay: It is often observed that the *vilambit vistar* involves realization of long stable notes. The intensity has a gradual decay in these regions, possibly due to maintaining pitch perfection for long breaths.

- Duration and intensity act as complementary effects: Followed from the last observation, the realization of emphasis is observed to be either in intensity being high for a short note segment or duration being long for a low intensity segment.

- Intensity vs. accent: The lyric plays a major role in the computation of intensity. Followed from above, the short notes of high intensity are often realized with short vowels or consonants that creates a jump in intensity contour indicating accented syllables.

Continuing from the duration study on the effect of focal note on phrase shape, we carry out measurements to understand if the realization of emphasis is also encoded in the intensity of the phrase. The constituent svara intensities (R & G in raga Bhupali, same as above) vs. the normalized tala cycle index is shown in Figure 6.12. The phrases at the end of concerts are phrases from antara vistar, the overall intensity of these phrases are observed to be high which is also captured in the R intensities. The non-focal region G notes are observed to have higher intensity dispersion. The initial part of the concert the G intensity is high due to emphasis on the mukhda phrases. At the end of the concert (antara phrases), there exists a peak, but with less than -12 dB (unlike R). This may be due to physiological limitation to render an extended note at higher pitch with high intensity.

We present a few pointers that can be interpreted from the combination of duration and intensity together. We suspect that intensity independently is not a great predictor, given the limitations of robust measurement techniques from a polyphonic audio of vocal concerts, but these help explain some events in duration-based measurements.

- In the initial part of the Bhupali concerts (as above), all GRS phrases have both duration and intensity of G note high. This matches with the observation (from listening) that musicians intended to emphasize on the G svara for the mukhdas of initial cycles.

- The median of the intensity distribution of the focal note regions is less than the non-focal regions for both R and G notes. This indicates that the duration and intensity plays a complementary role.

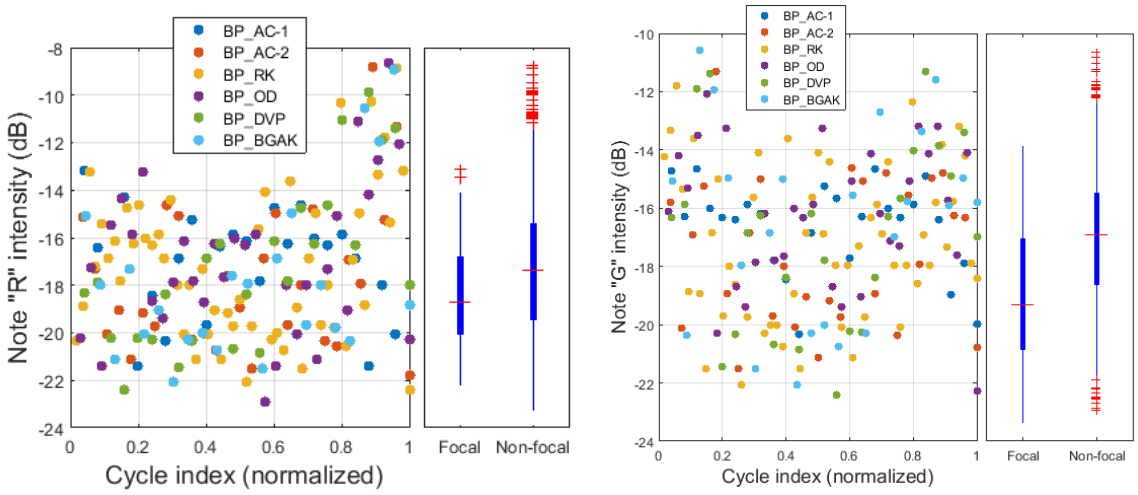

Figure 6.12: Feature correlations among predictors: R (left) and G (right) svara intensity with normalized cycle indices.

- The intensity peaks at the end of concert which correspond to antara vistar (higher octave renditions) where the voice is brighter. These points (specially for R note) can explain the peak observed in the duration plot.

6.4 MIR applications

We present experiments that help us identify the aspects of optimal representation, that discriminate the two ragas maximally based on our labeled dataset of 12 concerts across two ragas. Our common approach for both attributes is to use unsupervised clustering (k-means with k=2) of the feature vectors and optimize the separation between the clusters over the considered choices for feature implementation. This is to simulate what we had attempted for pitch distributions [65]. If the representation is robust, it should give out homogeneous clusters of each raga.

Performance in unsupervised clustering can be measured via the 'cluster purity' as described in Chapter 5. To compare the predictive powers of the measured acoustic features, we perform 'Feature Selection' using Weka [7] datamining tool. We use the "InfoGainAttributeEval" as the attribute selector that evaluates the value of an attribute by measuring the information gain with respect to the class, in conjunction with the "Ranker" search method that ranks attributes by their individual performances. We use the svara segmentation method outlined in 4.3.2 to obtain the components of the phrase shape corresponding to the sequence of stable svaras as well

[7] http://www.cs.waikato.ac.nz/ml/weka/

as the transient regions. In this section we present a statistical description of the features corresponding to the different events. We also compare the discrimination powers of the different features via a clustering experiment.

6.4.1 Feature selection and evaluation

We construct a feature vector for each instance of the GRS (Or PDS) phrase with 5 $Duration$ features, one for each event, and $Intonation$ and $Slope$ features corresponding to each of the three stable svaras. Of these 11 features, we obtained the most significant features in terms of predictive power as the following: (i) R/D $Duration$, and (ii) G/P $Intonation$, with the third feature in the list placed considerably lower. This outcome is consistent with the raga grammars where these two aspects are considered distinctive properties of raga Deshkar.

Next, we perform an unsupervized clustering of the 298 GRS phrases into two classes using the Euclidean distance between the 2-element vectors of the two selected features. The achieved cluster purity value is 0.99 (i.e. only 5 instances of the 298 are misclassified). As a next step, we investigate whether duration normalization is helpful. Given that overall phrase duration is correlated with tempo [191], it is natural to expect that the variability of phrase event durations may be reduced by normalization by the overall phrase duration. However, it turned out that the cluster purity with the duration-normalized $Duration$ (of R svara) feature coupled with the previous $Intonation$ (of G svara) feature reduced to 0.94 (i.e. 19 instances were misclassified). This indicates that musicians interpret the raga grammar in terms of raw durations rather than relative to the tempo.

Similarly for the PDS phrase category, (i) D $Duration$, and (ii) D $Intonation$ qualify as selected features, the achieved cluster purity value is 0.98 (i.e. only 5 instances of the 201 are misclassified). Cluster purity with the duration-normalized $Duration$ (of D svara) feature coupled with the previous $Intonation$ (of D svara) feature reduced to 0.92 (i.e. 16 instances were misclassified).

6.4.2 Comparing phrase representations

The state-of-the-art systems for the task of melodic similarity, involve the continuous pitch representation with vector measures for distance computation. Figure 6.13 shows the feature representation in the continuous space (110 Deshkar phrases (left) and 188 Bhupali phrases

(right)), G onset till S onset as used in our previous measurements, interpolated to a constant length (Int_len) of 4 sec (average duration over the dataset) and time-aligned with DTW.

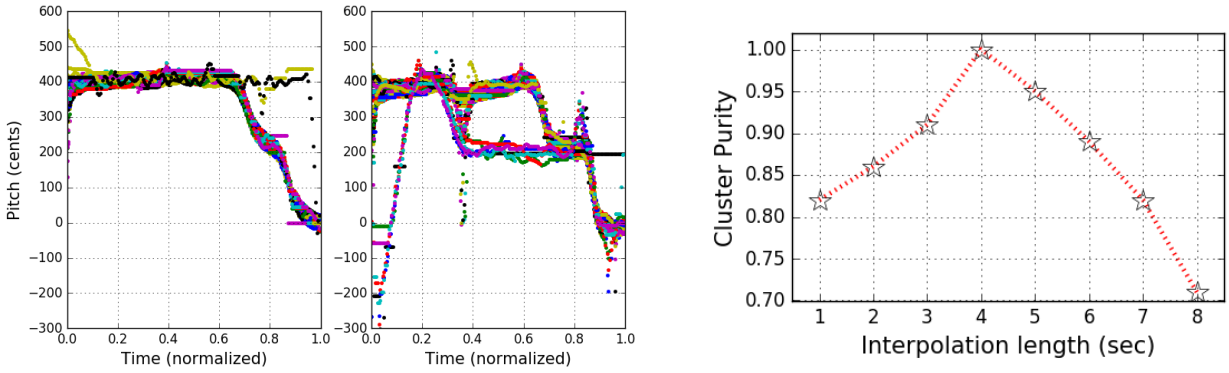

Figure 6.13: [Left] The ground truth GRS phrases of the two ragas Deshkar (left) and Bhupali (right) after DTW alignment to Int_len 4 sec. [Right] Cluster Purity values obtained for different Int_len for the GRS phrase category during unsupervised clustering.

The aim of this experiment is to compare the discrimination performance of the parametric space, as above, with the continuous space, through the same experiment i.e. unsupervised clustering of the 298 phrases into two classes using the DTW distance between the (time-aligned) continuous melodic contours as input feature. From the phrase dataset (refer to Table 3.2), we create pairs of each phrase with every other phrase for the GRS phrase category, i.e. we obtain 110x188 (=20680) pairs for the GRS phrases. The evaluation measure is the same, we plot the cluster purity values for the unsupervised clustering into two classes, for different interpolation lengths as a parameter. Figure 6.13 (right) shows that, for GRS phrase category, the highest cluster purity (=0.99, i.e. only 5 instances of the 298 are misclassified) is obtained for an Int_len of 4 sec. The discrimination performance of the continuous space at this operating point is comparable to that of the parametric space. However, the parameter Int_len is observed to have a high sensitivity on cluster purity – the values drop on either side of the optimal tuning of Int_len.

6.5 Summary

In summary, our findings suggest that important predictors of duration and melodic shape of a raga phrase are: (i) location in a concert (cycle index), (ii) proximity of the approaching sam in the tala cycle, (iii) focal note (nyas svara) in a local context, and (iv) octave in which the phrase is sung. At the same time, duration and melodic shape are independent of: (i) the tala (e.g.

teental or ikvai (both 16 beats) does not matter as long as the tempo is comparable), (ii) tonic of the concert (e.g. the same artist singing with an offset of a semitone in some other concert. Mukhda phrases are observed to show less variability relative to the non-mukhda instances of the same phrase. The focal note in course of the melody evolution influences the duration of the constituent notes within a phrase. Temporal variability is more predictable than the pitch variations (alankars). The alankars, if any, do not show similarity across occurrences of the phrase. However due to proximity of sam, often gamak/tan-like movements are observed. Hence these variations are partially predictable via temporal cues. We also observe that the parametric space, despite being devoid of the actual pitch samples, is capable of capturing the relevant information about the pitch shape that discriminate the two allied ragas. This result is promising, not only because this improves computational complexity but this can be posed as a cognitively-based model for melodic (dis)similarity. We shall see results of behavioral experiments with human experts in Chapter 7, with human experts to tally the found trends and validate through expert listening. This would in-turn improve the computational model, and possibly adapt the tuning of hyperparameters, specific to Hindustani repertoire.

We carried out measurements of pitch and duration on a set of GRS and PDS phrases in two allied ragas Deshkar and Bhupali. The major findings that we can generalize to other ragas/concerts are discussed as below.

- **Predictability:** The 'improvisation' in Hindustani raga performance has a strong underlying structure and is predictable to a certain extent. However there are certain degrees of freedom which the improvised phrase instances are independent of. Between the two major dimensions (temporal and pitch variability), temporal adjustments/distortions are found more predictable. It is the time warping that broadly explains the phrase variability, the pitch variations (alankars like kan) are impromptu and hence unpredictable. However, a typical ornament 'gamak' is often predictable via temporal cues. If the sam (or any other important rhythm marker) is in close proximity, artists often use a tan-like oscillatory movement to adjust the timing.

- **Relative variability:** There is a major difference in behavior of improvisation depending on whether a phrase is a mukhda or a non-mukhda phrase. The mukhda phrase is often constrained with other factors (e.g. same lyrics, # beats allocated, proximity of the sam, the octave in which it is sung) and hence the scope of variations are observed to be limited

compared to a non-mukhda phrase. The same phrase (e.g. the GRS phrase) behaves differently in different ragas depending on whether it is a raga-characteristic phrase (in Deshkar) or is just a part of the ascending/descending melodic progression (raga Bhupali). The time-warping across instances of the same phrase within a concert is observed to be non-linear. Much of the elongation/compression is absorbed by the stable svara regions, while the transient svara-segments are observed to preserve their shape. Global behaviors are observed to translate to local behaviors. E.g. the focal note in a broad region also influences the nyas svara within a phrase in the same region.

- **Prototype:** From all instances of the same phrase, a grand average is indicative of a 'prototype' or exemplar form of the phrase. This hypothesis helps explain the embellishments (alankars) as an additive superposition onto the canonical shape. This also facilitates exploring context-dependence for melodic variability.

Chapter 7

Perceptual similarity of raga phrases

7.1 Introduction

Musicians are trained to produce and recognize raga phrases. An interesting analogy would be to imagine a phrase as a spoken word in a language that musicians understand. We want to present a musician with many acoustic versions (each slightly modified to a different extent from the "canonical" form, e.g., what might be stored in their long-term memory). We would like to know whether they are sensitive to the differences and measure how the physically measured acoustic signal differences relate to perceived differences. To answer the question how this would be useful for us, we expect music learners to make mistakes akin to the deviations in

[0]This chapter is largely drawn from the following papers:

- K. K. Ganguli and P. Rao. "On the perception of raga motifs by trained musicians," in Journal of the Acoustical Society of America (JASA), 145(4), pp. 2418-2434, April 2019. [66].
- K. K. Ganguli and P. Rao. "Exploring melodic similarity in Hindustani classical music through the synthetic manipulation of raga phrases," Cognitively-based Music Informatics (CogMIR) workshop, August 2016, New York, USA. [60]
- K. K. Ganguli and P. Rao. "Imitate or recall: How do musicians perform raga phrases?" in Proc. of the Frontiers of Research on Speech and Music (FRSM), December 2017, Rourkela, India. [62]
- K. K. Ganguli and P. Rao. "Perceptual anchor or attractor: How do musicians perceive raga phrases?" in Proc. of the Frontiers of Research on Speech and Music (FRSM), November 2016, Baripada, India. [61]
- K. K. Ganguli and P. Rao. "Discrimination of Melodic Patterns in Indian Classical Music," in Proc. of the National Conference for Communications, February 2015, Mumbai, India. [59]

certain melodic aspects. If we can predict how a good musician responds to such stimuli, we can give proper feedback to the learner (correct/slightly incorrect/very wrong etc.). We hypothesize that a trained Hindustani musician performs a memory abstraction for the raga characteristic phrases. Our recent work [60] investigated, through acoustic measurements followed by behavioral experiments through listening, the possibility of a canonical form or "prototype" of a raga characteristic phrase. In our context, a prototype may be considered as the phrase that serves to establish the raga around the initial phase of the performance. The case study was conducted for a characteristic phrase DPGRS in raga Deshkar. We first determine all the distinct independent dimensions of actual physical variability by observing actual instances from concerts. We would like to verify whether the existence of a "prototype" only applies to raga-characteristic phrases or it extends to any melodic pattern. The chief objective of the current chapter is to investigate via perception experiments whether a non-characteristic melodic shape behaves like a prototypical melodic motif.

Prototype and exemplar models are the two predominant approaches to modern conceptualizations of perceptual categorization, each providing a different assumption about the nature of category membership [7]. Prototype models attribute categorization to the comparison of incoming stimuli with internal prototypes, which are some form of averaged or ideal category representations. Exemplar models propose that experiences (exemplars) are stored in memory and that categorization is determined by the set of exemplars elicited by the incoming stimulus. Exemplars for a given category thus are specific instances of a stimulus, rather than a single, averaged representation of experienced or idealized stimuli [7].

Our recent study [56] used standard music information retrieval (MIR) tools to explore melodic structures in a data-driven way, to validate certain musicological hypotheses. Judicious use of both data and knowledge can lead to building a cognitively-based computational model that could simulate the human-judgment of melodic similarity. Apart from its pedagogical value, this approach has potential applicability as a compositional aid.

The structure of this chapter is as follows. We first review previous studies on human similarity-judgment in speech and music literature. Next we discuss relevant music concepts to decide on a suitable raga characteristic phrase for a case study. The following section discusses the preparation of suitable stimuli for a set of behavioral rating experiments and results. Finally, we summarize our findings and propose planned future work.

7.2 Literature review

The reported works in literature revolves around speech perception, specifically phoneme categories. A related field in sequence perception is in speech prosody where existence of categorical perception in boundary tones are examined. The past investigations in speech-like stimuli include chord perception in Western music. For all of the above, the baselines have been carefully chosen in terms of familiarity to the stimulus, e.g. native speakers of different languages, persons with/without absolute pitch ability, musicians versus non-musicians etc. To the best of our knowledge, there has been no reported studies on Indian art music.

[73, 110, 135] have reported categorical perception (hereafter CP) in the perception of phonemes. CP is the experience of percept invariance in sensory phenomena that can be varied along a continuum. CP is revealed when an observer's ability to make perceptual discriminations between things is better when things belong to different categories rather than the same category, controlling for the physical difference between the things. Sensory signals that could be linearly related to physical qualities are warped in a nonlinear manner, transforming analog inputs into quasi-digital, quasi-symbolic encodings. CP is an important phenomenon in cognitive science because it involves the interplay between humans' higher-level conceptual systems and their lower-level perceptual systems [73]. The generation of CP (enhanced within-category similarity and enhanced between-category differences) by perceptual learning has been described as the 'acquired similarity (or difference) of cues' but no mechanism has been proposed to explain how or why it occurs [87]. In short, the same entity gets a different identity when the perspective changes [55]. CP's origins are also tied to the 'motor theory' of speech perception, according to which the similarities and differences between speech sounds are determined for our ears by how our mouths would have produced them [1]. This is to say that what we utter is determined by our hearing mechanism in reference to the articulatory mechanism in terms of how the input speech / sound could have been produced at some internal representation – as in a kind of 'silently-talking listener' process. Categorical perception has been of major importance in speech-perception research since it was first named and systematically studied by Liberman et al. (1957) [110]. He demonstrated CP by generating a continuum of equally spaced consonant vowel syllables with endpoints reliably identified as /be/ and /ge/, using voice onset time (VOT) as the feature by varying the second formant transition. As the same concept is intuitively valid for most speech-like stumuli, musical pitch categories may be thought of

as being examples of CP effects that arise primarily as a result of learning. Music indeed is a very interesting test case, because it is less likely than speech to have inborn feature detectors already 'prepared' by evolution [86]. Yet there exist parallels; absolute pitch, rhythm and harmony are among the variables to investigate [1]. Instrumental tone recognition by timbre perception also involves pattern matching of templates and is unanimously agreed to be categorical in nature [158]. We shall now briefly look at the previous attempts made to explore CP in music.

A strong motivation for studying CP in pattern recognition problems is that the irrelevant variations within clusters can be greatly deemphasized. It has been argued that experimental demonstrations of CP are strongly influenced by the way a task is presented [168]. The two main tasks involved in CP experiments in speech that are applied to music are Identification and Discrimination. Among the first attempts of studying CP in music, [28] exploited ascending melodic intervals to design a three-category identification problem with an experimental setup that closely matches speech perception experiments. Additionally in the discrimination task, the subjects were presented two successive melodic intervals and were asked to judge which interval was wider. In a similar study of CP of tonal intervals, [172] remarked that "musicians can't tell sharp from flat": even trained musicians who are great experts in identifying standard musical manipulations (tone, duration) on a pattern, perceive a melodic entity holistically and become deaf to minute changes. All the above cases involved repetition and randomization of stimuli in different trial blocks to ensure consistency of the subjects' response. In the study of CP in musical patterns, [53] argued that there can be only three decision categories defining the degree of "difference" of any inter-pattern relationship: (i) same, (ii) derived, and (iii) distinctly different. The above study addressed a subjective test for the discrimination task by incorporating tonal-duration manipulation on 40 pairs of tonal-rhythmic patterns of electronically synthesized tones with flute-like timbre. Note that there is a difference in the time-scale of the stimuli in this survey as opposed to the aforesaid ones. To summarize, there have been attempts of imparting 'warping' in both pitch and temporal dimensions to the original stimulus in order to observe the perceptual-intellectual demands of the listener in better understanding and recognizing musical patterns.

Several authors have noted top-down effects of musical expectancy interacting with lower perceptual processes. [6, 7, 16, 118] found that in the case of major chords, musical expectancy actually narrows a category, i.e. discrimination becomes sharper near the prototype (it acts

like a perceptual anchor). This brings us to an interesting theory called the perceptual magnet effect (hereafter PME), where a prototype is expected to act either as a perceptual attractor or an anchor. This means that the sensitivity of a listener to discriminate between stimuli is either decreased (attractor) or enhanced (anchor) around a prototype, i.e., the perceptual space is warped with distinct behaviors in regions around prototypical (P) shapes and non-prototypical (NP) shapes. The question to ask for the case of raga phrases is whether the prototypes act like attractors or like anchors.

Prosodic phrases are also shown to be categorical in nature [157, 166]. Authors have taken an analysis-by-synthesis framework to generate synthetic stimuli and perform perception experiments to investigate a possible presence of a 'prototype' representation of a prosodic phrase. Pierrehumbert and Steele [140] presented short phrases differing in peak delay to their subjects. The task was to imitate the presented stimuli as closely as possible. If the presented stimulus continuum represents a gradual change of one specific pattern, then subjects are expected to imitate the continuum correctly. Authors remark that when perception is categorical, the subjects' repetitions should fall into two clearly separable categories with respect to the manipulated feature. Furthermore, the imitations should show an explicit category boundary between the first and the second category. Authors found hints for CP of two peak categories in American English. Several other researchers demonstrated that this method seems to be quite successful for confirming the existence of intonational categories. Redi [155] reported that the subjects did not repeat all points in the continuum but they seemed to have two distinct peak delay categories in mind, which they reproduced. She also argued that this imitation task is a much better design to test for the perception of intonation than the classical one used to confirm CP, because the subjects' intuition is tested in an indirect way, without explicitly asking to identify or discriminate stimuli. This, therefore, seems to be a more natural way to receive information about the behavior and the perception of intonational categories.

7.2.1 Literature summary and proposed method

The works based on representation-cum-similarity measure aimed to develop a computational model informed by human similarity-judgment. There is a commonality among researchers [121, 122, 124, 126, 195, 196] who had explored this view-point that they have recorded subjective similarity-ratings to tune the parameters of a computational model, i.e. the empirically derived (dis)similarity measure used the human ratings in the form of a linear regression. The principle

ideas we learn from their work is: (i) how to choose a set of suitable melodic 'predictors' based on the music repertoire concerned, (ii) how to systematically encode the predictor values while generating variants of the standard melody, (iii) how to judge the significance of each of these predictors and use the values to weight the regression model, and finally (iv) how to evaluate the proposed model by comparison with state-of-the-art distance measures.

The reviewed works on cognitive musicology discuss different memory aspects concerned with music perception and cognition. The main highlights of the concepts we learn include: (i) perceptual attributes of musical dimensions (e.g., pitch, rhythm, contour, timbre, loudness etc. that enable extraction of the 'right' acoustic feature from a music signal), (ii) how humans 'gist'/'chunk'/'group' music information (this reflects perceptually relevant melodic features that help better represent the melody line the way humans memorize it), (iii) short- and long-term memory of music and their mutual relationship (this lead us towards finding a canonical form or a template of a melodic phrase and improvisations thereof), and finally (iv) Gestalt principles applied to music. Though many of these literature talk about high-level cognitive phenomena, we try to adopt the methodological aspects of behavioral experiments. We choose to take up the "recognition paradigm" [125] and "similarity-judgment paradigm" for for designing the experiments.

We plan to carry out three experiments to investigate musical expectancy and the effects of short- and long-term memory across musicians trained in different music repertoires. The two paradigms under study are:

Recognition paradigm

The "recognition paradigm" [125] helps develop the model for phrase similarity as a function of several semi-independent predictor parameters. The ratings are recorded on a n-point Likert-type scale encoding subjects' decision and their judgmental confidence. The idea behind the recognition paradigm is that correct memorization of melodic features should result in the ability of subjects to detect possible differences between the reference and the test phrases. The motivation behind such an experimental design is to judge the long-term memory components of a trained musician or a trained listener.

Similarity judgment paradigm

The "similarity judgment" paradigm involves measurement of melodic distance in an AX phrase-pair configuration. Given the resynthesized melodies A and its variant X, subjects (Hindustani vs. Carnatic vs. Western musicians vs. Non-musicians) are asked to mark their judged similarity on a nn-point Likert-type scale wherein 0 corresponds to least similar and nn corresponds to highly similar. As this paradigm presents a stimulus-pair, we expect subjects to use their short-term memory for the rating task.

7.3 General method

The two main experiments involve two main hypotheses: (i) PME involves one category and a prototype and non-prototype acoustic realization of it, (ii) CP involves two categories and the acoustic boundary between them. Therefore the perception experiments required are: (i) given an acoustic realization, label it as good example of Deshkar through a goodness rating experiment on these and find the prototype and non-prototype member, this is done at individual subject level, and then P and NP are chosen by statistical averaging, (ii) create a continuum of stimuli in equal steps from P and NP, and use these in a AX discrimination test to establish PME. Later, CP identification and discrimination experiments are carried out to examine categorization and the precise location of category boundaries in the perceptual space.

7.3.1 Choice of the phrase: relation to acoustic measurements

We exploit musicological knowledge to choose the stimuli for this experiment. We choose the DPGRS phrase in raga Deshkar where the 'R' is a small step within the glide between 'G' to 'S', but always present. Thus this phrase is a good choice for a controlled experiment. Though one might argue that the typicality of raga Deshkar lies in the GRS portion of the DPGRS phrase, but the context of DP provides the identity to be independently qualified as a raga Deshkar phrase. With our experience of observing raga phrases, raga Deshkar seems to be the most invariant across different artists which makes it a good choice for case study. We extract many instances of the GRS phrase from real-world concert audios by eminent Hindustani vocalists to study the systematic melodic variations.

Figure 7.1 shows a computed prototype shape of the GRS phrase in raga Deshkar. The

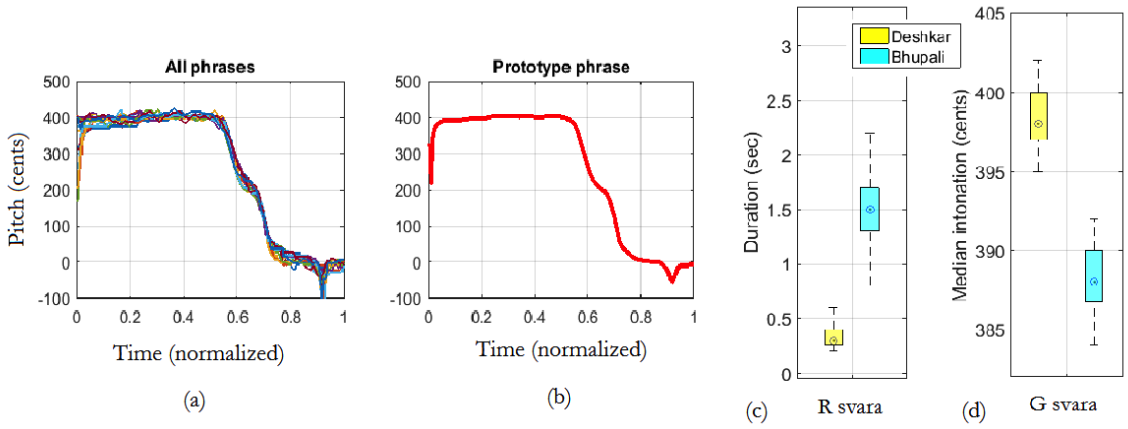

Figure 7.1: Corpus-based approach to derive a prototypical shape for the GRS phrase in raga Deshkar: (a) time-aligned ground truth phrases, (b) computed prototype shape by taking a grand average. Next, data-driven distribution of GRS phrase descriptors for the raga Deshkar and its allied raga Bhupali: (c) duration of R svara, and (d) median intonation of G svara.

audio recordings used in this study partially belongs to the Hindustani music dataset from the Dunya [1] corpora. We considered only alap and slow development of the bandish for the estimation of canonical phrase shape. The annotator (a trained Hindustani musician) is instructed to mark the phrase from the onset of the note G till the offset of the note S in Praat [2] software, however we do not stress on the precision of the boundary markings. A total of 110 GRS phrases are time-aligned using dynamic time-warping (DTW) (Figure 7.1(a)) and a grand average (using Euclidean distance between time-aligned pitch samples) is computed (Figure 7.1(b)). The second set of annotations was the instances of the GRS phrases in raga Bhupali which is an allied raga of raga Deshkar. Shyamrao [102] remarks that, notably the key-phrases of a raga do not consist of any ornamentation but unmistakably consist only of notes. The author takes examples of the ragas Deshkar and Bhupali to state that delineating the fundamental phrases becomes quite complex when the ragas in question comprise the same notes. The GRS phrase is a common phrase between the two ragas and it happens to mark the end of the descending line of the scales.

The phrase is a sequence of svaras whose melodic realization includes specific intonations and transitions to/from neighboring svaras [102]. While computational models for measuring melodic similarity between phrases have employed distance measures between time-series of pitch values of the phrase segments, we might expect that a more discriminative representation

[1] https://dunya.compmusic.upf.edu/Hindustani

[2] http://www.fon.hum.uva.nl/praat/

is possible by explicitly incorporating features that contrast the two ragas. Distinctive features suggested by the comparison are: (i) durations of each of the stable svara regions, (ii) the durations of the glides connecting the svaras, and (iii) the pitch interval of the svara G. The implementation of these features would involve decisions on segmentation of stable svaras, and determining the pitch interval value from the pitch continuum in the region. Figure 7.1(c) shows the distribution of segmented R duration from 110 phrases from raga Deshkar and 188 phrases from raga Bhupali. We observe that there is a region of non-overlap between the two distributions which is of interest to us, and to be investigated further.

The variation of intonation of stable svaras are presented in terms of the distribution of the median values. We observed that the stable svaras (e.g. G in raga Bhupali) can be rendered with embellishments like *kan* (touch note). Trained musicians can focus on a pseudo-steady perceived pitch in presence of the embellishment. If the number of kans is more, the mean pitch of the G segment is biased while median is not so much. Hence we chose to use median to be a metric that, we believe, is a better perceptual correlate of perceived pitch. Figure 7.1(d) shows the distribution of median intonations of the same set of GRS phrases across the two ragas. We observe non-overlapping distributions here as well, which demands further investigation in both the dimensions to finding the technical boundary between the two allied ragas.

7.3.2 Stimulus creation

[98] reported a new methodology to collect human similarity ratings on Flamenco a-cappella recordings. After evaluating various experimental designs (e.g. pair-wise comparison), authors decided to collect the similarity ratings in a free sorting task. Using the sonic mapper [3] software [164], subjects were asked to create groups of similar interpretations, leaving the number of groups open. The participants were explicitly instructed to focus on the melody only, neglecting differences in voice timbre, lyrics, percussion accompaniment and sound quality. Nevertheless, in order to isolate the melodic line as a similarity criterion, the experiment had also been conducted with the synthesized versions of the excerpts.

We incorporate systematic distortions along the extrapolated trend found by the acoustic analyses. As there is a chance that the irrelevant variations get magnified and affect the correct perception of the phrase, stylizing the pitch contour is important. Moreover, by stylizing we get rid of artist-dependent vocal variations. Thus combining all of the above aspects, the stylization

[3] http://www.music.mcgill.ca/ gary/mapper

steps include: (i) segment the phrase into steady and transient segments, and (ii) add measured noise to the steady segments to simulate vocal jitter, replace the transients with a 3^{rd} degree polynomial, retain a few boundary samples for ensuring continuity.

We adopt the reverse strategy to define a time-warping path and generate a time-warped phrase from the given pitch contour. The resultant phrase is obtained by a point transformation of the input pitch data with respect to the given function. The following equation shows the mapping function, where a_i denotes the slope of the i^{th} segment extending from x_i to x_{i+1} and y_{x_i} is the intercept at the coordinate (x_i, y_i).

$$f(x, y) = a_i(x - x_i) + y_{x_i} \quad ; \qquad 0 \leq i \leq n - 1 \tag{7.1}$$

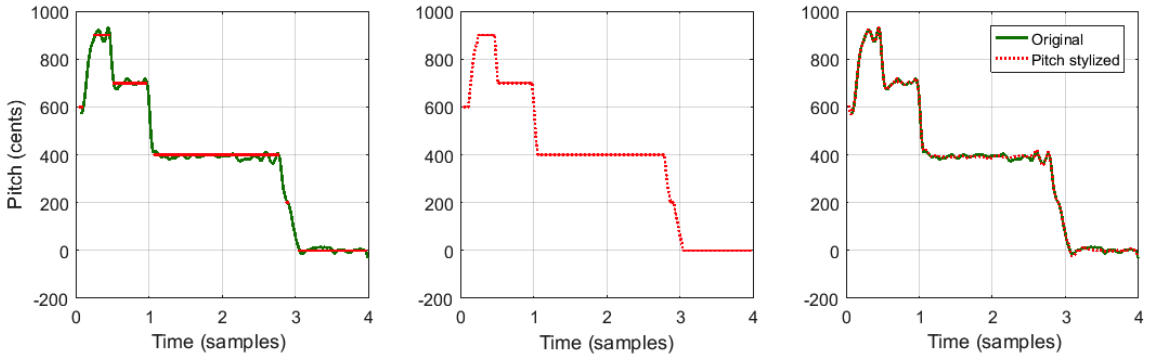

Figure 7.2: Signal processing steps for designing the stimuli from audio.

The values of the slopes a_i are chosen arbitrarily to expand (or contract) a certain segment with the factor a_i ensuring that $f(x, y)$ is a strictly monotonically increasing function, to guarantee possibility of inverse mapping. It follows from equation (7.1) that a_n (path slope of the last segment) cannot be a user input; hence there is a constraint while choosing the reference phrase and the concerned segment for time warping. But here we take a slightly different strategy to maintain the onset location of the last note 'S', hence the 'G' note is compressed accordingly. We define two types of stimuli, viz., (i) Type $\mathcal{A}$: stimuli generated by varying the 'R' duration in the model space, (ii) Type $\mathcal{B}$: stimuli generated by varying the 'G' intonation in the model space. The range and spacing of the physical continuum is discussed in the respective experiments sections.

Another important aspect is synthesis of the stimuli. To remove bias of the artists' voice, we synthesize a voice-like tone with a constant timbre consisting of five harmonics. The relative harmonic weights with respect to the fundamental are empirically chosen as 0, 3, 5, -6 and -20

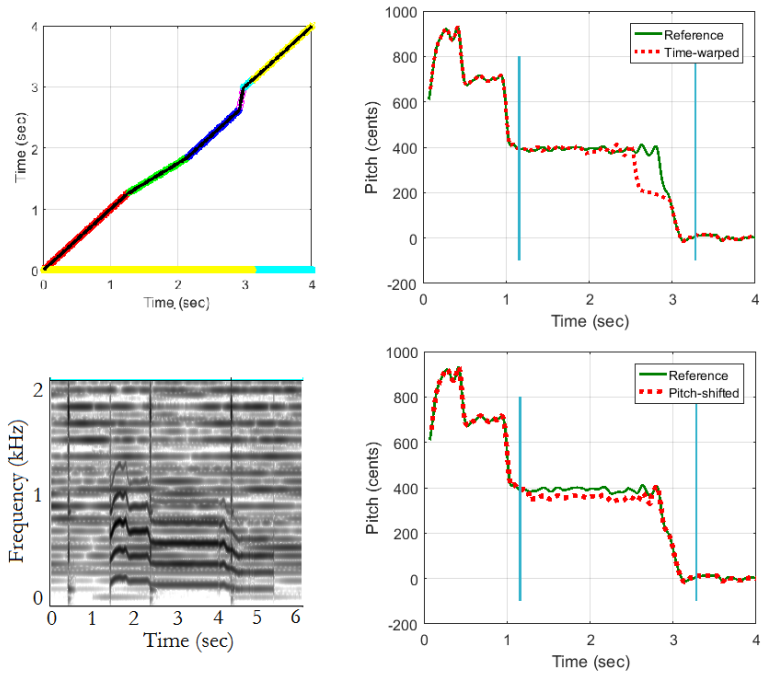

Figure 7.3: Signal processing steps for designing the stimuli from audio.

dB respectively. 5 point median-filtered vocal energy contour is used for the synthesis to retain the natural intensity dynamics. As there is no one-to-one correspondence of a melodic phrase to a raga without the tonic, we add a good quality recorded tanpura track in the background. Based on the feedback obtained from a pilot survey to judge the quality and intelligibility of the synthesized phrase, we add a longer 'tanpura pause' at the start and metronome at 'G' and 'S' note onsets to provide melodic as well as rhythmic anchor to the listener.

7.3.3 Participant summary

We had four disjoint subjects' group based on their discipline of training.

- Twenty seven Hindustani musicians (13 female) with average age of 32 years (SD=4.6) and average years of training of 15 years (SD=2.7) constituted the Hindustani subjects' group, including ten instrumentalists. Seven of the subjects are active performers (4 of vocal music and 3 instrumentalists), five of the subjects (including 2 of the performers) have more than 8 years of teaching experience. All of the subjects had learned raga Deshkar, eight of them performed it at least twice. Four elderly subjects were out of touch from performing for more than 15 years, they were considered as trained listeners.

- Twelve Carnatic musicians (8 female) with average age of 22 years (SD=0.7) and average

years of training of 7 years (SD=1.1) constituted the Carnatic subjects' group. These subjects were postgraduate students at the Department of Music at Madras University in Chennai, India. All of the subjects had exposure to Hindustani ragas, but none of them performed them.

- The Western subjects' group comprises eleven musicians (3 female) with average age of 19 years (SD=0.9) and average years of training of 5 years (SD=1.4). Seven of the subjects are of European origin and their instruments of expertise include piano, cello, and guitar. Four of them (including one from European origin) had heard Hindustani raga performance snippets at least once, but are not regular listeners of the repertoire.

- For the NonMusicians subjects' group, fifteen students (6 female) volunteered who had no formal training in music. These subjects seldom listened to Bollywood (Indian film) music, but did not have any exposure of Hindustani raga performances.

- Apart from these, we had a special subjects' group for Experiment 3 whom we call the IndiPop subjects' group. Twenty four subjects (14 female) with average age of 21 years (SD = 1.2) with no formal training in Hindustani music, but adept in singing commercial Indi-pop (e.g. Bollywood) songs. This category of subjects is relevant to us, because their expertise in imitating a song which is the core task for Experiment 3.

Note

To critically elaborate the nuances of the experimental paradigms, Sections 7.4 through 7.7 would contain discussions only on Type $\mathcal{A}$ stimuli for Trained (Hindustani) musicians and Non-musicians subjects' group [66]. Furthermore, Section 7.8 would include discussion on other subjects' groups for Type $\mathcal{A}$ stimuli as well as all experiments for Type $\mathcal{B}$ stimuli.

7.4 Experiment 1: testing for the perceptual magnet effect

With Experiment 1 we investigate the hypothesis of a Perceptual Magnet Effect (PME) in raga phrase perception. This requires first identifying a prototype (P) and a non-prototype (NP) exemplar, around each of which perceptual discrimation can then be tested. We consider the DPGRS phrase of raga Deshkar with the specific melodic feature of R note duration. Subjects

Index \ Subject	Trained musician					Non-musician
	Hindustani		Indi-pop	Carnatic	Western	
	Vocal	Instrumental				
Expt. 1a	16	7				
Expt. 1b	18	9		12	11	15
Expt. 2a	16	7				
Expt. 2b	18	9				15
Expt. 3	18	6	24			

Table 7.1: Summary of participants for the proposed experiments. The numbers corresponding to Experiments 1 and 2 are for both Type $\mathcal{A}$ and Type $\mathcal{B}$ stimuli; Experiment 3 is conducted with only Type $\mathcal{A}$ stimuli. The sub-experiments (e.g. Expt. 1a) are not indexed as is in the following section headings, but are self-explanatory.

are asked to rate the ”goodness” (i.e. belongingness to raga Deshkar) of each stimulus within the range of R duration variation. Of the stimuli that are thus determined to be associated with raga Deshkar, we identify the best and worst rated ones. These are used in the construction of stimulus pairs representing a range of physical separations with respect to either the P or the NP stimulus.

7.4.1 Method

The Perceptual Magnet Effect paradigm consists of a goodness rating and a subsequent discrimination task with subject participation as provided in Table 7.3.3.

7.4.1.1 Goodness rating

The subjects” task was to evaluate the quality of the presented stimulus with reference to the raga Deshkar motif on a predefined scale from -3 (very bad exemplar) to 3 (very good exemplar). The question posed is “How good is the phrase as an example of raga Deshkar?” Note that, the textual description were provided for only two extremes (‘-3: very bad’ and ‘3: very good’) and the mid-point (‘0: neutral’) of the rating scale.

During the test the 13 stimuli were repeated twice in randomized order within a trial block.

Stimulus no.	Scale factor w.r.t. reference	Absolute R duration (sec)
1	0.01	0.003
2	0.25	0.07
3	0.5	0.15
4	0.75	0.22
5	1	0.3
6	1.5	0.45
7	2	0.6
8	2.5	0.75
9	3	0.9
10	3.5	1.05
11	4	1.2
12	5	1.5
13	6	1.8

Table 7.2: Stimulus description in terms of index, scaling factor and absolute duration of the R note (Type $\mathcal{A}$ stimuli) for Experiment 1. All stimuli from 1 through 13 are used in Experiment 1a and the stimuli 5 through 11 are used in Experiment 1b.

2 trial blocks were presented to each subject with a minimum gap of 1 hour between blocks. Listeners had to select one of the rating values before they could proceed to the next stimulus. Subjects were explicitly asked to use the full range of the rating scale as far as possible. Each stimulus is 4 sec long, added with the tanpura background amounts to 6 sec. Each rating takes no more than 10 sec (assuming single play, but complexity of the Likert-type scale over a binary choice). This accumulates to 30*10 = 300 sec $\sim$ 5 minutes per trial block. If the number of plays is more (upto 3 or 4 for confusing stimuli), the total time taken for a trial block is no more than 12 minutes. The 2 trial blocks are separated by at least 1 hour.

The goal of this experiment was to identify a prototype P and a non-prototype NP representation of the raga Deshkar motif from among the presented stimuli. The P would correspond to the token receiving the highest goodness rating for fitting into the category while NP would be one that can be considered to belong to raga Deshkar but receives the lowest rating.

7.4.1.2 PME discrimination

In the discrimination task, a pair of stimuli, was presented to elicit the judgement of whether the two stimuli were the same or different. One member of each pair was one of the P or NP stimulus determined via the goodness rating experiment. The other member of the pair was chosen from the stimulus set indicated by the middle box of Table 7.4, so that the acoustic difference within the pair was up to to six steps. Using the convention of an AB pair comprising 2 stimuli in increasing order of indices, we create 6 AB and 6 BA pairs out of the 7 stimuli with reference to the P, and the same number with reference to the NP. A matching total number of identical pairs (12) is then included for each of P/NP to avoid any bias. This gives us a total of 192 pairs in the context of 2 repetitions per trial block and 2 blocks per subject.

In the interest of curtailing test duration, only a subset of 12 trained musicians was administered the full set of 192 stimulus pairs. The remaining participants (of the 27+15 as in Table 7.3.3) received a balanced mixed of AB/BA and AA pairs with random selection as executed by the Sonic Mapper software, for a limited set of 96 stimulus pairs.

Each stimulus is 4 sec long, added tanpura background results in $\sim$ 6 sec. So each pair takes 12 sec + inter-stimulus interval (ISI) 0.5 sec = 12.5 sec. If each pair is listened only once, the time taken for each pair is 15 sec. (from start of play to the start of next pair). Thus the whole set takes 48*15 = 720 sec or 12 minutes. If the number of plays is more (usually upto 4 times for confusing pairs), the total time taken for a trial block is less than 20 minutes. The 2

trial blocks are separated by at least 1 hour. The non-control subjects take less time.

7.4.2 Results and discussion: goodness rating

Most subjects used at least 6 of the proposed 7 points of the rating scale. However, four subjects used only 4 values of the scale. Nevertheless, all subjects used rating values from the upper half of the scale (1 to 3) showing there were enough stimuli perceived as good example of raga Deshkar.

Figure 7.4 shows grand average of the ratings (blue) from 23 Hindustani musicians in 2 trial blocks. Thus mean for each of the 15 stimulus is contributed from 23 * 4 = 92 ratings. Red curve shows the response time (in terms of number of replays of the prompt). The results confirmed the membership of the selected stimuli to the Deshkar category. Then, a prototype P as well as a non-prototype NP had to be calculated. For each stimulus, the rating value averaging over all participants is calculated. The stimulus with the highest rating then corresponds to P and the stimulus with the lowest rating corresponds to NP. This procedure however does not take the individual differences into account. As we observed only few individual differences, we chose to take average ratings to determine the P/NP candidates.

The hypothesized prototype (Scaling factor = 1) is observed to have the highest rating, this is chosen as the P. The extension of the stimulus continuum toward the lower R duration resulted in a decreasing goodness rating, supporting the P location. To the right end of the scale, choosing a rational NP is a crucial step. We had intentionally expanded the continuum to include higher R durations. This was validated by the negative goodness ratings. However, as the NP should have a lower goodness rating yet a valid instance of the category, the stimulus with scaling factor 4 (stimulus no. 11) was chosen as the NP given that there is a steep drop in the ratings after this. The goodness ratings for this stimulus range are between 0 and -1 which serves as a good reason for choosing a NP. Also, the response time plot (orange curve: in terms of no. of plays) shows a peak in this region, indicating increased confusion in Hindustani musicians in deciding the goodness. Taking both goodness rating and the response time curves, we pin down to this stimulus as the NP of the raga Deshkar category. Table 7.4 shows the model space for the discrimination test, presented next. Post-experiment comments by some of the subjects indicated that the higher extreme stimuli actually evoked the sensation of raga Bhupali.

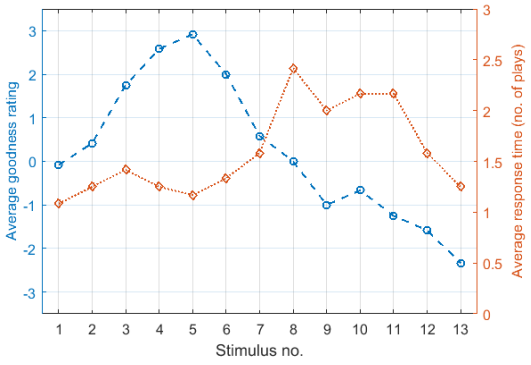

Figure 7.4: Average across listeners and trials of the goodness ratings (blue) from 23 Hindustani musicians for Type $\mathcal{A}$ stimuli. Orange curve shows the response time (in terms of number of repetitions for each stimulus).

7.4.3 Results and discussion: PME discrimination

It is important to consider the effect of order of presentation in the context of stimulus pairs. In perceptual discrimination studies with speech stimuli, pairs with the more prominent stimulus occurring second in the pair are more discriminable, where prominence is defined as produced with more effort [17]. Order of presentation dependence was studied with our control subjects who did both AB and BA versions of every pair of distinct stimuli. We found that while the means of the BA (i.e. the prototype comes second in the pair) were the same or marginally higher than those of AB for the corresponding condition for the musicians, none of the differences was significant in any subject category ($p > 0.01$ in a 2-sample T-test [204]). This phenomenon is illustrated in Figure 7.5.

Given this, we combined the AB and BA pairs for a given stimulus pair to obtain averaged results. Figure 7.6 (a) shows the discrimination performance with increasing acoustic difference from either P or NP as selected from the goodness rating experiment by the Hindustani musicians' group. The error bars are contributed from all ratings (27 subjects * 2 trial blocks * 2 repetitions) per stimulus.

However, the discrimination in the vicinity of the NP was higher for the same acoustic difference compared to the P surrounding as illustrated in Figure 7.6. The neighbors of P were always discriminated significantly worse that the respective neighbors of NP. This indicates that the warping of the perceptual space was found not only for the immediate neighbors of P versus NP, but for the subsequent neighbors too. This is exactly what PME hypothesis is – perception is warped around the prototype, but not around the non-prototype. Thus, the results discussed here show clear evidence for a Perceptual Magnet Effect in the category of low R duration,

corresponding to raga Deshkar. From the non-musicians' response as shown in Figure 7.7, salient finding is that the discrimination was better in the stimulus difference of 4 – 5 steps from P centre. This maybe because of the lower R duration in the reference (P). For the same acoustic difference from NP, the reference shape has a larger R duration which may have affected the discrimination performance.

Expt / Subject	Control subjects				Non-control subjects			
	#AB	#BA	#AA	Total	#AB	#BA	#AA	Total
PME	6	6	12	192	3	3	6	96
CP	10	10	20	160	5	5	10	80

Table 7.3: Stimulus count for the PME discrimination task for control versus non-control subjects for Type $\mathcal{A}$ stimuli.

Stim. diff. / Context	1	2	3	4	5	6
Prototype	0.085	0.092	0.062	0.012	0.021	0.028
Non-prototype	0.084	0.081	0.069	0.076	0.031	0.028

Table 7.4: Statistical significance for the differences of average discriminability between AB and BA pairs (order effect) for prototype and non-prototype vicinity for Type $\mathcal{A}$ stimuli.

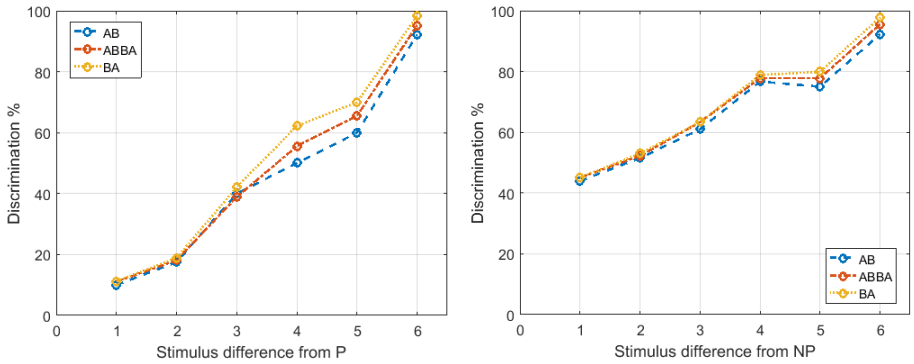

Figure 7.5: Order-of-presentation effect for Hindustani subjects for PME discrimination experiment.

7.4.3.1 Discrimination sensitivity

We observe that with increasing distance from either of P or NP, the average discrimination performance improves, as expected. The discrimination 1 stimulus step away from P is poorer

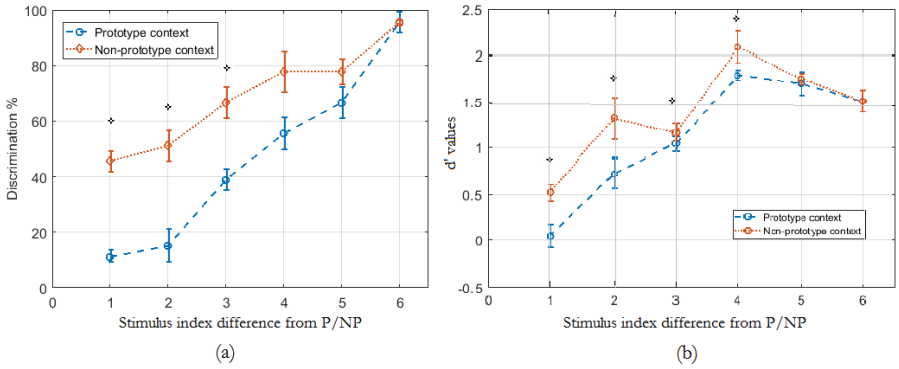

Figure 7.6: (a) Percentage discrimination and (b) d' values, averaged over participants and trials in the vicinity of P/NP for the trained musicians' group; * indicates that the difference between the corresponding P and NP contexts is significant at threshold of 0.01; p = ($6 * 10^{-16}$, 0.0001, 0.009, 0.048, 0.063, 0.1) for discrimination scores, and ($3 * 10^{-8}$, 0.0001, 0.009, 0.0006, 0.049, 0.1) for d'.

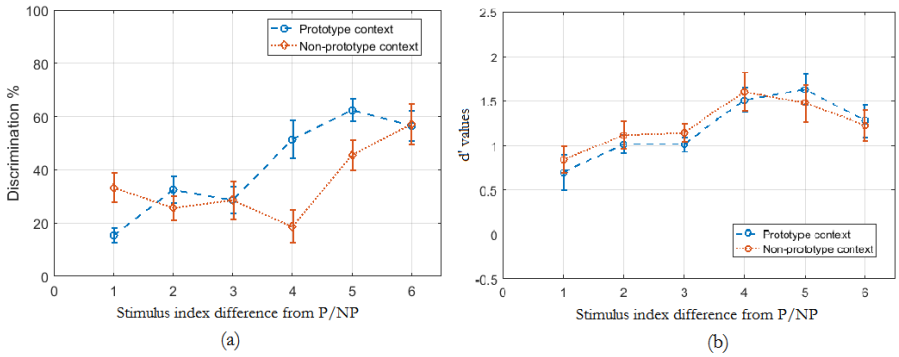

Figure 7.7: (a) Percentage discrimination and (b) d' values, averaged over participants and trials in the vicinity of P/NP for the non-musicians' group. None of the differences between the corresponding P and NP contexts were found to be significant at threshold of 0.01. (p = (0.065, 0.081, 0.1, 0.034, 0.068, 0.1) for discrimination scores and (0.029, 0.029, 0.021, 0.05, 0.03, 0.069) for d' values at 5% significance level).

(10%) than that of NP (50%) and the difference between the two contexts gradually decreases. There is significant difference between P and NP neighborhoods upto 4 stimulus steps. However the percentage discrimination is influenced by two factors, viz. the subject-dependent response bias and the sensitivity or perceptual distance between stimuli. It is the latter represented by d' [205] that we are interested in comparing across the different conditions.

The decision whether a stimulus-pair are same or different, can be explained with the Signal Detection Theory (SDT) [51] that provides tools for analyzing decisions in the presence of possible uncertainty. In perception experiments, subjects almost always differ from each other in identification and/or discrimination performance. According to SDT, listeners who share the same perceptual precondition (identical auditory threshold) can produce different results in a perception test. Equally good listeners can produce different results for the same stimuli, the reason being they might react differently to different features of the signals.

SDT attributes responses to a combination of sensitivity index d' and response bias λ. On any trial in a discrimination task, the rating is "different" when the evidence for the signal is larger than the individual response criterion λ_{center}, and "same" when it is smaller. This implies that the number of hits and false alarms depends on this criterion. The sensitivity index d' provides the separation between the means of the signal and the noise distributions, compared against the standard deviations thereof. The value of d' does not depend upon the individual criterion, but instead it is a true measure of the internal response. The d' is calculated by taking the z-transformations (Gaussian distribution) of the hit rates (h) and the false alarm rates (f) by the following equation.

$$d' = Z(h) - Z(f) \tag{7.2}$$

where h is the hit rate and f is the false alarm rate. However, there is an issue adapting this formula to behavioral experiments. There are no z-transformation values for results with exactly 100% or 0% performance, because z-transformation is a narrowing function in which these ideal cases are not reachable. But our experiments involves ratings of 100% or 0% discrimination performance. However, to ensure the z-transformation values to be closer to the real values, we decided that all values above 99.9% were decreased to 99.9% and all values below 0.1% were increased to 0.1%. With these enlargements of the data range, it is now possible to correctly analyze the results of the experiments.

Figure 7.6(b) shows the distribution of d' values averaged over musician subjects and

Stim. diff. / Subject	1	2	3	4	5	6
Musician	<<0.0001	0.0001	0.009	0.048	0.063	0.1
Non-musician	0.065	0.081	0.1	0.034	0.068	0.1

Table 7.5: Statistical significance (p-value) between discrimination scores of Musicians' and Non-musicians' response for PME discrimination task for Type $\mathcal{A}$ stimuli.

Stim. diff. / Subject	1	2	3	4	5	6
Musician	<<0.0001	0.0001	0.009	0.0006	0.049	0.1
Non-musician	0.029	0.029	0.021	0.05	0.03	0.069

Table 7.6: Statistical significance (p-value) between d' values of Musicians' and Non-musicians' response for PME discrimination task for Type $\mathcal{A}$ stimuli.

trials. There is significant difference between the means for the P/NP context upto 4 stimulus steps. Our results thus support the hypothesis that the trained musicians' perception is warped around the P, but not around the NP. Thus the results indicate Perceptual Magnet Effect in the category of the low R duration, corresponding to raga Deshkar. Discrimination being low around the Prototype supports the perceptual attractor property [157]. The d' profile for the non-musicians does not support the PME hypothesis as the values for P/NP context are not significantly different.

7.5 Experiment 2: testing for categorical perception

Post-experiment feedback by the subjects indicated that the stimuli towards the end of the expanded continuum in the goodness rating task of Table 7.4 actually evoked a sense of raga Bhupali. This suggests the existence of a category at each end of the R-duration continuum. This motivates the next experiment to test for the Categorical Perception (CP) of the DPGRS phrase shape. The second category in question is raga Bhupali; the low R duration corresponds to raga Deshkar while the high R duration corresponds to raga Bhupali.

The stimulus continuum is derived from the PME goodness rating stimuli, with adjusted end-points and resampling the space with equal intervals between the P of raga Deshkar (scale factor = 1) to the stimulus no. 13 (scale factor = 6). This was chosen as the hypothesized category centre for the Bhupali category, adhering to the corpus-based observations. This range was then divided into 11 equally-spaced steps in an arithmetic progression with a step size

of 0.15 sec, causing a continuum from raga Deshkar to Bhupali. The model space of this continuum is illustrated in Table 7.5 with the exact ratio and absolute values.

Stimulus no.	Scale factor w.r.t. reference	Absolute R duration (sec)
1	1	0.3
2	1.5	0.45
3	2	0.6
4	2.5	0.75
5	3	0.9
6	3.5	1.05
7	4	1.2
8	4.5	1.35
9	5	1.5
10	5.5	1.65
11	6	1.8

Table 7.7: Stimulus description in terms of the stimulus indices, scale factor of R duration (Type $\mathcal{A}$ stimuli) with respect to the reference phrase, and the absolute R duration for Experiment 2 and 3.

7.5.1 Method

The Categorical Perception paradigm also consists of an identification and a subsequent discrimination task. Twenty three trained Hindustani musicians participated in on a voluntary basis. Additionally for the discrimination task, the non-musician subjects' group participated.

Identification: During the identification test, subjects had to listen to all 11 stimuli which differed only in the R duration. Each stimulus within a trial block was repeated 2 times and presented in a randomized order to receive a reliable set of results which could be analyzed statistically. Another trial block was presented with the same set of stimuli, but with a different randomized order. Thus each subject had to listen to 44 stimuli to decide whether the stimulus belonged to raga Deshkar or Bhupali. The answers were given by a labeling task, labeled with "Deshkar" or "Bhupali" bins. As discussed earlier, each identification rating takes no more than

8 sec (assuming single play). This accumulates to 22*8 = 176 sec $\sim$ 3 minutes. If the no. of plays is more (upto 3 or 4 for confusing stimuli), the total time taken for a trial block is no more than 8 minutes. The 2 trial blocks are separated by at least 1 hour.

Discrimination: During discrimination, pairs of stimuli which consisted of either identical (AA pairs) or of different (AB) stimuli were presented. For "different" pairs, the stimulus had to be immediate neighbors in the continuum. Control subjects (12) were given both AB and BA pairs. For the non-control subjects, a balanced mix of AB and BA pairs were presented. All pairs were repeated 2 times in each trial block. To exclude any bias towards the same/different choice, a comparable number (40) of AA pairs was added. In total, 80 stimulus pairs presented in randomized order had to be evaluated. Therefore, each pair takes 12 sec + ISI 0.5 sec = 12.5 sec. If each pair is listened only once, the time taken for each pair is 15 sec. (from start of play to the start of next pair). Thus the whole set takes 80*15 = 1200 sec. or 20 minutes. If the no. of plays is more (usually upto 3 times for confusing pairs), the total time taken for a trial block is less than 25 minutes. The 2 trial blocks are separated by at least 1 hour.

Inside each of the hypothesized raga categories identification values were hypothesized to be high. A steep crossover was expected between these raga categories. Furthermore, for the confirmation of CP, a discrimination peak should correlate with the category crossover from the identification, thereby supporting the existence of these two raga categories in Hindustani musicians' perception.

7.5.2 Results and discussion: identification

The identification results averaged over all subjects showed a clear s-shaped curve. Figure 7.8 shows a steep crossover between raga Deshkar versus Bhupali locations around a particular stimulus range 4 – 7. A crossover from one category to the other is indicated by a fall from more than 70% to less than 30% identification score of the respective category. However, there were individual differences in the exact location of the category crossover occurring between 4 and 7 as illustrated in Figure 7.9.

However, the individual results fulfilled the classical CP definition which expects a steep crossover between the categories under investigation. Therefore we examine whether the individual correlations between the identification crossover and the discrimination peak is significant enough to decisively confirm presence of CP between the two raga categories.

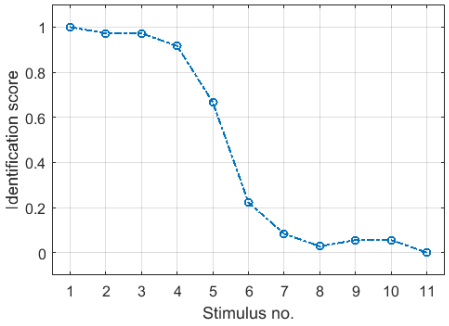

Figure 7.8: Deshkar identification scores by Hindustani musicians versus R duration.

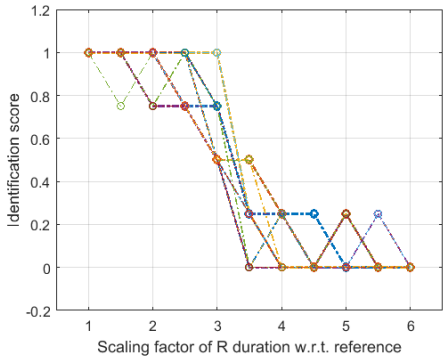

Figure 7.9: Musicians' individual response obtained for the Identification task. The means corresponds to 4 ratings (2 repetitions in 2 trial blocks) per stimulus. The linewidth is proportional to overlapping responses, e.g. the green line at step no. 2 (mean 0.8) is contributed by only 1 subject (thin line) whereas the crossover region (thick lines) is shared by multiple subjects.

$$Crossover = \frac{50\% - Stim_1\%}{Stim_2\% - Stim_1\%} + StimNum_1 \tag{7.3}$$

The results revealed that within each category identification was at a consistent level. The identification values for stimulus 1 to 4 as belonging to the Deshkar category ranged between 95% and 100%, whereas stimulus 7 to 11 received the label Bhupali in 90% to 99% of the cases. These results show that the subjects were able to assign these stimuli definitely to one of the two raga categories. The exact location of the category crossover can be calculated using Equation 7.3 [166]. This equation determines the precise crossover point by interpolating between the values of the immediate stimuli. The crossover corresponds to the point in the continuum where the identification function passes the chance level (50%), at which both categories are equally likely. Averaging over all subjects, the crossover between the categories was found to be between stimulus no. 5 and 6 at exactly 5.13, which corresponded to a R duration of $\sim$ 0.91 sec. Thus a GRS phrase presented with a R duration of 0.91 sec is equally likely to be interpreted as raga Deshkar or Bhupali, making its labeling is undecidable.

7.5.3 Results and discussion: CP discrimination

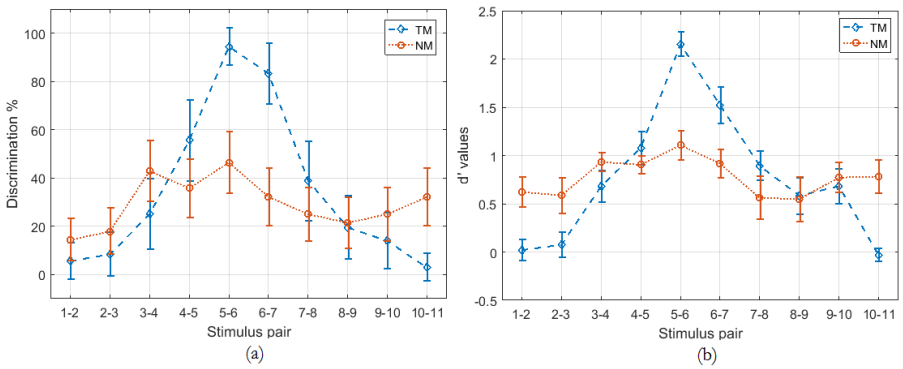

Figure 7.10: Mean and standard deviation of trained musicians' (TM) and Non-musicians' (NM) (a) discrimination scores and (b) d' values, averaged over participants and trials for the CP discrimination task for the characteristic DPGRS phrase.

7.5.3.1 Characteristic phrase context

Figure 7.10 shows the average discrimination function of the same 23 Hindustani musicians. A discrimination peak was anticipated between stimulus 5 and 6. Averaging over all subjects and trials, the discrimination peak between the categories laid between stimulus 5 and 6 at exactly

5.18, which also corresponded to a R duration of ~ 0.91 sec. The maximum discrimination accuracy for each subject is often 100%, except 3 subjects here who reached 75%. Discrimination peak may be distributed across consecutive pairs, (as seen from the plateau in Figure 7.11), we take the mean of these maxima locations for the particular subject.

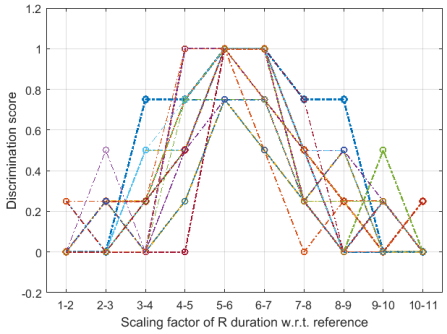

Figure 7.11: Individual discrimination functions for 23 musicians. The means corresponds to 4 ratings (2 repetitions in 2 trial blocks) per stimulus..

The discrimination results, however, revealed some issues in the data for a few participants. Two non-musician participants correctly discriminated almost all stimulus pairs, i.e. no discrimination peak but a high discrimination plateau was found. These two subjects' nearly perfect discrimination might result from perfect melody perception, but hard to explain with their training background. One further non-musician participant marked all stimuli-pairs as the same in the discrimination task. The ratings of this subject had to be excluded from further analyses. Therefore, the number of participants for non-musician subjects for the CP task was reduced to 14. Figure 7.10 also shows the average discrimination function of these 14 subjects, in contrast to the 23 musician subjects.

Pairs	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11
p-value	0.015	0.055	0.071	0.051	0.1	0.038	0.1	0.003	0.064	0.032

Table 7.8: Statistical significance (p-value) between average discriminability of AB and BA pairs (order effect) for CP discrimination task by trained Hindustani musicians for Type $\mathcal{A}$ stimuli for the characteristic DPGRS phrase.

Following the theory of CP, the individual crossover points were correlated to the individual discrimination peaks. For each of the 23 participants the exact crossover was computed using Equation 7.3. A regression analysis with the individual crossover points as the predictor variable and the individual discrimination peaks was performed, which led to a significant ($p < 0.0001$) correlation, illustrated in Figure 7.13. It is observed that the correlation between

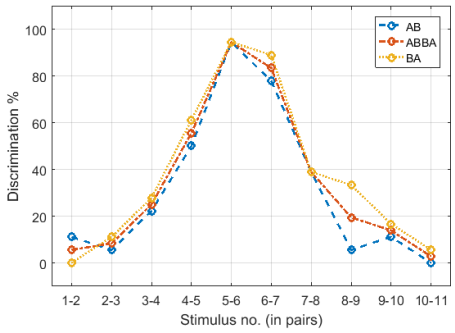

Figure 7.12: Order-of-presentation effect for Hindustani subjects for CP discrimination experiment.

the individual crossover points and discrimination peaks is 0.74, with a slope of regression line 0.53 with an R^2 value of 0.55.

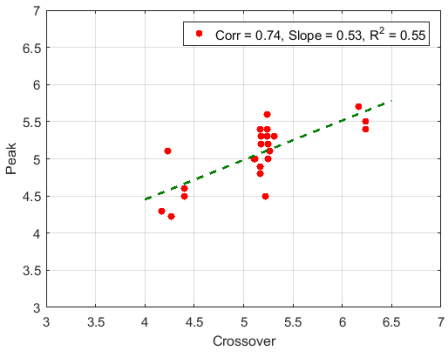

Figure 7.13: Correlation between individual crossover and discrimination peaks for 23 Hindustani musician subjects.

The results may, nevertheless, be partly influenced by the phrase shape used in this experiment, which is discussed later in Section 7.7.

7.5.3.2 Non-characteristic phrase context

A subset of the Hindustani subjects' group consisting of sixteen Hindustani musicians (subset of the participants of Experiment 2b) participated in this experiment. The experimental procedure exactly mimicked the CP discrimination paradigm, with the non-characteristic context of the GRS phrase as below. The reference stimulus for this experiment is a descending melodic sequence DPMGRS which is not a characteristic phrase of any particular raga. Figure 7.14 illustrates its comparison with the DPGRS reference phrase. We simulated this non-characteristic context by recording the DPMGRS phrase by a trained Hindustani musician. This step was necessary, because we wanted the melodic shape of the GRS portion to exactly match with that of

the characteristic DPGRS phrase of raga Deshkar. The same procedure of stylization was followed, the onset of G and S almost coincided, as observed. Both the phrases are of the same duration.

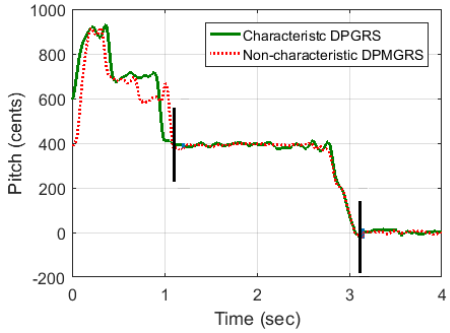

Figure 7.14: Melodic contours of the two stimuli groups. The GRS portion (from the onset of the G svara) of the two stimuli almost coincide upon stylization, the only difference lies in the pre-context DPM versus DP.

Figure 7.15 shows the discrimination performance for the same stimulus index as Experiment 2, for both characteristic DPGRS (orange) and non-characteristic DPMGRS (blue) phrases. The discriminability is observed to be high for the left-most and right-most columns, nullifying the presence of possible categories. The response dips for the intermediate columns, but the four musicians (we studied individual responses of a subset of four musicians) show difference in the locations of the dips. Upon interviewing the musicians, one striking find was that each of them assumed the non-characteristic DPMGRS phrase to belong to a particular raga (either of Shuddh Kalyan, Yaman, Maru Bihag, or Vachaspati). This is an interesting phenomenon, because the non-characteristic DPMGRS was recorded not to be typical of any particular raga (a neutral descending sequence with a high overlap with the characteristic DPGRS phrase). Musicians seem to have anchored to one particular raga (closest to each individual's opinion) and their perception was guided by this assumption.

7.6 Experiment 3: production based perception testing

Experiment 3 belongs to the 'Listen and Imitate' paradigm which was tested by [166] in context of CP. The task is to imitate the stimulus as closely as possible. The general idea for an imitation test is that speakers listening to the stimuli representing a continuum between extreme values of one specific pattern, are expected to produce instances of two discrete sets of stimuli when they perceive two distinct categories. The aim of this paradigm is to test whether subjects

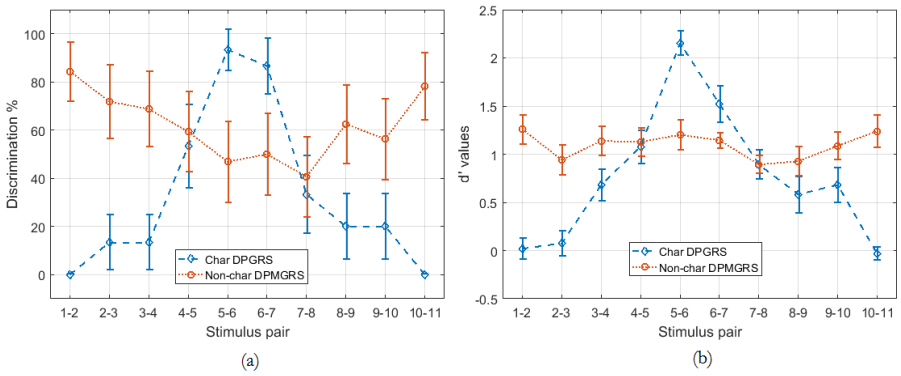

Figure 7.15: Mean and standard deviation of trained musicians' (a) discrimination score and (b) d' values averaged over participants and trials for the CP discrimination task for the non-characteristic DPMGRS phrase.

memorize the stimulus and tend to recall from their working memory or whether they interpolate the stimulus to the closest prototype template already stored in their long-term memory and reproduce the prototype. Several other researchers demonstrated that this method seems to be quite successful for confirming the existence of intonational categories.

7.6.1 Method

This experiment is conducted for both characteristic DPGRS and non-characteristic DPMGRS phrases. To recapitulate, the GRS segment for both phrase categories is almost coinciding and are of same duration, they differ only in the pre-context of the GRS segment – DP for characteristic DPGRS phrase and DPM for non-characteristic DPMGRS phrase, with coinciding G and S onsets.

As the name "listen and imitate" suggests, the experiment involves recording of participants' vocal rendition of the stimuli. Subjects were instructed to repeat the stimulus prompt as closely as possible, no other information about the stimulus was revealed. A total of 22 stimuli (11 each from the two phrase categories) were presented, each followed by recording of the subject's corresponding sung response, in each of the two trial blocks. The stimuli within each trial block were repeated twice and presented in a randomized order. After each stimulus was played, a longer pause was provided with the tanpura drone continuing in the background. Instructions were given about keeping overall length similar to that of the stimulus. While the stimuli consisted of a metronome, no metronome was played in the subsequent pause to ensure that subjects are not burdened with an additional constraint of aligning the G and S svara onsets with the metronome beats.

The recordings were carried out in a quiet environment, on a high fidelity digital recorder (Edirol R-09H) with an audio encoding of 44.1 kHz sampled 16-bit mono PCM (.wav) format. Subjects listened to the stimulus through an over-ear headphone (Sennheiser HD-180) with a moderate volume level. After the stimulus region, the background tanpura played only in the headphone, hence was not recorded on the recorder, ensuring we recorded a clean audio for better melody extraction and further analyses. One interesting finding was that all subjects attempted to maintain the tempo of the prompt, though was not instructed to. In fact, most of the responses were rendered as a continuum (of tempo) of the stimulus prompt.

We segmented (by the same algorithm as stated earlier [56]) the constituent svaras from the pitch contours of the recorded phrases and thereby carrying out acoustic measurements (duration, intonation) on each svara. We store the duration of the stable svaras. Figure 7.16 and 7.17 show the correlation of the model space duration and the sung duration of svara R for both phrase categories and subjects' groups.

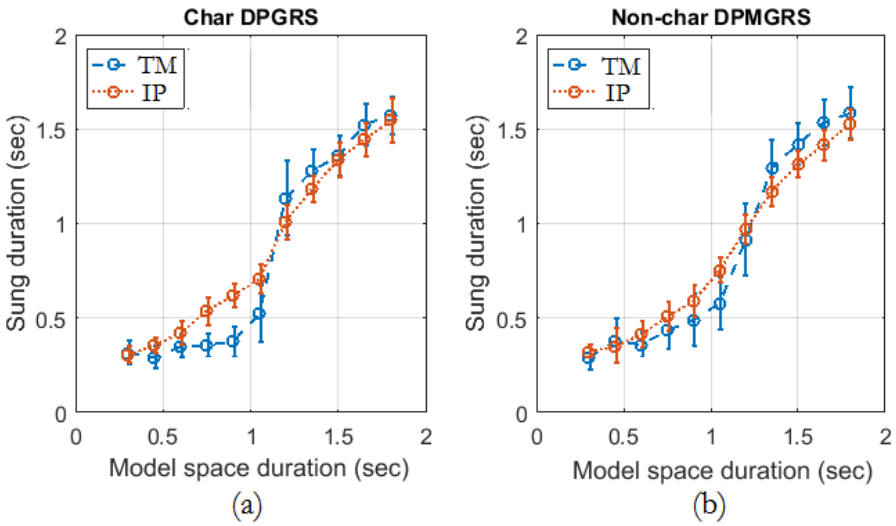

Figure 7.16: Sung duration mean and standard deviation, across participants (TM: trained musicians, IP: Indi-pop singers) and trials, versus prompt duration of R-note in the context of the (a) characteristic phrase, and (b) non-characteristic phrase.

7.6.2 Observations

7.6.2.1 Characteristic DPGRS

The hypothesis is that if a subject listens and exactly imitates, the correlation between model space durations and sung durations would be high and hence the points would lie along the diagonal of the plot. Observation on the data-points for the R svara for 'Trained musicians'

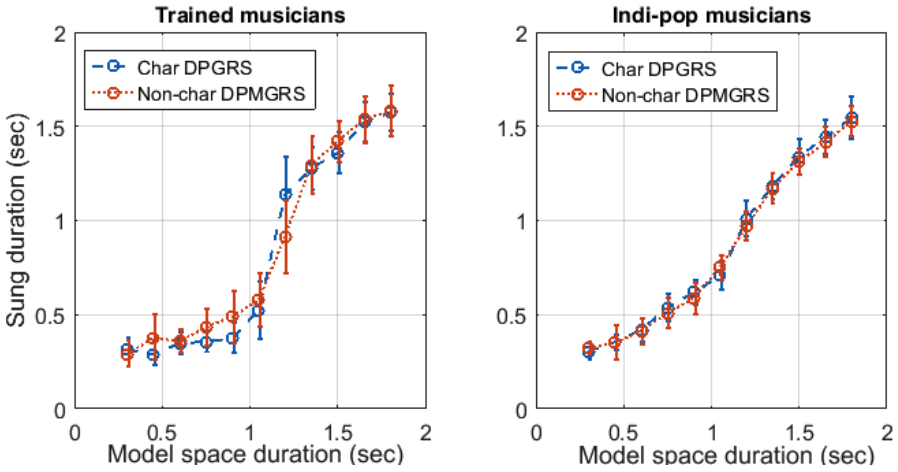

Figure 7.17: Correlation of model space versus sung duration of R svara in Trained (left) and Indi-pop (right) musicians' group between characteristic DPGRS and non-characteristic DPMGRS phrase categories.

subjects' group indicates a proper categorization in the sung duration with a sharp category boundary around the stimulus index 6 (absolute duration of 1.05 sec). The consistent short duration of the R svara for the stimuli indices 1 through 6 indeed corresponds to the prototype shape of raga Deshkar GRS phrase. However, the other data-points (corresponding to stimuli indices 7 through 11) shows more dispersion, resulting in higher slope of the fitted logistic function. The distribution of the sung duration of R svara (not shown in plot) is bimodal with a sharp peak at .4 sec and a flatter long tailed peak 1.3 sec. This indicates that trained musicians, who are familiar with the raga which the stimuli belonged to, used their stored prototypes, though instructed to imitate a given stimulus.

Indi-pop musicians, on the other hand, followed the melodic shape almost exactly – this is evident from the diagonal nature of the fitted curve. The subjects were also observed to adjust the G and R svara durations to keep the overall duration between the G and S svara onsets constant. Notably, all subjects, though not instructed to, tapped their hands at the given tempo of the metronome beats within the stimuli and tried to maintain it in their renditions of the phrase. We define a metric, γ as the ratio of sum of G and R durations in the model space and the musician-rendered phrases. The distribution of γ (mean=.94, SD=.3) indicates that subjects intended to maintain the tempo of the stimuli.

7.6.2.2 Non-characteristic DPMGRS

The case for non-characteristic DPMGRS is interesting, because there is no expected categorical context for this phrase for the trained musicians also. As expected, for Trained musicians subjects' group, the curve tends to graze the diagonal, the contrast with respect to the characteristic phrase is evident from visual comparison (refer to Figure 7.16(left) between blue and orange curves. The interesting observation is the high dispersions of the R svara along the mid-range of the model space. These large error-bars resulted from individual differences. All of these subjects disclaimed that they assumed the phrase to belong to a particular raga (e.g. Yaman, Sudh Kalyam, Vachaspati, and Maru Bihag) as also observed from the listening experiment of these stimuli group. The distribution of γ (mean=.96, SD=.2) indicates that the phrase is more closely imitated. The Indi-pop musicians' response is no different (evident from almost overlapping curves) from the characteristic phrase context, as speculated.

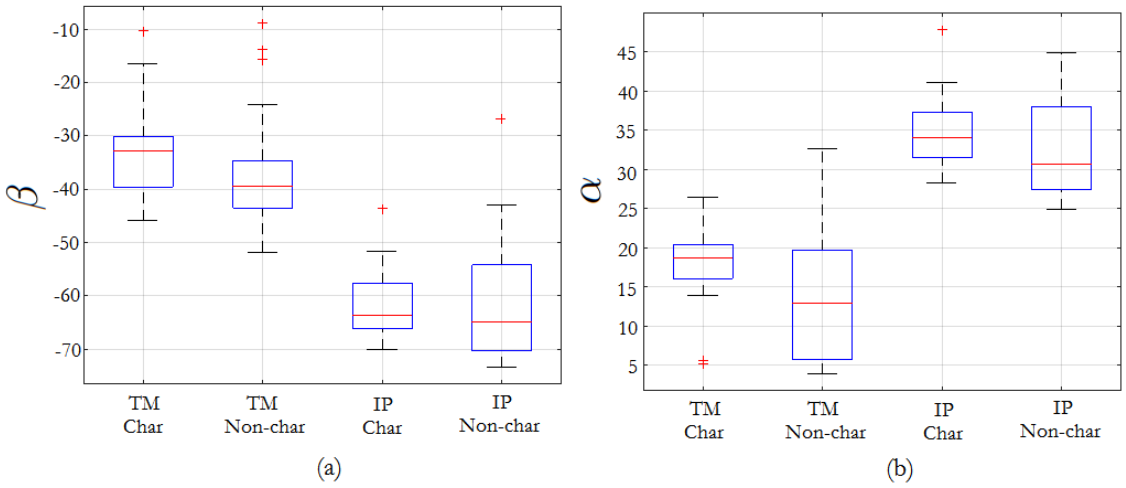

Figure 7.18: Boxplots of individual participant's (a) β and (b) α values for characteristic DPGRS and non-characteristic DPMGRS phrases for trained musicians (TM) and Indi-pop singers (IP) participant groups.

7.6.3 Statistical model fit

Logistic regression fits a model with a bias intercept and stimulus-tuned slope coefficients, but it differs from linear regression in that the model is fitted in a logged odds space. The logged odds transformation converts proportions in the range 0 to 1 into logits in the range $-\infty$ to $+\infty$. Logit values from the fitted model can be converted to probabilities so that fitted curves in the logged odds space become sigmoidal curves in the probability space. A logistic regression

model was fitted to each participant's sung R duration in the char DPGRS as well as non-char DPMGRS phrase. The model included a bias coefficient, and a duration-tuned coefficient, as shown in Equation 7.4.

$$p(R|Bhupali) = \frac{1}{e^{-(\alpha+\beta*R_{duration})} + 1} \tag{7.4}$$

where $p(R|Bhupali)$ is the expected R duration raga Bhupali category, α is a the intercept i.e. displacement along x-axis, and β is the regression coefficient [120] i.e. the growth rate.

One of the inputs to the model is the sung (raw) durations of the segmented R svara. First, the array of 44 R duration values for the stimulus continuum (indices 1 to 11, repeated 2 times for each phrase category in 2 trial blocks) is min-max normalized between [0,1]. The second input to the model is a categorical array of the stimulus continuum. We provide a categorization of 5 Deshkar + 6 Bhupali candidates as obtained from the CP results. The inferences from Figure 7.18 is as follows.

- On a broad level, the distributions for trained musicians and Indi-pop musicians are separated. The median values within each category are comparable. The absolute value ranges of β for trained musicians category corresponds to a step-like sigmoid function indicating categorization in the response. The same for the indi-pop category resembles a more linear response grazing the diagonal, indicating more exact imitation.

- The median for trained musician for the non-char phrase category (2^{nd} column) is lower than that of the char phrase category. This indicates that the response is less categorical, which is also observed from the raw durations. The dispersion (including the outliers indicated by red $+$) is also larger with respect to the characteristic phrase category, indicating more individual differences.

- Response from Indi-pop musicians for the char phrase category (3^{rd} column) has the lowest dispersion. This information, along with the low median value, indicates that the imitation task was very closely followed.

- The Indi-pop musicians for the non-characteristic phrase category (4^{th} column) has a comparable median for β but larger dispersion. This may have been caused by the inclusion of the extra M svara which could be difficult to adopt quickly by this musician category since the density of svaras. Because the participants roughly maintained the

overall phrase duration, they made the R duration smaller to make up for the extension caused by the D, P, M svaras. This gave rise to the higher β values (in the range of -40 to -50).

The distribution means for trained subjects' group are significantly different from that of the Indi-pop subjects' group, however means within groups are not significantly different. For the trained musician category, between characteristic versus non-characteristic phrase, $p = 0.18$. For the characteristic phrase category, between trained and untrained group, $p = 2.14 * 10^{-16}$.

Figure 7.17 (left) shows the correlation of model space versus recorded space for the Trained musician subjects' group. We consider the array of average values (indicated by circles) as the representative of the two phrase categories. The β value obtained for the characteristic DPGRS and non-characteristic DPMGRS phrases are -37.9 and -42.8 respectively. This result may appear counter-intuitive, but the reader should note the high dispersion of the (orange) curve corresponding to non-characteristic phrases.

In general, the imitation task design which does not tell the listener what has to be imitated and how closely, seems to be a good alternative to a standard CP test design, because the results seem to be achieved in a more natural situation for the subjects. Our findings suggest that trained Hindustani musicians perceive melodic phrases categorically, with less sensitivity to small changes around the acquired prototype region. Overall, the results indicate presence of CP in trained musician's perception and that they perform a memory abstraction in the long-term memory (LTM) for the melodic shape of raga characteristic phrases. Skilled vocalists, yet formally untrained in the repertoire, performed the imitation task with high precision. The stimulus prompt length was kept short to conform to the human acoustic memory. Also, it may be assumed that the skill of vocalizing a phrase is a part of the procedural memory for these participants. So, the results should apparently provide a true reflection of the perception of these melodic phrases and their association in the LTM.

One interesting finding was that all subjects attempted to maintain the tempo of the prompt, though not instructed to. In fact, most of the responses were rendered as a continuum (of tempo) of the stimulus prompt. Similar to the perception about belongingness of the non-characteristic DPMGRS phrase to a possible close raga where it fits, trained musicians again showed the same phenomenon in the imitation task. This indicates that musicians tend to interpolate any melodic pattern to a familiar raga where it 'fits' in and thereby their perception is not acoustically, rather

semantically, driven. With the convergence of the current findings with our previously reported ones, we believe that indeed there is a categorization in the musicians' perception and they, consciously or unconsciously, recall the memorized prototype while rendering a raga characteristic phrase in isolation.

7.7 General discussion

The broad goal of the work is to develop a computational model to measure phrase-level melodic similarity in Hindustani raga phrases. A set of acoustic measurements were carried out on a number of musician-annotated raga phrase (GRS characteristic phrase in raga Deshkar) in a supervised manner, to study the dimensions of variation. Extracted from a data-driven method, we proposed a canonical or 'prototype' melodic shape model for the GRS phrase in raga Deshkar. To validate our finding, we created artificial stimuli by varying the learned dimensions in the model space. Resynthesized melodies (with uniform timbre across stimuli) were rated by trained Hindustani musicians to record the 'goodness rating' and find prototypical and non-prtotypical shapes corresponding to the phrase under study. Next, we carried out an AX differential discrimination experiment to estimate the sensitivity to small changes in stimuli across raters (trained Hindustani and Carnatic musicians, Western musicians with no exposure to Hindustani music). The findings suggest that trained Hindustani musicians perceive melodic phrases categorically, with less sensitivity to small changes around the prototype region, indicating that prototypes work as a perceptual attractor and that musicians tend to perceive melodic phrases holistically. The same set of experiments were carried out with the same GRS melodic shape with a different pre-context where it was not characteristic for any particular raga. Also, individual differences among musicians' responses gave insights towards interpreting the possible perceptual mechanisms at play.

Though the subjects are not given any prior information about the stimuli (except a short practice session to habituate with the setup), all subjects rate the same stimuli differently. This indicates that subjects are able to create a mental mapping of the dimension of variation across stimuli within a trial block. E.g. for the non-characteristic phrase for Hindustani musicians, we do not have any hypothesized prototype. If the model space variation was the only predictor, we would have expected the same prototype stimulus (from the characteristic phrase category) would also show low discriminability which is not the case. The possible reason is explained

as follows. Trained subjects show link to learned categories for identification task, which is used further for the discrimination task. Responses show that subjects incline to link between the stimuli and some learned category while discrimination task, which ideally should involve short-term memory of acoustic (dis)similarity and not a learned category. It is interesting to observe individual differences which resulted in different locations of low discriminability.

The test phrase may be syntactically biased in favor of raga Deshkar mode because of the higher G intonation (as used in raga Deshkar) used for all stimuli. However, it should be taken into account that this syntactic bias may or may not have been compensated for by the fact that there might be an inherent hierarchy of dimensions that trained subjects pay attention to. Future experiments should therefore use a hybrid stimulus design that avoid such biases. In sum, Categorical Perception was found for the categories of raga Deshkar and Bhupali for trained Hindustani musicians. The broader crossover obtained from the individual identification function correlates with the individual discrimination peak.

In regard to control subjects, we observed high inconsistency within a trial block (for the repeated stimulus-pair) in the non-musician subjects, hence it was not reliable to compare across trial blocks. However, the musician subjects were consistent within a trial block, hence we examined the "order of presentation" effect across trial blocks.

7.8 Extended experiments

While we described all experimental paradigms and associated results in the previous sections for the R-duration manipulation (Type $\mathcal{A}$- stimuli) for trained Hindustani musicians and Non-musicians. In this section, we dicsuss the remaining stimulus type and subject groups.

7.8.1 Type $\mathcal{B}$ stimuli

We first present the Experiment 1a (goodness rating) counterpart for the Type $\mathcal{B}$- stimuli. The aim of this experiment was to find a prototype (and a non-prototype) of the specific class (belongingness to raga Deshkar). The x-axis labels indicate the offset of the G svara, e.g. abscissa 0 refers to the reference phrase. The shift ranged from -35 to +35 cents with a step of 5 cents, making a total of 15 stimuli to be rated. As discussed earlier, each trial block had 2 repetitions of the stimuli in a randomized order.

Figure 7.19 For Type B stimuli, the reference phrase shape (x-axis value = 0) is observed

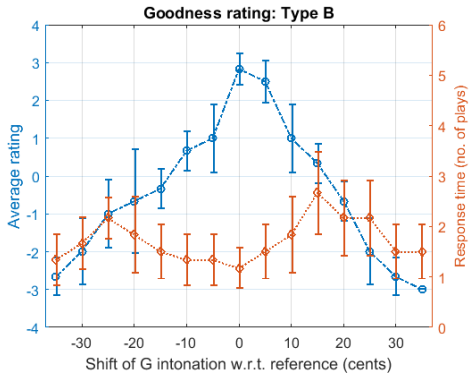

Figure 7.19: Grand average of the ratings (blue) from 23 Hindustani musicians in 2 trial blocks for Type $\mathcal{B}$ stimuli. Red curve shows the response time (in terms of no. of repetitions each stimulus was played).

to have the highest rating. There is a drop in goodness rating towards both sides of this stimulus index. All ratings for shifts beyond ± 30 cents lie below -2. Hence we truncate the stimulus continuum at ± 25 cents shifts with respect to the reference. We observe that mean reaction time (in terms of number of plays) around the reference phrase shape is lower than the vicinity towards both sides of the continuum.

Next, we divide the acoustic difference between the extremums into 11 equi-spaced intervals (5 cents each) for the discrimination experiment. The model space for Type $\mathcal{B}$ stimuli is presented in Table 7.8.1.

Experiment 1b involves measurement of melodic distance in an AX phrase-pair configuration in a differential discrimination setup. The objective of this experiment was to verify the perceptual magnet effect (PME), i.e. whether prototypes act as perceptual attractors or anchors. According to perceptual attractor hypothesis of the PME, discriminability near the prototype centre is poor compared to away from the prototypical region. We hypothesized that musicians would be less sensitive to the dissimilarities between the AX pair if either of them is one of the prototypes (or close to the prototype centre). Similarly the AX pair away from the prototype centre was speculated to be more sensitive to small differences in musicians' rating. However, as we did not have a clear hypothesis about any possible prototype and a clear non-prototype was not found from the goodness rating, we conducted an all-versus-all configuration. 11 unique stimuli were obtained for each stimulus type, making a total of 110 pairs ($\binom{11}{2} + 55$) to be rated either same or different. To capture the 'order of presentation' effect, additional pairs were required making it 220 pairs. Randomization was incorporated in each trial block. The ratings were taken in two trial blocks from 27 Hindustani musicians, 12 Carnatic musicians, 11

Stimulus no.	Offest (cents) w.r.t. reference	Absolute G intonation (cents)
1	-25	369
2	-20	374
3	-15	379
4	-10	384
5	-5	389
6	0	394
7	5	399
8	10	404
9	15	409
10	20	414
11	25	419

Table 7.9: Stimulus description in terms of the stimulus indices, offset of G intonation (Type $\mathcal{B}$ stimuli) with respect to the reference phrase, and the absolute G intonation for Experiment 2.

Western musicians, and 18 NonMusicians subjects.

The results of the differential discrimination experiment is shown in the form of "proportion of 'different' responses" for closely spaced stimuli in the model space. The interpretation of the figures (e.g. Figure 7.20) is summarized as follows: for each column on x-axis (e.g. first column $\{$1-2,2-3,1-3$\}$), the y-axis distribution corresponds to the proportion of the stimulus-pair marked as 'different' across all pairwise comparisons. The median (marked in red) being close to 0 indicates poor discriminability and vice-versa, statistical outliers are marked by $+$ symbol.

For Indian musicians (Hindustani and Carnatic), there is an observed asymmetry which is interesting. The mean intonation of the G segment in the reference phrase (394 cents) is higher than the 'Just Intonation' (JI) tuning of G intonation (386 cents). Therefore, upon lowering the intonation of G around stimuli range $\{$3-4,4-5,3-5$\}$, this resembles the JI tuning which is consonant in presence of the tanpura. This could be one possible reason of musicians overlooking the differences and listening holistically. Whereas a positive shift of intonation (stimuli range $\{$7-8,8-9,7-9$\}$) causes it sound 'out-of-tune' (as described earlier) that might have resulted a sharp change in discriminability. Figure 7.20 shows the average response for different subjects'

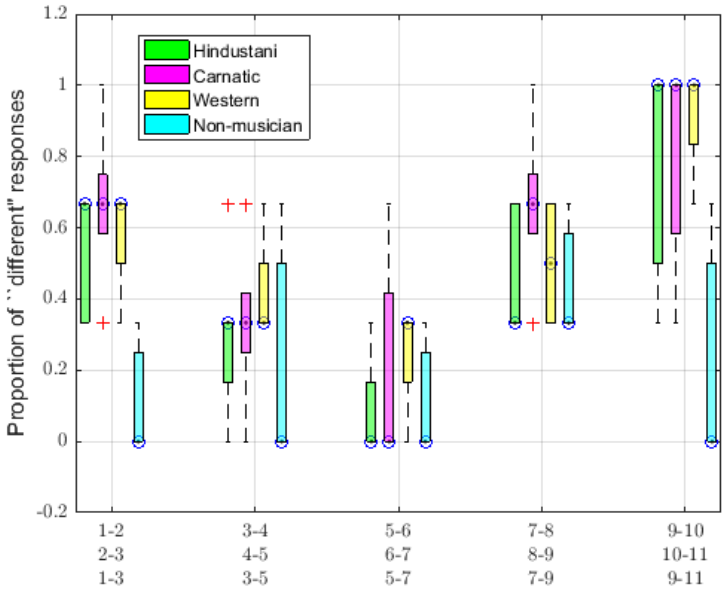

Figure 7.20: Aggregate "proportion of 'different' responses" for closely spaced Type $\mathcal{B}$ stimuli by the (A)Hindustani, (B) Carnatic, (C) Western musicians, (D) Non – musicians subjects' groups. Each trial block consisted of pairwise comparison of 11 equi-spaced stimuli in the model space.

group.

- For the Hindustani subjects' group, we observe the median to be close to 0 for the middle column ({5-6,6-7,5-7}) and gradually increasing to the left and right. The low discriminability around the {5-6,6-7,5-7} pair indicates presence of a $\mathbb{P}$ centre, musicians discrimination around the prototype is poor. There is an asymmetry in Hindustani subjects' ratings (lowering the G up to 10 cents had similar effects to the prototype), indicating the width of the prototype is large.

- For the Carnatic subjects' group, a similar trend to that of Hindustani subjects' group is observed. The median values, however, had a sharper change relative to the $\mathbb{P}$ centre ({5-6,6-7,5-7}). The asymmetry is observed here as well – the stimuli with positive offset in G intonation were highly discriminable compared to the ones with a negative offset value, making the $\mathbb{P}$ region wider.

- For the Western subjects' group, ratings show that the subjects are sensitive to small pitch offsets, discrimination in the range around the hypothesized prototype is better than that of Hindustani musicians. However the asymmetry exists here too. This might be due to the fact that G intonation around the Just Intonation (386 cents: stimulus index 4) major

third with the background tanpura creates consonance. Discrimination is observed to be poor around this stimulus range.

- For the NonMusicians subjects' group, there is no observed trend (like Type A stimuli), except the region of stimuli range {7-8,8-9,7-9} where a better discrimination is observed, which may have the possible explanation. The mean intonation of G note for these phrases are very close to the equi-tempered G location (400 cents) and may have caused a roughness in presence of the tanpura. This might have created a disturbance, causing the subjects to mark the pairs as "different".

In sum, discrimination results for Type $\mathcal{B}$ stimuli across different subjects' groups lead to more of an acoustic similarity rather than a learned schema for intonation. Hindustani musicians' response show that the perception of a melodic phrase is holistic – lower sensitivity is shown (see [172]) for even larger steps of intonation difference while a pilot experiment showed that these musicians are sensitive to small intonation differences when isolated notes were presented. Only out-of-tune intervals showed highest sensitivity – this is true for the presence of tanpura due to the psychoacoustic effect of dissonance. Carnatic musicians showed similar intonation perception which is reasonable given the similarity in the pedagogy. For Western musicians, one major factor was that they were not accustomed to hearing phrases with a drone in the background, however, in general, these musicians showed the highest sensitivity to small differences. For Non-musicians, there was no specific trend found while they could not make out any difference other than the out-of-tune interval, which might have resulted from a disturbance caused due to biological hearing and has nothing to do with their music perception at all.

7.8.2 Equivalent discussion on type $\mathcal{A}$ stimuli

For the integrity of description of experimental methods, we had restricted our discussion of Type $\mathcal{A}$ stimuli to only trained Hindustani musicians and Non-musicians in Sections 7.4 through 7.7 in this chapter. This was also based on the rationale that existence of a prototype is not a fair assumption for Carnatic or Western musicians. Here we revisit the discrimination results for all subjects' groups for the same all-vs-all configuration, with a view to finding any pattern in discriminability and not concluding on PME or CP effects.

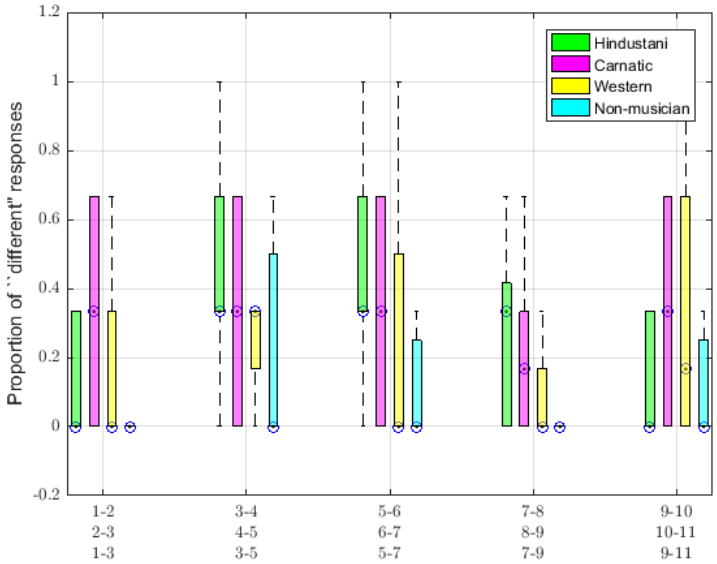

Figure 7.21: Aggregate "proportion of 'different' responses" for closely spaced Type $\mathcal{A}$ stimuli by the (A)Hindustani, (B) Carnatic, (C) Western musicians, (D) Non − musicians subjects' groups. Each trial block consisted of pairwise comparison of 11 equi-spaced stimuli in the model space.

Figure 7.21 shows aggregate response for the all subjects' groups. The interpretation is as follows.

- We observe the median to be close to 0 for the left-most column. The low discriminability around the {1-2,2-3,1-3} pairs indicates presence of a $\mathbb{P}$ centre, musicians discrimination around the prototype is poor. The median for the subsequent columns increase gradually, but again decreases at the last column {9-10,10-11,9-11}. This indicates possible presence of another $\mathbb{P}$ around {9-10,10-11,9-11} which was hypothesized to be a $\mathbb{NP}$ centre. This is indeed true with the region corresponding to the GRS melodic shape that is characteristic of a different raga.

- In contrast to the Hindustani subjects' group, the median for the first column ({1-2,2-3,1-3} pairs) for the Carnatic subjects' group is higher than the other columns, with a decreasing trend toward the right. This is indicative of a $\mathbb{P}$ centre around {9-10,10-11,9-11} which was hypothesized to be a $\mathbb{NP}$ centre. The Carnatic subjects' group were not familiar with raga Deshkar, but were exposed to Hindustani raga performances. The presence of a prototype (though the median is not 0) indicates that the GRS melodic shape is characteristic of some raga that the subjects are familiar to.

- For the Western subjects' group, the ratings show a low discriminability around the {1-

2,2-3,1-3} pair, but high discriminability at the other end. The poor discrimination around the hypothesized $\mathbb{P}$ centre was not speculated for this subjects' group, as they were not exposed to raga Deshkar before. From interviewing it is suspected that they seem to have attended to the variations in the R duration only when its absolute duration exceeded 1 second. Towards the right the discrimination was good, indicating that the subjects used the acoustic differences to rate the similarity. For that matter, the high discriminability around the $\mathbb{NP}$ centre might not have arised due to a perceptual anchor, for the same reason. Difference between 0 and 50% in the first sample.

- NonMusicians' responses are not consistent or stable, this tallies with the findings of [124]. Upon interviewing, some subjects expressed that they did not find any difference at all while others said all pairs sounded different. We had carried out stimuli synthesis by controlling timbral factors and only melodic variations were incorporated. But it is not clear from the ratings which aspect of the stimulus the subjects attended to.

7.8.3 Visualization of perceptual space

To visualize the mapping between the model space and the perceptual space, we carried out multidimensional-scaling (MDS) on the ratings of Type $\mathcal{A}$ stimuli by Hindustani subjects' group. Figure 7.22 shows the 1-D MDS which is similar to the model proposed by [135]. The order of stimuli in the model space is preserved in the perceptual space, with non-linear spacing between them. This confirms the hypothesis that the perceptual space is warped. Again, the proposed $\mathbb{P}$ centre acts as a perceptual attractor, where the discriminability is poor.

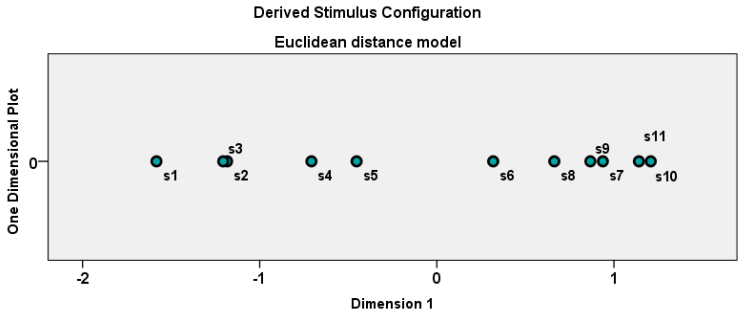

Figure 7.22: Perceptual space: multidimensional scaling (1-D) of the (dis)similarity matrices of the aggregate response for Type $\mathcal{A}$ stimuli by the Hindustani subjects' group.

The number of dimensions in MDS is a parameter to be tuned. While projecting the ratings on a single dimension showed the warping in the perceptual space, the stimulus order was also

preserved, indicating the model space variation is rightly captured in the single dimension. We wanted to visualize the MDS in 2-dimensions and validate if the same grouping is observed. First, we verified that we had enough data-points for a valid 2-D MDS via the equation $J = 40D/(I-1)$ (where J: minimum no. of subjects, D: expected dimensionality, and I: no. of stimuli used in the experiment) proposed by [72]. With D=2 and I=11, the minimum no. of subjects required J=8, we have 16 subjects in the Hindustani musicians subjects' group. Figure 7.23 shows the 2-D MDS of the ratings for Type $\mathcal{A}$ stimuli (left). We observe that in one of the dimensions, the order of the stimuli is preserved (in opposite direction) with respect to the model space. There are three clusters observed, two of the $\mathbb{P}$ regions corresponding to ragas Deshkar and Bhupali and the third with the stimuli between the two $\mathbb{P}$ centres along the R svara duration continuum. However, the other dimension (y-axis) is difficult to interpret.

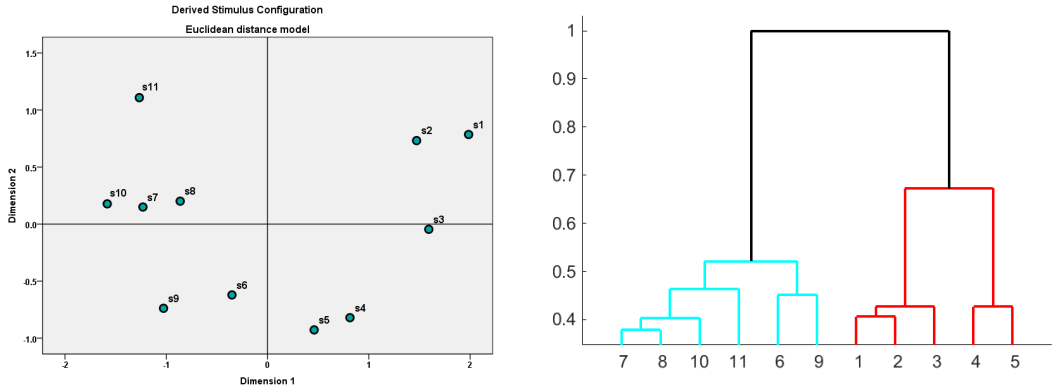

Figure 7.23: Two visualizations of the aggregate response for Type $\mathcal{A}$ stimuli by the Hindustani subjects' group. Left: multidimensional scaling of the (dis)similarity matrix in 2 dimensions. Right: dendrogram of the same (dis)similarity matrix.

Figure 7.23 also shows another way of visualizing the groupings in the perceptual space, the dendrogram (right). This is a hierarchical clustering where all data-points are initialized as individual clusters and they are iteratively merged to form a single cluster. The dendrogram visualization shows two major groups (in different colors) that preserves the stimuli order and also indicates two $\mathbb{P}$ regions corresponding to the two ragas Deshkar and Bhupali.

7.9 Concluding remarks

The broad goal of the work is to develop a computational model to measure phrase-level melodic similarity in Hindustani raga phrases. As described in the previous study [67], a set of acoustic

measurements were carried out on a number of musician-annotated raga phrase (GRS characteristic phrase in raga Deshkar) in a supervised manner, to study the dimensions of variation. Extracted from a data-driven method (vector quantization), we proposed a canonical or 'prototype' melodic shape model for the GRS phrase in raga Deshkar. To validate our finding, we created artificial stimuli by varying the learned dimensions in the model space. Resynthesized melodies (with uniform timbre across stimuli) were rated by trained Hindustani musicians to record the 'goodness rating' and find prototypical and non-prtotypical shapes corresponding to the phrase under study. Next, we carried out an AX differential discrimination experiment to estimate the sensitivity to small changes in stimuli across raters (trained Hindustani and Carnatic musicians, Western musicians with no exposure to Hindustani music). The findings suggest that trained Hindustani musicians perceive melodic phrases categorically, with less sensitivity to small changes around the prototype region, indicating that prototypes work as a perceptual attractor and that musicians tend to perceive melodic phrases holistically. The MDS obtained from the disparity matrix from the musicians' response show that the perceptual space is warped, dense for a prototypical context and sparse for non-prototypical context. The same set of experiments were carried out with the same GRS melodic shape with a different pre-context where it was not-so-characteristic for any particular raga. Also, individual differences among musicians' responses gave insights towards interpreting the possible perceptual mechanisms at play. Some of the take-aways are listed as below.

1. Same material is rated differently by different categories of musicians

   The same model space (both Type $\mathcal{A}/\mathcal{B}$) with same acoustic differences in the same dimension were rated differently by subjects of different training background whose ratings were significantly different (specific results discussed later). This shows that different subject (musician) categories had developed different mental (dis)similarity models in due course of training.

   According to Signal Detection Theory (SDT), listeners who share the same perceptual precondition (identical auditory threshold) can produce different results in a perception test because they use response criteria of different sizes [166, 167]. Hence we carried out discrimination sensitivity tests via d-prime and lambda-center to support the conventional metric "proportion of 'different' responses". A combined observation of both indicates the different locations and widths of the prototype and non-prototype regions for different

musician categories.

2. Same subject rates differently for different test scenarios

   This involves two scenarios: (i) when the model space is different (Type $\mathcal{A}$ vs. Type $\mathcal{B}$), (ii) when the model space is same, but the phrase pre-context is different (char DPGRS vs. non-char DPMGRS). (i) For this case, the prototype stimulus is exactly the same, but the other stimuli vary in either duration of R (Type $\mathcal{A}$) or intonation of G (Type $\mathcal{B}$). Though the subjects are not given any prior information about the stimuli (except a short practice session to habituate with the setup), all subjects of all categories rate them differently. This indicates that subjects are able to create a mental mapping of the dimension of variation across stimuli within a trial block. (ii) For this case (limited to only Hindustani subjects), we do not have any hypothesized P for the non-char DPMGRS phrase. If the model space variation was the only predictor, we would have expected the same $\mathbb{P}$ stimulus would also show low discriminability which is not the case. The possible reason is explained as below. However, Non-musicians' response show a similar behavior alike char DPGRS phrase.

3. Trained subjects show link to learned categories for identification task

   The PME discrimination task for the non-char DPMGRS phrase (limited to Hindustani musicians only) showed that subjects incline to link between the stimuli and some learned category while discrimination task, as per literature, involves short-term memory of acoustic (dis)similarity. It is interesting to observe individual differences of raga-mapping (also confirmed by interviewing), which resulted in different locations of low discriminability (assuming that the corresponding stimulus acted as $\mathbb{P}$). Thus, for this specific case-study, the aggregate distribution of responses are not conclusive.

4. Subjects show different behavior along the continuum of stimuli

   The same acoustic difference near the prototype and non-prototype region is perceived differently by all musician categories. In general, discriminability is significantly better near the $\mathbb{NP}$ region (chance level) compared to the $\mathbb{P}$ region (10% level).

   Readers should note that the assignment of hypothesized $\mathbb{P}$ and $\mathbb{NP}$ is with respect to Hindustani musicians. For Type $\mathcal{A}$ stimuli, responses between Hindustani and Carnatic musicians are significantly different for stimuli differences 2 – 6 from the P centre; be-

tween Hindustani and Western musicians are significantly different for stimuli differences 3 – 6 from the $\mathbb{NP}$ centre; between Hindustani and non-musicians are significantly different for stimuli differences 2 – 5 from the P centre and 4 – 10 from $\mathbb{NP}$ centre.

5. Certain contradictory cues can affect the rating which is beyond our control

   Though we aim to capture the perceptual space of subjects, certain psychoacoustic phenomenon can influence 'listening' and as a result ratings. This was specially observed for Type $\mathcal{B}$ stimuli where an interaction of the intonation of elongated G note in presence of the tanpura (providing the tonic) played a major role. When the intonation of G was shifted higher from the P stimulus, a certain range of stimuli (indices 7–9) had the G intonation around the equal tempered tuning that created a 'roughness'. Non-musicians' discrimination, that was otherwise having higher variance, became consistently high around this range. The roughness might have disturbed their biological hearing and hence they might have prompted to mark these pairs as 'different'.

6. Existence of PME does not necessarily ensure categorization

   The PME hypothesis states that the perceptual space is warped, with distinct behaviors in regions around prototypical vs. non-prototypical shapes. The prototypes can act like either attractors (decreased discriminability) or like anchors (increased discriminability). For the case of Type $\mathcal{A}$ stimuli, beyond the hypothesized NP showed decreased discriminability (see Figure 7.21), which may be indicative of prototype of another category. Hence we carried out CP tests for Type $\mathcal{A}$ stimuli, which showed anticipated identification and discrimination functions, confirming presence of categorization.

   On the other hand, the Type $\mathcal{B}$ stimuli did not show decreased discriminability anywhere else other than the hypothesized $\mathbb{P}$ region, except stimuli range 3 – 5 (the Bhupali intonation region, see Figure 7.20) where the discrimination was better than chance level for all subject categories. One possible reason for this is the high consonance of the G interval (resembling just intonation tuning) which made all subjects deaf to the small acoustic differences. However the case for Hindustani musicians demands some more discussion. Acoustic measurements [67] showed that this intonation range qualifies raga Bhupali GRS phrase, though there is a significant difference in the R duration with respect to Deshkar phrases. The non-existence of a second prototype (as in case of Type $\mathcal{A}$) indicates that G intonation cannot be a sole predictor for categorization, as the R duration

in all Type $\mathcal{B}$ stimuli were short (Deshkar prototype). This also supports the previous observation of R duration being topmost discriminative feature in classification via 'feature selection'. To resolve the contribution (and priority) of the R duration and G intonation dimensions, one could investigate the same experimental paradigm with hybrid stimuli. This is posed as a future scope of our work.

7. Order of presentation effect depends on stimulus index in the model space

   The order of presentation effect is important to report in pairwise comparison studies. We captured this effect in two ways in our experiments: (i) from control subjects who were presented both AB and BA pairs within the same trial block, (ii) from other subjects who were presented AB and BA pairs in two independent blocks. Type $\mathcal{A}$ stimuli did not show any significant order of presentation effect, whereas for Type $\mathcal{B}$ stimuli it was not easy to interpret because of no clear presence of prototypes or category boundaries.

In sum, we aimed to impart music knowledge into a data-driven computational model towards modeling human-judgment of melodic similarity. This cognitively-based model can be useful in music pedagogy, as a compositional aid, or building retrieval tools for music exploration and recommendation.

Chapter 8

Retrieval applications

8.1 Motivation

Time-series pattern matching methods that incorporate time warping have recently been used with varying degrees of success on tasks of search and discovery of melodic phrases from audio for Indian classical vocal music [38, 82, 128, 151]. While these methods perform effectively due to the minimal assumptions they place on the nature of the sampled pitch temporal trajectories, their practical applicability to retrieval tasks on real-world databases is seriously limited by their prohibitively large computational complexity. While dimensionality reduction of the time-series to discrete symbol strings is a standard approach that can exploit computational gains from the data compression as well as the availability of efficient string matching algorithms, the compressed representation of the pitch time series itself is not well understood given the pervasiveness of pitch inflections in the melodic shape of the raga phrases. We propose methods

[0]This chapter is largely drawn from the following papers:

- K. K. Ganguli, A. Rastogi, V. Pandit, P. Kantan, and P. Rao. "Efficient melodic query based audio search for Hindustani vocal compositions," in Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR), October 2015, Malaga, Spain. [68]
- K. K. Ganguli, A. Lele, S. Pinjani, P. Rao, A. Srinivasamurthy, and S. Gulati. "Melodic shape stylization for robust and efficient motif detection in Hindustani vocal music," in Proc. of the National Conference for Communications, March 2017, Chennai, India. [57]
- A. Lele, S. Pinjani, K. K. Ganguli, and P. Rao. "Improved Melodic Sequence Matching for Query Based Searching in Indian Classical Music," in Proc. of the Frontiers of Research on Speech and Music (FRSM), November 2016, Baripada, India. [105]

that are informed by domain knowledge to design the representation and to optimize parameter settings for the subsequent string matching algorithm. The methods are evaluated in the context of an audio query based search for Hindustani vocal compositions in audio recordings via the mukhda (refrain of the song). We present results that demonstrate performance close to that achieved by time-series matching but at orders of magnitude reduction in complexity.

8.2 Background

The automatic detection of the repetitive phrase, or motif, from the audio signal would contribute to important metadata concerning the identity of the bandish. The *mukhda* (refrain of the composistion) is recognised by the lyrics, location in the cycle and its melodic shape. While these are in order of decreasing ease in terms of manual segmentation of the mukhda, the melodic shape characterized by a pitch contour segment is most amenable to pattern matching methods [82, 151]. The challenge here arises from the improvisatory nature of the genre where the raga grammar allows for considerable variation in the melodic shape of any prescribed phrase. Previous work has shown that the variability in the mukhda across the concert, similar to that of other raga-characteristic phrases in a performance, can be characterized as globally constrained non-linear time-warping where the constraint appears to depend on certain characteristics of the underlying melodic shape [150, 151, 161]. A dynamic time-warping (DTW) distance measure was used on the time-series segments to model melodic similarity under local and global constraints that were learned from a raga-specific corpus[151]. More recent work has also validated the DTW based similarity measure in the context of melodic motif discovery but the high computational costs associated with time-series search limited its applicability [38, 82, 128]. Given that DTW based local matching, with relatively minimal assumptions, on the pitch time-series derived from the audio is largely successful in modeling the relevant melodic variations, we focus on targeting similar performance with greatly reduced complexity. Computationally efficient methods to search and localize occurrences of the mukhda in a concert, given an isolated audio query phrase, have the following potential real-world applications: (i) automatic segmentation of all occurrences of the mukhda provided one manually identified instance, with a goal to reduce manual effort in the rich transcription of concert audio recordings, and (ii) retrieving a specific bandish from a database of concert recordings by querying by its mukhda provided either by an audio fragment or by user singing.

As mentioned earlier, DTW can be used in an exhaustive search across the concert of the sampled pitch time series to find the optimal cost alignment between the query and target pitch contours at every candidate location. We see therefore that any significant computational complexity reduction can only come from the reduction of dimensionality of the search space. An obvious choice is a representation of the melodic contour that uses compact musical abstractions such as a sequence of discrete pitch scale intervals (essentially, the note sequence corresponding to the melody if there was one). String-matching algorithms can then be applied that find the approximate longest common subsequence between the query and target segments of discrete symbols. Krannenburg [193] used this approach on audio recordings of folk songs to establish similarity in tunes across songs. Each detected pitch value was replaced by its MIDI symbol and the Smith-Waterman local sequence alignment algorithm was used on the resulting strings. Note however that there was no reduction in the size of the pitch time-series. If the pitch time-series is segmented into discrete notes, a far more compact string representation can be obtained by using each symbol to represent a tuple corresponding to a note value and duration. In this case, a number of melodic similarity methods based on the alignment of symbolic scores become available [9, 74, 119, 189, 193]. The effectiveness of this approach, of course, depends heavily on the correspondence between the salient features of the pitch contour and the symbol sequence. A specific challenge in the case of Hindustani vocal music is that it is characterized just as much by the precisely intoned raga notes as it is by the continuous pitch transitions and ornaments that contribute significantly to the raga identity, motivating a more careful consideration of the high-level abstraction [144, 152, 169]. We also investigate the symbolic representation of continuous pitch transients that, when combined with stable notes, leads to the more complete representation of a raga-phrase. Our approach is to stylize the continuous pitch curve with a low-order representation and obtain an inter-symbol distance measure for the ensuing codebook of shapes. Another important predictor in melodic similarity paradigm is 'duration'. E.g. the same sequence of notes with different relative durations should be tractable and be penalized by the algorithm, more so because Hindustani music contains many such examples where melodic phrases from two different ragas bear the same note sequence but with different relative durations (refer to [55] for further discussion).

The main contributions of this work are (i) a study of the suitability of two distinct high-level abstractions for sequence representation in the context of our melodic phrase retrieval task, and (ii) using domain knowledge for the setting of various representation and search parameters

of the systems. In the next section, we describe our test dataset of concerts with a review of musical and acoustic characteristics that are relevant to our task. This is followed by a presentation of our melodic phrase retrieval methods including approaches to the compact representation of the pitch time-series and discussion of the achievable reduction in computational complexity with respect to the baseline system. A description of the experiments follows. Finally the results are discussed with a view to providing insights on the suitability of particular approaches to specific characteristics of the test data.

8.3 Dataset and proposed systems

The test dataset comprises 75 commercial CD-quality concert audio recordings by 23 eminent Hindustani vocal artists. The accompaniment consists of tanpura (drone) and tabla, along with harmonium or sarangi. The concerts have been chosen from a large corpus[176] in a deliberate manner so as to achieve considerable diversity in artists, ragas and tempo. We restrict our analysis to the vilambit (slow tempo) and madhyalaya (medium tempo) sections of these concerts for the current task. Drut (fast tempo) sections are excluded because their mukhda phrases contain a considerable amount of context-dependent variation and hence melodic similarity is not as strongly preserved. Table 8.3 summarises our dataset where 59 concerts are of vilambit laya and the remaining 16 are madhyalaya. The average duration of a vilambit bandish is 17 minutes and contains an average of 20-25 mukhda instances that occur once each in a rhythmic cycle.

# Song	Dur (hrs)	# GT	Dur (hrs)	Ratio	# Unique	
					Raga	Artist
75	22:18	1754	2:49	13%	55	23

Table 8.1: Description of the test dataset. We name this as the 'mukhda dataset', this is disjoint from the datasets discussed in Chapter 3.

Manual annotation of the mukhda segments with start and end boundaries was carried out by a musician and validated by a second very experienced musician. Mukhdas are most easily identified by listening for the lyrical phrase that occurs about the first beat (sam) of the rhythmic cycle as evidenced by the accompanying tabla strokes. The mukhda is labeled together with its boundaries as detected from the onsets of the lyric syllables. These annotations serve as the

ground truth (GT) for the evaluation of the different systems under test which exploit only the similarity of melodic shape to that of the audio query. The query thus could be an instance extracted from the audio track, or it could be a sung or hummed likeness of the melodic phrase generated by the user.

We consider various approaches towards our end goal which involves searching the entire vocal pitch track extracted from the audio recording to identify pitch contour sub-segments that match the melodic shape of the query. We present a discussion of the different systems in terms of algorithm design and complexity.

8.3.1 Baseline system

Our baseline method is the "subsequence DTW", an adaptation of standard DTW to allow searching for the occurrence and alignment of a given query segment within a long sequence[127, 186]. Given a query Q of length N symbols and a much longer sequence S of length M (i.e. the song or concert sequence in our context) to be searched, a dynamic programming optimization minimizes the DTW distance to Q over all possible subsequences of S. The allowed step-size conditions are chosen to constrain the warping path to within an overall compression / expansion factor of 2. No further global constraint is applied. The candidate subsequences of the song are listed in order of increasing DTW distance to which a suitable threshold can be applied to select and localize the corresponding regions in the original audio. The time complexity of subsequence DTW is $O(MN)$ where $N(M)$ is the number of pitch samples corresponding to the query (song) duration (i.e. 50 pitch samples per second of the time series duration, given that the pitch is extracted at 20 ms intervals)[34, 127, 190]. We see that the time-series dimensions contribute directly to the complexity of the search. Our goal is to find computationally simple alternatives to DTW by moving to low dimensional string search paradigms. This requires principled approaches to converting the pitch time-series to a discrete symbol sequence, two of which are presented next.

8.3.2 Behavior based system

This refers to the behavioral sequence modeling as discussed in Chapter 4 (Section 4.4.2.1). A melodic phrase can be viewed as a sequence of musical gestures by the performer, with a behavioral symbol then potentially corresponding to a single (arbitrary movement) in pitch

space. A sequence of symbols would serve as a sketch of the melodic motif. The database is pre-processed and the symbol sequence representation of each complete concert recording is stored. When a query is presented, it is converted to its symbol sequence (which currently depends on the song to be searched) and an exact sub-sequence search is implemented on the song string. The choice of the fixed parameters: window duration, hop duration and number of subsegments within a window turn out to heavily influence the representation. The window duration should depend on the time scale of the salient features (movements in pitch space). The subsegments must be small enough to retain the melodic shape within the window. The hop of the sliding window compensates for alignment differences of the different occurrences of the template in the pitch time-series of the song. We present "parameter settings" for two configurations.

Version A: Fixed parameter setting (window = 126 samples, hop = 5 samples, # subsegments per window = 3)

Version B: Query dependent setting (window = (0.5 * N) samples, hop = 5 samples, # subsegments per window = 4)

8.3.3 Pseudo-note system

This refers to the svara segmentation and labeling as discussed in Chapter 4 (Section 4.3.2). An approximation to staff notation can be achieved by converting the continuous time-series to a sequence of piece-wise flat segments if the section pitches are chosen from the set of discrete scale intervals of the music. If the achieved representation indeed corresponds to some underlying skeleton of the melodic shape of the phrase, we could anticipate obtaining better matches across variations of the melodic phrase. The database is pre-processed and the svara sequence representation of each complete concert recording is stored. When a query is presented, it is converted to its symbol sequence and an approximate sub-sequence search is implemented on the concert string based on an efficient string matching algorithm with parameter settings that are informed by domain knowledge as described next. The similarity measurement of the query sequence with candidate subsequences of the song is based on the Smith-Waterman algorithm, widely used in bioinformatics but also applied recently to melodic note sequences[175, 193]. It performs the local alignment of two sequences to find optimal alignments using two devices. A symbol of one sequence can be aligned to a symbol of the other sequence or it can be aligned to a gap. Each of these operations has a cost that is designed as follows.

Substitution score: In its standard form, the Smith-Waterman algorithm uses a fixed positive cost for an exact match and a fixed negative score for symbol mismatch. In the context of musical pitch intervals, we would rather penalize small differences less than large differences. We present alternate substitution score functions that incorporate this.

Gap Function: This function deducts a penalty from the similarity score in the event of insertion or deletion of symbols during the alignment procedure. The default gap penalty is linear, meaning that the penalty is linearly proportional to the number of symbols that comprise the gap. Another possibility, that is more meaningful for the melody context, is the affine gap function where the gap opening cost is high compared to the cost incurred by adding each successive symbol to the gap[76]. This is achieved by a form given by $mx + c$ where x is the length of the gap and m, c are constants. Intuitively, increasing c will penalize gap openings to a greater extent, while increasing m will have a similar effect with regard to gap extension. We present different designs for the relative costs motivated by the musical context.

With variations in each of the above two controls of the Smith-Waterman algorithm, we obtain the following three distinct versions of the pseudo-note system.

Version A: This setting is similar to the default Smith-Waterman setting, with a distance-independent similarity function that assesses a score of +3 for symbol match and -1 for a substitution. Gap function is linear, with penalty equal to symbol length of gap.

Version B: Substitution score that takes pitch difference into account, i.e. Score of +3 for a match, 0 for symbols differing by upto 2 semitones, -1 for substitution, and an affine gap penalty with parameters m = 0.8, c = 1.

Version C: Query dependent settings where we use the settings of B as default with the following changes for particularly fast varying and slowly varying query melodic shapes as determined by a heuristic measure of ratio of squared number of symbols to query duration. We have the following parameter settings. (i) fast varying: Substitution score of +1 to symbols differing by upto 2 semitones. Gap penalty is affine with parameters m = 1, c = 0.5, and (ii) slowly varying: Similarity score of -0.5 to symbols differing by upto 3 semitones. Gap penalty is affine with parameters m = 0.5, c = 1.5.

Finally, the Smith-Waterman algorithm has a time complexity given by $O(MN^2)$ where N is the query length in symbols and M is the song length[175]. By constraining the allowed gap length to be no longer than that of the query itself (N), justified by the musical context, we achieve a complexity reduction to $O(MN)$.

8.4 Experimental results

We present experiments that allow us to compare the performance of the different systems on the task at hand, namely correctly detecting occurrences of the mukhda in the audio concert given an audio query corresponding to the melodic shape of the mukhda phrase. The queries are drawn from a set of 5 mukhda's extracted from the early part (first few cycles) of the bandish. The early mukhda repetitions tend to be of the canonical form and hence correspond well with an isolated query that a musician might generate to describe the bandish. For the investigation of a given method, we process the database to convert each concert audio to the pitch time series and then to the corresponding string representation. Next, the query is converted to the string representation and the search is executed. The detections with time-stamps are listed in order of decreasing similarity with the query as determined by the corresponding search distance measure. A detection is considered a true positive if the time series of the detection spans at least 50% of that of one of the ground-truth labeled mukhdas in the song. An ROC (precision vs recall) is obtained for each query by sweeping a threshold across the obtained distances. The ROC for a song is derived by the vertical averaging (i.e. recall fixed and precision averaged) of the ROCs of the 5 distinct queries[52]. The performance for each song is summarized by the following two measures: precision at 50% recall and the equal error rate (EER) (point on the ROC at which false acceptance rate matches false rejection rate). We further present performance of the best performing pseudo-note system on song retrieval in terms of the mean reciprocal rank (MRR)[84] on the dataset of 50 concerts as follows. We use the set of the first occurring labeled mukhda of each song to form a test set of 50 queries. Next for each test query, every song is searched to obtain a rank-ordered list of songs whose first 5 detections yield the lowest averaged distance measure to the query.

Table 8.5.1 compares the performances of the various systems on the task of mukhda detection in terms of the average EER and average precision at a selected recall across the 50 songs where each song is queried using each of the first five mukhdas. We also report the computational complexity reduction factor over that of the baseline method (given by the square of the dimension reduction factor). To obtain more insight into song dependence, if any, we show the distribution of the precision values for the 50 songs set in the bar graphs of Figure 8.1, one system for each category, represented by the best performing one.

From Table 8.5.1, we observe that the baseline system represented by subsequence DTW

Method (version)		Mean EER	Prc at 50% Rec		Reduc.
			Mean	Std.	
Subseq DTW	—	0.33	0.73	0.18	1
Behavior based system	(A)	0.47	0.56	0.26	100
	(B)	0.41	0.61	0.25	
Pseudo-note system	(A)	0.47	0.61	0.19	2500
	(B)	0.42	0.64	0.19	
	(C)	0.41	0.65	0.18	

Table 8.2: Comparison of the two performance measures and computational complexity reduction factor across the baseline and proposed methods.

on the pitch time-series performs the best while the pseudo-note methods achieve the best computation time via a reduction proportional to the square of the reported dimension reduction factor (i.e. 50). We will first comment on the relative strengths of these two systems, and later discuss the behavior based system. We observe an improvement in performance of the pseudo-note system with the introduction of domain knowledge and query dependent parameter settings for the subsequence search algorithm. From Figure 8.1, we see that the subsequence DTW has a right-skewed distribution indicating a high retrieval accuracy for a large number of songs. However we note the presence of low performing songs too which actually do better with the pseudo-note system. Closer examination of these songs revealed that these belonged to ragas characterized by heavily ornamented phrases. In the course of improvisation, the mukhda was prefaced by rapidly oscillating pitch due to the preceding context. This led to increased DTW distance between the query and mukhda instances. The oscillating prelude was absent in the pseudo-note representation altogether leading to a better match.

The behavior based system was targeted towards capturing salient features of the melodic shape of the phrase in a symbolic representation. The salient features should ideally include steady regions as well as specific movements in pitch space that contribute to the overall melodic shape. As such, it was expected to perform better than the pseudo-note method which retains relatively sparse information as seen from a comparison of the two representations for an example phrase. However, the selection of the duration parameters required for the time-series

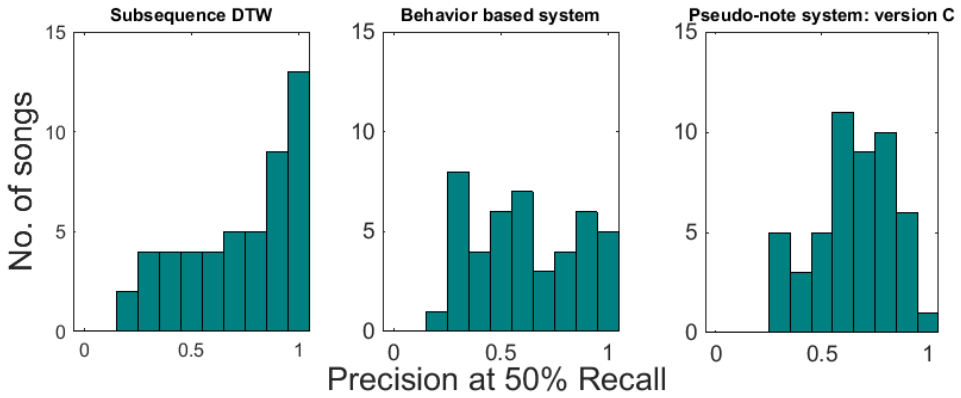

Figure 8.1: Histogram of the measure 'Precision at 50% Recall' across the baseline and proposed methods.

conversion turned out to be crucial to the accuracy of the system. Shortening the window hop interval contributed to reduced sensitivity to time alignment differences but at the cost of reduced compression and therefore much higher time complexity. Further, the data dependence of symbol assignment requires the query to be re-encoded for every song to be searched, and further if query dependent window length is chosen, the song must be re-encoded according to the query. Future work should target obtaining a fixed dictionary of symbols to pitch movement mappings by learning on a large representative database of concerts. Figure 8.1 also seems to be a direct outcome of the resolution / accuracy of 'tokenization' (of the pitch contour and durational information) which in turn seems to reflect in the way the 'semantic gap' (perceptual closeness) is closed and hence manifest in the different precision-recall characteristics observed.

Top 'M' hits	Correct songs	Accuracy
1	41 / 50	0.82
2	45 / 50	0.90
3	48 / 50	0.96

Table 8.3: Results of the song retrieval experiment.

Finally, we note the song retrieval performance of the pseudo-note version C in Table 8.3. The mean reciprocal rank (MRR) is 0.89. The top-3 ranks return 48 of the 50 songs correctly. The badly ranked songs were found to be narrowly superseded by other songs from the same raga that happened to have phrases similar to the mukhda of the true song. This suggests the potential of the method in the retrieval of "similar" songs where the commonality of raga is

known to be an important factor.

In summary, the melodic phrase is a central component for audio based search for Hindustani music. Given the improvisational nature of the genre as well as the lack of standard symbolic "notation", time-series based matching of pitch contours provides a reasonable performance at the cost of complexity. The conversion to a relatively sparse representation by retaining only flat regions of the pitch contour and introducing domain driven cost functions in the string search is shown to lead to a slight reduction in retrieval accuracy while reducing complexity significantly. The inclusion of further cues such as the lyrics and rhythmic cycle markers to mukhda detection is expected to improve precision and is the subject of future research.

8.5 Extension of the best performing proposed system

We take the version C (the then best performing pseudo-note system) that used query dependent preset parameters. For completeness we recall two relevant parameters: (i) substitution score and (ii) gap penalty. In its standard form, the Smith-Waterman algorithm uses a fixed positive score for an exact match and a fixed negative cost for symbol mismatch. In the context of musical pitch intervals, we penalize small differences less than large differences. The two parameter presets are (i) fast varying: substitution score of +1 to symbols differing by upto 2 semitones ('Close'), gap penalty is affine with parameters $m = 1$, $c = 0.5$; and (ii) slowly varying: substitution score of -0.5 to symbols differing by upto 3 semitones ('Close'), gap penalty is affine with parameters $m = 0.5$, $c = 1.5$.

DURATION / TRANSIENT	WITHOUT	WITH
WITHOUT	[R,P,R,S], [N,D]	[(R,390),(P,720),(R,550), (S,270)], [(N,4550),(D,280)]
WITH	[5,R,4,P,3,R,2,S], [3,N,2,D,1]	[(5,1070),(R,390),(4,910),(P,720), (3,1040),(R,550),(2,280),(S,270)], [(3,690),(N,4550),(2,440), (D,280),(1,270)]

Figure 8.2: Proposed schemes of melodic phrase representation (symbols and duration (ms) information) applied to the pitch contour.

Our proposed system partially uses the pseudo-note system with the introduction of symbols for the modelled transients. The main contribution, here, lies in the modified scoring scheme which is discussed next. The notations of interest are as follows. $Z = \max(\frac{T_{Query}}{T_{Candidate}}, \frac{T_{Candidate}}{T_{Query}})$,

where T_{Query} and $T_{Candidate}$ are the durations of the note in the query and candidate respectively (refer to the parameter 'Fraction'). $X1$: length of gap (no. of notes), $X2$: length of gap (no. of notes/transients), $Y1$: duration of gap (sec), $Y2$: weighted duration of gap (sec). The optimal values for the slopes and intercepts, as obtained from an empirical observation, are: $m1 = 3$, $m2 = 0.02$, $c1 = 1.5$, $c2 = 1.2$.

8.5.1 Scoring scheme

The modification of the scoring strategy involves two new parameters. The first, incorporating a parameter Z into the substitution score (schemes 2 and 4) that attenuates the positive score (cases: Same and Close) and amplifies the negative score (case: Far). This is because, by definition $Z \geq 1$ (we assign a lower bound of 1.25 chosen empirically) and a higher value of Z suggests a high mismatch between the query and candidate durations which should be compensated in the substitution score. Secondly, we incorporate a factor called 'weighted duration' $Y2$ for the transient segments in order to balance their importance with respect to the pseudo-notes. The weight is a linear combination of pre- and post-context of transient durations (with an empirically chosen weighting factor of 0.8) for each pseudo-note duration. The proportion of duration contributed by pseudo-notes and transients is very high which is compensated by introduction of the $Y2$ parameter. Figure 8.3 shows the parameters involved. Schemes 3 and 4 has two values for substitution scores (comma separated) which stand for pseudo-notes and transients respectively.

TRANSIENT \ DURATION	WITHOUT	WITH
WITHOUT	1 Same: +3 Close: +1 Far: -1 -m1*X1 + c1	Same: +3/Z Close: +1 Far: -1*Z -m2*X1*Y1 + c2 2
WITH	Same: +3, +1 Close: +1, +0.33 Far: -1, -0.5 -m1*X2 + c1 3	Same: +3/Z, +1/Z Close: +1, +0.33/Z Far: -1*Z, -0.5*Z -m2*X2*Y2 + c2 4

Figure 8.3: Proposed schemes of melodic phrase retrieval systems. The scheme indices are marked in red. The parameters and corresponding values for substitution score and gap penalty are presented.

The experiments allow us to compare the performance of the different schemes on the task at hand, i.e., detecting occurrences of the mukhda in the audio concert given an audio query corresponding to one instance of the mukhda phrase. In the previous experiment, we had the

assumption that early mukhda repetitions tend to be of the canonical form that a musician might generate to describe the bandish; here instead we consider all annotated mukhdas as queries. We process the database to convert each concert audio to the pitch time series and then to the corresponding stylized representation. Next, the query is converted to the corresponding representation and the search is executed. The detections with time-stamps are listed in the order of increasing distance with the query as computed by the corresponding search distance measure. We disallow the list to grow further twice the number of ground truth mukhdas since this would correspond well to the maximum number of mukhdas expected in the concert given its duration. A detection is considered a true positive if the time series of the detection overlaps at least 50% of that of one of the ground-truth labeled mukhdas in the song. A receiver operating characteristic curve (precision vs recall) is obtained for each query by sweeping a threshold across the obtained distances. The performance for each song is summarized by the value corresponding to the percentage of queries (i.e. mukhdas) that result in at least $n\%$ recall with respect to all the remaining mukhdas in the song. We term this "goodness %" of the song. We report performance for different choices of n.

[Precision, Recall] pairs for optimal [Precision, Recall]			
Scheme 1	Scheme 2	Scheme 3	Scheme 4
[1.00,0.57]	[1.00,0.91]	[1.00,0.60]	[1.00,0.60]
[1.00,0.57]	[1.00,0.91]	[0.58,0.96]	[0.61,0.96]

Table 8.4: Evaluation metrics in terms of optimal [Precision, Recall] pairs.

Threshold (n)	Average "goodness %" (# songs)			
	Scheme 1	Scheme 2	Scheme 3	Scheme 4
10%	0.56 (63)	0.64 (71)	0.68 (71)	0.77 (71)
30%	0.35 (21)	0.38 (40)	0.39 (44)	0.42 (58)
50%	0.38 (2)	0.35 (9)	0.38 (12)	0.24 (25)

Table 8.5: Average "goodness %" across song (# songs with at least one good query) for different thresholds for the 4 schemes.

From Table 8.5.1 we see that there is improvement of recall (with fixed precision) after including the duration information. Figure 8.4 shows an improvement of recall in each column

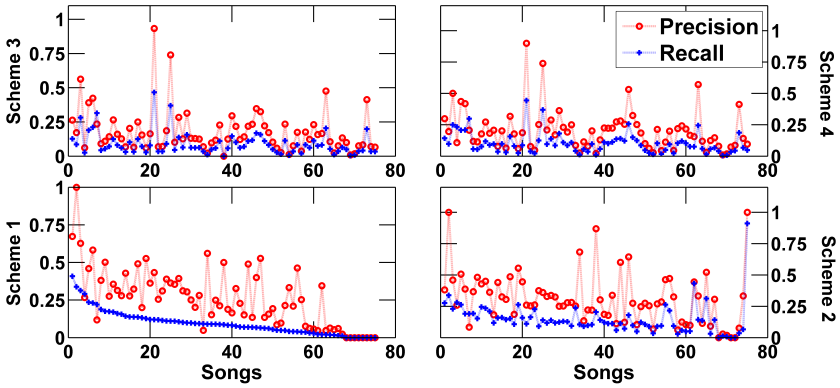

Figure 8.4: Comparison of evaluation metrics Precision and Recall for all 4 schemes. The song indices for Scheme 1 are obtained from descending order of recall, the same indices carry over to the other schemes.

from bottom to the top row, i.e. Scheme 1 to 3 and Scheme 2 to 4. Given that we attain consistent improvement in recall (while precision is marginally improving too), the proposed measure fraction of "good query" shows an improvement in Schemes 1 through 4. Table 8.5.1 shows the average fraction of good query for different threshold values (at least 10%, 30% and 50% recall). The number of songs where at least one 'good query' is found, improves in Schemes 1 through 4, irrespective of the threshold. On a rigorous error analysis, we find that the rootcause of the songs not having a single good query either belongs to slow tempo (vilambit) compositions where mukhdas are reasonably long or the mukhdas are performed with heavy embellishments that resulted in a long string sequence. Long queries might have resulted in amplification of negative score, a length normalization scheme could be useful to compensate this which is posed as a future work.

8.6 Discussion

We highlight the main contributions and summarize our work in terms of the design considerations at each stage. Refer to Section 4.4.2.2 for the computational models, this section discusses the insights obtained from the associated experiments for the retrieval task at hand.

(i) We use a representative corpus (diverse in terms of artists and ragas) to train a codebook of melodic transient shapes. The test set for the evaluation task is completely disjoint from the training corpus.

(ii) The choice of codebook size is evaluated from a very small (2) to a large (100) value. It is observed that beyond a certain size (8), there is no improvement implying that redundancy

is introduced into the codebook vectors. This indicates about a possible universality in the basic melodic movements between notes, though the same transient shape between the same pair of notes in two different ragas sounds perceivably different just due to the time-scaling (slowness) of the transient. Refer to the plot of the 8 transient shapes in Figure 4.11.

(iii) The k of k-means is chosen from the hierarchical clustering view. We carried out statistical methods (finding the elbow of the mean squared error curve iterated over k, gap statistics with varying k, observing the density of obtained clusters after truncating the dendrogram at different levels etc.) to conclude that $k = 7$ to 9 is reasonable. Hence we empirically chose $k = 8$.

(iv) There may be a criticism of the fact that the redundancy for a large codebook size might be resulting from a low (3^{rd}) degree polynomial. But observations confirm that a 3^{rd} degree polynomial is good enough to capture the trend. The residue is suggestive of a possibly superposed vibrato-type oscillation ('gamak'). A higher degree polynomial would have the danger of overfitting the transient segments.

Our previous works [68, 105] had shown to have improved retrieval accuracy by incorporating query-dependent preset parameter settings. In line with the same philosophy, we plan to propose task-dependent preset parameter settings for the same retrieval framework in the symbolic measure paradigm. The stylization, per se, smoothes the contour by discarding perceptually irrelevant fast pitch oscillations and also disregards measurement errors in F0 extraction (e.g. any dip in the melodic contour resulted from an unvoiced consonant). In the current experiments we needed only the string sequence (and not the continuous contour), but the stylized contour could find its use in other MIR application such as synthesis or perception experiments [61].

It would be ideal to be able to tackle the two broad tasks in melodic similarity paradigm: (i) categorization to a phrase family [202] and (ii) characterization of a phrase [150], by the same algorithm. Wherein the first case needs a relaxed distance measure to maximize detections, the second case demands the distance measure to be sensitive enough to disambiguate even minute differences. Though in both of the above works, the objective is the same i.e. to discriminate based on the family, this is a classic example of how a generic algorithm can robustly handle different hyperparameter presets for varying tasks at hand.

Chapter 9

Computational musicology applications

9.1 Motivation

Indian art music is quintessentially an improvisatory music form in which the line between 'fixed' and 'free' is extremely subtle. In a raga performance, the melody is loosely constrained by the chosen composition but otherwise improvised in accordance with the raga grammar. One of the melodic aspects that is governed by this grammar is the manner in which a melody evolves in time in the course of a performance. In this work, we aim to discover such implicit patterns or regularities present in the temporal evolution of vocal melodies of Hindustani music. We start by applying existing tools and techniques used in music information retrieval to a collection of concerts recordings of alap performances by renowned khayal vocal artists. We use svara-based and svara duration-based melodic features to study and quantify the manifestation of concepts such as vadi, samvadi, nyas and graha svara in the vocal performances. We show that the discovered patterns corroborate the musicological findings that describe the "unfolding" of a raga in vocal performances of Hindustani music. The patterns discovered from the vocal

[0]This chapter is largely drawn from the following papers:

- K. K. Ganguli, S. Gulati, X. Serra, and P. Rao. "Data-driven exploration of melodic structures in Hindustani music," Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR), August 2016, New York, USA. [56]

- S. Gulati, K. K. Ganguli, S. Gupta, A. Srinivasamurthy, and X. Serra. "Ragawise: A Lightweight Real-time Raga Recognition System for Indian Art Music," Late breaking demo at the 16th International Society for Music Information Retrieval Conference (ISMIR), October 2015, Malaga, Spain. [78]

melodies might help music students to learn improvisation and can complement the oral music pedagogy followed in this music tradition.

9.2 Background

The rules of the raga grammar are manifested at different time scales, at different levels of abstraction and demand a different degree of conformity. A number of textbooks and musicological studies exist that describe different improvisatory aspects of melodies in IAM [14, 15, 24, 33, 37, 55, 93, 191]. These works also attempt to uncover some of the implicit aspects of raga grammar.

A majority of the studies mentioned above are musicological in nature. These typically involve either a thorough qualitative analysis of a handful of chosen musical excerpts or a compilation of expert domain knowledge. Though these studies often present interesting musical insights, there are several potential caveats in such works. Some of these caveats are summarized below:

- Small repertoire used in the studies challenge the generalizability of the proposed musical models
- Bias introduced due to the subjectivity in the analysis of musical excerpts
- Absence of concrete quantitative evidences supporting the arguments
- The kind of analysis that can be done (manually) is limited by human capabilities, limited memory (both short- and long-term)
- Difficulty in reproducibility of the results

Several qualitative musicological works bring out new musical insights but are prone to criticism of not having supported their findings using a sizable corpus. Contrary to that, quantitative computational studies manage to scale to sizable data sets, but fall short of discovering novel musical insights. In the majority of cases, computational studies attempt to automate a task that is well known and is fairly easy for a musician to perform. There have been some studies that try to combine these two types of methodologies of working and corroborate several concepts in musical theories using computational approaches. In Chinese opera music, [156] has performed a comparison of the singing styles of two Jingju schools where the author exploit the potential of MIR techniques for supporting and enhancing musicological descriptions.

Autrim[1] (Automated Transcription for Indian Music) has used MIR tools for visualization of Hindustani vocal concerts that created a great impact on music appreciation and pedagogy in IAM. We find that such literature pertaining to melodic structures in Indian art music is scarce.

In this work, we perform a data-driven exploration of several melodic aspects of Hindustani music. The main objective is to use existing tools, techniques and methodologies in the domain of music information retrieval to support and enhance qualitative and descriptive musicological analysis of Hindustani music. For this we select five melodic aspects which are well described in musicological texts and are implicitly understood by musicians. Using computational approaches on a music collection that comprises representative recordings of Hindustani music we aim to study these implicit structures and quantify different melodic aspects related with them. In addition to corroborating existing musicological works, our findings are useful in several pedagogical applications. Furthermore, the proposed methodology can be used for analyzing artist or gharana-specific[2] melodic characteristics.

9.2.1 Breath-phrase segmentation

There are different types of unvoiced segments in the predominant melody. While some of these segments are musically a part of a melodic phrase (short-pauses), some others delineate the boundary between consecutive melodic phrases. A distribution of the duration of all the unvoiced segments for the entire music collection revealed that their type can be identified based on the duration of the pause. For identifying intended breath pauses that separate melodic phrases we empirically set the duration threshold to be 500 ms. The duration of the intra-phrase pauses is considerably smaller than this threshold. Melodic contours of all the intra-phrase breath pauses (i.e. with duration smaller than 250 ms) are interpolated using a cubic spline curve. We shall refer to a breath-phrase as BP hereafter.

9.2.2 Svara duration distribution

We consider the distribution of the svara duration for each BP. Figure 9.1 shows a stacked bar graph of the sorted durations of the svaras for each BP for a case-study concert in raga Todi by Ajoy Chakrabarty. We observe that the cumulative duration of the transcribed svaras range from approximately 1 to 8 seconds. An important point to note here is that there is a difference

[1] https://autrimncpa.wordpress.com/

[2] Refers to a lineage or school of thought.

between the absolute duration of a BP and the height of the stacked bar (in Figure 9.1). This is caused by the transient pitch segments that we ignored in our representation. Readers must note that the stable svara transcription, therefore, has an implication for the further analyses.

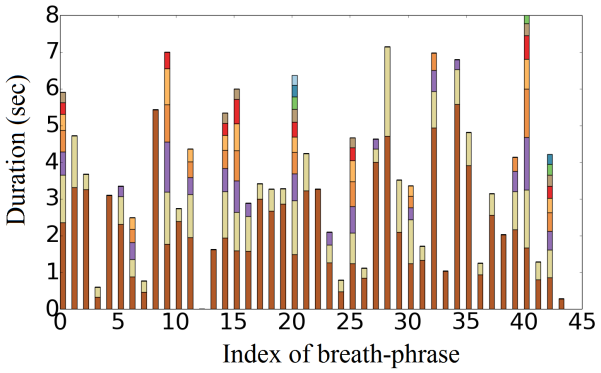

Figure 9.1: Bar graph of svara duration stacked in sorted manner for each breath-phrase for the case-study concert in raga Todi by Ajoy Chakrabarty. We observe that breath-phrases often comprise one long nyas svara and several other svaras of less duration.

To capture the changes in the svara pattern at a broader time-scale, we time-average the pitch histogram across ten BPs with a hop of one BP. This is followed by tracking of the most salient (in terms of accumulated duration) bin across the smoothened histogram. Finally, the obtained contour is further median filtered with one preceding and succeeding BP. We refer to this contour as the *evolution contour* (hereafter EC). Figure 9.2 shows the time-averaged histogram superimposed with the EC for the same concert.

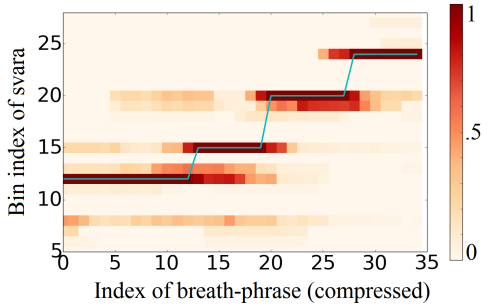

Figure 9.2: Time-averaged pitch histogram superimposed with the evolution contour for the case-study concert in raga Todi by Ajoy Chakrabarty.

9.2.3 Melodic evolution contour feature extraction

We would like to compare the ECs of different concerts to explore whether there is any common temporal trend. To normalize the time-scale and pitch range of the EC, we normalize each EC

within a unit range in both temporal and pitch dimensions. Thus a *modified evolution contour* (hereafter MEC) is obtained as:

$$MEC = \frac{EC - \min(EC)}{\max(EC) - \min(EC)} \tag{9.1}$$

with 100 equispaced temporal samples of values between [0,1].

We extract a collection of heuristic features (slope-based, duration-based, jump-based and level-based) from the MEC. A few important features are: slope between the MEC value of 0^{th} frame and the frame index where MEC reaches the value 1 for the first time (referred to as Slp), proportion of duration spent on each svara (referred to as Pro), centroid (considering salience of the bins as the weights in the centroid computation) of each svara (referred to as Cen), starting and ending svaras, (second) longest svara and proportion of its duration, magnitude of lowest/highest jumps between consecutive levels etc.

9.3 Dataset and analyses

The selected music material in the given collection is diverse in terms of the number of artists (40), recordings (mostly live concerts of both male and female renowned musicians from the last 6 decades) and the number of unique compositions (67). In these terms, it can therefore be regarded as a representative subset of real-world collections. Our collection includes a total of 75 concerts from 10 widely used ragas (8 pieces per raga on an average) that are diverse both in terms of the number of svaras and their pitch-classes (svarasthanas). All the concerts belong to either madhya or drut laya (and non-metered alap). The pitch range of the recordings spans approximately two octaves (middle octave and half of the lower/upper octave). All of the concerts comprise elaborations based on a bandish.

The scope of the study is limited to only the alap and vistar (barhat) [24, 33, 37, 93, 191] sections of the concert. Almost all of the concerts continue to subsequent fast improvisatory section (tan) after rendering the vistar. The melodic cue where the antara ends and the rhythmic cue where there is a slight increase in tempo just after, is quite universal and musically unambiguous. We employ a performing Hindustani musician (trained over 20 years) to annotate the time-stamps where the vistar ends. As the said annotation can be considered an obvious one (there is a less chance of getting subjective biases), we limit the manual annotation to one subject only. After cutting the concerts to the musician-annotated time-stamp, the average duration per concert is 16 minutes making a total of 20 hours of data.

We choose certain music concepts which are widely discussed among musicians and musicologists, for which there has not yet been an objective way of interpreting them from the audio. We cite the concepts (or knowledge-hypotheses, referred to as 'K') and discuss how a data-driven approach can help us validate them.

9.3.1 K1: evolution of melody in time

The barhat of a raga performance refers to the gradual "unfolding" of a raga by building on the svaras with a progression in a broad time-scale. But it is not very clearly illustrated in the musicology literature what the precise duration of each svara in course of this progression. Figure 9.3 shows the MECs of 37 (50% randomly chosen from our collection) concerts. We observe that the MECs, in general, start from a lower svara and gradually reach the highest svara in a step-like manner. The slope Slp of MEC, is quite consistent (mean = 1.3, standard deviation = 0.34) over the whole corpus. This gives an important insight that irrespective of the raga and concert-duration, artists take the same time to explore the melody and hit the highest svara. This also reinforces the nature of the time-scaling of a performance: for either a short 20 minute- or a long 50 minute-concert, the melodic organization bases more on relative and not absolute time. We also observe a sharp fall of the MEC at the end of the many concerts, this reflects how artists come down to a lower svara to mark an end to the vistar (this coincides with the musician's annotation). This phenomenon has a high resemblance with the time evolution of melody in course of the vistar, as shown in Figure 11 in [207]. Refer to Figure 5.6 for the construction of a similar idea, the difference being the time-scale of rhythmic cycle as opposed to BP here.

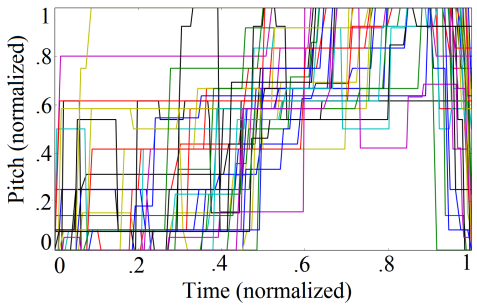

Figure 9.3: Modified evolution contours for 37 concerts in our music collection. The certain concerts that do not show the normal trend are either short in duration (less than 12 minutes) or madhyalaya concerts.

9.3.2 K2: transitional characteristics of nyas svaras

Raga guidelines mention about allowed svara sequences within a melodic phrase but it would be interesting to see if artists maintain any specific order in choosing the nyas svara across BP's or take liberty to emphasize any other svara. This is to be captured from the granularity of BPs and not in the time-averaged MEC. We generate a svara-transition matrix and populate it with a uniform weight for all transitions of the salient bins across BPs. Figure 9.4 shows the salient svara-transition matrix where the diagonal elements refer to self transitions. As indicative from wide steps of the MECs, there are quite a few self transitions but to our interest the salient transitions across BPs also follow a pattern similar to the allowed svara-transitions within a melodic phrase. This is not a trivial event. We compute a feature to measure the steadiness quotient Stq of the transition matrix, defined as the ratio of the trace of the svara-transition matrix to the sum of all bins. We observe a very low standard deviation (0.23) across our music collection which conforms to the fact that artists 'focus' on a nyas svara for consecutive BPs to establish that svara.

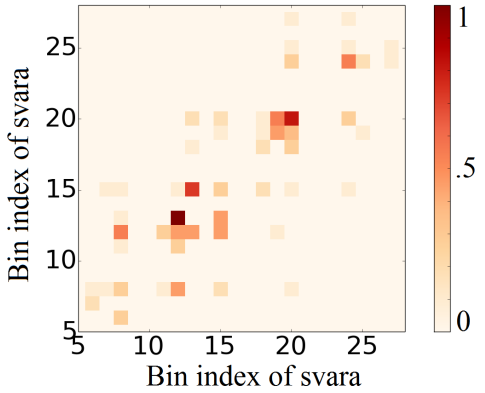

Figure 9.4: Svara-transition matrix of salient svaras of each breath-phrase for the case-study concert in raga Todi by Ajoy Chakrabarty. Intensity of each bin is proportional to the number of transitions taken from the svara of bin index on x-axis to the svara of bin index on y-axis.

9.3.3 K3: relationship between functional roles of svaras and their duration in melody

We discussed about functional roles of vadi/samvadi svaras, but it is not explicitly known how their 'prominence' is defined. Earlier work [36] use histogram and show that they are one of the most used tonal pitches. But it is not evident from a pitch histogram whether the peak heights

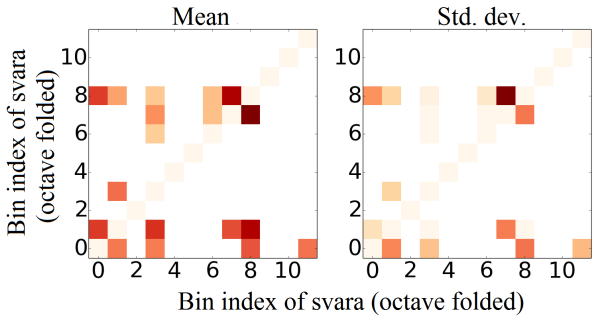

Figure 9.5: Functional roles of each svara (octave folded) for the case-study concert in raga Todi by Ajoy Chakrabarty. Mean (left) and standard deviation (right) of svara durations where each svara along x-axis is the salient svara of a breath-phrase.

are contributed by a large number of 'short' svara segments or a fewer 'long' svara segments. Figure 9.5 shows the mean (left) and standard deviation (right) of all svaras (octave folded) for each svara along x-axis being the salient svara in a BP. We observe that the role of each svara is defined in terms of their duration in context of a nyas svara. This also reconfirms the well-defined structure of the melodic improvisation that any svara cannot be stretched arbitrarily long, the nyas svara of the BP and the phrase itself decides how much variance all other svaras can incorporate. This also brings out a question whether there is any special functional role of vadi/samvadi svara in terms of their duration in a specific BP. One observation is that the vadi/samvadi svara, while a salient svara in the respective BPs, constrain the duration of all other svaras, making its relative prominence higher.

9.3.4 K4: duration and position of svaras in melody

In music theory, vadi and samvadi svaras are among the concepts which are often discussed. But we do not have an objective measure to observe how these svaras are different from other svaras in a raga. It is also interesting to know whether the vadi or samvadi svara always takes a focal role irrespective of their location in a BP and the overall position in a performance.

Position in melody: An important feature of the MEC is the Cen. We observe that the Cen is a raga dependent feature. E.g., an uttaranga vadi raga would have its vadi centroid in the latter half of a concert. This is supportive of the fact that the vadi is explored in due course of melodic improvisation adhering to the trend observed as in Section 9.3.1. The musicological hypothesis that these are the focal svaras of a raga does not necessarily imply that these svaras are explored from the very beginning of the concert. Rather the performance starts from a lower svara (graha svara) and reaches the vadi in course of the gradual development of the melody.

Duration in melody: We compute the average duration of all salient svaras per BP in two groups of svaras: (i) group 1: vadi/samvadi, and (ii) group 2: rest. It is observed on the whole corpus that Pro of group 1 is higher than the group 2 for all ragas. This reinforces the fact the term 'focus' or 'shine' (that qualifies vadi) is manifested in the temporal dimension. This also brings out a question whether we can predict the vadi/samvadi of a raga from the melody by data-driven features. From the overall pitch histogram it is difficult to infer, but from our designed features, we observe an accuracy of 83% while predicting the vadi/samvadi of a given raga.

9.3.5 K5: presence of possible pulsation in melody

There has been a discussion among musicians and musicologists whether there exists a pulsation [3] in the melody of an unmetered alap in Hindustani music. Musicians agree there is an implicit pulsation present, but quantification is left to subjects. At the same time, the subjective bias only results in an octave difference, i.e., there is a octave relation among the pace in which the subjects tap to the melody. We propose a measure, through our data-driven approach, to estimate a possible metric for the pulsation. We assume that the pulse obtained from the melody would correlate to the percussive pulsation anticipated during the metered bandish section. We compute the ratio of inter-onset-interval of the most salient svaras across BPs. Figure 9.6 shows a pulsation at 0.8 seconds and its harmonics which correspond to 75 beats per minute (bpm) and the percussive tempo of the concert is approximately 40 bpm. The noise in the estimate may also follow from a few short BPs (e.g., BP index 3, 7 etc.) as observed in Figure 9.1. This measure, therefore, needs further investigation before we generalize over the corpus.

9.4 Summary

We performed a data driven exploration of implicit structures in melodies of Hindustani music. We outlined the motivation and relevance of computational approaches for quantitatively studying the underlying musical concepts. We computed musically relevant and easy-to-interpret acoustic features such as svara frequency and svara duration histogram. For computing these features we primarily used existing tools and techniques often used in information retrieval of

[3] Pulsation is typically what listeners entrain to as they tap to a piece of music, colloquially termed the 'beat' or more technically the 'tactus'. Even a person untrained in music, can generally sense the pulse and may respond by tapping a foot or clapping.

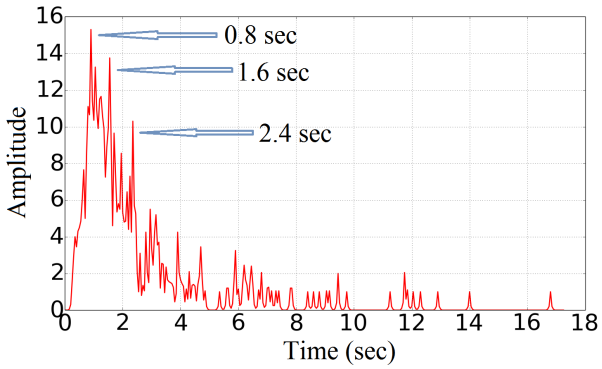

Figure 9.6: Ratio of inter-onset-interval of salient svaras across breath-phrases for the case-study concert in raga Todi by Ajoy Chakrabarty. We see a tatum pulse (peak) at 0.8 seconds and its harmonics.

Indian art music. We performed a quantitative analysis of 75 music concerts in Hindustani music in 10 different ragas. With that we showed how the musical concepts are manifested in real-world performances and experimented with several ways to quantify them. With this we also corroborate some of the interesting music concepts and discover implicit relationships between svaras and duration in the temporal evolution of a raga performance. In the future, one possible research direction would be to use these findings for characterizing artist-specific aspects and highlighting different nuances across gharanas in Hindustani music. Finally, we present an application that corroborates different ideas presented in the thesis to come up with an end-to-end system interface for real-time raga recognition running on a web browser.

9.5 Ragawise: A lightweight real-time raga recognition system for IAM

We demonstrate a web-based lightweight real-time melodic analysis and visualization system for IAM. Our system uses pitch class profiles, pitch transitions and melodic phrases for melodic characterization and raga recognition. For each raga we store a dictionary of its svaras (notes), svara transitions, and typical melodic phrases. We process the input vocals in real-time to estimate pitch, and subsequently perform melody transcription. The likelihood of each raga is updated in real-time based on the transcribed melody. In order to highlight the melodic events that are characteristic of a raga, we perform a dynamic visualization of the evolution of the likelihood of all the ragas for the sung melodic excerpt. The likelihood is computed as a weighted score of three important melodic aspects that characterize a raga, namely (i)

svaras, (ii) svara transitions, and (iii) raga-characteristic motifs. The dictionary is populated with these metrics as learnt from our datasets, and is verified by the author who is a performing vocalist of Hindustani music. The dynamic visualization refers to the raga salience distribution which is constructed as a probability distribution function where the area under the histogram sums to unity. Thus upon each update of likelihood, the salience of each raga get dynamically redistributed. The scoring scheme involves transcription of stable svaras and structural attributes as discussed in Chapter 6.

We demonstrate a lightweight real-time raga recognition system for Hindustani music that combines all three characterizing aspects of ragas for its automatic recognition, namely svaras, svara transitions, and characteristic phrases. The input to the system is a live audio stream of monophonic singing voice, and the output is a dynamic visualization of the evolution of the computed raga salience. The proposed system is fully functional on modern day web-browsers. Apart from efficient raga recognition, with this system, we can further explore musically interesting relationships between ragas through a gradual unfolding of the svaras. The proposed system would also find its use among advanced students of IAM in music pedagogy. Figure 9.7 shows the block diagram of the proposed approach and the modules at play.

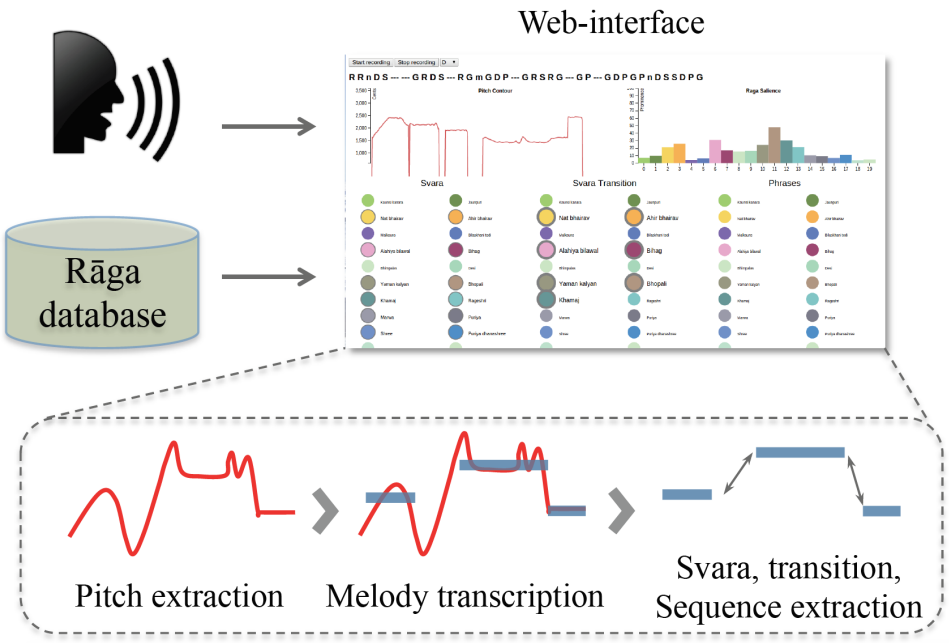

Figure 9.7: Block diagram of the proposed approach for Ragawise [78].

We indexed the raga grammar using a dictionary of svaras, the allowed melodic transitions, and characteristic melodic phrases. For each raga, its grammar also specifies the weights

of the different melodic events in its characterization. The raga grammar was specified by a professional Hindustani musician who has more than 20 years of formal music training. Using the output of the melodic transcription, the stored dictionary, and the weights for different melodic events, we compute and update the raga salience every time a svara is detected.

The real-time visualization on the web interface has four components: (i) pitch contour, (ii) transcribed svara sequence, (iii) raga salience, and (iv) raga space for different melodic events. The pitch contour for the past few seconds of audio is shown in cents, with the sequence of transcribed notes above it. The raga salience distribution is shown as a bar chart, and updated every time a svara is detected. The raga space has separate visualizations for each of the computed melodic events – current svara, the melodic interval from the previous svara, and the occurrence of sequence of three/four svaras. At the onset these melodic events, the ragas that allow the occurrence of these events are highlighted.

Our current system performs real-time pitch tracking and melody transcription of singing voice on a web browser along with dynamic raga recognition and visualization based on a stored dictionary of raga grammar. A simple heuristic based transcription is found to be sufficient for Hindustani music, a formal evaluation of which remains to be done in the future. The system is able to capture and highlight the aspects of the sung melody and identify the corresponding raga. The dynamic visualization also enables the realization of insightful relationships between ragas using the melodic aspects computed by the system. A computational profiling of the system showed that the entire processing can be done at real-time on a personal computer.

Chapter 10

Conclusion

10.1 Summary of contributions

Our analyses of performance audio use computational models to achieve an understanding of: (i) how the mandatory grammar is satisfied in performance, (ii) whether there are other consistent practices observed for a raga apart from what is specified by the grammar, and (iii) how the artiste improvises, i.e. uses the stock knowledge of the raga lexicon/grammar in "new" ways. We choose well-known allied raga-pairs for an empirical study of computational representations for the distinctive attributes of tonal hierarchy and melodic shapes. Raga notes, their relative saliences and intonation is supported by the concert-level continuous pitch histograms and discrete svara histograms. The frequency of occurrence of characteristic motifs is supported by relative counts of the characteristic motifs within performances. Melodic shape measurements with regard to duration are consistent with specified aspects, if any, of the melodic shape, i.e. the variability is found to be least where the raga lexicon/grammar has a firm prescription. Elsewhere, the variability is somewhat predictable from context information. We also got certain insights on similarity and differences across performances of a raga. The more "specified" a raga, the less variable it is across performances. At least one prominent aspect of "improvisation" is the course of svara-level exploration as the concert progresses in time, i.e. what is "spontaneous" is the decision to use a particular phrase or sequence of phrases at a given moment in the performance. Experienced artists use the melodic phrases in flexible ways without violating the prescriptions. "Ungrammaticality" refers to extending the flexibility so far as to tread on another raga. By comparing the surface realization of a phrase with the melodic shape of the same phrase in the allied raga, we observe that this is avoided by eminent performers.

Insights into the practical realization of the musicological concepts of raga delineation and melodic progression at the concert time scale are obtained. This points to the future possibility of developing the proposed methods for large-scale comparative studies in musicology which also ensures reproducibility. Although not the main focus of this work, the obtained outcomes can also be applied to the general raga recognition task, given the performance demonstrated on the relatively challenging sub-problem of discriminating allied ragas. A more direct validation of the proposed computational model would involve relating the predicted ungrammaticality of a performance to the ungrammaticality actually perceived by expert listeners.

Motivated by the parallels between musical structure and prosodic structure in speech, the identification and discrimination experiments are performed that explore the differences between trained musicians' and non-musicians' perception of ecologically valid synthesized variants of a raga-characteristic phrase, presented both in and out of context. Trained musicians are found to be relatively insensitive to acoustic differences in the vicinity of a prototypical phrase shape while also clearly demonstrating elements of categorical perception in the context of the boundary between ragas. Finally, combining all attempted paradigms, it would be of great interest to investigate the relative weighting of the different raga attributes for an overall rating of grammaticality, possibly at the different time scales, based on observed expert judgments.

10.2 Concluding remarks

Music, after all, is a performing art form. An individual may have different perception of the same music excerpt depending on context, training, or cultural background. To summarize, there exists a grey area between the actual signal and the perceived content. Moreover for MIR research, the challenge is to obtain a universal ground truth. There is no evidence of a 'true' notion of melodic similarity across subjects' judgments in IAM. It is very difficult to claim that all musicians would unanimously agree on the musical treatment of the embellishments – and may map it to different terminlogies. Then, is it that there is no 'universality' in human similarity-judgment? No, the situation is not that bad either, thanks to the holistic aspect of human perception that prefers to 'gist' a melodic entity and does not pay attention to the whole of it. This gives a feeling that stylizing the melodic contour is a good choice we made. It would be rather interesting to explore whether musicians from different 'gharana' (school of thoughts) perceive the same melody differently, following the way they have been trained to.

Most importantly, there is a visible 'gap' between the description of a melodic phrase in the musicological text and the actual music in performance. It is, therefore, customary to consult experienced performing musicians to validate our ground truth annotations, but we often come across comments related to 'musicality' (in an abstract manner) to explain these gaps. Our primary goal was to study such cases more closely and formulate quantitative measures to convey such 'musical' ideas.

There are few methodological considerations as well. The research in categorical perception (CP) was initiated by speech domain researchers and later adopted in the realm of music research. But the core methodology followed in speech research, might not translate directly to music CP. But there have been rarely any newly proposed experimental methodologies in this domain. We would say, these are rather scientific methods, that should be adapted to the concerned paradigm, when ported to. So did the researchers, as discussed in [121, 195], who had borrowed certain aspects of the methodology to obtain 'predictor' coefficients (melodic similarity measure in a Likert-type scale) from a perception experiment and later used in a multiple regression to propose a new empirical distance measure. While 'linear regression' is just another way of quantifying human similarity judgment, we think, we should critically question the methodological aspects like: what is the significance of using a 5-point Likert-type scale? What if the number of levels were less or more? If we use a 10-point scale (which is very common in music-emotion recognition scenario in the 2-D valence-arousal space [162]), what precisely is the difference in perception when a subject chooses a 7 or an 8. Is it the 'true' resolution that (s)he has perceived and 'correctly'[1] quantified? There are even crucial considerations like primacy and recency effects, e..g. order of presentation effect that we touched upon in our experimental setup.

We might, following the recent technology trends, like to propose a weighted combination of data-driven and knowledge-driven research. With the advancement of computational resources, a bunch of sophisticated machine learning models have been introduced, such as DNN, RNN, LSTM, end-to-end systems, attention-based and adversarial frameworks. These models are capable of producing competitive performance even without labeled data through

[1]Heuristics & Biases are the two main elements of human judgment under uncertainty. In fact, representativeness (probability of an object A belonging to a class B) is one of the main heuristic factors (mental shortcuts that usually involve focusing on one aspect of a complex problem and ignoring others [5]) that depends on an individual. It would be interesting to see whether there is a 'true' notion of similairty among Indian musicians.

deep learning and data-mining techniques. It would be interesting to investigate whether an intermediate representation of a melodic segment in such systems, e.g. a hidden DNN-layer, matches the abstraction level of human musical memory as engineered from our handcrafted features. Another important aspect is computational complexity. In the modern fast-paced world, everyone is fascinated about real-time applications that run on mobile platforms. To make some market-impact in such an exacting scenario, sincere efforts should be offered towards optimising on the computational front. This is a potential avenue for future research, building on the proposed models.

Finally we are excited to note that all of the above concerns boil down to the neuropsychological aspects of melodic similarity, i.e. what do humans 'understand' by 'similarity' – a classical problem of 'music genre recognition' is an easy way to conceive this idea. May it be the effect of training in music perception, or the process of 'gist'-ing music information in human memory, or storing a representative/exemplar 'template' for a melodic phrase in human brain, or modelling musical 'knowledge'; study of cognitive musicology is the best pathway. Thus, given the available resources (paradigms like: music technology, cognitive musicology, music performance etc.) of an interdisciplinary research, though limited, we pursued our humble endeavor towards finding a computational measure for adequately modeling human judgment of melodic similarity.

Appendix A

Behavioral experiment details

A.1 Subjects

Here is the demographic details as obtained from the participants from different subjects' groups for the behavioral experiments. The summary, for a quick reference is as follows.

- Twenty seven Hindustani musicians (13 female) with average age of 32 years (SD=4.6) and average years of training of 15 years (SD=2.7) constituted the Hindustani subjects' group, including ten instrumentalists.

- Twelve Carnatic musicians (8 female) with average age of 22 years (SD=0.7) and average years of training of 7 years (SD=1.1) constituted the Carnatic subjects' group.

- The Western subjects' group comprises eleven musicians (3 female) with average age of 19 years (SD=0.9) and average years of training of 5 years (SD=1.4).

- For the NonMusicians subjects' group, fifteen students (6 female) volunteered who had no formal training in music. These subjects seldom listened to Bollywood (Indian film) music, but did not have any exposure of Hindustani raga performances.

- In the IndiPop subjects' group, twenty four subjects (14 female) with average age of 21 years (SD = 1.2) with no formal training in Hindustani music, but adept in singing commercial Indi-pop (e.g. Bollywood) songs were present.

Username	Name	Age	Gender	Years of training	Subject
H_AL	Ashwin Lele	22	M	14	Flute
H_CJ	Chaitanya Joshi	24	M	17	Violin
H_PG*	Partha Ganguli	56	M	25	Sitar
H_SG*	Subhra Ganguli	54	F	20	Vocal
H_SS	Saketh Sharma	26	M	15	Vocal
H_RA	Ritvij Athvale	25	M	10	Vocal
H_UC*	Uttara Chousalkar	45	F	20	Vocal
H_P	Prajakta	23	F	14	Vocal
H_VG	Vaidehi Gokhle	33	F	16	Sitar
H_SP	Swati Purohit	24	F	12	Vocal
H_SR	Savani Ravindra	23	F	10	Sitar
H_AB*	Archi Banerjee	25	F	17	Vocal
H_NC	Nandini R Chowdhury	25	F	15	Vocal
H_SC*	Shankha Chatterjee	29	M	12	Vocal
H_AG	Anwesha D Gupta	23	F	17	Vocal
H_HM	Hirak Modi	27	M	15	Harmonium
H_MI	Mugdha Indurkar	26	F	15	Vocal
H_AR	Adwitiya Rao	22	M	7	Flute
H_AJ*	Ankita Joshi	29	F	20	Vocal
H_PM	Prajakta Marathe	25	F	18	Vocal
H_VS	Vaishali Sinha	23	F	12	Vocal
H_SR2	Shoumi Roy	24	F	14	Sitar
H_RG*	Ramakant Gaikwad	29	M	22	Vocal
H_AM*	Aditya Modak	28	M	21	Vocal
H_DM	Deepak Mallya	25	M	15	Vocal
H_VH*	Vinayak H	26	M	17	Vocal
H_AS	Arjav Shah	20	M	9	Vocal
H_SS	Srivatsan Sridhar	22	M	13	Violin
H_RA	Rohit M A	23	M	12	Violin
H_AK	SAsawari Keskar	18	F	8	Sitar

Table A.1: Details of the Hindustani subjects' group. The subjects with (*) mark have teaching experience. The age column corresponds to paericipants' age as per 2017. This list covers all Hindustani trained musicians who participated in either of the listening and the singing experiments.

Username	Name	Age	Gender	Years of training	Subject
C_TS	T V Sukanya	25	F	10	Veena
C_SN	Surabhi N	22	F	5	Vocal
C_RB	R Brindha	23	F	10	Vocal
C_TK	T Kalaimagan	25	M	16	Vocal
C_RN	R S Nivaythaa	25	F	5	Vocal
C_GJ	G Josephie	22	F	15	Vocal
C_AS	Achala G S	22	F	12	Veena
C_AT	Anjo P Thomas	28	M	7	Vocal
C_AP	Aiswaria P S	23	F	8	Vocal
C_AK	Apoorva Krishna	22	F	17	Violin
C_VI	Vignesh Ishwar	28	M	20	Vocal
C_GP	Govind P	27	M	15	Vocal

Table A.2: Details of the Carnatic subjects' group.

Username	Name	Age	Gender	Years of training	Subject
W_SS	Sertan Senturk	30	M	14	Guitar
W_G	Giuseppe	21	F	7	Piano
W_JP	Jordi Pons	23	M	10	Guitar
W_GD	Georgi D	24	M	12	Piano
W_RC	Rafael Caro	26	M	10	Cello
W_CS	Cicelia S	21	F	8	Vocal
W_RG	Rong Gong	24	M	10	Guitar
W_DS	Divyansh S	24	M	12	Guitar
W_RS	Rhythm Saw	22	M	15	Guitar
W_AS	Anushree Singh	22	F	10	Vocal
W_BS	Baishali Sarkar	24	F	12	Vocal

Table A.3: Details of the Western subjects' group.

Username	Name	Age	Gender
N_NS	Niramay Sanghvi	23	M
N_KS	Kamini Sabu	25	F
N_SP	Saurabh Pinjani	22	M
N_AK	Avinash K	26	M
N_SS	Shruti S	23	F
N_RD	Ranjan Das	25	M
N_NK	Nitin Kashyap	27	M
N_RG	Rashmi Gupta	23	F
N_CV	Charvi Vittal	21	F
N_NS	Nataraj K S	32	M
N_JP	Jayneel Parekh	20	M
N_SH	Sree Harsha	23	M
N_SS2	Shweta Sawant	28	F
N_VM	Vidya Menon	23	F
N_SS3	Siddhartha Saha	27	M

Table A.4: Details of the NonMusicians subjects' group.

A.2 Interface

The Sonic Mapper [164] interface was used for recording the subjective ratings via listening experiments. The software takes care of the randomization and outputs a text file with all ratings with respect to the stimulus continuum. A Matlab program parsed the data into matrix form and further computations were performed. Figures A.1 and A.2 shows the interface for the goodness rating and PME discrimination tasks.

The Cool Edit Pro software was used to play stimulus prompts for the listen & imitate experiment. As shown in Figure A.3, the stimulus prompt was played and the participant had to record the response in an open track with the tanpura background. This software was used specifically to use the session features; however the indexed audio files recorded on the portable recorder (as described in Chapter 7) were used for further processing.

Username	Name	Age	Gender
IP_RD	Riju De	27	M
IP_AK	Atulit Khanna	24	M
IP_SD	Saloni Deodhar	21	F
IP_KL	Kaivalya Lal	20	M
IP_AP	Aneesh Pithuriya	21	M
IP_PP	Pushkar Pandit	23	M
IP_SJ	Shalmali Joshi	25	F
IP_PV	Prajakta V	23	F
IP_SP	Sunayna Purohit	22	F
IP_SS	Swati Sirshant	25	F
IP_AM	Avantika Mathur	26	F
IP_SC	Sanjana Chakraborti	23	F
IP_KG	Kasturi Ganguli	27	F
IP_AP	Aishwariya P	24	F
IP_VG	Bhargavi Ganesh	24	F
IP_HB	Hirak Banerjee	28	M
IP_MP	Mugdha Pandya	23	F
IP_PB	Poornima V B	25	F
IP_AT	Ankita Tiwari	22	F
IP_SM	Swarangi Marathe	27	F
IP_VS	Vaishali Sarkar	25	F
IP_TD	Tushar Datta	30	M
IP_AA	Ankur Apte	26	M
IP_MM	Maitreyee Mordekar	28	F

Table A.5: Details of the $\mathsf{Indi-pop}$ subjects' group.

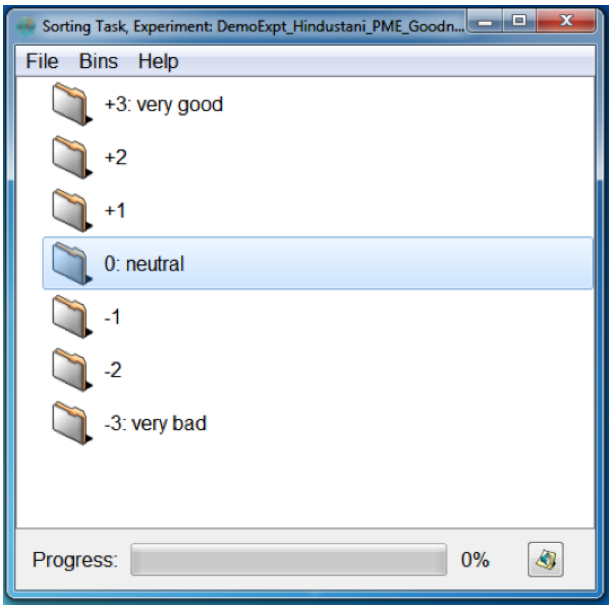

Figure A.1: The identification interface; shown is the goodness rating window for Experiment 1a (refer to Chapter 7).

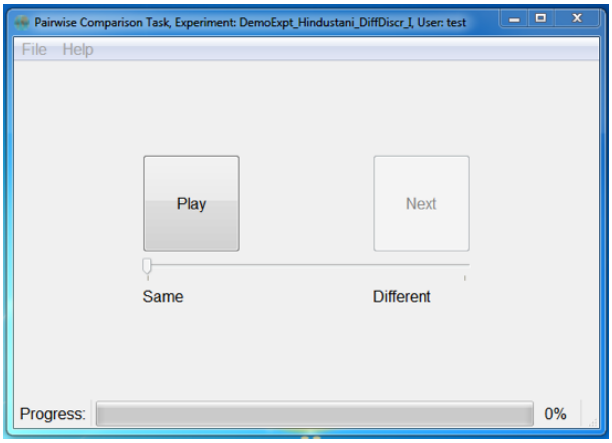

Figure A.2: The discrimination interface; shown is the PME discrimination window for Experiment 1b (refer to Chapter 7).

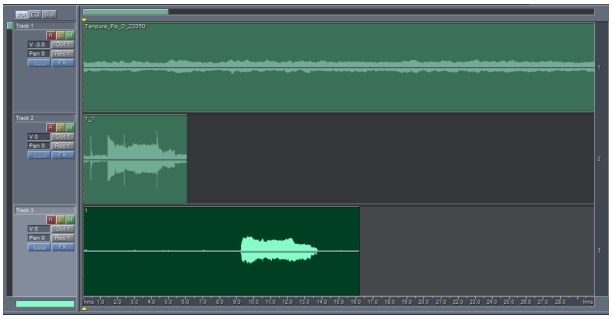

Figure A.3: The recording interface; shown is a sample recording session for a stimulus prompt for Experiment 3 (refer to Chapter 7).

Appendix B

Detailed review of some relevant literature

B.1 Algorithmic- versus human-judgement

B.1.1 A melodic similarity measure based on human similarity judgments

The main contribution of the paper [195] lies in designing an empirically derived measure (EDM) based on multiple regression with five predictor features, from a behavioral experiment of human judgment. Finally the authors compare their proposed similarity measure with the state-of-the-art string edit-distance Mongeau-Sankoff measure (MSM) and show the superiority of the proposed approach. The authors point to two main applications of melodic similarity based algorithms: content-based music search (cover song detection, music plagiarism detection etc.), and music composition tool that takes user-preference (in terms of genre, rhythm etc.). Literature on comparison of human similarity judgment vs. algorithmic judgments is mentioned, authors remark that most of melodic similarity-based MIR applications are devoid of cognitive approaches informed by human judgments. One interesting point is highlighted that edit distance measurements with a rich symbolic representation compares well to human similarity judgments according to a prior reference [122]. In this work, they are questioning this point. This study aims to extend beyond the mere comparison of empirical vs. algorithmic judgments of music similarity. The two ways of implementing a method based on human similarity judgment, as mentioned by the authors, are: within an existing application as is, or added on top as a higher level similarity measure, prioritized under certain conditions. The similarity measure is a function (multiple linear regression) of five high-levels quasi-independent musical features, chosen as predictors. These are: pitch distance, pitch direction, rhythmic salience,

melodic contour, and tonal stability. The choice of manipulations had referred to short-term memory theory (Murdock Jr., 1962).

For designing the empirically derived measure (EDM), a listening test is conducted (after a training session) with 22 participants (15 female, 7 males) in two trial blocks. Each block consisted of 96 stimuli (4 standard melodies * 12 variants of each * 2 transposed versions). Each standard melody is an eight-note melody (isochronous), while the comparison melody had only one note change at either 4^{th} or 5^{th} position. The stimulus tone-material is either C-major or C-minor scale and is accompanied by percussive rhythm spanning three bars (standard melody, followed by a silence bar and then the comparison melody). Listeners of the final experiment (after a training phase) marked the similarity on a Likert-type scale from 1 (least similar) to 5 (most similar). The encoding standards of each predictor (as mentioned above) are intuitive and some of them followed Narmour's implication-realization model (1990). A mixed design analysis-of-variance (ANOVA) is performed on all variants of stimuli, except between-subject transposition set. For non-transposed melodies, stepwise regression showed strong correlation with human similarity judgments for three predictors: pitch distance, pitch direction, and melodic contour. Whereas, tonal strength and melodic contour came out to be statistically significant predictors for transposed melodies. The observations are supported by the theory of local vs. global encodings of melody, proposed by Dowling (1978).

A similar strategy (as mentioned earlier) with 8 participants (5 male, 3 female) is used for the evaluation set, containing 12 comparison melodies. The objective distance measure for these melodies are chosen to be computed from the Mongeau-Sankoff algorithm. The distance measure is converted to a similarity value called Mongeau-Sankoff measure (MSM) through simple heuristics. Then both EDM and MSM are normalised to a scale between 0 and 1 for better comparison. The performance measure to compare the two, is the mean total error. From the performance values, authors claim that EDM clearly performed better than MSM. Though the authors wisely mention that the power of MSM might have been compromised, as their stimuli did not comprise any time-warping what Mongeau-Sankoff algorithm is widely known for. The authors have convincingly formulated an empirical measure for determining music similarity between two melodies differing by one note. The physical significance of the predictors were justified in regard to their importance in the perception of melodic structure. It is shown, in gen-

eral, that the empirical method (based on human similarity judgment) outperforms the chosen standard algorithm. The future direction, as mentioned by the authors, include experimentation with a more-than-one note changes in the comparison melodies. Authors also concern about the need for considering additional cognitive parameters such as primacy and recency effects of altered pitches, combined effect of musical predictors for each altered note etc.

Implications for us

We agree about the importance of a cognitive basis for melodic similarity. Edit-distance or DTW applied to pitch time-series cannot be expected to correlate highly with human similarity-judgement. Any reasonable distance measure must be informed by human similarity judgement experiments. The Likert scale seems suitable for similarity rating. A number of independent 'predictors' of melodic distance that are musically relevant for Indian music should be identified (e.g., pitch shift of the steady note segment, time-warping of a note-segment etc.). Then stimuli should be generated with the appropriate modifications of the given melodic shape. In the above case case, the various predictors are different aspects of the same modification, e.g., by changing pitch of a selected note. We need to design our 'predictors' in a similar way, i.e., we need to identify whether one selected modification influences several musical aspects.

B.1.2 An empirically derived measure of melodic similarity

The new contribution of this paper [196] lies in the addition of two configurations in the experiment design. The authors, in the previous work, had mentioned the need for considering additional cognitive parameters such as primacy and recency effects of altered pitches. The current work includes an experiment to observe the primacy and recency effects (referred to as PR). The second new experiment involves two-note change in the comparison melody. One additional detail that was not mentioned in the previous paper is that authors conducted a sub-experiment to record goodness-of-fit ratings from 9 participants where they had to judge how well a probe tone fitted within the context of the melody. In experiment 2, three different PR functions were used, viz. linear (PR-L), quadratic (PR-Q), and quadratic function with limits (PR-Q2), to better account for explained variance. The implementation was in terms of weights to the predictor values in the regression model. The weights depended on the location of the note-change: if change occurred in the beginning or end of the melody, the predictors would

have a greater contribution to the similarity model. For non-transposed comparisons, planned contrasts revealed that the original model performed worse overall than those models incorporating primacy and recency. Experiment 3 involved two-note change in the comparison melody, either in the form of an outer set (positions 2 & 7), or an inner set (positions 3 & 6).

The evaluation of the EDM was also done on the melodies from RISM A/II database, the melodies were chosen with certain criteria that matched the standard tones as composed by the authors. One of the concluding remarks was that greater familiarity with a melody enables more information pertaining to that melody to be stored in brain. For these melodies, listeners could get access to deeper information such as contour and rhythm. Whereas for unfamiliar melodies, listeners are likely to rely more on surface level information for similarity judgments. Results showed that similarity-judgments of one-note changes were modelled better after incorporating a limited form of primacy and recency effects as a non-linear function. Another important finding of the study is that for multiple-note changes, average models performed better than the additive models. However for this case, incorporating PR effects deteriorated the performance. One important future advancement that authors mention is that effects of note-duration should be incorporated to the model. Authors also propose to modify the hierarchy of tonal stability or to introduce external weighting schemes in order to overcome the limitations such as out-of-key and harmonic changes.

B.1.3 Measuring melodic similarity: human versus algorithmic judgments

This paper [122] is one of the first of its kind, [195] has stressed on few of the findings of this work. The organisation of the paper is excellent: authors have isolated the backgrounds of each subdiscipline (music psychology and mathematical music theory) and described the two-fold aim of the study. Methods include mathematical systemization and psychological rating test for melodic similarity. Results showed that there is a 'true' notion of melodic similarity across subjects' judgments. Apart from showing the effectiveness of edit distance or n-gram based similarity measures, authors highlighted an important fact that weighting of pitch data according to their duration appeared as an useful option for most measures. It is remarked that for different rating tasks and scenarios, subjects alter their rating strategy and pay attention to different dimensions. The authors concern about the abundance of melodic similarity algorithms that the measures that are cognitively most adequate need to be determined. The

processing stages of most of the similarity measures are generalized as follows. The first stage (representation) includes projection, composition and differentiation. The second stage (main transformation) focuses on rhythmic weighting, fuzzification, and contourization. The third and final stage is computation of a similarity value. Authors have classified this into three broad categories: vector measures (e.g., DTW), symbolic measures (e.g., string-based distances), and musical (mixed) measures.

Three rating experiments were done in a test-retest-design, where the 2^{nd} and 3^{rd} experiments served as control experiments. The inter- and intrapersonal judgments of the selected subjects showed very high correlations on various measures such as coefficient Cronbach's alpha. Authors assumed that melodic similarity would work on five dimensions: Contour information, interval structure, harmonical content, rhythm, and characteristic motives. The conclusion of the paper states that melodic similarity is a well-defined notion, at least for the group of western music experts. Authors also remark that for different tasks, music experts seem to have different strategies for similarity ratings. These strategies can be modeled optimally with compound measures that incorporate information from different sources. Another important point to note in the current context, is about what the authors remarked a decade back about the research trend in the then period. "A current development is to base similarity measurement not on entire melodies, but on phrases that make up longer melodic sentences. This will enable similarity comparisons between long reference melodies and short excerpts, as it is the case in many applications in music information retrieval and in search for 'motivic quotations'."

B.1.4 Perceptual evaluation of music similarity

The perceptual evaluation as discussed in this paper [131] consisted of intra-subject consistency and inter-subject concordance test. Significant differences between musicians and non-musicians, and between subjects being familiar or non-familiar with the music were found few cases. Multidimensional scaling revealed a proximity of songs belonging to the same genre, congruent with the idea of genre being a perceptual dimension in subjects' similarity ranking. The proposed theoretical models underline the multidimensional nature of perceived music similarity and stress the importance of the perceptual weighting of the various musical dimensions in this respect. Authors mention about Deliege's approach being an extension of the Gestalt theory applied to music. The fact that a listener uses his/her prior experience to segment the

musical piece and extracts features from every part, is highlighted. Thus a weighted comparison between the features extracted from different parts could tell whether two parts are similar and in which respect. Authors feel that, research published so far appears very fragmented or too specific for general applicative interest based on a global representation of similarity. As for the applications, few algorithms developed for assessment of music similarity between songs are linked to human perception. The open questions posed, are: do subjects have a common perception of music similarity? What are the principal perceptual features used by subjects in rating similarity? Which features are most relevant? Is there a significant influence of musical experience (musicians/nonmusicians, familiar/non-familiar musical material)? The authors advance the hypothesis that musicians, who perform more concordantly on a fewer set of triads than nonmusicians, have a more common approach to music interpretation than do non-trained listeners. For the case of music familiarity, a possible conclusion might be that subjects not familiar with the music (who show better concordance on few triads) make similarity judgments based more on the surface musical audio signal rather than on associated experience factors. Authors claim to be able to represent subjects' "perceptual genre space" in 3-dimensions.

B.2 Cognitive perspectives on melodic similarity

Insofar we have only looked at literature in the domain of acoustics and psychoacoustics. Our endeavor would remain incomplete if we do not touch upon at least some of the cognitive aspects of melodic similarity. One may genuinely argue that behavioral experiments fit well in the cognitive psychology domain, we would further like to inspect its neuropsychological facet. To summarise, we want to address the neurophysical means that a melodic segment is stored in human memory and how it is retrieved. Further, we would like to know about the cognitive notion of 'similarity', i.e., to estimate the allowed tolerance band that a 'similar' melody gets the same identity [55].

Whenever we talk about psychomusicology, the names that come to our minds are the eminent authors like Aniruddh D. Patel, Daniel J. Levitin and so on. Some of their books [108, 109, 137] are the finest of the references, but loaded with neuroscientific studies (analysing the human brain) and hence is out of the scope of this report. Sill, this report would be incomplete if we miss to mention their approach very briefly. For e.g., Tillman [185] in his review

of Patel's book [137] points to a phenomenal question: are cognitive and neural correlates domain-specific or common to music and language processing? Though a large segment of the literature emphasizes on the differences thereof, Patel's work however, stresses on the commonalities. Kranenberg [193] mentions about an underlying problem of folk-song retrieval from a cognitive point-of-view: how is a melody encoded in human memory and how is it transformed into an audible song instance during performance? Author also remarks that such a knowledge can be used to discriminate between stable and unstable elements of melodies in oral transmission.

Another interesting theory in cognitive sciences is the Gestalt psychology, the central principle of which says that human mind forms a 'global whole' with self-organizing tendencies [4]. The key principles of the Gestalt theory, as listed in [3], are: proximity, common fate, similarity, continuity, closure, and symmetry principles. Though very common in vision, researchers like Bregman [27] have addressed the issues of organization, grouping, and segmentation in the auditory domain as well. The auditory analogues of Gestalt priciples in audition is manifested in terms of differences and similarities in loudness, pitch, and timbre of sounds. In the literature of melodic contour retrieval, investigations of Gestalt principles are not that many. But we cite a paper [23] where the author challenges the Gestalt principles in memory-based models of melodic analysis.

We shall look at few of the relevant literature that would enable us design the retrieval algorithms coherent with human-thinking. The fisrt work we discuss, is a book chapter by Daniel J. Levitin. This gives a very comprehensive idea about musical memory as follows.

B.2.1 Memory for musical attributes

The author [107] addresses an immediate question whether different types of memory, as psychologists tend to make conceptual distinctions among, are for conceptual conveniences or whether they have an underlying neural basis. Author further comments, some of the conceptual labels for memory systems, such as 'procedural memory', actually encompass somewhat independent processes that are conveniently categorized together (for pedagogical reasons), but do not necessarily activate a single distinct brain structure. Different entities of memory, such as iconic memory, echoic memory, semantic memory, long-term memory (LTM), short-term

memory (STM), and working memory are described with interesting examples. The context of musical memory is next addressed where the author remarks about what chunking have to do with music. It is highlighted as a common intuition that the sole function of memory is to preserve the details of different experiences we've had. But there is a large body of research showing that our memory for details is actually pretty poor. For evidence, people tend not to have a very good memory for the exact words of a conversation, but instead remember the 'gist' of the conversation. What is the function of memory, then, if not to remember events accurately? We question, how does this concept translate to musical memories? In fact, there is a great deal of evidence that memory does preserve both the details and the 'gist' of experiences, and we are usually able to access information at the appropriate level.

The author comments that Objects in the visual world have six perceptual attributes: size, color, location, orientation, luminance, and shape. What do we mean by 'object'? Author proposes that an object is something that maintains its identity across changes (or transformations) in these attributes. In other words, as we move an object through space, it is still the same object, i.e., its identity is preserved. A performance of music contains the following seven perceptual attributes: pitch, rhythm, tempo, contour, timbre, loudness, and spatial location. The author further remarks that the term contour refers to the shape of a melody when musical interval size is ignored, and only the pattern of 'up' and 'down' motion is considered. Each one of these attributes can be changed without changing the others. With the exception of contour, and sometimes rhythm, the recognizability of the melody is maintained when each of these attributes is changed. In fact, for many melodies, even the rhythm can be changed to some degree and the melody will still be recognizable (White, 1960). To elaborate further, a melody is an auditory object that maintains its identity under certain transformations, just as a chair maintains its identity under certain transformations, such as being moved to the other side of the room or being turned upside down. A melody can generally retain its identity with transformations along the aforementioned dimensions. It is to be noted that one of the reasons we are able to recognize melodies is that the memory system has formed an abstract representation of the melody that is pitch-invariant, loudness-invariant, and so on. We take for granted that our memory system is able to perform this important function.

Some evidences of memory of contours is presented next. At first, author states that the

idea of contour being an important attribute of melody seems counterintuitive. Contour is a relatively gross characterization of a song's identity. However, its utility has been shown in laboratory experiments. There is evidence that for melodies we do not know well (such as a melody we have only heard a few times), the contour is remembered better than the actual intervals (Massaro, Kallman, and Kelly, 1980). In contrast, the exact interval patterns of familiar melodies are well remembered, and adults can readily notice contour-preserving alterations of the intervallic pattern (Dowling, 1994). Infants respond to contour before they respond to melody; that is, infants cannot distinguish between a song and a melodic alteration of that song, so long as contour is preserved. Only as the child matures, he/she is able to attend to the melodic information. The explanation of why the contour of a melody might be more readily processed is that it is a more general description of the melody, and it subsumes the interval information. It is only with increasing familiarity, or increasing cognitive abilities, that the intervallic details become perceptually important. The most relevant discussion for our work is presented next. The author questions, to what extent do our memories of music retain perceptual details of the music, such as the timbre, pitch, and tempo of songs we have heard? Do we remember all the details of the piece, even details that are not theoretically important? Specifically, since melody is defined by the relation of pitches and rhythms, it would be easy to argue that people do not need to retain the actual pitch and tempo information in order to recognize the song. However, the music theorist Eugene Narmour (1977) argued that listening to music requires processing of both absolute information (schematic reduction) and relative information (irreducible idiostructural), so the question is whether both types of information reach long-term memory.

The modern view is that memory is distributed throughout various parts of the brain, and that different types of memory engage separate neural structures. Memory for music, just like memory for prose or pictures, probably comprises different cognitive subsystems to encode the various aspects of music. There is a growing consensus that memory serves a dual function: it abstracts general rules from specific experiences, and it preserves to a great degree some of the details of those specific experiences.

B.2.2 Working memory in music: a theoretical model

The author [21] claims that music psychologists have accepted a dual memory system similar to the model of Atkinson and Shiffrin (1968), in which there are separate short- and long-term

storage components of information. Also, many psychologists have addressed the concern that short-term memory (STM) may consist of both storage and processing components, the composite unit is often labeled as working memory. The author remarks that two broad issues must be addressed: the nature of short-term storage of music versus other phonological (primarily verbal/textual) information, and actions that justify the existence of a processing component in STM in music. These two areas follow the two components of the phonological loop: storage and articulatory control. It is highlighted that Baddeley (1990) provides the following phenomena as evidence of the phonological loop: acoustic similarity, word-length effect (capacity), articulatory suppression, and perhaps most importantly for this study, unattended speech effect. Another primary consideration is the interaction between STM and long-term memory (LTM), especially as related to using LTM strategies to improve STM performance (e.g., chunking, rehearsal). The author states that modality, suffix, and recency effects are considered to be important characteristics of short-term auditory memory. In a study of melody recognition in STM, Dowling (1973) found a "J-shaped" serial position curve, typical of STM for verbal material. It is worthy of noting that highly trained musicians demonstrated a recency effect for both visual and auditory presentation when tested on immediate serial recall of notes; moderately trained musicians demonstrated this effect only for the auditory presentation. In addition, recency effects were noted for both auditory and visual presentation of music, with the explanation that musicians form both visual and auditory representations for written music. Roberts (1986) found that modality (advantage of aural over written presentation) and recency effects, when linked together, differ between music and language.

A number of studies have investigated perception and memory of melodies that are altered in some fashion, thereby increasing the processing component of working memory. There are two separate discriminations of melody: contour and intervallic content. An important finding is that untrained subjects do not find contour recognition much more difficult than do trained subjects (Burns & Ward, 1978; Dowling, 1978), however training seems to be important in interval recognition (Cuddy & Cohen, 1976). Perhaps trained subjects are able to draw on a richer LTM, allowing more efficient LTM strategies to be applied in order to chunk information so that storage can be increased. The author also remarks that tonal structure is certainly an organizational element that influences melodic recall. In fact, tonality might be one the greatest of all syntactical organizers in Western music. This would seem to suggest a processing of

information held in STM with structures kept in LTM and would seem to strongly support the model of working memory. As music is heard, the listener would attempt to place the information in some kind of organizational framework. A broad discussion of "unattended music effect" is presented and the final remarks by the author is as follows. Training seems to improve subjects' abilities in many memory tasks, especially in the ability to chunk information; they are able to draw on more better LTM traces and strategies. Individual differences portrayed in some music aptitude tests may then represent not talent or musical intelligence but ability, reflecting differences in working memory capacity.

B.2.3 Scale and contour: two components of a theory of memory for melodies

As the title suggests, the study [44] focuses on two main concepts: scale and contour. Authors develop a model of how melodies are stored in long- and short-term memory. It is claimed that the contour is an abstraction from the actual melody that can be remembered independently of pitches or interval sizes. Evidence is presented supporting the lifetime stability of scales and the fact that they seem to have a basically logarithmic form cross-culturally. The second component, melodic contour, is shown to function independently of pitch interval sequence in memory. The author remarks that remembering melodies is a basic process in the music behavior of people in all cultures. This behavior may involve production, as with the singer performing a song for an audience or the participant in a significant social event trying to remember the appropriate song. Or it may involve recognition, as with the listener whose comprehension of the later developments in a piece depends on memory for earlier parts. About the two components addressed, author's view is as follows. First, there is the melodic contour, the pattern of ups and downs, that characterizes a particular melody. Second, there is the overlearned musical scale to which the contour is applied and that underlies many different melodies. It is as though the scale constituted a ladder or framework on which the ups and downs of the contour were hung. The author has assembled evidence that memory for the contour can function separately from memory for exact interval sizes. That is, the contour is an abstraction from the actual melody that can be remembered independently of pitches or interval sizes. This is true for melodies in both short-term and long-term memory. Further, it is commented that the contours of brief, novel atonal melodies can be retrieved from short-term memory even when the sequence of

exact intervals cannot. Finally, instead of just holding a melodic phrase in short-term memory, subjects in Dowling's (1972) study on melodic transformations had to turn it upside down, backwards, or both. Recognition followed that order of increasing difficulty. Of interest here is the fact that subjects were able only to recognize the melodic contour when so transformed and not the exact interval sizes. Author also remarks about long-term memory. In addition to being preserved in short-term memory, melodic contours also seem to be retrievable from long-term memory independently of interval sizes. It is remarked that melodic contours of familiar tunes can be recognized even though the distortion of interval sizes is very great. The final commentary of the author is rather interesting to us. Author speculates, what we would expect from the present theory is that variants of a tune would share certain similarities, namely, those that arise from very similar contours being hung on the underlying scale framework in various ways. This variation might occur for three principal reasons: forgetting of the original intervals, the desire to create interesting innovations by manipulating interval relationships, and changes of instruments and scale systems that necessitate transformations of melodies (as in the case of adapting a non-Western melody to a Western scale).

B.2.4 Modeling memory for melodies

The aim of the presented study [125] was to understand structural descriptions of melodies that influence recognition memory for melodies. A further aim was the exemplary comparison of statistical modeling approaches for data from psycho-musicological experiments. The variables used to predict the subjective judgments comprised data about the subjects' musical experience, features of the original melody and its position in the music piece. The task followed the "recognition paradigm", widely used in memory research (e.g., Dowling et. al., 2002). The ratings were done on a six-point scale encoding the subjects' decision and their judgmental confidence in three levels ("very sure no", "sure no", "no", "yes", "sure yes", "very sure yes"). The idea behind the recognition paradigm is that correct memorization should result in the ability to detect possible differences between the melody in the song and the test melody. The authors remark that it is interesting to see that in all applied models the two measures of melodic similarity and structure similarity are the variables with the largest explanatory potential. From a viewpoint of a cognitive memory model this means that the structural relation and the quantifiable differences between melody in the song and single line test melody is more decisive for memory performance than are experimental parameters (like the position of the target melody in

the song or the duration of the different song parts) or information about the subjects' musical background. The authors finally remark that melodic features that may serve as further predictors are melodic contour, melodic and rhythmic complexity, coherence of melodic accents, and the familiarity of these features as measured by their relative frequency in a genre-specific database.

Implications for us

The implications of this work for us would be to explore more onto the recognition paradigm to record subjective similarity ratings and judgmental confidence against a suitable question in the domain of Hindustani classical music. This would help us develop the model for phrase similarity as a function of several semi-independent predictor parameters. But the even more important issue is the choice of baseline subjects. The widely used choices of baselines are musicians vs. non-musicians, musicians trained in different music repertoires etc. But we would rather like to investigate whether musicians develop an advanced perception in course of training that listeners are devoid of.

B.2.5 Memory-based models of melodic analysis: challenging the gestalt principles

The author [23] argues for a memory-based approach to music analysis which works with concrete musical experiences rather than with abstract rules or principles. Literature says that while listening to a piece of music, the human perceptual system segments the sequence of notes into groups or phrases that form a grouping structure for the whole piece. One of the main challenges in modeling musical segmentation is the problem of ambiguity: several different grouping structures may be compatible with a sequence of notes while a listener usually perceives only one particular structure. The author also states that most models of musical segmentation use the Gestalt principles of proximity and similarity (Wertheimer, 1923) to predict the low-level grouping structure of a piece, while some models also incorporate higher-level grouping phenomena, such as melodic parallelism and harmony. This paper thus proposes a supervised, memory-based approach to music analysis which works with concrete musical fragments rather than with abstract formalizations of intervallic distances, parallelism, meter, harmony or other musical phenomena.

Author also highlights that, recent psychological investigations suggest that previously heard musical fragments are stored in memory (e.g., Saffran et. al., 2000), and that fragments that are encountered more frequently are better represented in memory and consequently more easily activated than less frequently encountered fragments. The author remarks that it is important to study the merits of such a model so that its results may be used as a baseline against which other approaches can be compared. The paper considers the following memory-based parsing models from the literature: the Treebank grammar technique of Charniak (1996), the Markov grammar technique of Seneff (1992) and Collins (1999), and the Data-Oriented Parsing (DOP) technique of Bod (1993, 1998). The evaluation method employed the so-called 'blind testing method' which has been widely used in evaluating natural language parsers (Manning & Schutze, 1999). A widely used evaluation scheme in natural language parsing, the PARSEVAL scheme (based on the notions of precision, recall, and F-score (Black et. al., 1991)) was used. A qualitative analysis of the results indicated that there are grouping phenomena that challenge the commonly accepted Gestalt principles of proximity, similarity and parallelism. The author argues that music perception may be much more memory-based than previously assumed.

B.3 Concluding remarks

It is ruther the abundance (and not the scarcity) of available methodologies in the literature that concerns us to choose from. Researchers from different realms have proposed scientifically motivated methods and materials, though works in Indian music repertoires are rare. Thus we need to be consistent with the scientific standards while initiating an interdisciplinary approach for MIR in raga music, probably a collaborative effort with domain-experts would be a better choice to start with. We shall now summarise the findings from the reviewed literature that would be critical when comparing our results against.

B.3.1 Works based on representation-cum-similarity measure

The works under this facet aim to develop a computational model informed by human similarity-judgment. There is a commonality among researchers [121, 122, 124, 126, 195, 196] who have explored this view-point that they have recorded subjective similarity-ratings to tune the parameters of a computational model, i.e., the empirically derived (dis)similarity measure used

the human ratings in the form of a linear regression. The principle ideas we learn from their work is: (i) how to choose a set of suitable melodic 'predictors' based on the music repertoire concerned, (ii) how to systematically encode the predictor values while generating variants of the standard melody, (iii) how to judge the significance of each of these predictors and use the values to weight the regression model, and finally (iv) how to evaluate the proposed model by comparison with state-of-the-art distance measures. Regression is a popular model for data-fitting, though it being 'linear' sounds rather a strong assumption. Next we reviewed [131] that mentions about multidimensional-scaling (MDS) that models the underlying multidimensional nature of perceived music similarity and proposes a perceptual weighting of the various musical dimensions. This work also questions the influence of musical training and experience (musicians vs. non-musicians, familiar vs. non-familiar musical material) that motivated us to study the literature on musical memory under the realm of psychomusicology.

B.3.2 Works on cognitive music psychology

The reviewed works on this facet discuss different memory aspects concerend with music perception and cognition. The main highlights of the concepts we learn include: (i) perceptual attributes of musical dimensions (e.g., pitch, rhythm, contour, timbre, loudness etc. that would enable us extract the 'right' acoustic feature from a music signal), (ii) how humans 'gist'/'chunk'/'group' music information (this reflects perceptually relevant melodic features that would help us better represent the melody line the way humans memorize it), (iii) short- and long-term memory of music and their mutual relationship (this could lead us towards finding a canonical form or a template of a melodic phrase and improvisations thereof), and finally (iv) Gestalt principles applied to music (this would indicate the optimal time-scale that we should 'window' the music signal in the bottom-up analyses that correlates to human perception which, in contrast, is top-down). Though many of these literature talk about high-level cognitive phenomena, we try to adopt the methodological aspects of psychoacoustic experiments to be able to frame the 'right' question to extract the 'intended' human response that would indicate the variations in the 'correct' dimensions we expect to capture. The other important aspect that has a huge implication for us is to find a set of suitable melodic predictors for Hinduatani classical music. It is also necessary to pin down to fixed paradigms that is feasible to be experimented, given the limited choice of domains we have in our hands. Thus we decide to take up the "recognition paradigm" [125] and "similarity-judgment paradigm" for further investigation.

B.4 Literature on ERP studies

The following are the relevant literature on ERP studies in music to investigate musical expectancy in musicians vs. non-musicians. Only the last among the three works discusses on Hindustani music stimuli.

B.4.0.1 Besson (1993) [22]

Authors reported (based on a pilot study where only behavioral responses were recorded) that familiar musical phrases were better recognized by musicians than non-musicians. In their study, one of the three types of incongruities were each incorporated in the comparison set of melodies: (i) 'harmonic incongruity' where the end-note was made out of key (i.e., prohibited by the tonal structure), (ii) 'melodic incongruity' where the end-note was not out of key but different from either the composed tune (in case of familiar melodies) or the most expected note by the musical rules of closure (for unfamiliar melodies), and (iii) 'rhythmic incongruity' where the end-note was delayed by 600 ms. When categorization of the incongruity was independent of the specific musical knowledge (like the case of 'rhythmic incongruity') or it depended upon a episodic memory trace ('harmonic incongruity' for familiar melodies), non-musicians performed as good as musicians. In contrast, when musical knowledge was necessary (like the task of 'melodic incongruity' or 'harmonic incongruity' for unfamiliar melodies), musicians, as expected, outperformed non-musicians. Authors formulate four hypotheses from the above inferences. (i) brain responses for musicians must be higher than that of non-musicians, because musical expertise makes musical expectations clearer and more salient; (ii) such brain responses should be larger for familiar than for unfamiliar melodies because of comparatively more expectancy being generated for familiar melodies; (iii) brain waves for 'harmonic incongruity' should be larger than that of 'melodic incongruity', because disconfirmation of expectancies would be stronger in the former (that violates the musical rule of tonality) than in the latter (which are musically 'correct' but not the most expected note within the musical context); and (iv) no difference for the musicians and non-musicians is predicted for 'rhythmic incongruity', since no specific musical knowledge is necessary to detect a 600 ms delay between the last two notes. Additionally, authors remark on suitability of LPC-like analyses over N400 like ERP-data for music perception compared to linguistic processing in the brain.

B.4.0.2 Tervaniemi (1997) [184]

To reveal neurophysiological prerequisites of musicality, auditory event-related potentials (ERP) were recorded from musical and non-musical subjects, musicality being here defined as the ability to temporally structure auditory information. Instructed to read a book and to ignore sounds, subjects were presented with a repetitive sound pattern with occasional changes in its temporal structure. The mismatch negativity (MMN) component of ERPs, indexing the cortical pre-attentive detection of change in these stimulus patterns, was larger in amplitude in musical than non-musical subjects. This amplitude enhancement, indicating more accurate sensory memory function in musical subjects, suggests that even the cognitive component of musicality, traditionally regarded as depending on attention-related brain processes, in fact, is based on neural mechanisms present already at the pre-attentive level. Auditory stimulation in the Order-Change Condition consisted of a continuously looped stream of four 125 ms tones (...E–F–G–A–E–F–G–A...). To determine how accurately the temporal order of this tone pattern was automatically encoded by the auditory system, infrequent (P = 0.1) changes were embedded in the stream by occasionally reversing the order of two consecutive tones (E–G–F–A). In the Pitch-Change Condition, frequent tone sequences forming C major chords (...C–E–G–C...) were randomly (P = 0.1) replaced with sequences forming a C minor chord (...C–Eb–G–C...). There, a new pitch was introduced enabling us to determine the MMN to a relatively primitive stimulus change without the importance of temporal information. During the 'Discriminate Order-Change Condition', subjects' task was to indicate by a button press each time they detected any change in the stimulus. In conclusion, the present data showed that even in ignore condition the musical subjects' central auditory system responds more vigorously to temporal-order reversals in repetitive sound patterns than that of the non-musical subjects. This suggests that auditory-cortex based sensory memory encoded the auditory information structure more accurately in the musical than in the non-musical subjects. Thus, the structuring ability probed by the present musicality test relies on cortical functions which can be probed by ERP recordings even during a reading task. Consequently, the cognitive component of musicality, traditionally considered to depend on attentional high-level cognitive processes, in fact depends on brain mechanisms which operate already at the pre-attentive level. Further, since the MMN to pitch changes did not significantly differ between these 2 groups, the brain prerequisites of the cognitive and sensory components of musicality are, at least to a great extent, separate.

B.4.0.3 Hegde (in preparation)

Authors argue that the cognitive demand for musicians is different from non-musicians. Musicians often listen to music with a grammatical viewpoint that they attend the syntactic parameters while in attention mode. Initial experiments showed prominent P600 peaks in musicians' ERP that conform to the proposed hypothesis. The stimuli consisted of flute melody in a raga with notes of the natural major scale, flat sixth (komal dha) was incorporated at certain places as a melodic incongruity. The EEG signal was time-locked to the auditory stimulus and the amplitude of P600 component (related to syntactic incongruity) of the ERP was recorded at the points of incongruity. Prominence in the peak amplitudes is found in trained Hindustani musicians over non-musicians. Authors infer that musicians cognitive mode is such that they attend music grammatically and hence P600 was found to be very prominent whenever their expectancy was not met.

Glossary

Bibliography

[1] Advanced topics: Categorical perception (cp). last accessed: October 22, 2014.

[2] Cognitive Musicology. Wikipedia "http://en.wikipedia.org/wiki/Cognitive_musicology". last accessed: September 5, 2015.

[3] Gestalt Principles. Scholarpedia "http://www.scholarpedia.org/article/Gestalt_principles". last accessed: September 5, 2015.

[4] Gestalt Psychology. Wikipedia "http://en.wikipedia.org/wiki/Gestalt_psychology". last accessed: September 5, 2015.

[5] Heuristics in judgment and Decision-making. Wikipedia "https://en.wikipedia.org/wiki/Heuristics_in_judgment_and_decision-making". last accessed: September 5, 2015.

[6] B. E. Acker, R. E. Pastore, and M. D. Hall. Within-category discrimination of musical chords: Perceptual magnet or anchor? *The Journal of the Acoustical Society of America*, 95(5):2937–2937, 1994.

[7] B. E. Acker, R. E. Pastore, and M. D. Hall. Within-category discrimination of musical chords: Perceptual magnet or anchor? *Perception & psychophysics*, 57(6):863–874, 1995.

[8] C. R. Adams. Melodic contour typology. *Ethnomusicology*, 20(2):179–215, May 1976.

[9] N. Adams, M. Bartsch, J. Shifrin, and G. Wakefield. Time-series alignment for Music Information Retrieval. In *Proc. of Int. Soc. for Music Information Retrieval (ISMIR)*, pages 303–310, 2004.

[10] H. Allan, D. Müllensiefen, and G. A. Wiggins. Methodological considerations in studies of musical similarity. In *ISMIR*, pages 473–478. Citeseer, 2007.

[11] H. Allan, D. Mullensiefen, and G. A. Wiggins. Methodological considerations in studies of musical similarity. In *Proc. of Int. Soc. for Music Information Retrieval (ISMIR)*, pages 473–478, 2007.

[12] R. Asano and C. Boeckx. Syntax in language and music: what is the right level of comparison? *Frontiers in psychology*, 6:942, 2015.

[13] Autrim-NCPA. Music in Motion: The automated transcription for Indian music (AUTRIM) project by NCPA and UvA. url: https://autrimncpa.wordpress.com/, 2017. Last accessed: April 26, 2017.

[14] S. Bagchee. *Nād: Understanding Raga Music*. Business Publications Inc, 1998.

[15] S. Bagchee. *Shruti: A Listener's Guide to Hindustani Music*. Rupa Co, 2006.

[16] S. Barrett. The perceptual magnet effect is not specific to speech prototypes: new evidence from music categories. *Speech hearing and language: Work in progress*, 11:1–16, 1999.

[17] A. Batliner and L. Schiefer. Stimulus category, reaction time, and order effect-an experiment on pitch discrimination. *Proc. ICPhS (3)*, pages 46–49, 1987.

[18] S. Belle, R. Joshi, and P. Rao. Raga identification by using swara intonation. *Journal of ITC Sangeet Research Academy*, 23, 2009.

[19] A. Bellur, V. Ishwar, and H. A. Murthy. Motivic analysis and its relevance to raga identification in carnatic music. In *Proc. of the 2nd CompMusic Workshop*. Istanbul, Turkey.

[20] J. F. Bernabeu, J. Calera-Rubio, J. M. Inesta, and D. Rizo. Melodic identification using probabilistic tree automata. *Journal of New Music Research*, 40(2):93–103, 2011.

[21] W. L. Berz. Working memory in music: A theoretical model. *Music Perception*, 12(3):353–364, 1995.

[22] M. Besson, F. Faita, and J. Requin. Brain waves associated with musical incongruities differ for musicians and non-musicians. *Neuroscience letters*, 168(1-2):101–105, 1994.

[23] R. Bod. Memory-based models of melodic analysis: Challenging the Gestalt principles. *Journal of New Music Research (JNMR)*, 30(3), 2001.

[24] J. Bor, S. Rao, W. van der Meer, and J. Harvey. *The Raga Guide: A survey of 74 Hindustani ragas*. Nimbus Records with Rotterdam Conservatory of Music, 1999.

[25] B. Bozkurt. An automatic pitch analysis method for Turkish Maqam music. *Journal of New Music Research (JNMR)*, 37(1):1–13, 2008.

[26] B. Bozkurt, M. K. Karaosmanoglu, B. Karacali, and E. Unal. Usul and makam driven automatic melodic segmentation for turkish music. *Journal of New Music Research*, 43(4):375–389, 2014.

[27] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organisation of Sound*. MIT Press, second edition, 1999.

[28] E. M. Burns and W. D. Ward. Categorical perception – phenomenon or epiphenomenon: evidence from experiments in the perception of melodic musical intervals. *Journal of Acoustic Society of America (JASA)*, 63(2):456–468, Feb. 1978.

[29] J. J. Cabrera, J. M. Diaz-Banez, F. J. E. Borrego, E. Gomez, F. G. Martin, and J. Mora. Comparative melodic analysis of a cappella Flamenco cantes. In *Fourth Conference on Interdisciplinary Musicology (CIM)*, 2008. Thessaloniki, Greece.

[30] M. A. Castellano, J. J. Bharucha, and C. L. Krumhansl. Tonal hierarchies in the music of north India. *Journal of Experimental Psychology: General*, 113(3):394, 1984.

[31] S. H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.

[32] W. Chai. Melody retrieval on the web. Master's thesis, MIT, Sept. 2001.

[33] A. Chakrabarty. *Shrutinandan: Towards Universal Music*. Macmillan, 2002.

[34] A. Chan. An analysis of pairwise sequence alignment algorithm complexities. Technical report, Stanford University, 2004.

[35] P. Chordia and A. Rae. Raag recognition using pitch-class and pitch-class dyad distributions. In *Proc. of Int. Soc. for Music Information Retrieval Conf.*, pages 431–436, 2007.

[36] P. Chordia and S. Senturk. Joint recognition of raag and tonic in north Indian music. *Computer Music Journal*, 37(3):82–98, 2013.

[37] A. Danielou. *The ragas of Northern Indian music*. Munshiram Manoharlal Publishers, 2010.

[38] R. B. Dannenberg and N. Hu. Pattern discovery techniques for music audio. *Journal of New Music Research (JNMR)*, 32(2), 2002.

[39] A. K. Datta, R. Sengupta, N. Dey, and D. Nag. *Experimental analysis of Shrutis from performances in Hindustani music*. Scientific Research Department, ITC Sangeet Research Academy, 2006.

[40] A. K. Datta, R. Sengupta, N. Dey, and D. Nag. A methodology for automatic extraction of 'meend' from the performances in Hindustani vocal music. *Journal of ITC Sangeet Research Academy*, 21:24–31, 2007.

[41] A. K. Datta, R. Sengupta, N. Dey, and D. Nag. Automatic classification of 'meend' extracted from the performances in Hindustani vocal music. In *Proc. of Frontiers of Research on Speech and Music (FRSM)*, 2008.

[42] A. K. Dey. *Nyāsa in Rāga: The Pleasant Pause in Hindustani Music*. Kanishka Publishers, 2008.

[43] P. Dighe, H. Karnick, and B. Raj. Swara histogram based structural analysis and identification of Indian classical ragas. In *Proc. of Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 35–40, 2013.

[44] W. J. Dowling. Scale and Contour: Two components of a theory of memory for melodies. *Psychological Review*, 85(4):341–354, 1978.

[45] S. Dutta and H. A. Murthy. Discovering typical motifs of a raga from one-liners of songs in Carnatic music. In *Int. Soc. for Music Information Retrieval (ISMIR)*, pages 397–402, Taipei, Taiwan, 2014.

[46] S. Dutta and H. A. Murthy. A modified rough longest common subsequence algorithm for motif spotting in an alapana of Carnatic music. In *Twentieth National Conference on Communications (NCC)*, pages 1–6, 2014.

[47] S. Dutta and H. A. Murthy. Raga verification in carnatic music using longest common segment set. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, pages 605–611. Malaga, Spain, 2015.

[48] T. Eerola and M. Bregman. Melodic and contextual similarity of folk song phrases. *Musicae Scientiae*, 11(1_suppl):211–233, 2007.

[49] T. Eerola, T. Jaarvinen, J. Louhivuori, and P. Toiviainen. Statistical features and perceived similarity of folk melodies. *Music Perception: An Interdisciplinary Journal*, 18(3):275–296, 2001.

[50] T. Eerola and P. Toiviainen. MIR in Matlab: The MIDI Toolbox. In *Proc. of Int. Soc. for Music Information Retrieval (ISMIR)*, 2004.

[51] J. P. Egan. *Signal Detection Theory and ROC Analysis Academic Press Series in Cognition and Perception*. London, UK: Academic Press, 1975.

[52] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006.

[53] H. E. Fiske. Categorical perception of musical patterns: How different is 'different'. *Bulletin of the Council for Research in Music Education*, (133):20–24, 1997.

[54] T. Fujishima. Realtime chord recognition of musical sound: a system using common Lisp music. In *Proc. of International Computer Music Conference (ICMC)*, pages 464–467, 1999. Beijing. International Computer Music Association.

[55] K. K. Ganguli. How do we 'See' & 'Say' a raga: A Perspective Canvas. *Samakalika Sangeetham*, 4(2):112–119, Oct. 2013.

[56] K. K. Ganguli, S. Gulati, X. Serra, and P. Rao. Data-driven exploration of melodic structures in Hindustani music. In *Proc. of the International Society for Music Information Retrieval (ISMIR)*, pages 605–611, Aug. 2016. New York, USA.

[57] K. K. Ganguli, A. Lele, S. Pinjani, P. Rao, A. Srinivasamurthy, and S. Gulati. Melodic shape stylization for robust and efficient motif detection in hindustani vocal music. In *National Conference on Communications (NCC)*, 2017.

[58] K. K. Ganguli and P. Rao. Tempo dependence of melodic shapes in Hindustani classical music. In *Proc. of Frontiers of Research on Speech and Music (FRSM)*, pages 91–95, Mar. 2014. Mysore, India.

[59] K. K. Ganguli and P. Rao. Discrimination of melodic patterns in Indian classical music. In *Proc. of National Conference on Communications (NCC)*, Feb. 2015.

[60] K. K. Ganguli and P. Rao. Exploring melodic similarity in Hindustani classical music through the synthetic manipulation of raga phrases. In *Proc. of the Cognitively-based Music Informatics Research Workshop*, Aug. 2016.

[61] K. K. Ganguli and P. Rao. Perceptual anchor or attractor: How do musicians perceive raga phrases? In *Proc. of Frontiers of Research on Speech and Music (FRSM)*, Nov. 2016. Baripada, India.

[62] K. K. Ganguli and P. Rao. Imitate or recall: How do musicians perform raga phrases? In *Proc. of Frontiers of Research on Speech and Music (FRSM)*, Dec. 2017. Rourkela, India.

[63] K. K. Ganguli and P. Rao. Towards computational modeling of the ungrammatical in a raga performance. In *Proc. of the International Society for Music Information Retrieval (ISMIR)*, Oct. 2017. Suzhou, China.

[64] K. K. Ganguli and P. Rao. Validating stock musicological knowledge via audio analyses of contemporary raga performance. In *20th Quinquennial Congress of the International Musicological Society (IMS): Digital Musicology Study Session*, Mar. 2017. Tokyo, Japan.

[65] K. K. Ganguli and P. Rao. On the distributional representation of ragas: experiments with allied raga pairs. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 1(1):79–95, 2018.

[66] K. K. Ganguli and P. Rao. On the perception of raga motifs by trained musicians. *The Journal of the Acoustical Society of America (JASA)*, 145(4):2418–2434, 2019.

[67] K. K. Ganguli and P. Rao. A parametric approach to the structural representation of melody in Hindustani raga music. *Journal of New Music Research (JNMR)*, 2019. In preparation.

[68] K. K. Ganguli, A. Rastogi, V. Pandit, P. Kantan, and P. Rao. Efficient melodic query based audio search for Hindustani vocal compositions. In *Proc. of the International Society for Music Information Retrieval (ISMIR)*, pages 591–597, Oct. 2015. Malaga, Spain.

[69] C. Gaser and G. Schlaug. Brain structures differ between musicians and non-musicians. *The Journal of Neuroscience*, 23(27):9240–9245, 2003.

[70] A. C. Gedik and B. Bozkurt. Pitch-frequency histogram-based music information retrieval for Turkish music. *Signal Processing*, 90(4):1049–1063, 2010.

[71] P. K. Ghosh and S. S. Narayanan. Pitch contour stylization using an optimal piecewise polynomial approximation. *IEEE Signal Processing Letters*, 16(9), Sept. 2009.

[72] G. Giguere. Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using spss. *Tutorials in Quantitative Methods for Psychology*, 2(1):27–38, 2006.

[73] R. L. Goldstone and A. T. Hendrickson. Categorical perception. *Cognitive Science: Wiley Interdisciplinary Reviews*, 1(1):69–78, Feb. 2010.

[74] C. Gomez, S. Abad-Mota, and E. Ruckhaus. An analysis of the Mongeau-Sankoff algorithm for Music Information Retrieval. In *Proc. of Int. Soc. for Music Information Retrieval (ISMIR)*, pages 109–110, 2007.

[75] E. Gomez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, 18(3):294–304, 2006.

[76] O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.

[77] S. Gulati, A. Bellur, J. Salamon, H. G. Ranjani, V. Ishwar, H. A. Murthy, and X. Serra. Automatic tonic identification in Indian art music: Approaches and Evaluation. *Journal of New Music Research (JNMR)*, 43(1):53–71, 2014.

[78] S. Gulati, K. K. Ganguli, S. Gupta, A. Srinivasamurthy, and X. Serra. Ragawise: A lightweight real-time raga recognition system for Indian art music. In *Proc. of the International Society for Music Information Retrieval (ISMIR)*, Oct. 2015. Malaga, Spain.

[79] S. Gulati, J. Serra, K. K. Ganguli, S. Senturk, and X. Serra. Time-delayed melody surfaces for raga recognition. In *Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pages 751–757, Aug. 2016. New York, USA.

[80] S. Gulati, J. Serra, K. K. Ganguli, and X. Serra. Landmark detection in Hindustani music melodies. In *Proc. of Int. Computer Music, Sound and Music Computing*, pages 1062–1068, 2014. Athens, Greece.

[81] S. Gulati, J. Serra, V. Ishwar, S. Senturk, and X. Serra. Phrase-based raga recognition using vector space modeling. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 66–70, 2016.

[82] S. Gulati, J. Serra, V. Ishwar, and X. Serra. Mining melodic patterns in large audio collections of Indian art music. In *Int. Conf. on Signal Image Technology & Internet Based Systems (SITIS-MIRA)*, pages 264–271, 2014.

[83] S. Gulati, J. Serra, and X. Serra. Improving melodic similarity in Indian art music using culture-specific melodic characteristics. In *Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 680–686, 2015.

[84] Z. Guo, Q. Wang, G. Liu, J. Guo, and Y. Lu. A music retrieval system using melody and lyric. In *Proc. of IEEE Int. Conf. on Multimedia & Expo*, 2012.

[85] C. Gupta and P. Rao. *Speech, Sound and Music Processing: Embracing Research in India; CMMR 2011-FRSM 2011*, chapter Objective Assessment of Ornamentation in Indian Classical Singing, pages 1–25. Springer Berlin Heidelberg, 2012.

[86] S. Harnad. *Categorical Perception*, chapter Encyclopedia of Cognitive Science. Nature Publishing Group/Macmillan, 2003.

[87] S. Harnad, S. J. Hanson, and J. Lubin. Categorical perception and the evolution of supervised learning in neural nets. In *Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology*, pages 65–74, 1991.

[88] N. Hu, R. B. Dannenberg, and A. L. Lewis. A probabilistic model of melodic similarity. 2002.

[89] D. Huron. The melodic arch in Western Folksongs. *Computing in Musicology*, 10:3–23, 1995.

[90] D. Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, 2006.

[91] V. Ishwar, S. Dutta, A. Bellur, and H. Murthy. Motif spotting in an Alapana in Carnatic music. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, pages 499–504, 2013.

[92] ITC-SRA. Distinguishing between similar ragas. url: http://www.itcsra.org/Distinguishing-between-Similar-Ragas, 2017. Last accessed: April 26, 2017.

[93] N. A. Jairazbhoy. *The Rags of North Indian Music: Their Structure & Evolution*. Popular Prakashan, second edition, 2011.

[94] Z. Juhasz. A systematic comparison of different european folk music traditions using self-organizing maps. *Journal of New Music Research*, 35(2):95–112, 2006.

[95] G. K. Koduri, S. Gulati, and P. Rao. A survey of raaga recognition techniques and improvements to the state-of-the-art. *Sound and Music Computing*, 2011.

[96] G. K. Koduri, S. Gulati, P. Rao, and X. Serra. Raga recognition based on pitch distribution methods. *Journal of New Music Research*, 41(4):337–350, 2012.

[97] G. K. Koduri, V. Ishwar, J. Serra, and X. Serra. Intonation analysis of ragas in Carnatic music. *Journal of New Music Research*, 43(1):72–93, 2014.

[98] N. Kroher, E. Gomez, C. Guastavino, F. Gomez-Martin, and J. Bonada. Computational models for perceived melodic similarity in a cappella flamenco cantes. In *Proc. of Int. Soc. for Music Information Retrieval (ISMIR)*, Taipei, Taiwan, Oct. 2014.

[99] C. L. Krumhansl. *Cognitive Foundations of Musical Pitch*, chapter 4: A key-finding algorithm based on tonal hierarchies, pages 77 – 110. Oxford University Press, New York, 1990.

[100] C. L. Krumhansl and L. L. Cuddy. *A Theory of Tonal Hierarchies in Music*, pages 51–87. Springer New York, New York, NY, 2010.

[101] C. L. Krumhansl and E. J. Kessler. Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological review*, 89(4):334–368, 1982.

[102] S. Kulkarni. *Shyamrao Gharana*, volume 1. Prism Books Pvt. Ltd., 2011.

[103] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[104] V. Kumar, H. Pandya, and C. V. Jawahar. Identifying ragas in Indian music. In *International Conference on Pattern Recognition*, pages 767–772, 2014.

[105] A. Lele, S. Pinjani, K. K. Ganguli, and P. Rao. Improved melodic sequence matching for query based searching in Indian classical music. In *Proc. of Frontiers of Research on Speech and Music (FRSM)*, Nov. 2016. Baripada, India.

[106] D. J. Levitin. *Music, Cognition and Computerized Sound: An Introduction to Psychoacoustics*, chapter 23: Experimental Design in Psychoacoustic Research. MIT Press, 1999.

[107] D. J. Levitin. *Foundations of Cognitive Psychology*, chapter 13: Memory for Musical Attributes, pages 295–310. 2002.

[108] D. J. Levitin. *Foundations of Cognitive Psychology: Core Readings*. MIT Press, 2002.

[109] D. J. Levitin. *This is your Brain on Music: The Science of a Human Obsession*. Dutton, a member of Penguin Group (USA) Inc., 2006.

[110] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368, Nov. 1957.

[111] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proc. of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, New York, USA, 2003.

[112] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time-series with implications for streaming algorithms. In *Proc. of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2003.

[113] A. Mahajan. *Ragas in Hindustani Music: Conceptual Aspects*. Gyan Publishing House, 2010.

[114] A. Marsden. A study of cognitive demands in listening to Mozart's Quintet for piano and wind instruments, K.452. *Psychology of Music*, 15:30–57, 1987.

[115] A. Marsden. Representing melodic patterns as networks of elaborations. *Computers and the Humanities*, 35:37–54, 2001.

[116] A. Marsden. Interrogating melodic similarity: A definitive phenomenon or the product of interpretation? *Journal of New Music Research (JNMR)*, 41(4):323–335, 2012.

[117] G. McGuire. A brief primer on experimental designs for speech perception research. Online "`http://people.ucsc.edu/~gmcguir1/experiment_designs.pdf`", 2010. Methods in Speech Perception; last accessed: September 5, 2015.

[118] B. McMurray, J. L. Dennhardt, and A. Struck-Marcell. Context effects on musical chord categorization: Different forms of top-down feedback in speech and music? *Cognitive science*, 32(5):893–920, 2008.

[119] M. Mongeau and D. Sankoff. Comparison of musical sequences. *Computers and the Humanities*, 1990.

[120] G. S. Morrison and M. V. Kondaurova. Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. *The Journal of the Acoustical Society of America*, 126(5):2159–2162, 2009.

[121] D. Mullensiefen and K. Frieler. Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. *Computing in Musicology*, 13, 2004. Special edition on Music Query: Methods, Models, and User Studies.

[122] D. Mullensiefen and K. Frieler. Measuring melodic similariry: Human vs. algorithmic judgments. In *Proc. of Interdisciplinary Musicology*, 2004.

[123] D. Mullensiefen and K. Frieler. Optimizing measures of melodic similarity for the exploration of a large folk song database. In *Proc. of Int. Soc. for Music Information Retrieval (ISMIR)*, 2004.

[124] D. Mullensiefen and K. Frieler. Modelling experts' notions of melodic similarity. *Musicae Scientiae*, 2007.

[125] D. Mullensiefen and C. Hennig. *From Data and Information Analysis to Knowledge Engineering*, chapter Modeling Memory for Melodies, pages 732–739. 2006.

[126] D. Mullensiefen, G. Wiggins, and D. Lewis. High-level feature descriptors and corpus-based musicology: Techniques for modelling music cognition, 2008.

[127] M. Muller. *Information Retrieval for Music and Motion, Chapter 4: Dynamic Time Warping*, pages 69–84.

[128] M. Muller, N. Jiang, and P. Grosche. A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing. *IEEE Trans. on Audio, Speech & Language Processing*, 21(3):531–543, 2013.

[129] B. Nettl. Thoughts on improvisation: A comparative approach. *The Musical Quarterly*, 60(1):1–19, 1974.

[130] L. Nooshin and R. Widdess. Improvisation in Iranian and Indian music. *Journal of the Indian Musicological Society*, 36:104–119, 2006.

[131] A. Novello, M. F. McKinney, and A. Kohlrausch. Perceptual evaluation of music similarity. In *Proc. of Int. Soc. for Music Information Retrieval (ISMIR)*, 2006.

[132] A. Novello, M. M. McKinney, and A. Kohlrausch. Perceptual evaluation of inter-song similarity in western popular music. *Journal of New Music Research*, 40(1):1–26, 2011.

[133] V. Oak. 22 shruti. url: http://22shruti.com/, 2017. Lat accessed: April 26, 2017.

[134] G. Padmasundari and H. A. Murthy. Raga identification using locality sensitive hashing. In *Proc. of the 23rd National Conference on Communications (NCC)*, pages 1–6. IEEE, 2017.

[135] B. Pajak, P. Piccinini, and R. Levy. Perceptual warping of phonetic space applies beyond known phonetic categories: evidence from the perceptual magnet effect. *The Journal of the Acoustical Society of America*, 136(4):2107–2107, 2014.

[136] B. Pardo, J. Shifrin, and W. Birmingham. Name that tune: A pilot study in finding a melody from a sung query. *Journal of the Association for Information Science and Technology*, 55(4):283–300, 2004.

[137] A. D. Patel. *Music, Language, and the Brain*. New York: Oxford University Press, 2008.

[138] M. Pearce, D. Mullensiefen, and G. Wiggins. Melodic grouping in Music Information Retrieval: New methods and applications. In *Advances in Music Information Retrieval*, volume 274 of *Studies in Computational Intelligence*, pages 364–388. Springer, 2010.

[139] G. Peeters. Chroma-based estimation of musical key from audio-signal analysis. In *Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 115–120, 2006.

[140] J. B. Pierrehumbert and S. A. Steele. Categories of tonal alignment in english. *Phonetica*, 46(4):181–196, 1989.

[141] H. S. Powers and R. Widdess. *India, subcontinent of*, chapter III: Theory and practice of classical music. New Grove Dictionary of Music. Macmillan, London, 2nd edition, 2001. Contributions to S. Sadie (ed.).

[142] J. B. Prince. Contributions of pitch contour, tonality, rhythm, and meter to melodic similarity. *Journal of Experimental Psychology: Human Perception and Performance*, 40(6), 2014.

[143] L. Rabiner. Fundamentals of speech recognition. *Fundamentals of speech recognition*, 1993.

[144] D. Raja. *Hindustani Music: A Tradition in Transition*. D. K. Printworld, 2005.

[145] D. Raja. *The Raga-ness of Ragas: Ragas Beyond the Grammar*. D.K. Print World Ltd., 2016.

[146] R. Raman and W. J. Dowling. Real-time probing of modulations in south Indian classical (carnātic) music by Indian and western musicians. *Music Perception: An Interdisciplinary Journal*, 33(3):367–393, 2016.

[147] H. G. Ranjani, D. Paramashivan, and T. V. Sreenivas. Quantized melodic contours in indian art music perception: Application to transcription. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, pages 174–180, 2017.

[148] H. G. Ranjani, A. Srinivasamurthy, D. Paramashivan, and T. V. Sreenivas. A compact pitch and time representation for melodic contours in indian art music. *The Journal of the Acoustical Society of America*, 145(1):597–603, 2019.

[149] P. Rao and K. K. Ganguli. Linking prototypical, stock knowledge with the creative musicianship displayed in raga performance. In *Proc. of Frontiers of Research on Speech and Music (FRSM)*, Dec. 2017. Rourkela, India.

[150] P. Rao, J. C. Ross, and K. K. Ganguli. Distinguishing raga-specific intonation of phrases with audio analysis. *Ninaad*, 26-27(1):59–68, Dec. 2013.

[151] P. Rao, J. C. Ross, K. K. Ganguli, V. Pandit, V. Ishwar, A. Bellur, and H. A. Murthy. Classification of melodic motifs in raga music with time-series matching. *Journal of New Music Research (JNMR)*, 43(1):115–131, Apr. 2014.

[152] S. Rao, J. Bor, W. van der Meer, and J. Harvey. *The Raga Guide: A Survey of 74 Hindustani Ragas*. Nimbus Records with Rotterdam Conservatory of Music, 1999.

[153] S. Rao and P. Rao. An overview of Hindustani music in the context of Computational Musicology. *Journal of New Music Research (JNMR)*, 43(1), Apr. 2014.

[154] V. Rao and P. Rao. Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *IEEE Trans. on Audio, Speech & Language Processing*, 18(8), 2010.

[155] L. Redi. Categorical effects in production of pitch contours in english. In *Proceedings of the 15th International Congress of the Phonetic Sciences*, pages 2921–2924, 2003.

[156] R. C. Repetto, R. Gong, N. Kroher, and X. Serra. Comparison of the singing style of two jingju schools. In *International Society for Music Information Retrieval (ISMIR)*, pages 507–513, Malaga, Spain, Oct. 2015.

[157] J. Rodd and A. Chen. Pitch accents show a perceptual magnet effect: Evidence of internal structure in intonation categories. In *Speech Prosody 2016*, pages 697–701, 2016.

[158] J. G. Roederer. *The Physics and Psychophysics of Music: An Introduction*. Springer, fourth edition, 2008.

[159] J. C. Ross, A. Mishra, K. K. Ganguli, P. Bhattacharyya, and P. Rao. Identifying raga similarity through embeddings learned from compositions' notation. In *Proc. of the International Society for Music Information Retrieval (ISMIR)*, Oct. 2017. Suzhou, China.

[160] J. C. Ross and P. Rao. Detection of raga-characteristic phrases from Hindustani classical music audio. In *Proc. of 2nd CompMusic Workshop*, pages 133–138, 2012.

[161] J. C. Ross, T. P. Vinutha, and P. Rao. Detecting melodic motifs from audio for Hindustani classical music. In *Proc. of Int. Soc. for Music Information Retrieval (ISMIR)*, Oct. 2012.

[162] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[163] S. Sankaran, K. Sekhar, and H. A. Murthy. Automatic segmentation of composition in carnatic music using time-frequency cfcc templates. In *Proc. of 11th Int. Symp. on Computer Music Multidisciplinary Research (CMMR)*, 2015.

[164] G. P. Scavone, S. Lakatos, and C. R. Harbke. The sonic mapper: An interactive program for obtaining similarity ratings with auditory stimuli. 2002.

[165] M. A. Schmuckler. Melodic contour similarity using folk melodies. *Music Perception: An Interdisciplinary Journal*, 28(2):169–194, 2010.

[166] K. Schneider. The german boundary tones: Categorical perception, perceptual magnets, and the perceptual reference space. 2012. Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart, Germany.

[167] K. Schneider, G. Dogil, and B. Mobius. German boundary tones show categorical perception and a perceptual magnet effect when presented in different contexts. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[168] S. K. Scott and S. Evans. Categorizing speech. *Nature Neuroscience*, 13:1304–1306, 2010.

[169] K. Sekhar, V. S. Viraraghavan, S. Sankaran, and H. A. Murthy. An approach to transcription of varnams in carnatic music using hidden markov models. In *2017 Twenty-third National Conference on Communications (NCC)*, pages 1–6. IEEE, 2017.

[170] X. Serra. A multicultural approach to music information research. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, pages 151–156, 2011.

[171] X. Serra. Creating research corpora for the computational study of music: the case of the Compmusic project. In *Proc. of the 53rd AES Int. Conf. on Semantic Audio*, London, 2014.

[172] J. A. Siegel and W. Siegel. Categorical perception of tonal intervals: Musicians can't tell sharp from flat. *Perception & Psychophysics*, 21(5):399–407, 1977.

[173] N. A. Smith and M. A. Schmuckler. Pitch-distributional effects on the perception of tonality. In *Proc. of the International Conference on Music Perception and Cognition (ICMPC)*, pages 5–10, 2000.

[174] N. A. Smith and M. A. Schmuckler. The perception of tonal structure through the differentiation and organization of pitches. *Journal of Experimental Psychology: Human Perception and Performance*, 30(2):268, 2004.

[175] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[176] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra. Corpora for music information research in Indian art music. In *Proc. of Int. Computer Music Conf. / Sound and Music Computing Conf.*, Sept. 2014.

[177] A. Srinivasamurthy and X. Serra. A supervised approach to hierarchical metrical cycle tracking from audio music recordings. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5217–5221, 2014.

[178] SwarGanga. Swarganga music foundation. url: https://www.swarganga.org/, 2017. Last accessed: April 26, 2017.

[179] Y. Tanaka, K. Iwamoto, and K.Uehara. Discovery of time-series motif from multi-dimensional data based on MDL principle. *Machine Learning*, 58:269–300, 2005.

[180] Y. Tanaka and K. Uehara. Discover motifs in multi-dimensional time-series using the Principal Component Analysis and the MDL principle. In *Proc. of Int. Conf. on Machine Learning & Data Mining in Pattern Recognition*, pages 252–265, 2003.

[181] D. Temperley. What's key for key? the krumhansl-schmuckler key-finding algorithm reconsidered. *Music Perception: An Interdisciplinary Journal*, 17(1):65–100, 1999.

[182] D. Temperley and E. W. Marvin. Pitch-class distribution and the identification of key. *Music Perception: An Interdisciplinary Journal*, 25(3):193–212, 2008.

[183] M. Tervaniemi, M. Huotllainen, E. Bratiico, R. Ilmoniemi, K. Reinlkainen, and K. Alho. Event-related potentials to expectancy violation in musical context. *Musicae Scientiae*, 7(2):241–261, 2003.

[184] M. Tervaniemi, T. Ilvonen, K. Karma, K. Alho, and R. Naatanen. The musical brain: brain waves reveal the neurophysiological basis of musicality in human subjects. *Neuroscience letters*, 226(1):1–4, 1997.

[185] B. Tillman. Review of: Music, Language, and the Brain by Aniruddh D. Patel. *Psychomusicology: Music, Mind & Brain*, 20(1 & 2):180–185, 2009.

[186] P. Tormene, T. Giorgino, S. Quaglini, and M. Stefanelli. Matching incomplete time-series with Dynamic Time Warping: An algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine*, 45(1):11–34, 2008.

[187] R. Typke, M. Den H., J. De Nooijer, F. Wiering, and R. C. Veltkamp. A ground truth for half a million musical incipits. *Journal of Digital Information Management*, pages 34–39, 2005.

[188] R. Typke, F. Wiering, and R. C. Veltkamp. Transportation distances and human perception of melodic similarity. *Musicae Scientiae*, 11(1_suppl):153–181, 2007.

[189] A. Uitdenbogerd and J. Zobel. Melodic matching techniques for large music databases. In *Proc. of ACM Int. Conf. on Multimedia*, pages 57–66, 1999.

[190] A. Vahdatpour, N. Amini, and M. Sarrafzadeh. Towards unsupervised activity discovery using multi-dimensional motif detection in time-series. In *Proc. Of Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2009.

[191] W. van der Meer. *Hindustani Music in the 20th Century*. Martinus Nijhoff Publishers, 1980.

[192] W. van der Meer. Improvisation versus reproduction, india and the world. *New Sound: International Magazine for Music*, (32), 2008.

[193] P. van Kranenburg. *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*. PhD thesis, Oct. 2010.

[194] P. van Kranenburg, A. Volk, and F. Wiering. A comparison between global and local features for computational classification of folk song melodies. *Journal of New Music Research (JNMR)*, 42(1):1–18, 2013.

[195] N. N. Vempala and F. A. Russo. A melodic similarity measure based on human similarity judgments. In *Proc. of the Int. Conf. on Music Perception and Cognition*, July 2012.

[196] N. N. Vempala and F. A. Russo. An empirically derived measure of melodic similarity. *Journal of New Music Research*, 44(4):391–404, 2015.

[197] P. Verma, T. P. Vinutha, P. Pandit, and P. Rao. Structural segmentation of Hindustani concert audio with posterior features. In *Int.Conf. on Acoustics Speech and Signal Processing (ICASSP)*, pages 136–140. IEEE, 2015.

[198] A. Vidwans, K. K. Ganguli, and P. Rao. Classification of Indian classical vocal styles from melodic contours. In *Proc. of the 2nd CompMusic Workshop*, July 2012.

[199] K. G. Vijaykrishnan. *The Grammar of Carnatic Music*. De Gruyter Mouton, 2007.

[200] V. S. Viraraghavan, R. Aravind, and H. A. Murthy. A statistical analysis of gamakas in carnatic music. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, pages 243–249, 2017.

[201] V. S. Viraraghavan, R. Aravind, and H. A. Murthy. Precision of sung notes in carnatic music. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, pages 499–505, 2018.

[202] A. Volk and P. van Kranenburg. Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, 16(3):317–339, 2012.

[203] A. Volk, P. van Kranenburg, J. Garbers, F. Wiering, R. C. Veltkamp, and L. P. Grijp. A manual annotation method for melodic similarity and the study of melody feature sets. In *ISMIR*, pages 101–106, 2008.

[204] B. L. Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[205] T. D. Wickens. *Elementary signal detection theory*. Oxford University Press, USA, 2002.

[206] R. Widdess. Involving the performers in transcription and analysis: a collaborative approach to Dhrupad. *Ethnomusicology*, 38(1):59–79, 1994.

[207] R. Widdess. Dynamics of melodic discourse in Indian music: Budhaditya mukherjee's alap in rag puriya-kalyan. pages 187–224, 2011. Oxford University Press.

[208] R. Widdess. Schemas and improvisation in Indian music. *Language, Music and Interaction. London: College Publications*, 2013.

Selected Publications

Journal articles

- K. K. Ganguli and P. Rao. "On the perception of raga motifs by trained musicians," Journal of the Acoustical Society of America (JASA), 145(4): 2418-2434, 2019.
- K. K. Ganguli and P. Rao. "On the distributional representation of raga and melody," Transactions of International Society for Music Information Retrieval (TISMIR), 1(1): 79–95, 2018.
- K. K. Ganguli and P. Rao. "A parametric approach to the structural representation of melody in Hindustani raga music," manuscript for JNMR in preparation.
- A. Srinivasamurthy, A. Holzapfel, K. K. Ganguli, and X. Serra. "Aspects of tempo and rhythmic elaboration in Hindustani music: A corpus study," Frontiers in Digital Humanities (Digital Musicology), 4 (20), 2017.
- P. Rao, J. C. Ross, K. K. Ganguli, V. Pandit, V. Ishwar, A. Bellur, and H. Murthy. "Classification of melodic motifs in raga music with time-series matching," Journal of New Music Research (JNMR), 43(1), 2014.
- P. Rao, J. C. Ross, and K. K. Ganguli. "Distinguishing raga-specific intonation of phrases with audio analysis," Ninaad (Journal of ITC SRA), 26-27(1), 2013.

Conference papers

- K. K. Ganguli and P. Rao. "Towards computational modeling of the ungrammatical in a raga performance," in Proc. of the 18th International Society for Music Information Retrieval Conference (ISMIR), October 2017, Suzhou, China.
- K. K. Ganguli, S. Gulati, X. Serra, and P. Rao. "Data-driven exploration of melodic structures in Hindustani music," Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR), August 2016, New York, USA.
- K. K. Ganguli, A. Rastogi, V. Pandit, P. Kantan, and P. Rao. "Efficient melodic query based audio search for Hindustani vocal compositions," in Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR), October 2015, Malaga, Spain.
- K. K. Ganguli, A. Lele, S. Pinjani, P. Rao, A. Srinivasamurthy, and S. Gulati. "Melodic shape stylization for robust and efficient motif detection in Hindustani vocal music," in Proc. of the National Conference for Communications (NCC), March 2017, Chennai, India.

- K. K. Ganguli and P. Rao. "Imitate or recall: How do musicians perform raga phrases?" in Proc. of the Frontiers of Research on Speech and Music (FRSM), December 2017, Rourkela, India.

- K. K. Ganguli and P. Rao. "Perceptual anchor or attractor: How do musicians perceive raga phrases?" in Proc. of the Frontiers of Research on Speech and Music (FRSM), November 2016, Baripada, India.

- K. K. Ganguli and P. Rao. "Discrimination of Melodic Patterns in Indian Classical Music," in Proc. of the National Conference for Communications (NCC), February 2015, Mumbai, India.

Workshop and invited talks

- P. Rao and K. K. Ganguli. "Linking prototypical, stock knowledge with the creative musicianship displayed in raga performance," Invited talk at the Frontiers of Research on Speech and Music (FRSM), December 2017, Rourkela, India.

- K. K. Ganguli and P.Rao. "Validating stock musicological knowledge via audio analyses of contemporary raga performance," Invited talk at the 20th Quinquennial Congress of the International Musicological Society (IMS): Digital Musicology Study Session, March 2017, Tokyo, Japan.

- K. K. Ganguli and P. Rao. "Exploring melodic similarity in Hindustani classical music through the synthetic manipulation of raga phrases," Cognitively-based Music Informatics (CogMIR) workshop, August 2016, New York, USA.

Collaborative works

- S. Gulati, J. Serra, K. K. Ganguli, S. Senturk, and X. Serra. "Time-delayed melody surfaces for raga recognition," Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR), August 2016, New York, USA.

- A. Vidwans, K. K. Ganguli, and P.Rao. "Classification of Indian classical vocal styles from melodic contours," Proc. of the 2nd CompMusic workshop, 2012, Istanbul, Turkey.

- S. Gulati, J. Serra, K. K. Ganguli, and X. Serra. "Landmark Detection in Hindustani Music Melodies," Proc. of the Int. Computer Music Conf., Sound and Music Computing Conf., September 2014, pp. 1062–1068, Athens, Greece.

- J. C. Ross, A. Mishra, K. K. Ganguli, P. Bhattacharyya, and P. Rao. "Identifying Raga Similarity through Embeddings Learned from Compositions' Notation," in Proc. of the 18th International Society for Music Information Retrieval Conference (ISMIR), October 2017, Suzhou, China.

- T. P. Vinutha, S. Sankagiri, K. K. Ganguli, and P.Rao. "Structural segmentation and visualization of Sitar and Sarod concert audio," Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR), August 2016, New York, USA.

- J. C. Ross, R. Murthy, K. K. Ganguli, and P. Bhattacharyya. "Identifying Raga Similarity in Hindustani Classical Music through Distributed Representation of Raga Names," Proc. of the International Symposium on Computer Music Multidisciplinary Research (CMMR) 2017, Porto, Portugal.
- A. Lele, S. Pinjani, K. K. Ganguli, and P. Rao. "Improved Melodic Sequence Matching for Query Based Searching in Indian Classical Music," in Proc. of the Frontiers of Research on Speech and Music (FRSM), November 2016, Baripada, India.

Musicological articles

- K. K. Ganguli. "How do we 'see' and 'say' a raga: A perspective canvus," (invited article) Samakalika Sangeetham, 3(2), 2013.

Demonstrations

- S. Gulati, K. K. Ganguli, S. Gupta, A. Srinivasamurthy, and X. Serra. "Ragawise: A Lightweight Real-time Raga Recognition System for Indian Art Music," (late-breaking demo) in Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR), October 2015, Malaga, Spain.

Acknowledgments

There are so many people to thank for their sincere help, inspiration, company, and support for the last few years. The long journey of the PhD away from home for the first time – the journey being more enjoyable than the destination as I realize now – seems seamless with these great company. I shall try to cover as much without being my long winded self.

First and foremost, my sincere gratitude to thesis supervisor Prof. Preeti Rao for her guidance, encouragement, and patience over the tenure. Thank you so much for forcing me to think deeper, sometimes screaming, and most importantly bearing with me! You have been kind in accepting me as a research associate, keeping faith with my musical background (and a seemingly parallel career), and trusting my abilities to learn new things. Thank you for your tireless efforts on the journal publications; of course the conferences gave me the opportunity to network with peers and experts to broaden my horizon. Your support as a motherly figure and a Guru had an immense effect in the transformation of my current self. I am apologetic if I have been a difficult student to work with, or if I have unknowingly breached academic/professional ethics that hurt your sentiments as a guide.

I would like to thank my thesis committee (RPC): Prof. Bhaskaran Muralidharan, Prof. Azizuddin Khan, and Prof. Ajit Rajwade for their acceptance, appreciation, inspiration, and guidance. The annual progress seminars had really boosted my confidence and reassured the novelty and thoroughness of the work. My sincere gratitude to Prof. Pushpak Bhattacharyya, Prof. Soumyo Mukherji, Prof. Rajbabu, Prof. P. C. Pandey, and our HoD Prof. B. G. Fernandez for their kind gestures to make my stay at IIT Bombay a memorable one.

My humble regards to my Guruji Padmashri Pt. Ajoy Chakrabarty for laying the foundation of my musical journey. I owe all my musical concepts to him that has been the key to formulate relevant musical questions to address in the thesis. My sincere thanks to Prof. Xavier Serra for involving me in the CAMUT project, hosting me couple of times in his MTG research group in Barcelona (Spain), and entrusting my capabilities to let me arrange the all India tours for making contacts with musicians, institutions, content owners, and potential collaborators. Your role has been phenomenal in the global outreach I have been able to achieve so far.

I take great pleasure to thank my direct collaborators (and coauthors) Dr. Sankalp Gulati, Dr. Ajay Srinivasamurthy for their conceptual support, guidance in coding, and generous help during my stay in Europe; and interns Ashwin Lele, Saurabh Pinjani for being great friends and pursuing some of my desired yet unfinished

work. My endless thanks to labmates: Joe, Vinutha, Amruta, Shrikant, Hitesh, Ankita, Rohit, Nikhil, Sachin, Hirak, Kanhaiya, Nitya, Nataraj, Divyansh and many more whom I have to skip. Special mention to Dr. Tathagata Kar for being a great emotional support and confidant during my stressful times. Another special mention is a great friend Nandini Roy Choudhury who has not only collaborated in many annotations and listening experiments, but also gave insights from a humanities perspective.

Above all heavy-duty technical discourse, one great ventilation has been the music environment in and around the campus. My humble gratitude to all musician friends and seniors in Mumbai for giving me a chance to pursue my passion. Special thanks to the Roots club-members: Srivatsan, Deepak, Poornima, Mugdha, Soumi, Asawari, Ritvij, Digant for the unforgettable jamming sessions, organizing events and road-trips.

Last but not the least, my surrendering gratitude to my parents: Shri Partha Sarathi Ganguli and Dr. Subhra Ganguli for their unconditional love and support. Their quality time, generosity, sacrifice, and affection have been singular to make me feel safe, secure, and comfortable away from home.

Date: _____ Kaustuv Kanti Ganguli