

World Music Technology: Culturally Sensitive Strategies for Automatic Music Prediction

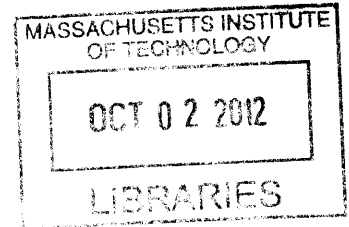
ARCHIVES

by

Mihir Sarkar

S.M. Media Technology, MIT (2007)

Diplôme d'Ingénieur ESIEA (1996)



Submitted to the Program in Media Arts and Sciences
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2012

© Massachusetts Institute of Technology 2012. All rights reserved.

Author
Mihir Sarkar
Program in Media Arts and Sciences
August 10, 2012

Certified by
Prof. Barry L. Vercoe
Professor Emeritus, Program in Media Arts and Sciences
Thesis Supervisor

Accepted by
Prof. Patricia Maes
Associate Academic Head, Program in Media Arts and Sciences

World Music Technology: Culturally Sensitive Strategies for Automatic Music Prediction

by

Mihir Sarkar

Submitted to the Program in Media Arts and Sciences
on August 10, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

Abstract

Music has been shown to form an essential part of the human experience—every known society engages in music. However, as universal as it may be, music has evolved into a variety of genres, peculiar to particular cultures. In fact people acquire musical skill, understanding, and appreciation specific to the music they have been exposed to. This process of enculturation builds mental structures that form the cognitive basis for musical expectation.

In this thesis I argue that in order for machines to perform musical tasks like humans do, in particular to predict music, they need to be subjected to a similar enculturation process by design. This work is grounded in an information theoretic framework that takes cultural context into account. I introduce a measure of *musical entropy* to analyze the predictability of musical events as a function of prior musical exposure. Then I discuss computational models for music representation that are informed by genre-specific containers for musical elements like notes. Finally I propose a software framework for automatic music prediction. The system extracts a lexicon of melodic, or timbral, and rhythmic primitives from audio, and generates a hierarchical grammar to represent the structure of a particular musical form. To improve prediction accuracy, context can be switched with *cultural plug-ins* that are designed for specific musical instruments and genres.

In listening experiments involving music synthesis a culture-specific design fares significantly better than a culture-agnostic one. Hence my findings support the importance of *computational enculturation* for automatic music prediction. Furthermore I suggest that in order to sustain and cultivate the diversity of musical traditions around the world it is indispensable that we design culturally sensitive music technology.

Thesis Supervisor: Prof. Barry L. Vercoe

Title: Professor Emeritus, Program in Media Arts and Sciences

**World Music Technology: Culturally Sensitive Strategies for
Automatic Music Prediction**


by

Mihir Sarkar

Thesis Committee

Advisor
Prof. Barry L. Vercoe
Professor Emeritus, Program in Media Arts and Sciences
Massachusetts Institute of Technology

Reader
Prof. Mitchel Resnick
LEGO Papert Professor of Learning Research
Academic Head, Program in Media Arts and Sciences
Massachusetts Institute of Technology

Reader . 
Prof. Christopher D. Chafe
Duca Family Professor
Director, Center for Computer Research in Music and Acoustics
Stanford University

Acknowledgments

Barry (Vercoe), thank you: for believing in me after our brief meeting in Bangalore, for welcoming me to the Media Lab, for your invaluable advice and unrelenting support (from near and far) and for the fun times—in Natick, Las Vegas, or Mumbai.

My other thesis committee members: Mitch (Resnick), for your help and support after Barry’s retirement from the Lab, for your trust and understanding, and for your feedback; Chris (Chafe), for your encouragements and for “being a fan of my work”—it means a lot to me—and for inviting me to spend time in Banff to brainstorm on the future of network music performance, as well as network and jam with other network jammers. (I missed the ski slopes though!).

Very special thanks to Sandy (Sener). You have been so helpful with each of my (often demanding) administrative requests, and taken care of every single detail during my time at the Lab. I will miss your kindness, and I wish you the very best.

My general exam committee members and master’s thesis committee members: Sandy (Pentland), Tod (Machover), DAn (Ellis), Miller (Puckette). Thanks for your insightful and probing questions, for making me explore new directions, and for helping me ask the right questions.

My incredible teachers at MIT, especially Mike (Bove), Roz (Picard), Patrick (Winston), and those before that, at ESIEA, and at La Source.

Fellow members and alums of the (now defunct) Music, Mind and Machine group: Wu-Hsi (Li), Anna (Huang), Owen (Meyers), Brian (Whitman), Judy (Brown). I thoroughly enjoyed our discussions, our demos, and our mango lassi fueled lunches!

My India Initiatives co-conspirators, in particular Andrea (Colaço), Rohit (Pandharkar), Joe (Paradiso) and the rest of the Design & Innovation workshop gang, Ramesh (Raskar), Paula (Anzer), Vinay (Gidwaney). Thanks for the wonderful times. I look forward to continuing on our path of striving to make a difference in the world.

Thank you, Joi (Ito), Frank (Moss), Joost (Bonsen) for supporting my initiatives, helping me make new connections, and sharing your insightful perspectives. And, more importantly, for enabling a Lab that has expanded my worldview and catered to my various interests: from computer music to technology for the developing world, and design-innovation-entrepreneurship...

Greg (Tucker), Kevin (Davis), Cornelle (King) from the facilities team for being awesome. And thanks so much, Greg, for helping fund my defense and my research in the last few months.

Thank you, Media Lab students, faculty, staff, NeCSys—too numerous to name here individually. You are an amazing community. It is hard to leave this place. But perhaps one never really leaves this place.

I would like to specially recognize Linda (Peterson) and her office who make things run against all odds!

All those who supported me in saving the 4th floor music studio from becoming yet another conference room (for a banking initiative on top of that)! ;-)

Special shout out to my \$100K partners in crime—Robin (Bose) and Sean (Leow).

Thank you to the MIT Legatum Center for Entrepreneurship and Development for supporting my final year of grad school.

Thanks to all my UROP and RSI students who helped me explore new grounds and make my projects a reality.

Outside the Lab, I would like to acknowledge Juan-Pablo (Caceres) for discussing and sharing ideas on the future of online musical collaboration. We haven't quite done enough together; I hope that will change. And Peter (Cariani) for being a great sounding board for bouncing ideas off of, especially on the perceptual-cognitive soundness of my approach. (Any faulty understanding is mine though.)

A mention to my first job fresh out of college in client-server software for banking and finance to help in making me realize that there was a better suited path for me that involved computers and music.

And, most importantly, my family: Amma and Baba, thank you for *everything*; my in-laws—Ma, Diai, Tom, Ishu, Asheema, Paro, Draffy—for your love and support, in intangible *and* tangible ways. Also my family in Europe, in India, and around the world, especially Jethumani, my late uncle, who would have been so proud, and Mammie; Pishi, my perimas; Dadamani and Mejda and my baudis and nephews who I haven't seen enough of in the past few years—I hope that will change now!

It is with great emotion that I dedicate this work to Anisha, Anjali, and Sharmila. To your love and affection, to the warmth and comfort you have brought to my life, to everything you have done and the sacrifices you have made to make this experience special for me. For your patience, for bearing with me through thick and thin, for egging me on when I was stuck. If this dissertation exists at all, it is thanks to you.

If I left your name out it is only on paper, not in spirit. Thank you to all those who have, in some way or another, directly or indirectly, contributed to this work.

*Some kind of sound called music appeals to everybody,
but has really great music a universal appeal?*

- Sri Aurobindo

Contents

1	World Music Technology	17
1.1	Musical Culture	17
1.2	Music Representation and Prediction	19
1.3	Motivation and Philosophical Thoughts	21
1.4	Scope and Definitions	23
1.5	Research Questions and Hypotheses	25
1.6	Contributions	27
1.7	Outline	29
2	Can Music Be Predicted?	31
2.1	An Information Theoretic Approach to Measuring Cultural Footprint in Music	31
2.2	Music as Information: Prior Art	33
2.2.1	Nature and Nurture in Musical Skill	33
2.2.2	Information Theory in the Context of Music	38
2.2.3	Entropy, Complexity, and Prediction	40
2.3	Musical Entropy	43
2.4	Musical Enculturation	48
2.5	Measuring Musical Culture	51
3	Culturally Inspired Models for Music Representation	53
3.1	Music Representation by People and Machines	53
3.2	Indian and Western Musical Traditions	58

3.2.1	Percussions	58
3.2.2	Winds	67
3.3	Culturally Appropriate Containers	75
3.4	Representation and Recognition of Musical Elements	76
3.5	Culturally Sensitive Design Principles	77
4	Automatic Music Prediction	79
4.1	Musical Expectation by People	79
4.2	Music Prediction by Machines	80
4.3	A Computational Model of Music Prediction	82
4.4	Machine Prediction and Human Expectation	86
4.4.1	Quantitative Evaluation	86
4.4.2	Qualitative Evaluation	87
4.5	Can Machines Predict Music?	89
5	Towards Culturally Sensitive Music Technology	93
5.1	From Music Representation to Music Prediction	93
5.2	Music Prediction in the Real World	94
5.3	Network Music Performance: a <i>Killer App</i> ?	96
5.4	Contributions	98
5.5	Future Work and Final Remarks	99
A	Glossary of Terms	101

List of Figures

1-1	Top-down and bottom-up processing of music	20
2-1	Depiction of melodic contour in Western music notation	46
2-2	Interval distribution for popular Indian music	49
2-3	Interval distribution for popular Western music (Kim et al., 2000) . .	49
2-4	Melody retrieval for 3-step, 5-step, and 7-step interval quantizations .	50
3-1	Tabla stroke waveform: na	59
3-2	Tabla stroke waveform: tin	60
3-3	Tabla stroke waveform: dha	60
3-4	Tabla stroke waveform: ga	61
3-5	Tabla stroke waveform: ka	61
3-6	Tabla stroke spectrogram: na	62
3-7	Tabla stroke spectrogram: tin	62
3-8	Tabla stroke spectrogram: dha	63
3-9	Tabla stroke spectrogram: ga	63
3-10	Tabla stroke spectrogram: ka	64
3-11	Tabla stroke recognition rate for varying FFT lengths	64
3-12	Tabla stroke recognition rate for varying dimensions	65
3-13	Tabla stroke recognition rate for varying k	65
3-14	Pitch tracking of note in Carnatic music: sa (ascending)	68
3-15	Pitch tracking of Carnatic music: ri (ascending)	68
3-16	Pitch tracking of Carnatic music: ga (ascending)	69
3-17	Pitch tracking of Carnatic music: pa (ascending)	69

3-18	Pitch tracking of Carnatic music: ni (ascending)	70
3-19	Pitch tracking of Carnatic music: sa' (ascending)	70
3-20	Pitch tracking of Carnatic music: sa' (descending)	71
3-21	Pitch tracking of Carnatic music: ni (descending)	71
3-22	Pitch tracking of Carnatic music: pa (descending)	72
3-23	Pitch tracking of Carnatic music: ga (descending)	72
3-24	Pitch tracking of Carnatic music: ri (descending)	73
3-25	Pitch tracking of Carnatic music: sa (descending)	73
3-26	Synthesis of Carnatic music: raga Hamsadvani ascending and descending	75
3-27	Synthesis-by-analysis via culture-specific containers for listening exper- iments	77
4-1	Automatic music prediction system block diagram	84
4-2	Tabla prediction error rate for <i>teental</i> , a 16 beat cycle	87
5-1	TablaNet: a real-time online musical collaboration system for Indian percussion (Sarkar and Vercoe, 2007)	96

List of Tables

1.1	Elements of music	19
2.1	Musical intervals	47
2.2	Musical entropy based on cultural priors	52
3.1	Confusion matrix for tabla stroke recognition	66
3.2	Swaras (note entities) in Carnatic music	74
4.1	Cultural plug-ins	86
4.2	Results of qualitative experiments on perceptual cultural fitness . . .	89

Chapter 1

World Music Technology

1.1 Musical Culture

Human beings are musical creatures. While every known society engages in music—whether as an art form, an accompaniment to dance, or a spiritual offering—each one of them has developed its own set of rules for music. Hence there exists a variety of musical genres, which have evolved by coming into contact with other musical traditions and new musical instruments.

In 1957 Max Mathews, then at Bell Labs, developed MUSIC, a computer program for music representation and synthesis, paving the way for computer music as a field of scientific and artistic inquiry. Computers are not simply tools for musicians; they are also composers, performers, and even listeners: they are musicians themselves.

To this day, however, computational models of musical elements have been biased towards Western music. They seldom account for the complexities of what is generally referred to as ‘World Music’ in the West (and which is actually a widely varied collection of distinct musical traditions that share little beside their ‘non-Western’ label). For instance MIDI (Musical Instrument Digital Interface) can only coarsely approxi-

mate the intricacies of *gamakas*, pitch contours that are an essential characteristic of musical notes in Carnatic (South Indian) music.

Current music technology, with its ‘one-size-fits-all’ approach, does not do justice to the variety of musical genres around the world. As with other forms of digital media, the creative affordances offered by digital music systems and tools are constrained by design decisions and engineering trade-offs. I would argue that the popularization of music technology that lacks cultural specificity has led to the emergence of commonalities and generalities in music production, and may have contributed to homogenous music listening patterns the world over.

Nothing may be more telling than the advent of musical ‘ringtones’ in India. In the late 1990s, mobile phone operators started providing the capability to customize one’s ringtone. The first popular ringtone was an excerpt of a patriotic song (‘Saare Jahan Se Accha’), but soon most of them were taken from ‘Bollywood’ films—a major source of popular music in India. In contrast with later MP3 playback that could accurately reproduce a piece of recorded music, the first ringtones were simply polyphonic—or even monophonic—versions of the original usually generated in realtime by an FM (Frequency Modulation) synthesizer and a MIDI sequencer. Anecdotal observations suggest that people didn’t mind the distortions introduced by the *quantization* (or approximation) in pitch and timbre introduced by MIDI and FM as long as the tune was recognizable, but market data suggests that the tunes were selected from simple popular songs, rather than the more complex ones influenced by classical music (with intricate pitch modulations and rhythms). Whether this was a case of unconsciously selecting the source to ensure optimal rendering, or of choosing the popular music of the day that was itself a product of studio technology influenced by FM synthesizers and MIDI sequencers with fixed quantization steps, the fact is that technology pervasively influenced the music that was produced and heard, and thereby the society’s cultural fabric.

I suggest that if we are to build computer music systems for music from around

Table 1.1: Elements of music

Physical	Perceptual	Cognitive
Frequency	Pitch, modulation	Melody, harmony, scale, mode, key
Amplitude	Loudness	Intensity, dynamics
Spectrum and amplitude shape	Timbre	Tone color, instrument identity
Simultaneity	Grouping	Note, chord, stroke, texture
Duration (short)	Beat	Rhythm, tempo, meter, time
Duration (medium)		Cell, figure, motif
Duration (long)		Phrase, form

the world then we must design technology that is sensitive to each of those musical traditions. In this dissertation I propose to define a measure of culture, develop a culturally sensitive representation of music, and design a model for music prediction based on that model.

1.2 Music Representation and Prediction

As individuals listen to sound and music, they develop mental primitives that determine pitch, beat, timbre, and learn higher-order constructs that identify scale, meter, instrument labels by extracting relationships and hierarchies from sequences of musical elements. Table 1.1 lists the elements of music. Top-down (culturally grounded) and bottom-up (biological) perceptual and cognitive processes lead to the acquisition of musical skill, as well as the ability to derive taste and emotion from music (see figure 1-1).

There is much evidence that points to musical expectation (the capacity to predict subsequent musical events) as a key to music understanding and emotion (Meyer, 1961). Musical expectation relies on the memorization and recall of musical structures, and consequently on the *exposure to prior music*. The role of enculturation in music understanding and music prediction is therefore of prime importance.

Memories of musical elements are regularized (i.e. rendered invariant to pitch

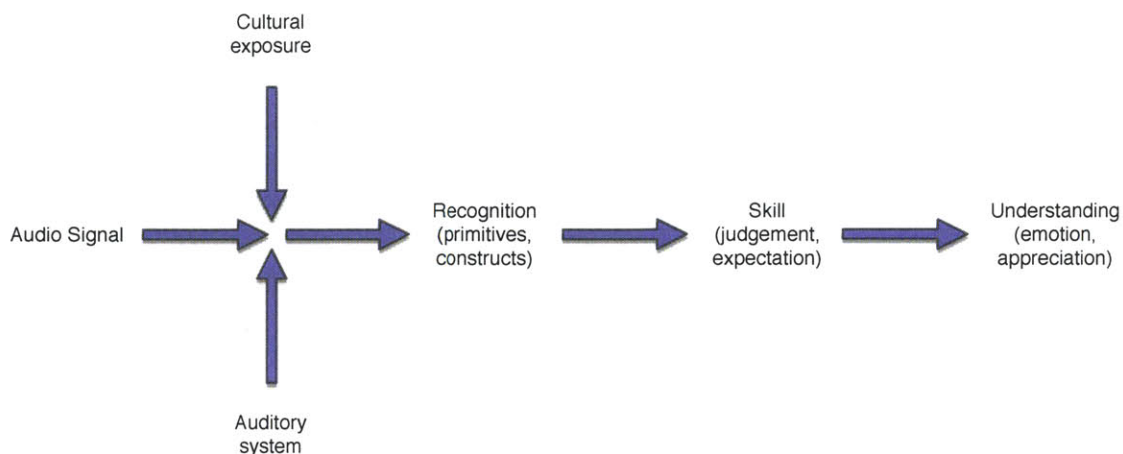


Figure 1-1: Top-down and bottom-up processing of music

transposition or tempo variation) and stored, and then retrieved on presentation of an auditory stimulus (whether from an external sound source, or from an internal recall or spontaneous generation). Context is important. Specifically, a particular individual’s mental representations depend on the auditory content that the person has been exposed to. In fact, even the experience of language can have an influence on the acuity of pitch perception: it has been shown (Wong et al., 2012) that speakers of a tonal language like Mandarin exhibit higher sensitivity to pitch variations and interval distances than speakers of a non-tonal language, thereby supporting the idea of “culture-to-perception influences.”

This supports the idea that cultural context needs to be considered when representing musical structures. Furthermore I suggest that the enculturation process needs to be taken into account when attempting to design machine listening systems that learn musical structures from audio data. However, where an individual would acquire auditory primitives and build musical constructs through a process of listening through mere exposure, it may not be practical or feasible to subject a machine to the same time-consuming process. In a computer music system one alternative could be to design *well-formatted containers* in order to embed *well-formed* cultural knowledge by bootstrapping internal music representations with models suitable to a particular culture in a way that simulates the enculturation and learning processes.

As suggested by Piaget’s constructivist theory of learning (1952), new knowledge arises from experiences through *assimilation* (by incorporating new experiences in an existing framework) and *accommodation* (by reframing our internal models when our expectations are violated). In addition, Piaget stresses the importance of the environment and of social interactions with knowledgeable people who give meaning and utility to symbolic structures (i.e. labeling). State-of-the-art machine listening systems (for Computational Auditory Scene Analysis for instance) incorporate some amount of contextual information (the equivalent of cultural knowledge) in the form of statistical models. I propose to provide cultural context as a starting point (with empty, but appropriate containers), emulate the process of *assimilation* through exposure to musical content, and continuously update internal representations by *accommodation* to new incoming information. Labeling is optional, but the grouping of similar elements is explicitly defined by supervised learning.

1.3 Motivation and Philosophical Thoughts

In *Why People Think Computers Can’t*, Minsky (1982) observes “how even the earliest AI programs excelled at ‘advanced’ subjects, yet had no common sense,” and that “much ‘expert’ adult thinking is basically much simpler than what happens in a child’s ordinary play!” In other words, machines can do well what humans can’t, and vice versa.

It follows that music prediction, which seems so natural, almost intuitive, to humans is a difficult task for machines. The current work strives to propose strategies to improve automatic music prediction. In fact, the engineering motivation behind this research is to propose design principles based on cultural relevance for the computational representation of music and its predictive-capability building. This work is also driven by an anthropological motivation, which is to shed light on the cultural aspects of music perception and cognition that contribute to musical expectation.

While similitudes and parallels have often been drawn between music and language, music is widely considered ‘universal’ because on the surface its tokens do not require translation. However, I would argue that a deeper understanding of music, one which leads to predictive capabilities and gives rise to emotions, is tied to a particular instantiation of culture as a genre. In fact most people would agree that a Western classical music aficionado would find it difficult to make sense of heavy metal, or an exclusive fan of hip-hop would find it challenging to understand, let alone appreciate Hindustani (North Indian) music—unless they had been informally exposed to or formally trained in the other genre.

Is music similar to language in that it has to be learnt and can indeed be translated from one genre to another? How much of Noam Chomsky’s theory on universal grammars can be applied to music? If music and language share an innate common grammar, is it only sufficient to learn its lexicon and grammatical constraints to understand and appreciate a specific genre of music?

The question of whether music can be predicted, and if so, by how much raises the question of a deterministic world (Hofstadter, 1979). As much as the complexity of the world precludes us from making such an assumption, experience suggests that the *short-term* future can be predicted to some extent. In the case of music prediction, this raises two questions: what can be considered short-term (i.e. how many milliseconds) and to what extent are predictions valid (how constrained, or how small, can the solution space be made)?

Generally, probabilistic models tend to provide ‘most likely’ solutions in stochastic environments. However, our concern here is with individual musicians’ responses to specific inputs (preceding musical events) with particular priors (cultural context and stylistic rules). Therefore approaches that tend to model the self-replicating patterns of music (articulations - tones - phrases - forms) with a dynamic rule-based system that is evolved from data are favored over those that learn probabilistically.

In some sense, the proposed system should convey a sense of musical intention

that meets listeners' expectations by generating well-formed constructs that 'make sense' in the musical idiom under consideration.

The current work is situated in the following fields of scholarship: music perception and cognition (psychology), computational ethnomusicology, machine listening (DSP, machine learning, AI), and communication theory (symbolic representations).

1.4 Scope and Definitions

The present work is concerned with the symbolic representation of music, rather than its sample-based digital form. While some music processing systems are culture-agnostic because they represent music as an audio signal in a communication model like that of Shannon-Weaver (1949)—see for example the PCM audio scheme or the MP3 compression format—most systems that use a symbolic representation for music tend to be biased towards a particular musical tradition, Western music in most cases.

A symbolic representation of music does not imply that there should be a score or any other form of notation. As it happens, most non-Western musical cultures are based on oral traditions, which by definition lack formalized or well-specified notational systems that could serve as a basis for their machine representation. Therefore improvised, or score-less, performances are considered as the source material for training rather than score following.

In the case of symbolic structures, appropriate representations are especially critical because the granularity of control is at the level of musical ideas, rather than at the sample level.

In anthropology, culture is defined as a combination of shared values (in terms of knowledge, belief, behavior, etc.) within a community. Individuals get acquainted with a particular culture through interactions with members of the community and exposure to artifacts and media such as music. The anthropological point-of-view

establishes culture as arising from people’s capacity for symbolic thought and social (or imitative) learning. These are the principles by which I aim to design computational representations of particular cultures. Furthermore, a certain culture can have multiple styles, movements, or schools for its artistic activity. Western tonal music for instance consists of forms like sonata, symphony, and concerto.

In this dissertation certain words that pertain to the subject matter appear with a certain frequency. Although some of them might sometimes be used interchangeably, I define them in appendix A with the meaning I intend them to have in this document.

When referring to music prediction, this dissertation actually concerns itself with melodic, or timbral (in the case of percussive instruments), and rhythmic prediction. To limit the scope of this work, and following a reductionist approach, other dimension of sound and music like amplitude and timbre (in the case of instruments that produce tones) are treated as independent variables. Furthermore only monophonic textures are considered to the exclusion of polyphonic and harmonic material.

In addition to purely auditory stimuli, other cues like the orchestration, the effects used (e.g. reverberation parameters), or the ‘sound’ of a piece of music set expectations. Context and metadata also inform what people hear. The current work does not take these into consideration.

North Indian music and South Indian music are studied as proxies for ‘world music’ and compared with Western music. The audio material used in this study has been sourced from personal recordings and downloaded from user-generated content on the internet. This work focuses on modeling existing forms of acoustic music, and although the system described here could lead to new types of music it is not a primary concern.

1.5 Research Questions and Hypotheses

Can music be predicted? On the one hand, this work can be described as an attempt to predict music in a way that, even if the estimations are off, they still make sense in the particular musical idiom under consideration. The system’s role, therefore, is not so much to make accurate predictions in short time windows as it is to convey a sense of the performer’s musical intentions at longer time scales. On the other hand, this work provides insights into musical creativity—what parts of music follow tradition, stylistic rules and known patterns, and how much does the performer shape? Music indeed strikes a balance between predictability and surprise, but how much of each is up to the composer or performer versus the rules of the genre?

What can be considered an appropriate representation of music? In order to accurately analyze, process, or synthesize music from around the world, computational representations that incorporate a model for each specific musical style under consideration are required. I propose in this thesis to develop and study music systems that take into account culture-sensitive features. These representations are meant to include a lexicon of musical events (e.g. notes, embellishments, percussive strokes, durations) and relationships between them (in terms of sequence and timing), as well as specific stylistic rules. Representations are initialized with assumptions and constraints from the culture under consideration, and get dynamically updated with the incoming stream of auditory events. The proposed musical representations are, to borrow a computer science paradigm, object-oriented data structures: they are the containers, the data, and the processes that apply to them.

How can music be predicted? What are the perceptual and cognitive cues that allow for musical expectation? How much should systems for automatic music prediction inspire themselves from biological and psychological processes versus being based on machine-specific models? And how much insight on human musical expectation can we glean from computational models for music prediction?

Based on the above, the research questions that I aim to tackle in this dissertation are:

- Can music be actually predicted, and if so, to what extent?
- Can musical traditions be quantified and compared?
- What is the role of musical enculturation in musical expectation?
- What is an appropriate representation of music?
- Can the human capacity for musical expectation be modeled with a machine-based prediction engine?

I hypothesize that in order to predict music, statistical analysis is necessary, but not sufficient. The machine learning algorithm development process requires the designer to select appropriate representations (data structures) similar to the possibly hardwired receptacles in the mind that get filled with musical structures through the process of enculturation. In addition, to establish an accurate representation of music, I posit that a symbolic system needs to extract ad-hoc primitives from audio because there are no currently suitable representations (e.g. MIDI, **kern).

The proposed model is trained from data. If the model is too general (in that it has too many degrees of freedom), prediction may be inaccurate because too little constrained. On the other hand, if the model is too specific, it may not account for surprises (outliers in a statistical model) or *creativity*. Therefore the model needs to find the right amount of complexity through selection of training data and selection of data structures—both of which are the responsibility of the system designer.

The experimental work described in this thesis arises from the following hypotheses:

- Music complexity (i.e predictability) is a function of musical culture.

- Computer music systems that incorporate a culturally appropriate symbolic representation provide a perceptually equivalent output to a signal-level representation of the musical source.
- Music prediction systems that are culturally specific are more robust and accurate than systems that generate symbolic musical data with no attempt to understand cultural content.
- The input-output behavior of a culturally sensitive design for automatic music prediction performs similarly to human musical expectation.

1.6 Contributions

No prior work on synthetic music systems has, to my knowledge, included models that learn from exposure to a particular culture. In some sense, such systems must convey a sense of musical intention by fitting to the musical form and its associated higher-level structures rather than to localized events: local prediction errors are to be expected, but prediction results should be judged on their ability to preserve higher-level constructs, like repetitions or surprising variations, especially as they relate to each other in a dynamic ecological setting by generating well-formed constructs.

Experimental results indicate that cultural learning is indeed an essential component of intelligent music systems that collaborate and participate in live interactions with musicians and listeners. It is thus a worthy, if challenging, enterprise to propose an analytical framework for culture in music. I propose a series of experiments on two sets of musical sources that will enable us to analyze and compare musical traditions. The learning algorithm lets us study how being exposed to one tradition can lead us to make correct judgments and predictions on music from other cultures. This method can also be used on an unfamiliar piece of music to compute how closely related it is to other known styles and devise a cross-culture transformation to better understand it.

The analytical framework to quantify musical culture is based on the principles of information theory—in particular, a measure of entropy for melody and rhythm. In this case, entropy, or information content, denotes the capacity for a listener, whether human or machine, to predict musical events as a function of prior musical exposure (i.e. enculturation): musical culture is to be understood here as an aggregate of the music collection that has been subjected to experience, either direct (by a particular individual) or indirect (by the community, resulting in a set of shared rules). This approach not only provides us with the means to compare the information content of various musical genres, but also allows us to investigate cross-cultural influences in musical styles.

The present work will result in the following artifacts and contributions:

1. An analytical tool to measure and compare musical cultures with a measure of musical entropy, from the perspectives of melody and rhythm, based on the principles of information theory.
2. Design principles and strategies for culturally sensitive music representation.
3. A software framework for music prediction based on machine listening and music synthesis that generates musical symbols corresponding to future estimates of the current musical input.
4. Cultural plug-ins for this framework that will include models for:
 - (a) tabla (North Indian percussion used in Hindustani music)
 - (b) drums (Blues)
 - (c) bansuri (North Indian transverse bamboo flute used in Hindustani music)
 - (d) flute (Blues)
5. A study that supports the importance of computational musical enculturation for music prediction.

6. I introduce the concept of *prelag* ('negative lag') when music prediction precedes musical expectation in the cognitive realm.
7. A series of computer music applications enabled by culturally sensitive automatic music prediction.

I expect that the work presented in this thesis will justify and enable the design of culturally sensitive music technology that will sustain and cultivate the diversity of musical traditions around the world, and especially empower underserved communities in the areas of creativity and education in ways that are relevant to them. The tools under development here can also serve to support ethnomusicology research.

1.7 Outline

Chapter 2 strives to answer the question of whether music can be predicted by taking an information theoretic approach to measuring musical complexity from a cultural referential. Measures of melodic and rhythmic entropy are introduced and computed. They represent the predictability of Indian and Western popular music based on prior exposure to music of either genres. This establishes an objective baseline for music prediction.

The following chapter is concerned with music representation. Culturally sensitive models are contrasted with culturally agnostic ones. Design principles for genre-specific containers for musical elements are introduced along with the notion of computational musical enculturation. Symbolic music data is extracted from audio and provides the content that informs as well as fills those containers.

In chapter 4 an automatic music prediction engine inspired by human musical expectation and cultural specificity is described. The proposed algorithm is based on a hierarchical grammar of genre-appropriate lexical tokens for melody or timbre, and rhythm.

Chapter 5 concludes this document by summarizing findings, presenting applications of the technology and concepts developed in the current work, and describing my contributions.

Chapter 2

Can Music Be Predicted?

2.1 An Information Theoretic Approach to Measuring Cultural Footprint in Music

In order to understand how people experience music, it is not only important to study the process by which the human auditory pathway analyzes musical acoustic input, but it is also essential to appreciate the role of prior musical exposure. Narmour (1990) distinguishes these two complementary processes, which he refers to respectively as bottom-up (biological) and top-down (cultural).

Listening to music from a young age (as early as one year old according to Morrison and Demorest, 2009) leads to an enculturation process that shapes how people interpret music that they later hear. This background knowledge sets the stage for composers to tickle the human cognitive capacity for anticipation, and either to fulfill people's expectations or to create surprise, and thereby generate tension, frisson, or even laughter. The mechanisms for human learning extract statistical regularities from the world around, including from music, and generate mental representations for structures like pitch sequences and rhythmic primitives.

The current chapter examines a method to quantify musical culture based on the principles of information theory. In particular, entropy, or information content, denotes the capacity for a listener, whether human or computer, to predict musical events as a function of prior musical exposure. Morrison and Demorest (2009) explain how “exploring music cognition from the standpoint of culture will lead to a better understanding of the core processes underlying perception and how those processes give rise to the world’s diversity of music forms and expressions.”

Musical structures are stored in memory as symbolic structures invariant to linear transformations (e.g. time stretching and compression, or pitch transposition) and robust to noise (Snyder, 2000). Mental representations of auditory data may be associated with other sensory inputs, or contextual metadata (e.g. visual or textual information). This way of describing enculturation matches the role of the environment in Darwinian literature, which specifies that human behavior is determined by heredity, learning, and environmental exposure (Huron, 2001).

As much as this combination of factors would make for highly individual responses to similar music stimuli, it has been shown that people from a similar cultural background understand and respond in similar ways to music (Cross, 2001). Therefore the importance of common references in making sense of music cannot be discounted. As essential and universal as music may be to the human experience and to human expression, it has evolved into a diverse set of musical practices. In spite of that, music theory has largely focused on tonal Western music. Nonetheless, various sub-fields of computer science, in particular artificial intelligence, machine learning, and information theory, are well suited to the study of music from around the world with models based on rules extracted from data rather than input manually.

The understanding and appreciation of particular musical styles through enculturation result in the creation and encoding of musical constructs and sequences in memory, which lead to an increase of musical skill and the ability to derive pleasure and emotion from music. Expectation, or how well a person can regularize and retrieve

musical constructs from memory, plays a crucial role. The tools of information theory and statistical learning applied to symbolic and non-symbolic (audio) representations of music can provide us with insights not only on the prediction mechanisms used by people, which can be leveraged for algorithm design, but also on the relationships between the stylistic aspects of music from different cultures, and thereby support the study of comparative music theory.

Although information theory as introduced by Shannon and Weaver (1949) has oft been explored as a tool for musical analysis, it has met with mixed results. I briefly survey the literature on using musical information theory to model musical expectation. I propose to use *entropy* as an information theoretic measure to analyze cultural content in music. This approach not only allows a comparison of the information content of various musical genres, but also enables an investigation of cross-cultural influences in musical genres.

2.2 Music as Information: Prior Art

2.2.1 Nature and Nurture in Musical Skill

In music cognition, like with other biological processes, there is an ongoing debate about the innate or learnt nature of many auditory phenomena. Huron (2006) shows that nature actually does not have this preoccupation. From a biological perspective, there is a decisive factor by which it is best for a behavior to be instinctive or learnt. The determining factor is the stability of the environment: when there is little environmental change, conditions favor instinctive or innate (or hardwired) behavior, which are usually fast and effective; on the contrary, when the environment changes quickly it is best to learn. Therefore, the difference between instinctive and learnt behavior is not that the first one is genetic and the second one, environmental—learning involves more genetic machinery than do instinctive behaviors. In fact, instincts re-

flect a longer and more profound interaction with the environment than does learning. This evolved capacity to learn is referred to as the *Baldwin Effect*.

For instance, some recent studies argue that rhythmic grouping may not be innate but rather a product of enculturation. Iversen et al. (2008) performed an experiment to compare the perceptual grouping of rhythmic patterns between English and Japanese native speakers. They found that the two groups came up with very different results. This finding extrapolates into melody, suggesting that the comprehension of melodic structures (segmenting a sequence of tones into motives and phrases) is a function of musical exposure. Povel and Okkerman (1981) previously described the rhythmic grouping of equitone sequences, but did not focus on the cultural aspect of the perceptual response.

On the other hand, Krumhansl (1995) performed a series of studies to test Eugene Narmour's implication-realization model (1990; 1991; 1992) for tone-to-tone melodic expectancies. This study provides insights into how our minds encode, organize, remember, and process musical information over time. Her conclusion is that perceptual organization is based on bottom-up processing alone, negating any role of musical training or enculturation. However, Larson and McAdams (2004) suggests that Narmour's theory and Krumhansl's experiments seem to work only for melodic continuation (i.e. the next tone), and not for melodic completion (i.e. the whole phrase). His experiments demonstrate that including the top-down component of Narmour's model allows it to generate entire completions.

Evolutionary musicology, which is part of the broader and relatively recent field of 'biomusicology,' has seldom been studied in the context of machine listening systems. However, as an increasing number of theories are formulated on the evolution of the mind, explaining the evolutionary significance of music and its adaptive function is a worthy enterprise.

The study of the origin of music is highly speculative. Although most scholars (traditionally ethnomusicologists) have focused on the psychological, social, and cul-

tural origins of music, many (cognitive psychologists, most recently) have also sought to explain music within an evolutionary framework. In fact there are archeological, anthropological, biological, and ethological indications that support the fact that music may result from evolutionary adaptation: the first complex human artifacts to be found have been musical instruments (including a bone flute), and “every known human society has what trained musicologists would [recognize] as ‘music’ ” (Blacking et al., 1995); moreover infants acquire early on naive musical abilities similar to adults, which may suggest an innate ability for music and an opportunity found through it to refine motor skills; finally the amount of resources (time, money) that is spent on listening to and making music in various cultures is considerable regardless of the society’s development level.

The main question, according to Huron, is: “What advantage is conferred on those individuals who exhibit musical behaviors over those who do not?”

Several theories have been exposed to explain the origin of music from an evolutionary perspective:

- i Pinker’s position is that music is “auditory cheesecake,” an unnecessary by-product of other adaptive functions (1997);
- ii Miller (2000), following some indications by Darwin, supports the view that music evolved on the basis of sexual selection;
- iii Dunbar and others (1996) argue that music allows for social cohesion and solidarity through mutual grooming and “muscular bonding” (e.g. synchronized dancing).

Pinker links music with other aspects of demonstrably evolutionary human experience (concerned with survival so having a clear functional role): language, auditory scene analysis, habitat selection, emotion, and motor control. In other words, music may be harmless to natural selection and therefore has not interfered with it. Cross

(1999) revisits Pinkers arguments from a musicologist's perspective: "our musicality is grounded in human biology as expressed in the evolution and development of our cognitive capacities and of our social and environmental interactions," thereby suggesting that music does not only result from, but also participates in individual and social development (through enculturation) while contributing to the evolution of our species. He emphasizes the importance of music in helping individuals acquire emotional intelligence: the ability to communicate ones feelings through one's face, voice, body, and decode them in others.

Miller supports his view on the role of music in sexual selection by submitting an anecdote of Jimi Hendrix and the sexual attraction he elicited on the opposite sex. Some authors have supported this position by recounting evidence of the correlation between interest in music (evidenced by time and money spent on music listening), which generally peaks during adolescence and as young adults, and sexual activity. Moreover Miller showed that creativity was valued more highly than wealth when looking for a potential mate; the former being considered as genetically superior for the purpose of child fathering, while the latter, considered as circumstantial, being viewed as simply adequate for child rearing.

A concurrent view to that of Dunbar is that of Kogan (1997) who sees music as adaptive through group selection: for him music promotes group morale and identity through common activities—leading to a common emotional response when moving rhythmically in a synchronous manner. This group behavior leads to a tension-releasing function for the individual without any destructive influence on social cohesion, and may actually contribute to group effort. This may have an evolutionary role: according to Cross, "music enables risk-free action and facilitates risky interaction."

A different view is taken by Mithen (2005) who describes human cognitive development as the capacity for information-processing mental modules (which perform specialized tasks rapidly and efficiently as the result of adaptation) to transfer competences across domains by the formation of a general representation. Mithen supports

the view that music and language co-evolved and first appeared as proto-music and proto-language. His outlook considers music as an adaptation contributing to perceptual development.

Levitin (2006) presents a summarized account of the various views held on the origins of music and notes that due to evolutionary lag our brains are currently adapted to the music and dance forms of around 50,000 years ago. At that time music was not just an aural commodity, but with no specialized role for music in society, the relation between music and action was much more explicit. In fact, Cross argues that music's adaptive role may have been through sound and movement (e.g. singing, playing instruments, dancing) and may have occurred by exercising some of the faculties that were necessary for survival (e.g. prosody, recognition of emotions from facial expression, periodicity of movement).

Patel (2010) doubts the adaptive role of music and assigns it a different category: that of a transformative technology (to exercise or stimulate our capabilities) rather than an innate and adaptive predisposition. Patel argues that we might be able to live without music, but we value it because it transforms our lives in ways we care about deeply—through emotional and aesthetic experience, identity formation, and social bonds.

The brain processes all sounds, but not equally. According to Minsky (1981) “much of sound is one-time (...) [which] is why we don't have ear-lids.” But musical sounds differ from other acoustic signals in that they convey human-induced meaning, in a mode of abstract communication. While the emotion that is conveyed in music is ‘deeper’ than the intellect (in an evolutionary sense and according to the cerebral areas that are solicited), Handel (1993) observes that the low-level primitives supplied by our auditory system clearly have an evolutionary function, in particular by alerting us of danger (as remarked earlier they are ‘always on’ even when we are asleep)—for instance the crack of a twig has high frequency content to which we respond much faster on a neurophysiological level than low frequency content.

In his book *The Musical Mind* Sloboda (1985) proposes a taxonomy of the world's music in order to understand the relation of music to culture. The parameters that are suggested include instruments, forms, scales and tuning systems, and social context. Sloboda explores the cultural and social factors that account for the wide differences between musical cultures. He suggests that the existence of a system for musical notation (versus an oral tradition), and more recently of recording technology, might play a significant role in the evolution of a musical culture.

2.2.2 Information Theory in the Context of Music

In the field of information theory, entropy is defined as a measure of uncertainty, or information content, associated with a random variable in a probabilistic model. Entropy quantifies the average number of bits per symbol needed to encode a message for a particular data source (Shannon and Weaver, 1949). In communications theory, entropy is used to specify the amount of lossless compression that can be applied on a particular transmission channel.

Assuming that a source of music can be modeled as a Markov process, for a zero-order source (i.e. each symbol independent from the preceding symbols) entropy is given as:

$$H(x) = - \sum_i p_i \log_2 p_i$$

where H is the amount of uncertainty, p_i is the probability of observing event i , and h is the information content of event i .

For a first-order Markov source (i.e. the probability of a symbol is dependent of the immediately preceding symbol), the entropy is:

$$H(x) = - \sum_i p_i \sum_j p_j(i) \log_2 p_j(i)$$

where $p_j(i)$ is the probability of observing symbol j given that the preceding symbol

is *i*.

As we increase the order of the Markov source, the computation of entropy becomes increasingly complex. Given that the entropy is maximal if all the output symbols are equally likely, such value would describe complete randomness in the data. However each musical tradition has constraints that distinguish it from others. Therefore we can assume that the entropy value for each musical style or genre is a distinct value that allows musical traditions to be compared to each other.

For music, information has been understood to refer to “the freedom of choice which a composer has in working with his materials or to the degree of uncertainty which a listener feels in responding to the results of a composer’s tonal choices” (Youngblood, 1958). Youngblood thereby suggests that information content could serve as a method to identify musical style. He showed that excerpts of Schumann’s music have slightly greater entropy than excerpts from Mendelssohn’s. But to this end he only uses one scalar entropy value.

In 2008, Margulis and Beatty modeled the dynamic adaptive expectation of melodies based on various pitch information sources. She notes that entropy has been intermittently pursued as a potential tool for musical analysis; however, significant problems have prevented it from leading to a fruitful theoretical approach. According to Margulis, three challenges have hindered the development of information theory applications in music: first, obscurities in the framing of the musical questions that information theory might answer; second, practical obstacles to tabulating the musical entities needed for information-theoretic analysis; and third, uncertainties regarding the type of musical entity that should serve as the unit of analysis.

Entropy can help explain and illustrate melodic structures: in its thermodynamic incarnation, entropy represents a state of little potential energy. It is common for melodic patterns to follow structures in the form of consonance (balance) – dissonance (tension) – consonance (cadence). Dissonance here suggests a state of high energy that leads to resolution. This characteristic of melodic phrases can benefit from being

quantized by its dynamic entropic value.

Knopoff and Hutchinson (1983) suggest that entropy could be used to study the distribution characteristics found in particular musical styles and identify the musical style. He views a music composition as a path in the sparsely populated space of music parameters for which the composer selects elements from several musical parameters.

2.2.3 Entropy, Complexity, and Prediction

In the case of melody, entropy can be defined as the ability for an informed listener to infer the next note in a phrase. This notion can be quantified by the number of bits required to specify the next note. (In a concrete representation of sound we would consider how many bits were needed to specify the next signal sample.)

Other dimensions of music, like timbre, necessitate additional encoding. However, this information can also be constrained by the instruments of the orchestra, or in the case of computer music, by the parameters of a particular synthesis method. Csound and its related NetSound, which is at the basis of the Structured Audio Orchestra Language in the MPEG-4 audio standard, encodes synthesized musical instruments in a surprisingly small, text-based, payload (Scheirer and Vercoe, 1999).

Speech recognition uses the concept of entropy to evaluate a string of potential phonemes by using machine learning techniques like Hidden Markov Models and the Viterbi algorithm for decoding.

It has been suggested that comparing, or even ranking, composers or whole genres or musical traditions by some metric—why not musical entropy—might be an interesting endeavor. The difficulty then seems to come down to identifying the proper dimensionality of the entropy vector and its adequate mapping to musical features in order to account for the multi-dimensional characteristics of music.

Dubnov et al. (2004) have studied how listeners react to an unfamiliar piece of

contemporary music over time by comparing their reaction to predictions based on automatic analyses of the audio signal. Subjects were asked to continuously rate their level of *Familiarity* (based on self-similar structures in the piece) and *Emotional Force*. For their part, the algorithms computed signal similarities and measured signal predictability over time. The results showed a correlation between signal properties and human reactions. The information theoretic analysis of signal predictability used low-order cepstral coefficients as feature vector to describe the evolution of the spectral envelope over time. To evaluate signal predictability, Dubnov introduced the notion of *Information Rate*—the “reduction of uncertainty that an information-processing system achieves when predicting future values of a stochastic process based on its past,” or in other words, “the additional amount of information that is added when one more [event] of the process is observed”. The information rate can be interpreted as the amount of information a signal carries into its future. Dubnov et al. (2006) shows the significance of this measure over music signals and natural sounds. As the number of events n grows large, the Information Rate becomes:

$$\rho(x) = \lim_{n \rightarrow \infty} \rho(x_1, x_2, \dots, x_n) = H(x) - H_r(x)$$

where x is the sequence of events, $H(x)$, the marginal entropy and $H_r(x)$, the entropy rate, with:

$$H(x) = - \sum_i p_i \log_2 p_i$$

and

$$H_r(x) = \lim_{n \rightarrow \infty} \frac{1}{n} H(x_1, x_2, \dots, x_n).$$

The information rate can hence be interpreted as the amount of information a signal carries into its future. In fact, Dubnov demonstrated the significance of this measure for music signals and natural sounds.

Cont (2008) modeled musical expectation in his doctoral work under Dubnov. According to Cont expectation characterizes information content and provides insights

into the stored musical structures.

As much as information theory seems to have fallen out of favor as a music-analytic tool, statistical learning has emerged in psychology and the learning sciences (for a review see Saffran, 2003). Statistical learning refers to the ability of learners to exploit statistical regularities in their environment to generate abstract structures. Infants have been shown to exploit such statistical properties to learn about their environments. New empirical evidence suggesting that humans can track complex statistical properties in numerous domains makes information theory newly interesting from the perspective of psychology. It follows that information theory could be useful not merely in characterizing styles, but also in addressing questions in music cognition.

In this context it makes sense to consider a musical entropy vector that takes into account contour, duration, and intervals. Simon (2006) has proposed multiple entropies to account for music's multi-dimensional characteristic (in the case of jazz), including melodic entropy, harmonic entropy, rhythmic entropy, and a composite musical entropy:

$$H = - \sum_{i=1}^k \frac{v_i}{t} \log_2 \frac{v_i}{t}$$

where v_i is the number of times the i^{th} melodic variation appears, t is the total number of melody notes and k is the total possible melodic variations.

Henry (2000) introduces the concept of genre entropy in his study of Indian folk songs where entropy represents the diversity of melodies used within a genre. Henry explains that with more melodies come less order and more uncertainty. In some societies in North India, particular genres of folk songs have few distinct melodic lines. Henry illustrates his point by comparing folk music, which is basically a carrier for the text in various rituals, to highly entropic entertainment music, which makes use of musical and poetic novelty and a blurring of the boundaries between genres.

2.3 Musical Entropy

Can music indeed be predicted? If so, to what extent?

How much information is required to identify and classify a piece of music in a particular culture? Can musical traditions be quantified and compared?

How do learnt musical structures influence predictions about musical events? What is the role of musical enculturation in musical expectation? Does culture-specific training improve predictability?

If we are to understand how people identify music pieces, we must determine the set of audio features required to do so. I propose to investigate the use of musical entropy to quantify cultural influence on music in light of the previous research questions.

Initial inquiries are constrained to the study of melody. I use a repository of MIDI files of popular Indian music and compare it to a study of popular Western music.

The experimental steps consist in:

1. Feature extraction: contour, duration, and interval distribution.
2. Non-supervised clustering to group pieces for ground truth classification, and comparison with human labeling.
3. Computation and comparison of melodic entropy for each tradition.

The current approach analyzes the complexity required of a relevant musical feature, namely melody contour, to retrieve a song from a catalog of 224 popular Indian instrumental songs. To achieve this goal, I developed a theory of musical features inspired by Ullman's theory for visual features (2002). I implemented a computer program that analyzes instrumental MIDI versions of the songs to extract relevant parameters. My program identifies music pieces based on a melody contour of varying

length and complexity. I experimentally verified the hypothesis that music pieces are recognized by a melody contour of intermediate complexity.

The entropy of melody contour is defined by the number of ‘quantization steps’: from 3 (starting from a base note, the next note goes up, goes down, or remains the same) to 7 (where harmonic intervals are taken into account). The musical interval cut-off point is determined by the analysis performed on the interval distribution in the song database. The Indian music files in my test set suggest that most of the information is located below a perfect fourth (5 semitones above and below the reference note).

A query program lets the user enter a melody contour with a 3-, 5- or 7-quantization step in order to retrieve songs; the system provides the track numbers where the pattern is found weighted by the number of occurrences in the tracks. Another program computes statistical scores by retrieving songs depending on the quantization step (3-, 5- and 7) and sample length (from 3 notes to 11 notes in steps of 2). Results indicate that for a sample length of 5 notes and above, the 5-step quantization model works as well as the 7-step quantization. This confirms that very little information is encoded in the higher musical intervals. This suggests that the optimal melody contour model needed to retrieve a song is of *intermediate* complexity.

This finding is a significant step in understanding how we recognize music and build a cognitive model of expectation.

People can recognize a piece of music by listening to just a few notes. Sometimes, based on context, as little as 2 or 3 notes are enough to identify a song. A minimum number of features are needed to retrieve a song from memory. Nevertheless, there is also evidence that too many features are detrimental to musical recall: a song never sounds exactly the same. Every time a song is played, its loudness, background noise, reverberation, even its interpretation, affect its ‘quality’. Moreover, if some parameters, such as its frequency response (for instance with bass enhancement), are modified, or the piece undergoes transposition, tempo variation, or remix, the song

generally remains recognizable to a large extent. Therefore I suggest that there are higher-level features that enable people to recognize and identify music from memory.

When we hear a song at a rock concert, it takes us no more than a few seconds to recognize it, even though the noise affects what we hear, and the version that is performed may be different from the one on the CD. This suggests that we extract regularities from the music we hear. At the lowest level we identify notes and timbres. At the highest level we identify rhythmic patterns and melody contour as an abstraction that is for the most part pitch-invariant, loudness-invariant, and timbre-invariant.

Literature indicates that music has the following perceptual attributes along different orthogonal dimensions:

- Pitch or melody contour: the most salient feature of a musical piece
- Rhythm
- Pitch: absolute pitch is not as important as relative positions
- Tempo
- Timbre
- Loudness
- Spatial location
- Reverberation

In this section we investigate musical source from a top-down perspective.

Most of the existing work on melody contour has been conducted on Western music (from various periods and genres). In order to generalize my theory I chose Indian music for my experimental data. My test set is representative of modern



Figure 2-1: Depiction of melodic contour in Western music notation

popular Indian music. Being modal, Indian music has features that are technically different from Western tonal music.

Melody contour is the identity of the melody, its general shape. This may seem like a coarse approximation because musical interval (the interval between consecutive notes) is not considered as significant as the shape of the melody. Although we gain more understanding of musical intervals with training, I show that it does not play a major role in music recognition.

In the mid-70s, Parsons, inspired by Barlow and Morgenstern's *Dictionary of Musical Themes* (1948), published *The Directory of Tunes and Musical Themes*, also known as the "up-down book". This volume contains 10,000 themes, which are uniquely identified by a maximum of 16 nodes that describe melody contour with three values: up, down, and same. This may seem like too simple a model for our brain because we need much less than 16 notes to recognize a melody. However my findings indicate that musical intervals may have less importance than our intuition suggests. This work also supports the importance of the relative rather than the absolute positioning of notes.

The current model uses 3 different quantization formats for melody contour, ranging from the simple "up-down" model, to a 7-step quantization, which takes into account limited knowledge of musical intervals.

I downloaded 227 MIDI files from various sources on the internet, out of which 3 were unusable. I performed little checking on the quality of the song transcription because I required only a suitable approximation of the songs for my model.

Table 2.1: Musical intervals

Interval	Quantization Steps
Minor third	± 3 semitones
Perfect fourth	± 5 semitones
Perfect fifth	± 7 semitones

A program computes the interval distribution of the downloaded songs. The program parses each instrument channel in each MIDI file, and compares each note with the previous one to determine the information content of the melody contours of this particular set of files. I do not consider chords (simultaneous notes), which do not provide melody contour information. This data is assumed to be buried in the noise.

Then the program drops all non-relevant information and creates strings of characters from the note sequence in each track. The basis for this program is the concept of musical idiom set forth by Jackendoff who suggests in his work on music cognition that the mental representation of high-level musical features is akin to grammar. Therefore I construct a string that can be interpreted as language tokens assembled according to the musical grammar used by the piece of music. The purpose of this step is to create a searchable database of songs containing only the relevant features. Each output file stores melody contour information for a particular instrument track in a song in three different formats: a 3-step quantization, a 5-step quantization, and a 7-step quantization. Running this module generated 1838 files corresponding to an average of 8 tracks per song.

A retrieval program allows users to enter a melody contour using textual input. Starting from a base note, users enter the value for the next note based on the following mapping (for a 3-step quantization):

/: go up
 \: go down
 0: stay the same

For a 5-step quantization, the parameters are the following:

- 0: stay the same
- /: go up by less than a perfect fourth (5 semitones)
- \: go down by less than 5 semitones
- +: go up by more than a perfect 4th
- : go down by more than 5 semitones

For a 7-step quantization, we introduce another limit at 7 semitones (above and below the reference note).

The retrieval program outputs song names, track numbers, and occurrences of the input sample in the song database.

2.4 Musical Enculturation

Figure 2-2 shows the interval distribution for popular Indian music. It indicates that 20% of the time, a note is played twice in succession. The graph also shows that around 50% of the information is gathered in the $[-5; 5]$ semitone interval. There is also a significant drop at 6 semitones (tritone), an interval rarely used in Indian music, but often heard in jazz. We notice another peak at -7 and 7 semitones, and then the curve drops significantly (higher intervals have little information to contribute).

We compare the interval distribution for Indian music with that of Western music (figure 2-3), which shows two significant peaks at -3 and 3 semitones (minor third), which is a common interval in Western music. Kim and his colleagues chose this interval to quantize their 5-step melody contour. The optimum choice of interval quantization is dependent on the content of the music. The appropriate interval for the melody contour model can be computed by analyzing the music source.

Figure 2-4 shows the results obtained by running a test over the complete database with random sampling (with at least one match). With each quantization step (3,

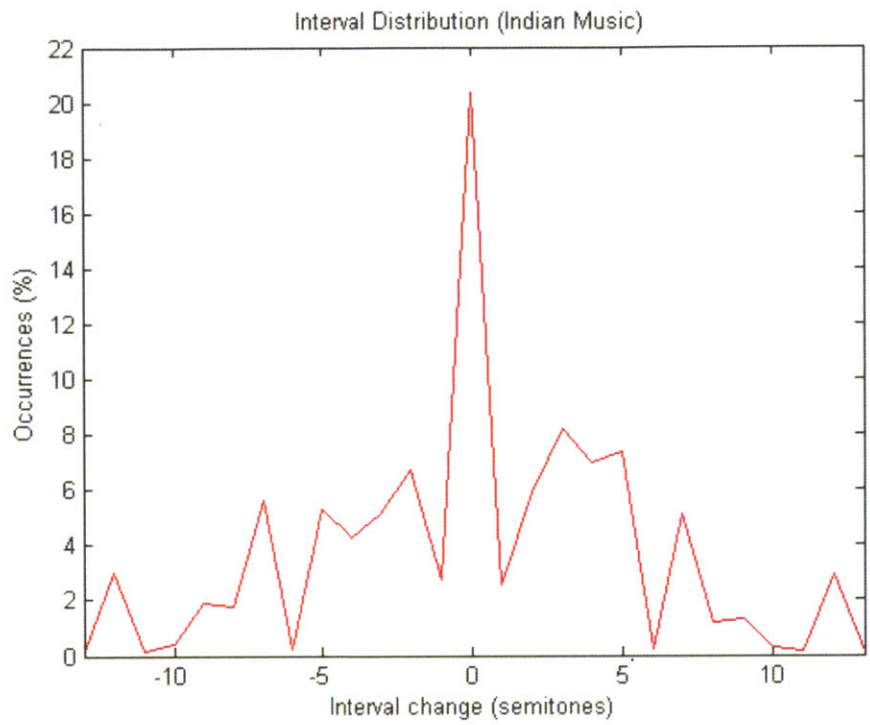


Figure 2-2: Interval distribution for popular Indian music

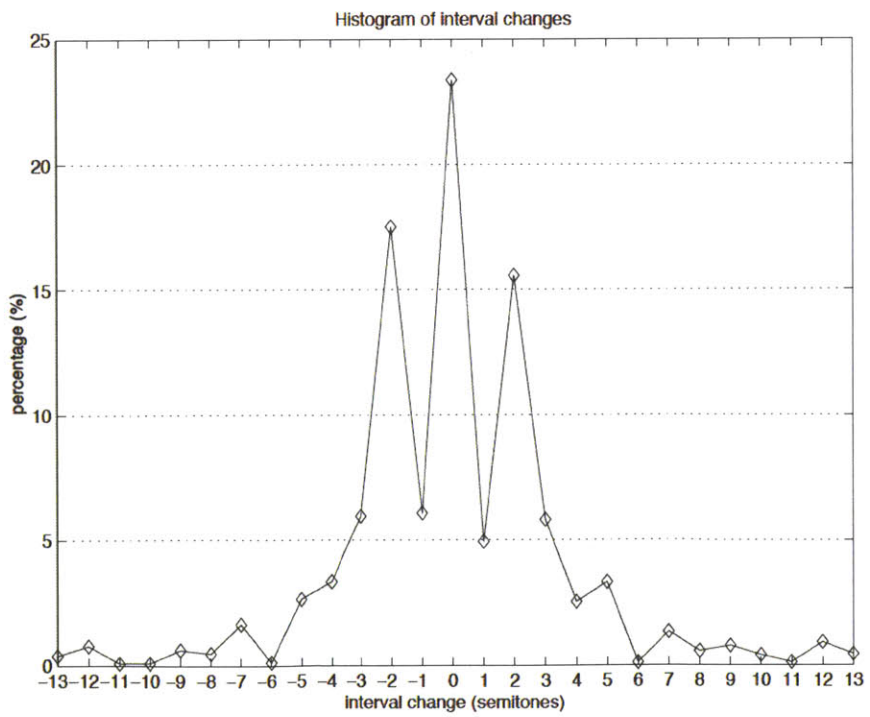


Figure 2-3: Interval distribution for popular Western music (Kim et al., 2000)

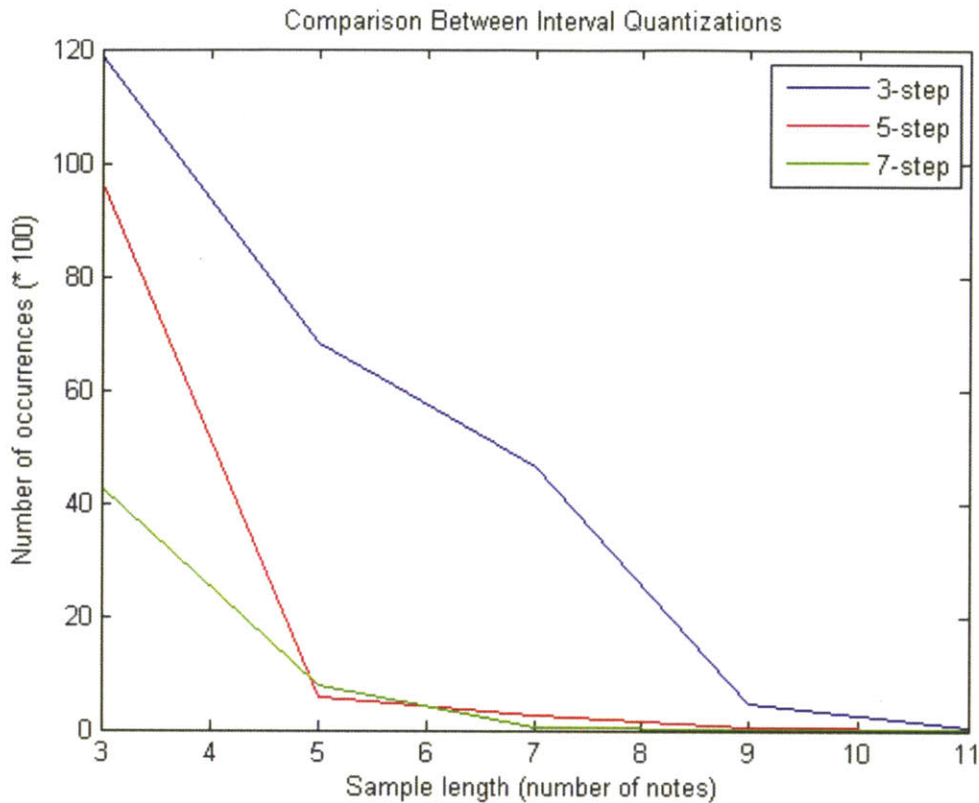


Figure 2-4: Melody retrieval for 3-step, 5-step, and 7-step interval quantizations

5 and 7), I used 25 samples of varying length (3 notes to 11 notes). The number of occurrences corresponds to the number of ‘hits’ in the song database of the sample melody under consideration.

We observe that for the three quantization steps an increase in the sample size (number of notes) results in a decrease of the number of occurrences in the database (we can presume that the system zeroes-in onto the right set of songs). At a sample size of 3 notes, the higher the quantization step, the better the model performs.

The important (and surprising) result here is that for a sample size of 5 notes, the 5-step quantization curve dips below the 7-step quantization curve. This strongly suggests that the best 5-step quantization model performs as well as the 7-step quantization model above sample sizes of 5 notes, which may be the average number of notes required to recognize a song.

This evidence points to the optimality of a model of intermediate complexity in melody contour for the purpose of music recognition.

This work is a significant step in the understanding of human melodic representation and recognition. It is a simplified model that takes into account only melody contour as a particular attribute of music. In reality rhythm and other dimensions also contribute in discriminating music.

2.5 Measuring Musical Culture

To ground this work on an analytical framework, I introduce a measure of entropy to compute culture footprint in melody. Entropy, or information content, in this context refers to the capacity for a listener to predict musical events as a function of prior musical exposure, or enculturation. Using this model we compare pieces of popular Indian music with popular Western music. We explore how learnt musical structures influence predictions about musical events, and how much information is required to identify and classify a piece of music in a particular culture.

The result of this experiment is the computation of an asymptotic entropy vector based on the level of enculturation or exposure, and hence on the level of generalization achieved by the system. The entropy vector describes how well the next event can be predicted by quantifying the information content of the music.

In this measure of entropy, priors play the role of cultural context: we train the model in culture A and predict in culture A, and then train in culture B and predict in culture A.

Table 2.2 reports on the entropy values for music from different cultures based on the enculturation model from priors.

Table 2.2: Musical entropy based on cultural priors

Source	Context	Entropy
Indian	Indian	0.37
Indian	Western	0.59
Western	Western	0.31
Western	Indian	0.65

These results suggest that Indian melody is more complex than Western melody. In linear note-to-note sequences, there is more uncertainty in Indian music than there is in Western music. From a musicology perspective this makes sense. Complexity in popular Western music is found in chords and harmonic progressions (the ‘vertical’ component) whereas Indian music, popular or traditional, is concerned with ‘horizontal’ intricacies.

Results from table 2.2 support our hypothesis when taking priors into account. We find that Western music is more predictable when trained on Western music than when trained on Indian music. And the same holds for Indian music. This suggests that enculturation does indeed create mental models that help with music prediction in a particular genre. *Music is not universal*: people best understand the music they have been exposed to.

We find that music can be predicted to some extent. Based on the entropy values we find, the amount of uncertainty is significantly lower than chance. By taking into account musical structure of varying length we suggest that it is possible to make informed decisions as to the continuation of melody and other musical constructs, based on appropriate priors.

These findings support the hypothesis that music complexity (i.e predictability) is a function of musical culture:

$$music_complexity = f(musical_culture)$$

and provide the rationale for designing a music prediction engine.

Chapter 3

Culturally Inspired Models for Music Representation

3.1 Music Representation by People and Machines

Even though a music representation language like Humdrum Toolkit's `**kern` (Huron, 1993) is significantly ahead of a digital music format like MIDI (Musical Instrument Digital Interface) in terms of expressivity and accuracy, it is biased towards the constructs of Western music. Although characteristics of non-Western music like pitch inflections found in Indian music can be approximated, both in `**kern` and in MIDI, they do not have a *native* and formal representation. On this basis I argue that a music representation model must be specified for a particular musical tradition.

In *You Are Not a Gadget: A Manifesto* (2010), Lanier tellingly says: “[MIDI] could only describe the tile mosaic world of the keyboardist, not the watercolour world of the violin.”

When represented as a signal, digital music is culture-agnostic. Pulse Code Modulation and perceptual coding (e.g. MP3) for instance treat music as any other sound

source. On the other hand symbolic representations like MIDI or MPEG-4's Structured Audio Orchestra Language (Scheirer and Vercoe, 1999), which may have been specified for music 'in all its forms' actually embed constructs specific to Western music. Prime among them is the concept of a note. It is generally assumed that a note can be described as a single pitch value sometimes with associated parameters for vibrato or ornamentations. It is common practice in Indian music however to consider a note as a *pitch contour* (called *gamaka* in South Indian Carnatic music).

Here we consider music as event-based, rather than sample-based. A format like `**kern` is a syntactic description language. For music from various traditions I propose a lexical representation that should automatically recognize tokens from audio and identify them as part of a specific genre. The tokens considered here are low-level elements such as notes, or drum strokes, and associated pitch inflections.

The computational representation template that I propose is inspired by the mental representation of musical structures, which is itself informed by the elements of the musical traditions one has been informally exposed to through *enculturation*.

In the process of enculturation, mental structures are formed according to the following process that starts when hearing musical material:

1. Segmentation and identification of musical primitives
2. Adaptation of container structure to capture musical primitives
3. Adaptation of hierarchical musical grammar to capture musical constructs (i.e. structured primitives)
4. Adaptation of higher-level cognitive structures for appreciation, expectation, skill, etc.

Rather than developing systems that learn musical representations from scratch by creating containers adequate for specific musical content, the present work concerns

itself with designing appropriate representations that can be used by machine listening systems to perform auditory and musical tasks like humans do.

Even though sound is a succession of discrete events in time, music is explicitly built with repetitive structures (e.g. a regular rhythmic pulse, a form that alternates verse, chorus, and bridge) to allow people to develop mental representations at different scales and to derive meaning from music. Minsky (1981) considers that a piece of music explores a musical space like mental map-making—in metaphorical terms, it starts at one location (theme), explores some (e.g. chorus), then heads back home (theme) before venturing back further (variation). Minsky distinguishes *knowing*, or memorizing, from *understanding* (i.e. representing a concept or an idea from different angles so they can be ‘thought’ about). Applied to music, this definition of ‘understanding’ leads to the storage of musical structures of “intermediate complexity”—similar to visual recall—like is described in chapter 2.

To clarify our taxonomy we define pitch, amplitude, timbre, localization, etc. as “low-level” musical primitives, and musical phrases, forms, and style as “high-level” constructs. Like other cognitive tasks, music functions by pattern recognition in the brain. The learning of patterns happens through enculturation, or informal exposure, but also depends on a proclivity to learn that may be, according to Huron, the result of an adaptive function (i.e. the Baldwin Effect).

Music is a source of pleasure and strength, of ritual, joy, and laughter. It is the carrier of emotion and moves us into various affective states. Bernstein (1976) attempted to transpose Chomsky’s universal grammar for language to music. While his results were mitigated—maybe partly because the idea of innate knowledge for language is more widely accepted than for music—Jackendoff and Lerdahl (1996) devised a generative theory for tonal music (GTTM) that illustrates the recursive and self-similar structures of Western tonal music. The genetic basis of the neural structures in the brain enables us to handle music, like we do language, and provides the motivational drive for it, in particular with its link with emotion. In contrast,

animals probably have the capability to perceive perceptual features like pitch, but may not have the motivation to do so, unless it is used for some survival task—information transmission, in the case of birdsong.

I would argue that the relevant parameters to distinguish the adaptive or otherwise functional roles of music are based on their processing level (physical to cognitive, see table 1.1) rather than their category. For instance, the physical parameters of frequency, amplitude, and spectrum, and their perceptual correlates as pitch, loudness, and timbre (perception of auditory objects through spectral fusion), may be derived from an evolutionary mechanism because they result from the processing of auditory events that are demonstrably required for survival. On the other hand, higher-level mental representations of these parameters as scales (and melodic contours), dynamics, and instrument identity are culturally motivated and not adaptively evolved. Similarly, timing as a low-level (and adaptive) construct leads to mid-level beats and tempo, and high-level meter, phrases, and forms.

This dichotomy hardly lets us speak of ‘adaptive’ versus ‘non-adaptive’ music; most music has components of both the adaptive and the contextual, or environmental. However some types of music may be approximated to one category or another. For instance, community drumming encountered in the drumming circles of Africa or Australia has many low-level features involved, and few high-level structures. The meaning conveyed by this type of music is often linked to its social role as mentioned previously (which plays, arguable, an adaptive function). On the other hand, music with little dynamic changes could be considered as ‘non-adaptive’ where the meaning of music is often linked to the emotional response it elicits from the listener, often putting him or her in a pleasurable state of relaxation or meditation.

Keeping some of our assumptions and approximations in mind, we find that there are indeed some musical forms that can be categorized as ‘adaptive’ and ‘non-adaptive’ based on the level of musical parameters they draw on. It is helpful in this context to look at how some authors define proto-music: multimodal (with sound and

movement), dynamic (temporal, rhythmic, melodic elements), mimetic (containing sound symbolism, gesture), holistic (with no segmented elements), and manipulative (capable of influencing emotional states, behaviors).

The brain does not store sounds. Instead, it interprets, distills, and represents sounds. It uses a combination of several underlying representations for musical attributes. But how does the brain know which representation to use? Huron (2006) suggests that expectation plays a major role. There exists evidence for a system of rewards and punishments that evaluates the accuracy of our unconscious predictions about the world. Mental representations are being perpetually tested by their ability to usefully predict ensuing events, suggesting that competing and concurrent representations may be the norm in mental functioning (Cont, 2008).

The issue of music representation is a complex and challenging one. While many elements of music, like rhythm and harmony, are mathematical constructs, which computer languages are well suited to represent, many other, less tangible, elements form other essential dimensions of music. Dannenberg (1993) argues that the knowledge we gain from specifying as completely as possible a language for musical description provides us with insights into music theory and music understanding by machines, but also into music cognition. Current music representations range from highly symbolic and abstract music notation schemes to concrete recorded audio signals. Intermediate to these are languages that explicitly describe a set of musical parameters, such as Csound or MIDI.

In spite of its shortcomings, MIDI is well suited to the study of melody. The main argument that favors it is the ready availability of musical pieces of various cultures for download on the Internet (albeit of varying quality). According to Dannenberg, one of the main limitations of MIDI is its lack of structural relationships between musical elements; in particular MIDI considers each note as independent. However, this is compensated for by the implicit grouping mechanism built in the 16 channels that segregate instruments and their control parameters. Moreover if we assume melody

to be a Markov chain with no priors, we do not need to take structural relationships into account.

3.2 Indian and Western Musical Traditions

In the present work we compare existing models suited to Western music with new ones that we develop for Indian music.

3.2.1 Percussions

In the case of percussion we analyze the tabla, a pair of hand drums from North India. In my master's thesis I write about the tabla: "They are played with the fingers and palms of both hands. The right drum (from a player's perspective) produces a high pitched sound, whose pitch can be tuned with the rest of the orchestra. The left drum produces bass sounds with varying pitches depending on the pressure applied on the drumhead with the palm of the hand. The tabla can play a variety of different sounds, both pitched and unpitched. Each of these sounds, called bol, has a syllable associated with it. Thus rhythmic compositions can be [vocalized], and transmitted in an oral tradition."

Some of the questions that are raised are:

- How well can different bols be discriminated by a human listener?
- How well can a machine automatically classify tabla strokes compared to a human musician?

Tabla stroke recognition can be interesting both for music information retrieval in large multimedia databases, or for automatic transcription of tabla performances. In my case, I look at it from the point of view of the representation of tabla strokes.

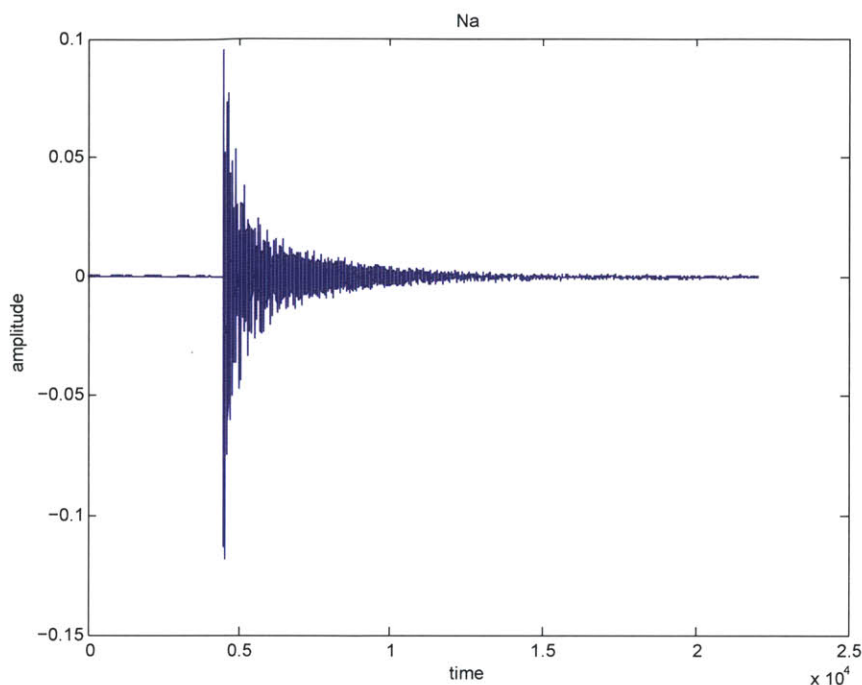


Figure 3-1: Tabla stroke waveform: na

Figures 3-1 to 3-5 show the time-domain waveforms of various strokes (more than 10 different strokes can be played). *Na* and *Tin*, which sound—and look—quite similar, are played on the right hand drum. *Dha* is played with both hands, a combination of *Na* and *Ga*. *Ka* is a damped sound played on the bass drum.

Spectrograms of the strokes are shown in figures 3-6 to 3-10. Additional information can be inferred from these: for instance the *Na* has a clear steady pitch, while *Ka* consists mostly of transitory noise.

In my previous work, I extracted Power Spectral Density (Welch’s method) features and reduced vector dimensionality using Principal Component Analysis for stroke recognition. Using a k-Nearest Neighbor algorithm I achieved above 90% recognition rate (as compared to 87% for human recognition with no contextual information surrounding the strokes).

The confusion matrix 3.1 for automatic recognition rate is interesting because it

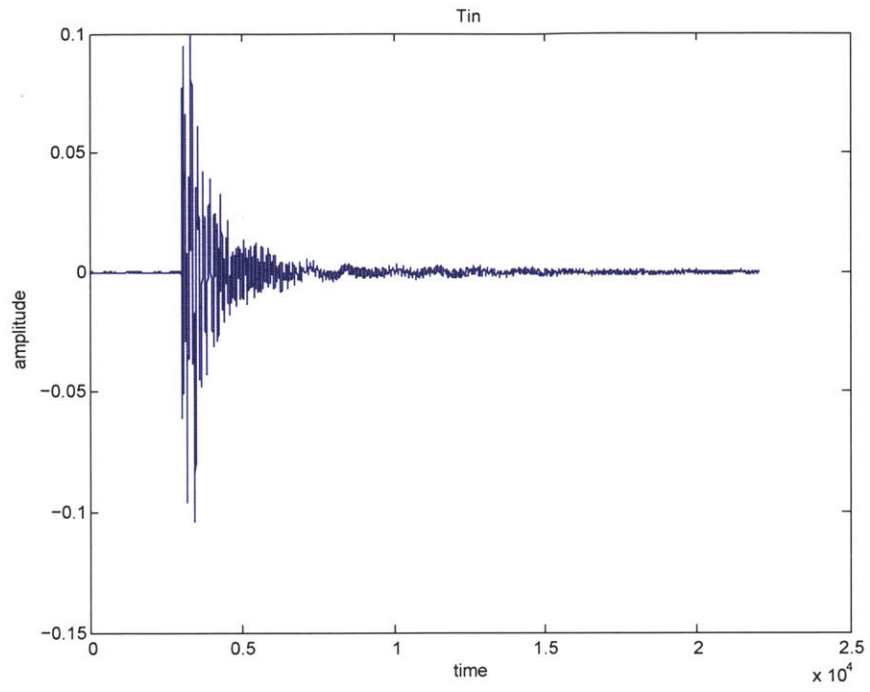


Figure 3-2: Tabla stroke waveform: tin

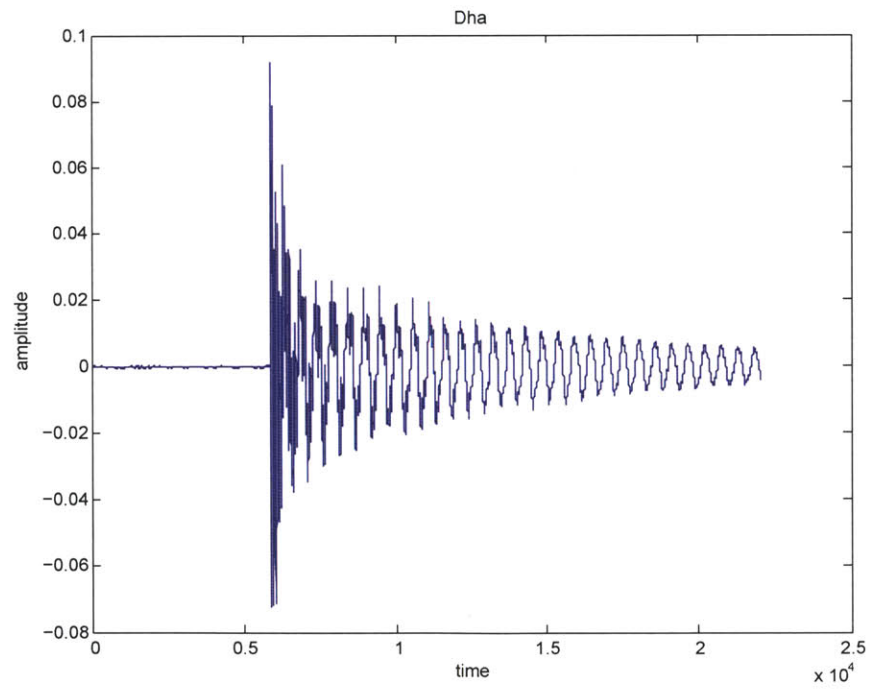


Figure 3-3: Tabla stroke waveform: dha

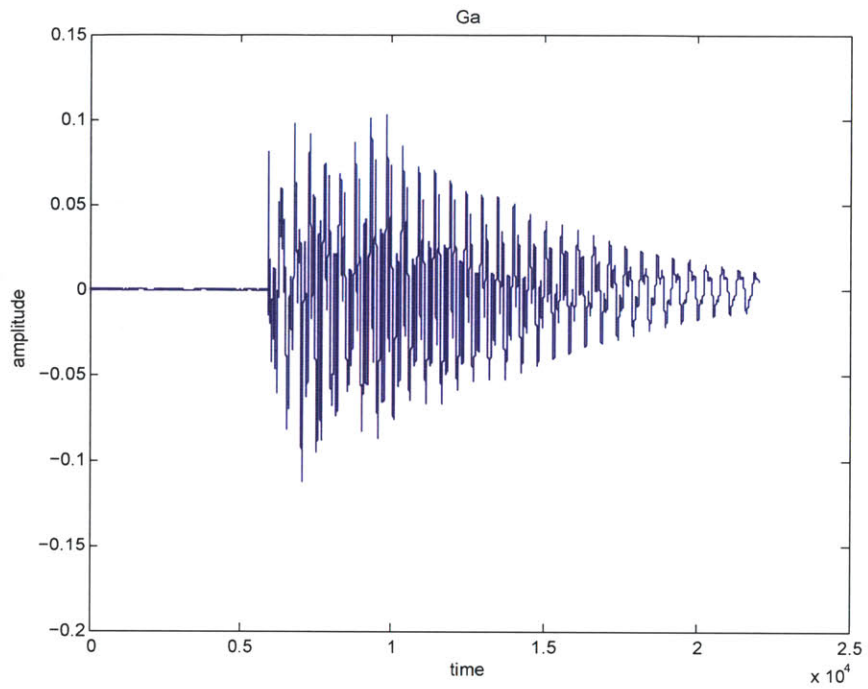


Figure 3-4: Tabla stroke waveform: ga

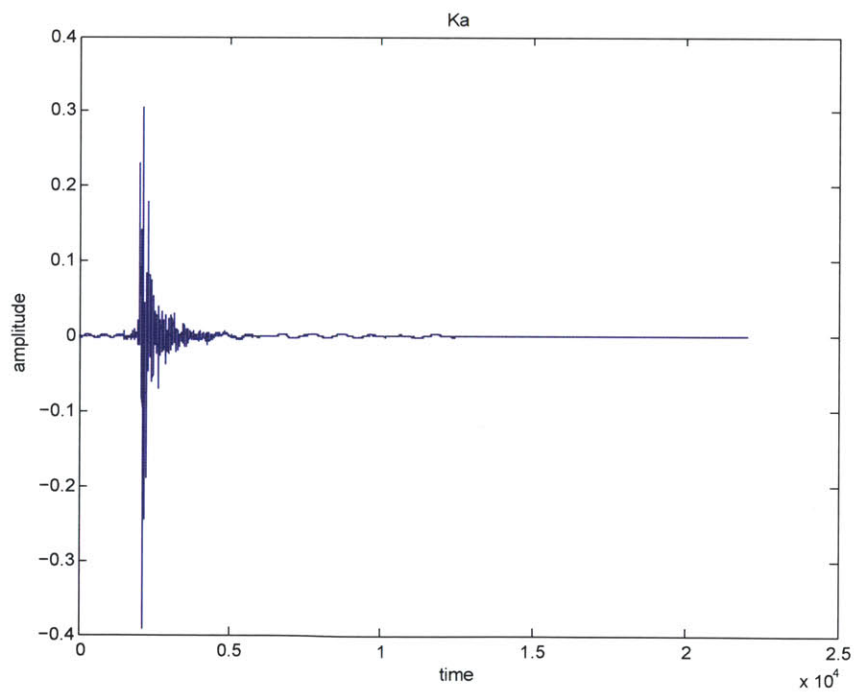


Figure 3-5: Tabla stroke waveform: ka

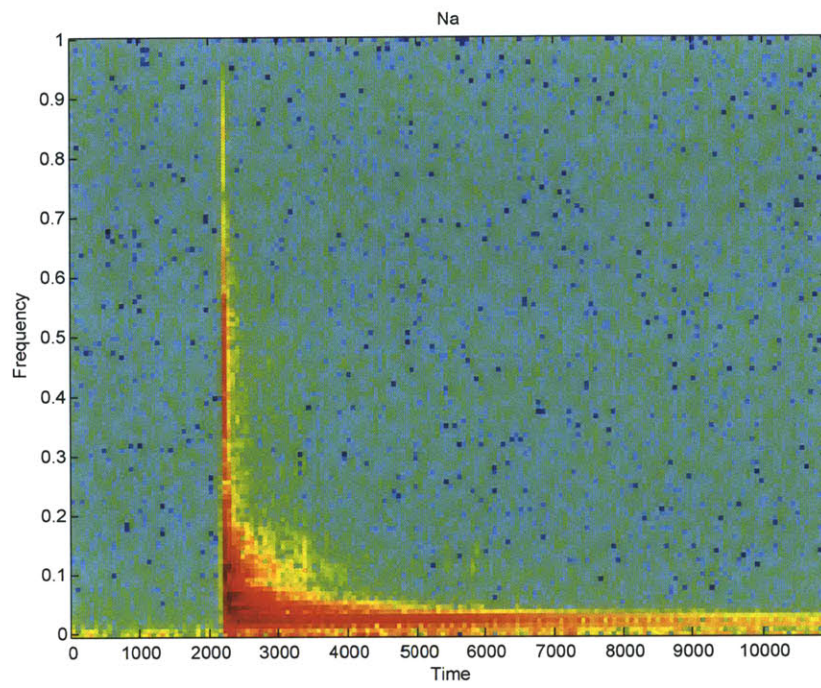


Figure 3-6: Tabla stroke spectrogram: na

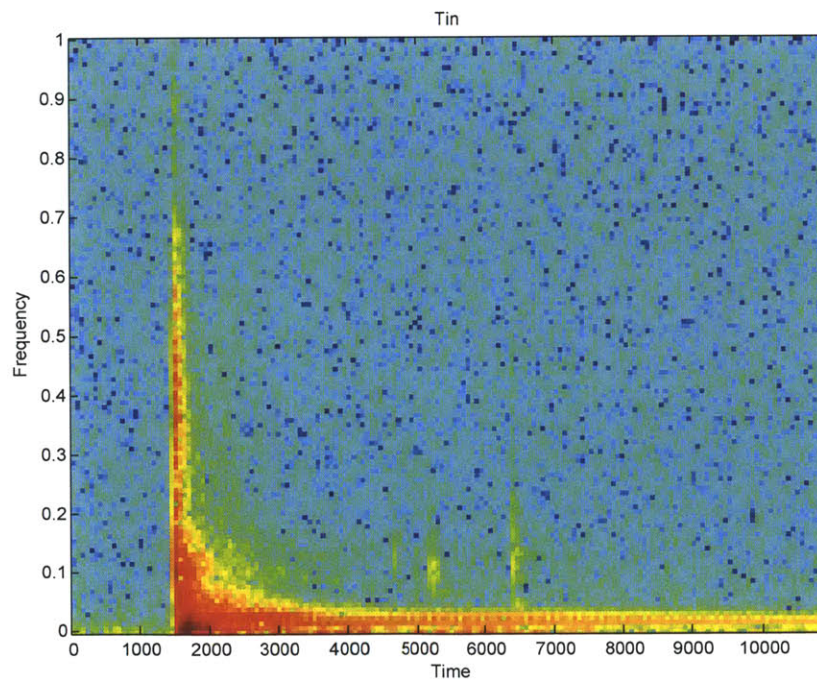


Figure 3-7: Tabla stroke spectrogram: tin

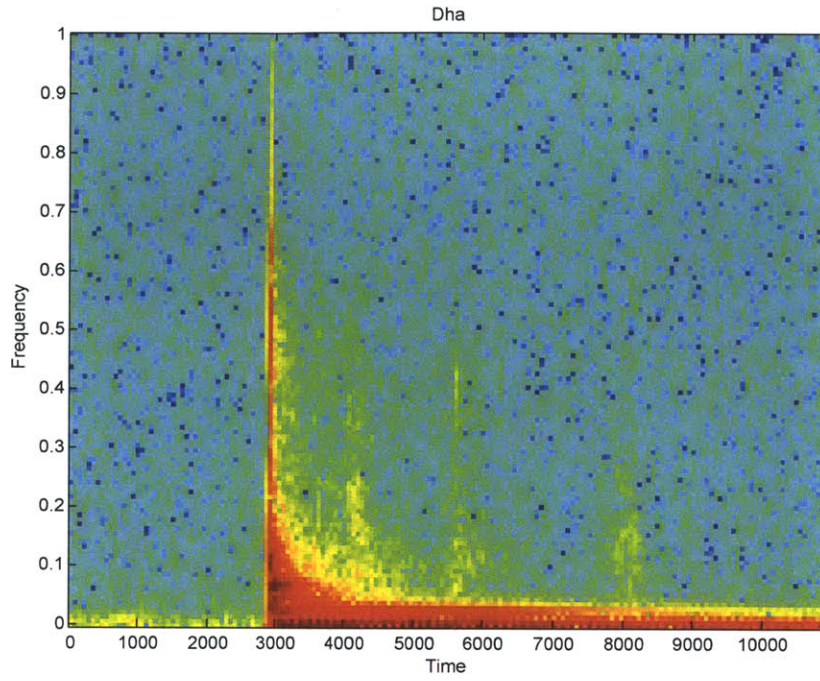


Figure 3-8: Tabla stroke spectrogram: dha

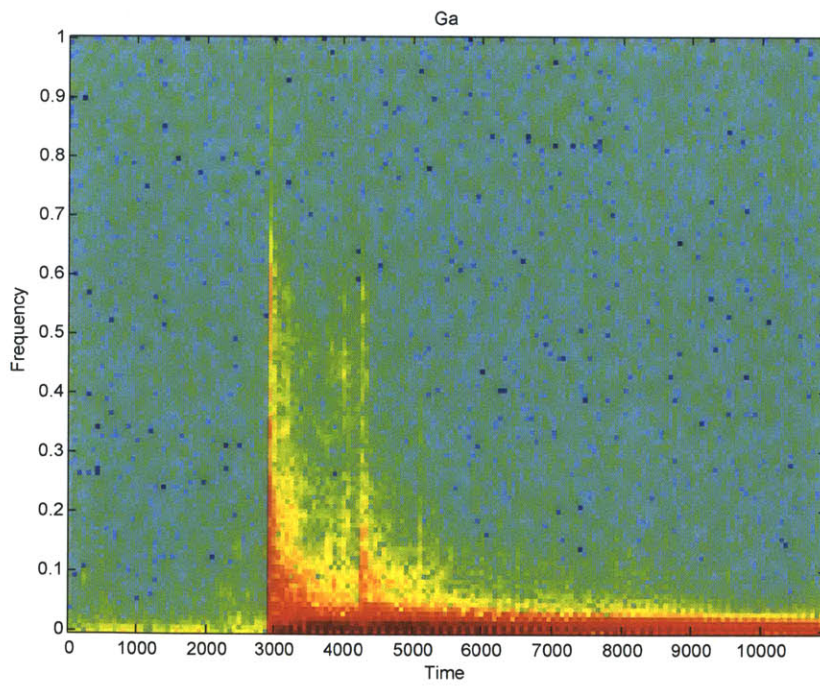


Figure 3-9: Tabla stroke spectrogram: ga

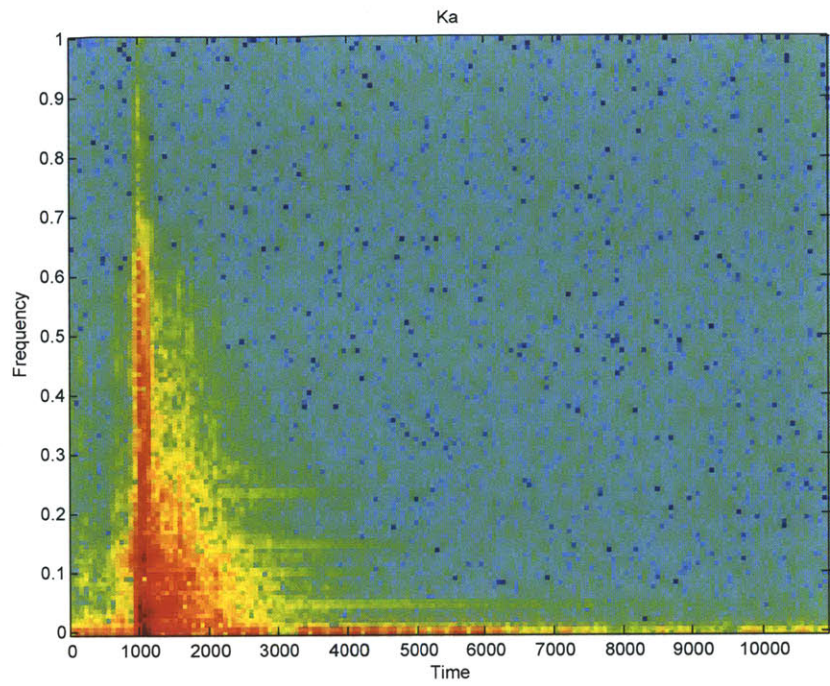


Figure 3-10: Tabla stroke spectrogram: ka

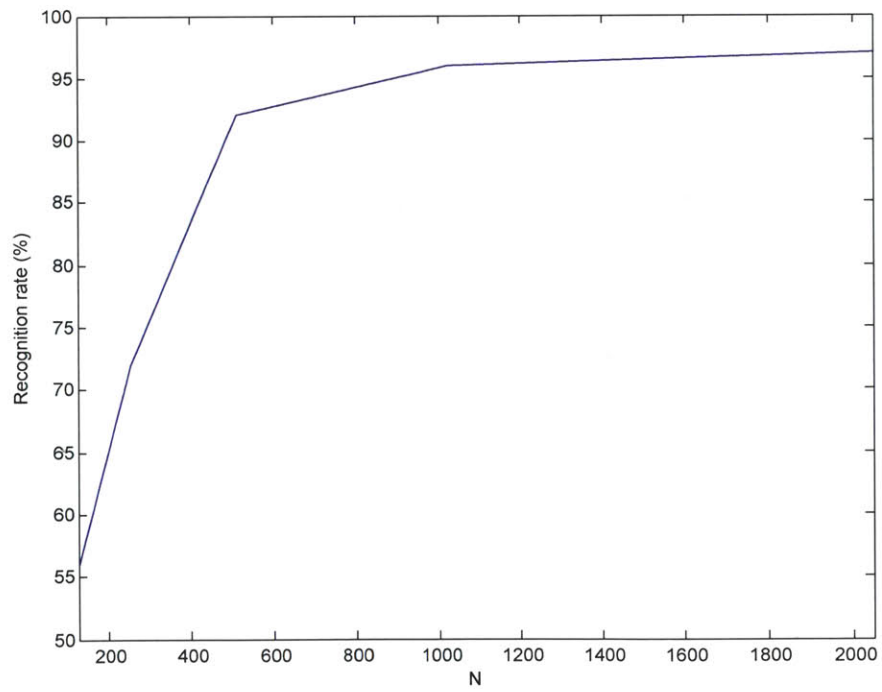


Figure 3-11: Tabla stroke recognition rate for varying FFT lengths

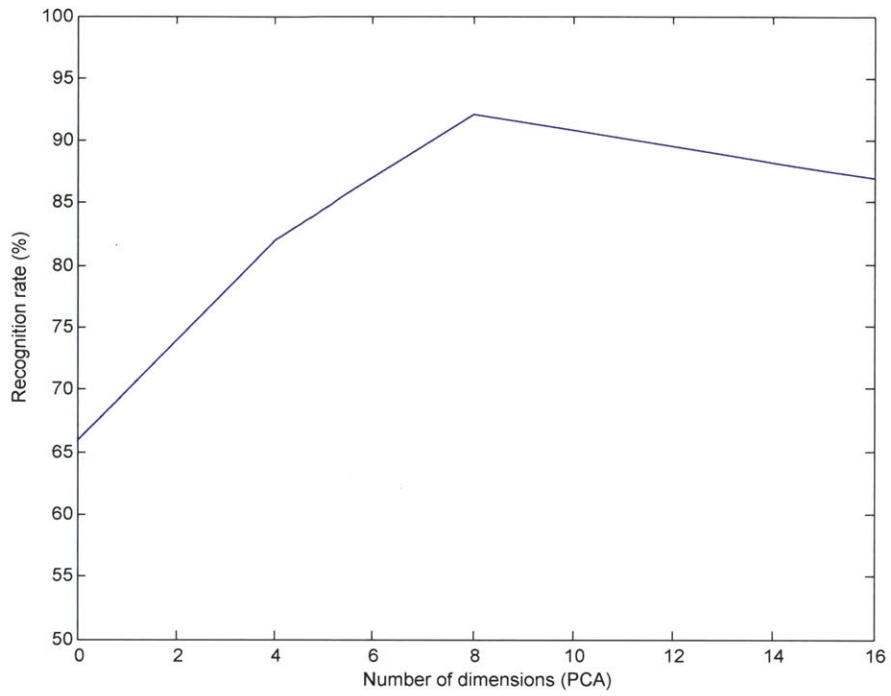


Figure 3-12: Tabla stroke recognition rate for varying dimensions

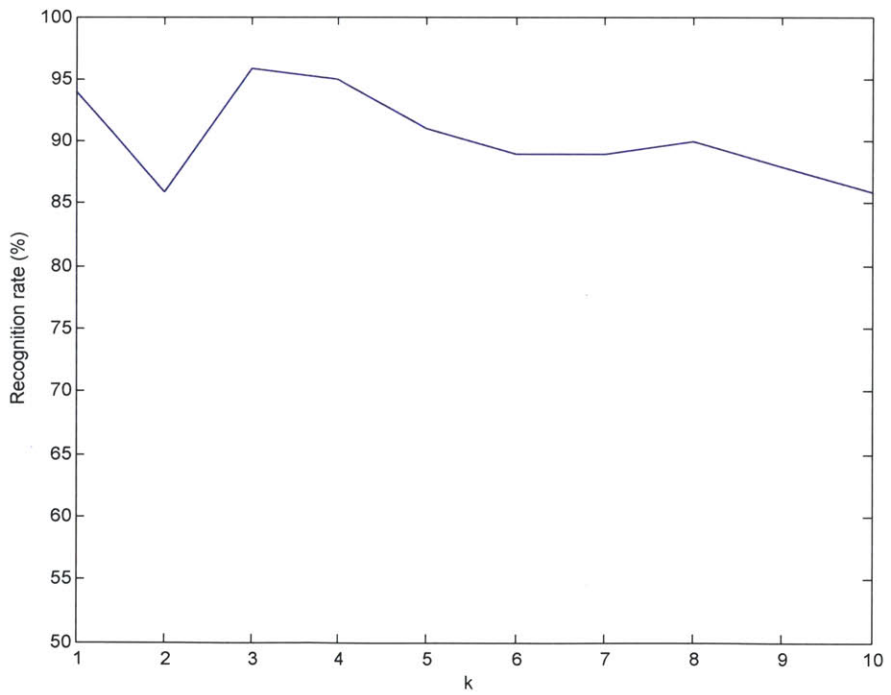


Figure 3-13: Tabla stroke recognition rate for varying k

Table 3.1: Confusion matrix for tabla stroke recognition

	Na	Tin	Ga	Ka	Dha	Dhin	Te	Re	Tat	Thun
Na	5	0	0	0	0	0	0	0	0	1
Tin	0	3	0	0	1	0	0	0	0	2
Ga	0	0	4	0	0	2	0	0	0	0
Ka	0	1	0	3	0	1	0	1	0	0
Dha	0	0	2	0	2	1	1	0	0	0
Dhin	0	0	1	0	1	4	0	0	0	0
Te	0	1	0	0	0	0	1	0	4	0
Re	0	0	0	0	0	0	1	4	0	1
Tat	0	0	0	0	0	0	2	1	3	0
Thun	0	1	0	0	0	0	0	0	0	5

matches the results for human stroke identification.

Two observations point to the unsuitability of a traditional computational model of Western drums to the tabla.

First, the window size used for detecting each stroke from the onset time is specific in that it has to account for strokes of longer durations due to pitch inflections (by applying continuously varying pressure to the drum head, usually on the bass drum) and by treating strokes with a quick succession of multiple onsets as one single stroke with its associated *bol*. Frame size has a considerable influence on recognition rates. In the case of a Western drumset individual drum sounds or strokes are of finite and limited duration. In the case of the tabla, varying the frame size increases recognition accuracy until a threshold (750ms in the dataset used) where the next stroke takes place. Therefore windows of varying lengths have to be implemented.

Second, is the presence of pitched sounds and pitch bends. Tabla sounds can therefore not be synthesized accurately with simple wavetable-type synthesis, like drum sounds often are. I use waveguide synthesis.

In contrast, drums used in blues music are modeled using a General MIDI synthesizer.

3.2.2 Winds

Indian music is based on the multi-dimensional concept of *raga*. Here I am especially concerned about the notion of *gamakas* that are embedded into them. Gamakas are ornaments (quite different from the embellishments of Western music) that form an integral part of Carnatic music. They are characterized by micro-tonal oscillations and variations.

A raga is a musical concept or system that comprises, among other things, a scale or mode, gamakas, characteristic phrases, and visiting notes. “Raga is a vast ocean of latent potential, waiting to be realized” (Allen, 1998).

Modeling gamakas are therefore an essential step in modeling ragas. It is important to note that there exists classification schemes for gamakas. Depending on the author, the number of categories ranges from 23 to 15 to 10 to 3 (Swift, 1990). The present work does not aim to provide fixed templates for gamakas, but rather to provide an extensible language to represent gamakas with the ability for users to add their own types.

I analyzed an excerpt from the recording of raga *Hamsadvani* sung in the context of the *Pagavari varnam* (singer Kamala Ramamurthy recorded by Richard Wolf in Madurai, Tamil Nadu India, 1985). Raga Hamsadvani comprises the following notes: do re mi sol ti do (both in the ascending and descending scale).

To analyze the recording, I created a pitch-tracker patch on Max/MSP using Jehan’s pitch object. The collected samples were then imported into Matlab for alignment and plotting. I performed an analysis on the pitch contour after having aligned each musical entity. The estimated pitch is scaled according to the MIDI pitch number to maintain linearity. Figures 3-14 to 3-25 plot the pitch contour of various notes of the raga. The y-axis represent pitch-equivalent MIDI note values (60 is the middle C). Gamakas are clearly visible.

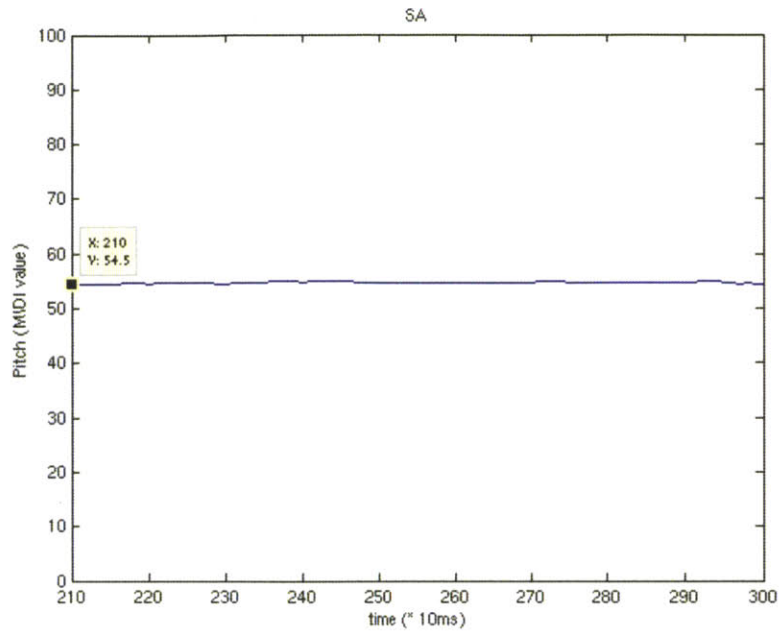


Figure 3-14: Pitch tracking of note in Carnatic music: sa (ascending)

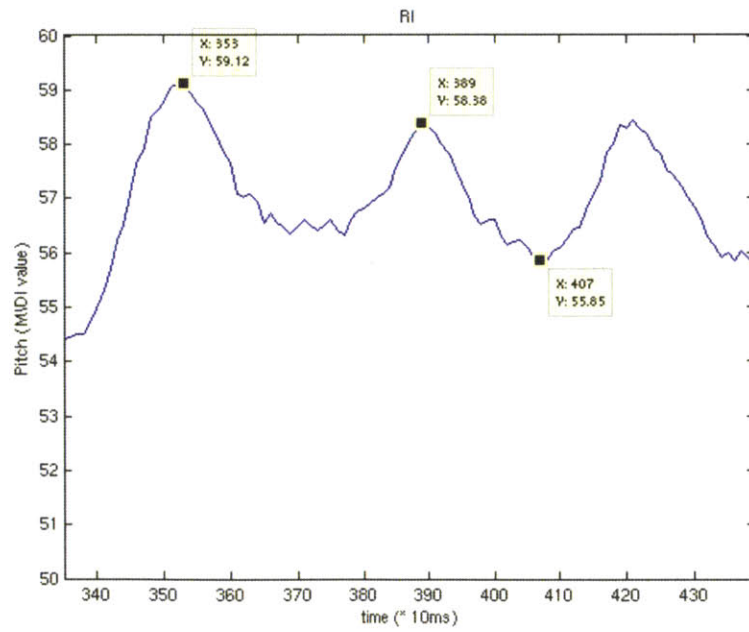


Figure 3-15: Pitch tracking of Carnatic music: ri (ascending)

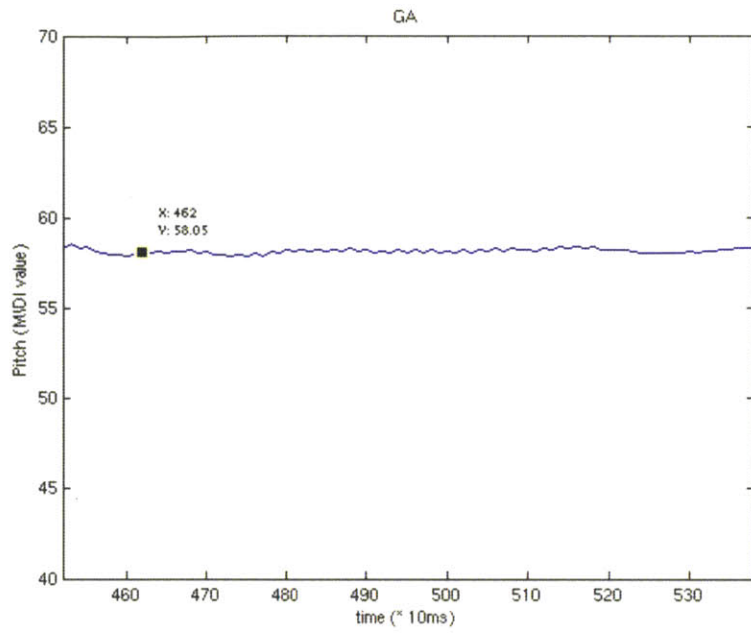


Figure 3-16: Pitch tracking of Carnatic music: ga (ascending)

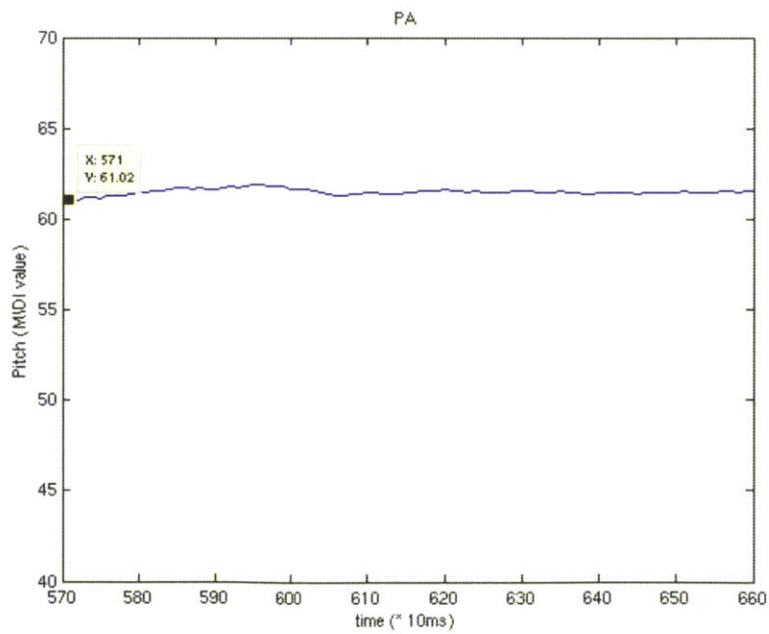


Figure 3-17: Pitch tracking of Carnatic music: pa (ascending)

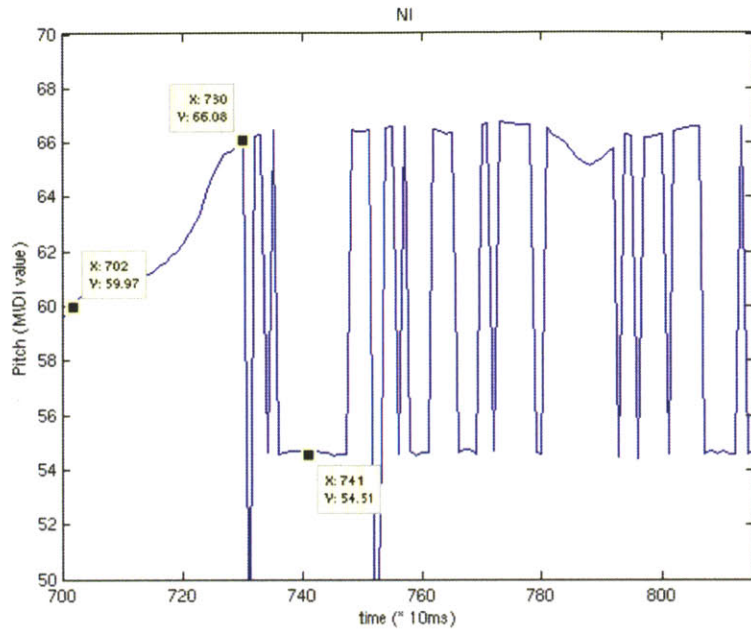


Figure 3-18: Pitch tracking of Carnatic music: ni (ascending)

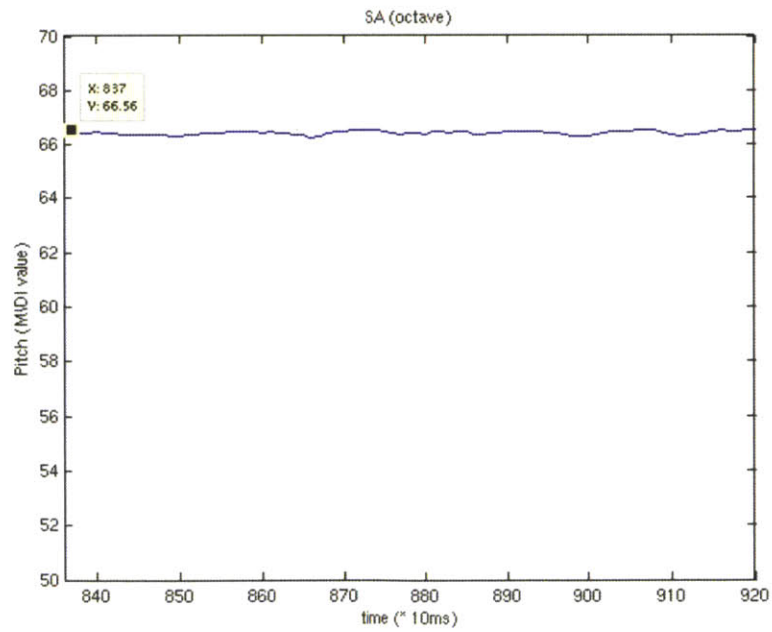


Figure 3-19: Pitch tracking of Carnatic music: sa' (ascending)

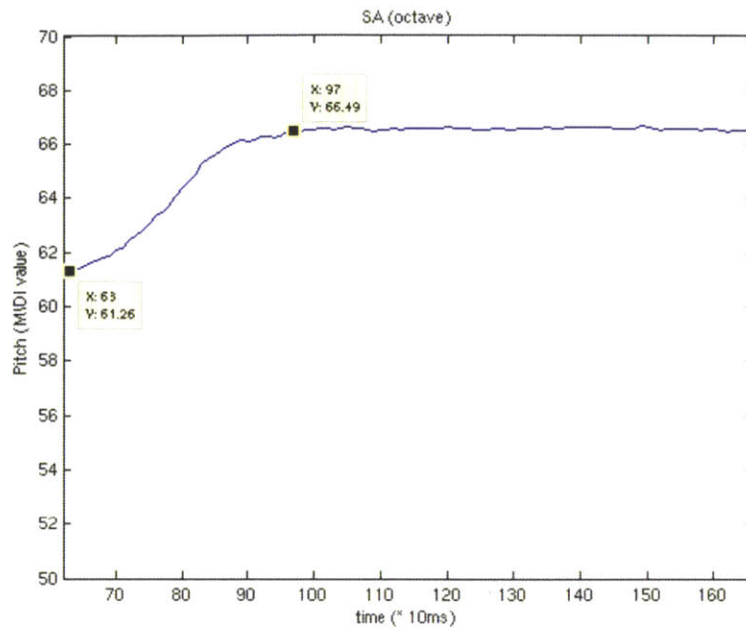


Figure 3-20: Pitch tracking of Carnatic music: sa' (descending)

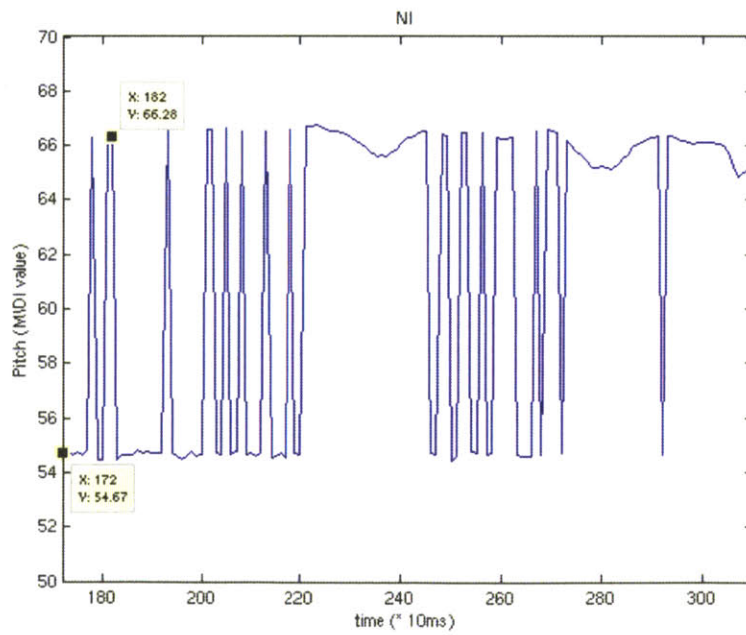


Figure 3-21: Pitch tracking of Carnatic music: ni (descending)

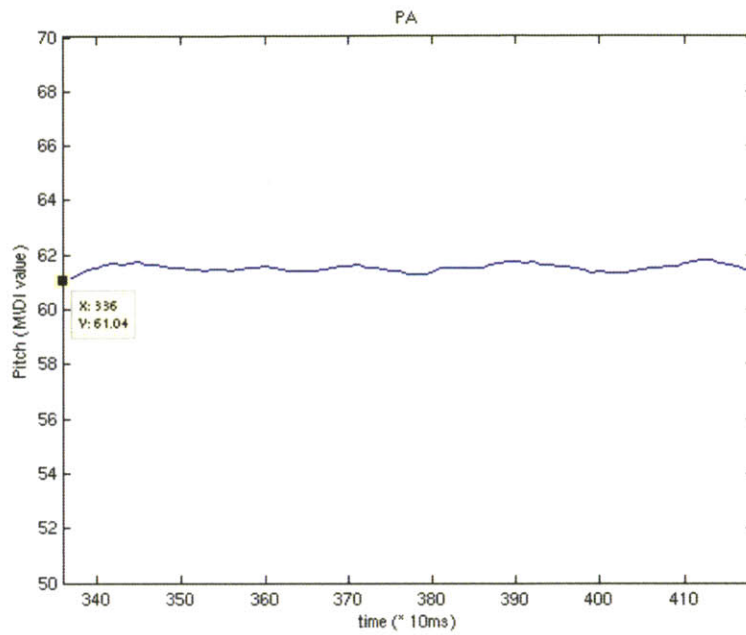


Figure 3-22: Pitch tracking of Carnatic music: pa (descending)

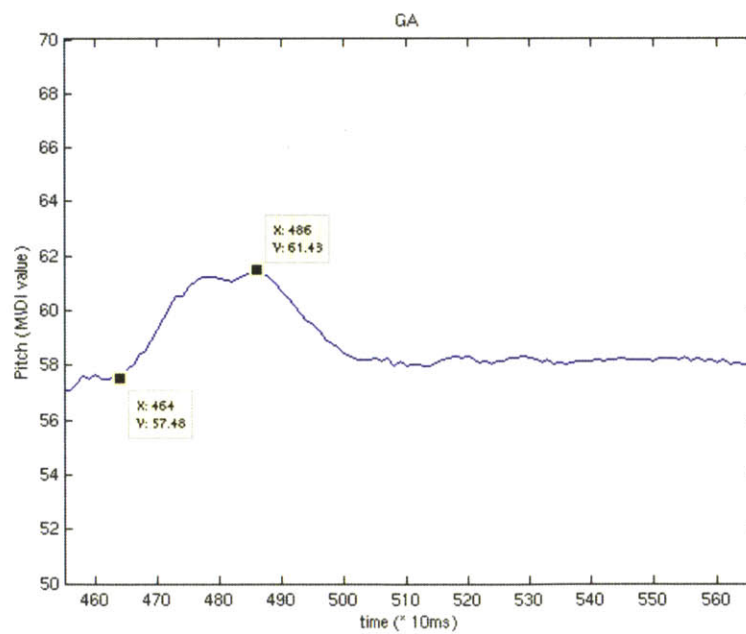


Figure 3-23: Pitch tracking of Carnatic music: ga (descending)

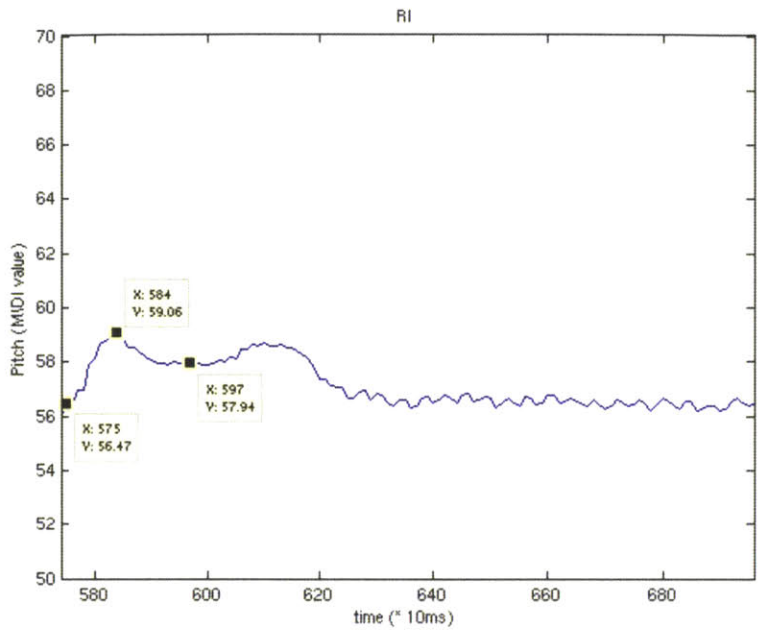


Figure 3-24: Pitch tracking of Carnatic music: ri (descending)

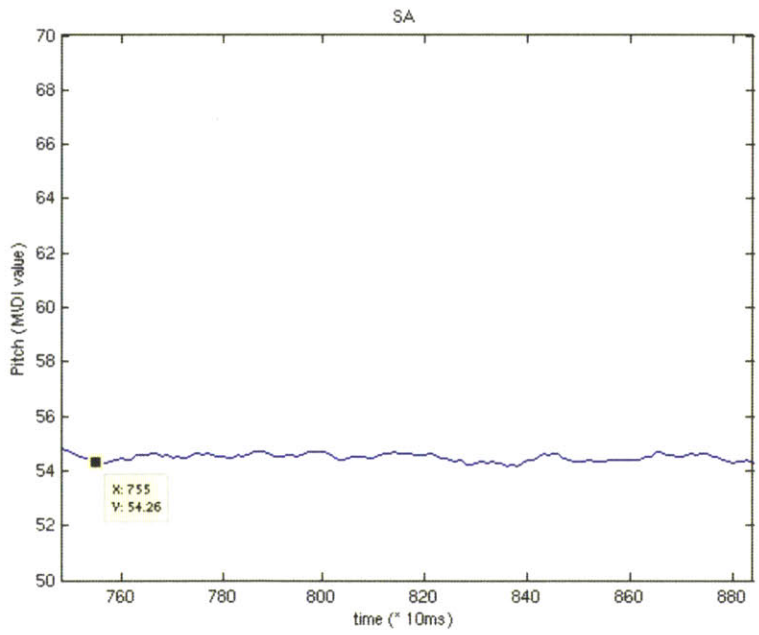


Figure 3-25: Pitch tracking of Carnatic music: sa (descending)

Table 3.2: Swaras (note entities) in Carnatic music

Note (Carnatic)	Note (Chromatic)	Frequency (Hz)	MIDI pitch
SA	C	261.626	54
RI	D	293.665	56
		311.127	57
GA	E	329.628	58
		349.228	59
		369.994	60
PA	G	391.955	61
NI	B	493.883	65
SA	C	523.251	66

It is important to observe that some notes (*swaras*) have a different pitch contour whether they are ascending or descending. The pitch tracker has trouble with the *Ni* and jumps an octave down during the analysis window. This can be controlled for.

To synthesize the South Indian bamboo flute (*Pulangoil*), I use the Csound programming environment. To model gamakas I took advantage of function-tables (i.e. look-up tables) to create cubic spline curves that match the analyzed pitch contour for each note. Each gamaka is considered as a separate instrument (in Csound terminology), and can therefore be triggered anytime in the score file. Synthesis is done using waveguide synthesis (adapted from Perry Cook’s instruments).

Table 3.2 describes various representations of the note entities used in the synthesis process.

Figure 3-26 shows the pitch tracking result of the synthesized version of raga Hamsadvani. Despite some artifacts gamakas can be observed.

Notes in Indian music (both Carnatic and Hindustani) have different instantiations on ascent and descent. Moreover, pitch inflections are an inherent part of their identity and need to be encoded along with pitch values. It follows that a representation of notes in Indian music has to have more than a pitch value. We introduce the idea of

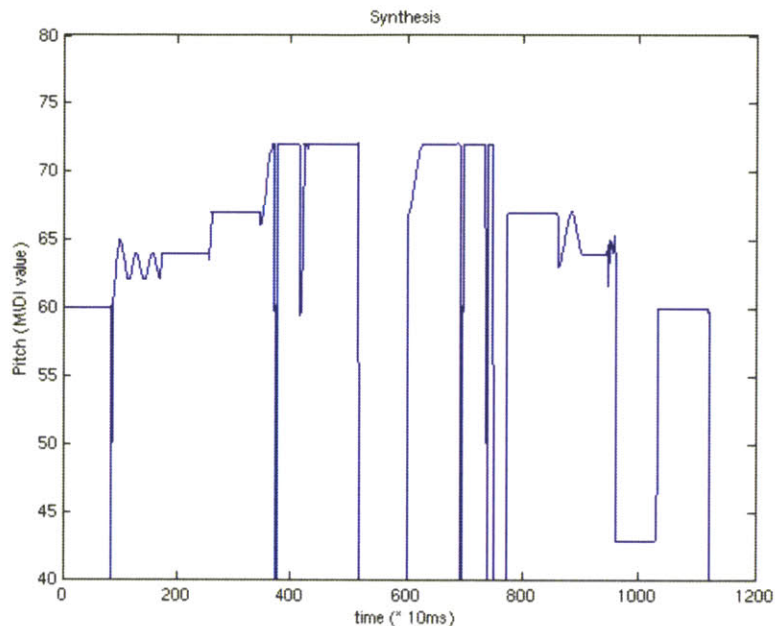


Figure 3-26: Synthesis of Carnatic music: raga Hamsadvani ascending and descending

a data structure for musical elements in order to represent musical entities that are not adequately captured with the models used in current music technology.

For the Western concert flute we use simple MIDI notes and a similar Csound instrument.

3.3 Culturally Appropriate Containers

What is an appropriate representation of music?

Based on the previous analysis, it may seem evident that a generic representation, or one that is biased towards one particular musical tradition, will perform poorly when capturing music from a vastly different tradition. We test this idea with listening experiments in section 3.4.

The representation template for musical elements is specified as an object-oriented

data structure that is designed specifically for a particular tradition.

For instance in the case of Carnatic music we propose the following structure for a note:

```
struct carnatic_swara {
    pitch_t pitch;
    raga_t raga;
    bool ascend;
    gamaka_t gamaka; // gamaka template
    gamaka_t transition_from;
    gamaka_t transition_to;
}
```

3.4 Representation and Recognition of Musical Elements

We test the hypothesis that computer music systems that incorporate a culturally appropriate symbolic representation provide a perceptually equivalent output to a signal-level representation of the musical source:

$$\textit{synthesis}(\mathbf{appropriate_symbolic_representation}) \equiv_{\textit{perception}} \textit{signal_representation}$$

The music prediction system is evaluated on the following musical data:

- Pulangoil in Carnatic style
- Western flute playing Blues

The data is sourced from user-generated websites like YouTube. Excerpts of music

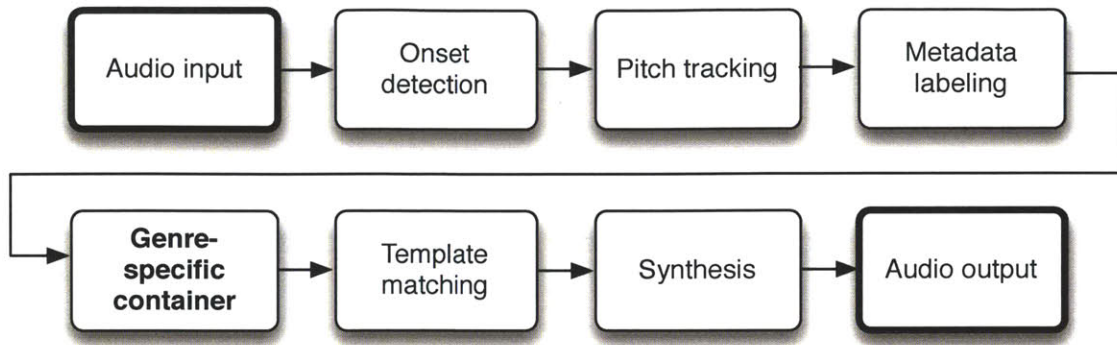


Figure 3-27: Synthesis-by-analysis via culture-specific containers for listening experiments

are pitch-tracked and stored in a Carnatic container (data structure with multiple features) or a Western container (MIDI note). Then the containers are used for sound synthesis. The containers are compared in a qualitative evaluation.

Music aficionados well-versed either in Indian music (Hindustani or Carnatic) or in Western music (Blues) were recruited for listening tests. (The MIT institutional review board approval of the research study on file.) The experiment evaluates the perceptual merit of one culturally relevant representation versus another non-culturally specific. In blind listening tests, participants are presented snippets of audio—sequences of musical primitives (tokens) assembled in musical phrases. Participants are asked to rate the subjective *well-formedness* of the phrases based on their familiarity with each musical style. Audio clips are randomly presented to participants.

3.5 Culturally Sensitive Design Principles

In 85% of the cases, participants judged the culture-appropriate container a better fit for the source material in terms of musical ‘well-formedness.’ This supports that a specific representation performs better than a generic one because it enables more accurate reproduction of a musical idiom.

This finding supports the idea that representation models are to be specifically designed to capture the complexity of a musical culture. The representation of a particular genre is not only evaluated for its intrinsic perceptual quality, but it is also evaluated in conjunction with a prediction algorithm where its suitability to generate well-formed musical constructs is analyzed quantitatively (chapter 4).

While in the present work, tokens are specified by a human designer, it would be worthwhile to study whether containers can be identified automatically with unsupervised learning, for instance with a clustering algorithm, and how well those would perform perceptually in comparison with human-designed ones. This approach might be especially relevant when the designer may not be familiar with a particular genre of music that the model should cater for.

Chapter 4

Automatic Music Prediction

4.1 Musical Expectation by People

No study of musical expectation can be complete without the mention of Meyer's *Emotion and Meaning in Music* 1961. In his book, Meyer, a Western music theorist, uses psychological insight (including Gestalt principles) to explain musical meaning and musical communication, and the rise of emotion from music.

Meyer suggests that emotion in music comes from the satisfaction of expectations and especially from the lack thereof: "The customary or expected progression of sounds can be considered as a norm, which from a stylistic point of view it is; and alteration in the expected progression can be considered a deviation. Hence deviations can be regarded as emotional or affective stimuli." This norm comes from enculturation, or exposure to previous music sharing similar stylistic rules.

Meyer further states that "because expectation is largely a product of stylistic experience, music in a style with which we are totally unfamiliar is meaningless." In other words, musical enculturation is key to music understanding. He defines musical styles as "more or less complex systems of sound relationships understood and used

in common by a group of individuals.” This is what I call culture in this dissertation.

Narmour (1990, 1991, 1992) built on Meyer’s work on musical expectation and created a model called Implication-Realization. Music theorists agree that Narmour’s model is complex to understand. In a *Music Perception* review from 1995, Ian Cross describes Narmour’s theory as ”treat[ing] melody primarily as a note-to-note phenomenon.” Narmour’s theory is interesting in light of the empirical studies that followed to test the theory (Krumhansl, 1995).

Huron (2006) has published much work recently on the nature of musical expectation from a cognitive psychology point-of-view. As for Patel (2010) he has been interested in drawing parallels between language and music in order to understand human cognition.

4.2 Music Prediction by Machines

Computational prediction of signals has a long history. Speech signals have used Linear Predictive Coding for efficient transmission by modeling the excitation source and the speech tract. At the far-end Finite Impulse Response filter coefficients are predicted based on past samples. Similarly Kalman filtering uses estimation and prediction for stochastic processes, for instance to track objects by predicting their future position based on their previous one.

Symbolic systems also have significant background. Most work on symbolic music processing systems has involved music composition and score following. Harry Olson’s early work at Bell Labs involved the generation of musical scores based on the analysis of songs by Stephen Foster. Following his footsteps, Lejaren Hiller and Robert Baker used Markov processes to create their *Computer Cantata* in 1963. Their statistical model defined the sequence of notes and durations by setting the next event, but did not take into account larger order structures of the piece of music. Hiller, who started

the Experimental Music Studio at the University of Illinois, further collaborated with John Cage on HPSCHD (1969), a piece that combined electronic sound tapes with solo compositions for harpsichord where the choices were made by throw of dice.

That same year, John Melby, composed *Forandrer: Seven Variations for Digital Computer* at the University of Illinois. Instead of using the university's ILLIAC computer solely for composing music and producing a score, he used it to synthesize sound. He later used Barry Vercoe's Music360 to compose *91 Plus 5* for IBM 360/91 and 5 brass instruments. In both cases the signal was first transferred to tape and later played back.

In the early days, while many US computer musicians were building on Max Mathews' work at Bell Labs and focusing their efforts on sound synthesis, many European composers were pushing the boundaries of computer-based music composition by creating new musical styles instead of mimicking existing ones. Among them were Gottfried Michael Koenig and Pierre Schaefer. And then there was Pierre Boulez's ever-important IRCAM in Paris, which at the time concentrated much of the work in computer music throughout the world.

Chomsky's work on universal grammar (1957, 1965) has proven to be almost as important a landmark in music theory as it was in linguistics. However the modeling of music with a grammar has remained disputed. While Roads and Wieneke (1979) supports the use of grammar as representation for music, Dempster (1998) refutes this idea by supporting that musical structure is not a genuine grammar because it does not encode meaning.

Composer David Cope from the University of California Santa Cruz has been interested in writing music in a specific style with his *Experiments in Musical Intelligence*. Drawing on Schenkerian analysis and on Chomsky's generative grammar of natural languages, Cope performed hierarchical analyses of the components of a composition and reassembled parts as to resemble the original style ("recombinant music"). He released a CD called *Virtual Mozart* in 1995 and other Chopin-like pieces

that received good reviews, including by Douglas Hofstadter who managed to fool a majority of faculty and students of the Eastman School of Music who could not tell apart a piece produced by the computer and an original piece of Chopin's.

Manzara et al (1992) studied cognitive models of melodic expectations with Bach Chorale melodies. Conklin (1990) followed this up with a prediction and entropy computational model for melodic continuation in Bach Chorales. In his work he finds that chorale melodies with low entropy achieve good continuation success.

Music theory has historically emphasized the study of Western tonal music. While ethnomusicology has catered to non-Western music, it has done so to a much lesser extent than traditional musicology. It is therefore of interest to design computational tools to help in the study and comparison of music from different cultures.

Generative music systems have often made use of probabilistic modeling tools like Hidden Markov Models. However, can a music sequence be likened to a Markov process? Many composition generators make that assumption, but I would argue that music is a hierarchical process that has its roots in a cultural body of knowledge, which does not only depend on immediately preceding musical events (at the very least far-order (n -order, $n \geq 1$) effects have to be considered for melodic and rhythmic continuation). In my view generative grammars on symbolic data are more likely to capture the essence of music, and perform adequate predictions, than statistical models. However, there may be musical traditions that a Markov process may accurately represent and therefore predict. The key is in choosing the right tool for a particular cultural context.

4.3 A Computational Model of Music Prediction

Can the human capacity for musical expectation be modeled with a machine-based prediction engine?

I aim to test the following two hypotheses:

$$error_rate(prediction_{culture_specific}) \leq error_rate(prediction_{culture_agnostic})$$

$$prediction_{culture_specific} \approx human_musical_expectation$$

In other words:

1. Music prediction systems that act on culture-specific containers are more robust and accurate than systems that generate symbolic musical data with no attempt to understand cultural content.
2. The input-output behavior of a culturally sensitive design for automatic music prediction performs similarly to human musical expectation.

The automatic music prediction system presented here is based on a hierarchical grammar model.

The choice of a data structure for prediction is based on the specifications of the musical style. Much music establishes a norm, then disturbs it, and finally resolves it. A tree structure is selected to represent the musical forms under consideration—North Indian and South Indian music, and Blues.

It is important to take into account self-similarity in music as well. Variations should be correctly classified either as identity-preserving embellishments, or as variations with significant dissimilarities. Similarly, the grouping of events in the rhythmic domain is culture-dependent.

I designed a software framework for the automatic prediction of musical events in the melodic (or timbral in the case of percussive elements) and rhythmic spaces. Figure 4-1 presents the software framework with its architecture for *cultural plug-ins*.

This system mediates musical communication between a performer and a machine

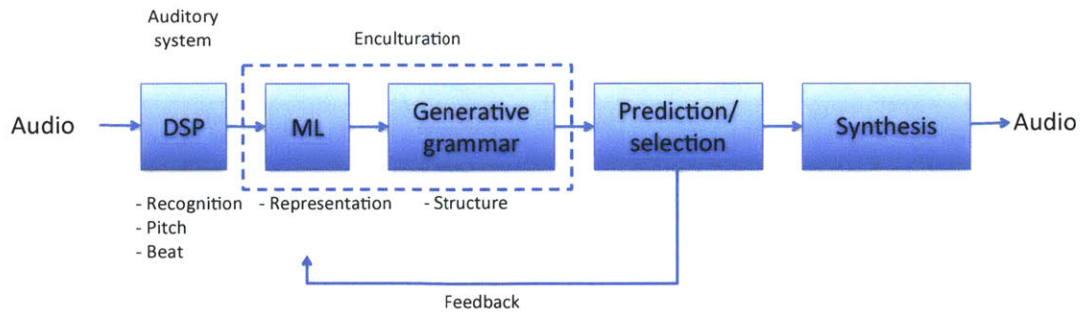


Figure 4-1: Automatic music prediction system block diagram

by incorporating an appropriate representation to the music prediction framework. Music prediction serves the dual purpose of evaluating the chosen representation scheme by comparing its output with the actual input and with listeners' expectation, and of providing for a variety of novel applications.

The system is evaluated on quantitative as well as qualitative grounds by a series of experiments. The representation models are evaluated on the accuracy of the prediction outcomes in relation to people's expectations. Quantitative measurements assess the prediction engine's error rate as related to the maturity of the underlying representation, in terms of enculturation.

The front end of the system listens to incoming audio, detects auditory events at their onset with an envelope follower and segments them, then applies machine listening techniques: pitch tracking in the case of musical notes, and feature extraction for classification purposes in the case of drum strokes. Each recognized note or percussive stroke is then labeled along with a timestamp and associated properties (e.g. ornamentation, articulation, dynamics). Additional machine listening techniques for beat detection and tempo estimation are applied on longer time windows.

A prediction model generation process captures symbolic information from the front-end and trains a grammar model from a lexicon of primary musical elements that are designed for the particular musical instrument and genre under consideration. Lexical tokens are assembled into grammar rules that follow a hierarchy, accounting

for multiple levels: linear sequencing of notes, assembly of notes and cells into figures, motifs, and phrases, and global structure, or form, of the piece. The grammar model is meant for a particular genre in that it takes into account its structure and characteristics.

The final stage is the prediction engine, which:

- Generates a symbolic output from the grammatical model. The output is in the form of a piano roll (time as x-axis, and pitch or timbre as y-axis). In addition, a tree of possible future outcomes with likelihood function is generated. As the incoming musical events unfold, the tree is pruned and a unique path, the most likely one, is traversed.
- Includes a feedback loop to compare predicted output with corresponding input to continuously learn from successes and errors.
- Plays back audio from a sound synthesizer for the most likely next event.

When the prediction engine receives an incoming event it performs a lookup through its grammar model and outputs a predicted symbol that triggers a sound synthesis engine for audio output.

The formal grammar model follows the rules of an unrestricted grammar in the Chomsky hierarchy. The hierarchy levels consist in phrases, cells, and notes.

The software framework provides an infrastructure for *cultural plug-ins*, which embed knowledge about the characteristic of a musical instrument playing in a particular musical tradition. Containers are designed explicitly in a way that is relevant to a particular genre (see chapter 3) and filled with content during the training, or *computational enculturation*, phase

Table 4.1 lists the cultural plug-ins that were developed for this study.

While Indian music, both Hindustani and Carnatic, have strong rules within which

Table 4.1: Cultural plug-ins

Instrument & Tradition	Percussion	Wind
Hindustani	Tabla	Bansuri
Carnatic	Mrindangam	Pulangoil
Western	Drums	Concert Flute

each performer expresses their individuality, Blues is more individualized with weak rules. One of the characteristics of Blues that is expressed in the temporal model is syncopation.

4.4 Machine Prediction and Human Expectation

The music prediction system is evaluated on 4 types of musical data (4 instruments belonging to 2 musical traditions):

- Tabla, a pair of North Indian hand drums (in the North Indian Hindustani tradition)
- Drums (Blues)
- Bansuri, a North Indian transverse bamboo flute (Hindustani music)
- Western Flute (Blues)

The data is sourced online from user-generated content on sites like YouTube, and manually annotated. Audio clips are randomly grouped in training, testing, and evaluation sets.

4.4.1 Quantitative Evaluation

The model is trained with musical data from culture A and tested with musical data from cultures A and B. Prediction output is compared with ground truth and averaged

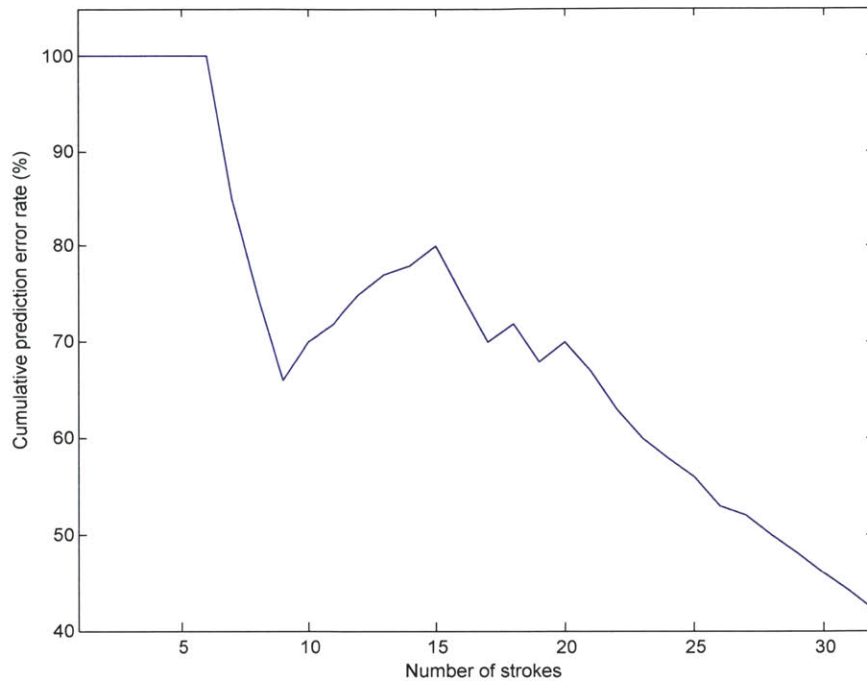


Figure 4-2: Tabla prediction error rate for *teental*, a 16 beat cycle

over several instances of varying lengths of sequences of musical events.

As an objective measurement, prediction error rate is compared to manually labeled audio input (ground truth).

It is futile to measure a static value for the error prediction rate as it continuously varies based on the input and the dynamically updating prediction model.

As mentioned in chapter 3 the stroke recognition rate is close to 95%. Based on figure 4-2, the phrase prediction error rate decreases from 100% to 40% within 30 strokes and continues to vary from there on.

4.4.2 Qualitative Evaluation

Qualitative studies are designed to compare the prediction output with people’s musical expectations.

I recruited participants (musicians familiar with Indian music or Western music) for perceptual tests. In these tests, the output of the prediction engine is evaluated for its intrinsic merit (subjective structural quality) and compared in blind tests with non-predicted data based on the original (through appropriate containers and generic containers for primary musical elements).

Each of the following blocks has 2 versions: one for Hindustani music and one for Blues. Both are for the flute (Bansuri or Western concert flute):

1. Source material
2. Container
3. Grammar model (prediction engine)

The output of the grammar model is synthesized and played to the listener.

The combinations are as follows:

1. Bansuri through Hindustani container with no prediction
2. Bansuri through Hindustani container with Hindustani prediction
3. Bansuri through Hindustani container with Western prediction
4. Bansuri through Western container with no prediction
5. Bansuri through Western container with Hindustani prediction
6. Bansuri through Western container with Western prediction
7. Flute through Western container and no prediction
8. Flute through Western container and Western prediction
9. Flute through Western container and Hindustani prediction

Table 4.2: Results of qualitative experiments on perceptual cultural fitness

Instrument	Container	Prediction	‘Western’	‘Indian’
Bansuri	Hindustani	no prediction	3.8	4.0
Bansuri	Hindustani	Hindustani	3.2	3.8
Bansuri	Hindustani	Western	2.6	2.8
Bansuri	Western	no prediction	3.0	2.6
Bansuri	Western	Hindustani	2.2	2.6
Bansuri	Western	Western	2.8	2.4
Flute	Western	no prediction	4.6	4.4
Flute	Western	Western	4.2	4.4
Flute	Western	Hindustani	3.2	3.0
Flute	Hindustani	no prediction	3.4	3.0
Flute	Hindustani	Western	3.0	3.0
Flute	Hindustani	Hindustani	2.2	3.2

10. Flute through Hindustani container and no prediction

11. Flute through Hindustani container and Western prediction

12. Flute through Hindustani container and Hindustani prediction

5 participants were recruited for this task. Feedback was collected in the form of a survey that was administered during listening. Experiments were conducted double-blind to avoid bias. Listeners were asked to rate the ‘well-formedness’ of each musical phrase on a scale of 0 to 5.

4.5 Can Machines Predict Music?

Results of the qualitative experiments are in table 4.2. The instrument are analyzed (pitch tracked) and placed into a container either suitable or unsuitable for its tradition. Then the container is used as an element to a prediction model that is trained either on a similar or on a dissimilar genre. ‘Western’ and ‘Indian’ correspond respectively to listeners familiar with Western music and Indian music.

Some observations can be noted:

- As expected an appropriate container with no prediction performs best for Indian and for Western music.
- Western flute through an Indian container performs relatively well because the container captures a superset of the parameters required of it.
- No prediction performs better than prediction indicating room for improvement in the prediction engine.
- Listeners familiar with Indian music were also mostly familiar with Western music so their ratings of the Western container and/or predictor are almost on par with Western listeners.
- I was expecting ‘Western listeners’ to generally rate the output higher than ‘Indian listeners’, maybe because of their familiarity with synthesized music, but it turns out that Indian listeners were more comfortable with the output (average of 3.26 versus 3.18).

It turns out that the system performs like an intermediate player. Although low-level predictions at the note and cell level are oftentimes error-prone, we observe that they are robust with respect to intermediate-level figures (or riffs) or motifs, and phrases—in most likelihood because they are constrained by higher-level rules of musical form (e.g. scale, meter, structure) and style. I posit that such music prediction constitutes a system of communication that conveys a sense of musical intention through musical ideas, rather than an objective transmission of the ‘implementation’ details. In other words it trades off accuracy of musical elements for musical meaning.

Possible latency compensation using this system is a function of tempo. In various realtime musical applications, especially online or where much computational resources are required, music is subject to *lag* between the incident musical gesture or event, and the corresponding response. For instance, in a network music performance setup, network latency introduces much of the lag, which is further compounded by algorithmic delays. It turns out that music prediction in the way that it is defined

here allows to submit a ‘negative lag’, which we call *prelag*. With prelag, people hear musical events ahead of their expectations. This is especially relevant for performers whose expectation model is tuned to their acoustic environment—it is either dependent on reverberation time or, in case of a remote interaction, learnt by adaptation to network delays.

Edmund Husserl, the founder of the philosophical school of phenomenology, has described the phenomenology of temporality in perception with three aspects: retention (memory), the present, and protention (expectation) (Merleau-Ponty, 1945). Protention corresponds to our perception of the next moment—a few milliseconds ahead of time.

Chapter 5

Towards Culturally Sensitive Music Technology

5.1 From Music Representation to Music Prediction

To design computational music systems for existing musical genres from around the world it is indispensable that we fully understand the role of culture in music perception and cognition, and incorporate mechanisms of musical enculturation in our designs. This dissertation discusses the importance of a culturally sensitive design for music representation and music prediction.

This dissertation introduces a measure of musical entropy for melodic contour that allows us to quantify the role of prior musical exposure. My findings suggest that musical complexity is a function of musical enculturation, in that predictability for a particular piece of music is enhanced when the entropy model is trained by music with shared features (i.e. in the same genre).

Current musical representations have inherent limitations due to design choices

that are based on Western music. This work introduces design principles for music representations that aims to cater to music from other traditions. Listening experiments suggest that there is no one representation that satisfies the requirements and salient features of all musical genres. Instead this work supports the design of specific representations for specific musical traditions.

A system for automatic prediction is then presented. The system predicts musical events described by the culture-specific representations introduced previously. Prediction accuracy is evaluated quantitatively, and more importantly by subjective listening tests. The goal of music prediction is to satisfy human musical expectation in a listening situation rather than ground truth from the musical source.

To conclude this work I present some scenarios that the design strategies introduced here can be applied to.

5.2 Music Prediction in the Real World

A system that implements culture-based computational music representation and automatic music prediction enables several applications such as:

- A network music performance system where music prediction compensates for network lag.
- A music generation system that ‘predicts’ an entire piece from an initial seed (i.e. the first note) and its learnt grammar.
- A music transcription system that generates a score from audio.
- An automatic feedback or accompaniment system for individual performers.
- A music ‘translation’ mechanism between musical instruments or genres.

A system that predicts music can serve to compensate latency in remote music collaboration by synthesizing estimated future events with no wait time. Such a system was demonstrated by Sarkar and Vercoe (2007).

The system presented in this dissertation has been tested for the prediction of one future event. Informal experiments have shown that the error rate quickly deteriorates when predicting more events in advance. However, if the generative grammar is modeled after a particular tradition, the output will still match its requirements and constraints. Nothing stops the system from predicting a whole piece of music from a seed feeding into its grammar model: it will produce music ‘in the style of’ whichever model it is based on.

The building blocks of the system presented here (e.g. pitch tracking, beat detection) enable monophonic music transcription. If we were to invent a notation system for non-Western oral traditions, we could imagine the system generating a score for a particular instrument in that tradition based on the salient parameters of its representation model.

Such a system could also provide feedback based on recognized primitives for a music learner to learn and improve musical skill.

Taking the idea of the network music system presented above and keeping it local, we could devise a system that would have a grammar where left input elements would correspond to the instrument played locally, and the right output elements would trigger another musical instrument. The system would then ‘accompany’ the local instrument with a synthesized instrument.

Or instead of accompaniment, we could think of a ‘translation’ mechanism between musical instruments or genres by having the grammar modified to output idioms corresponding to another tradition.

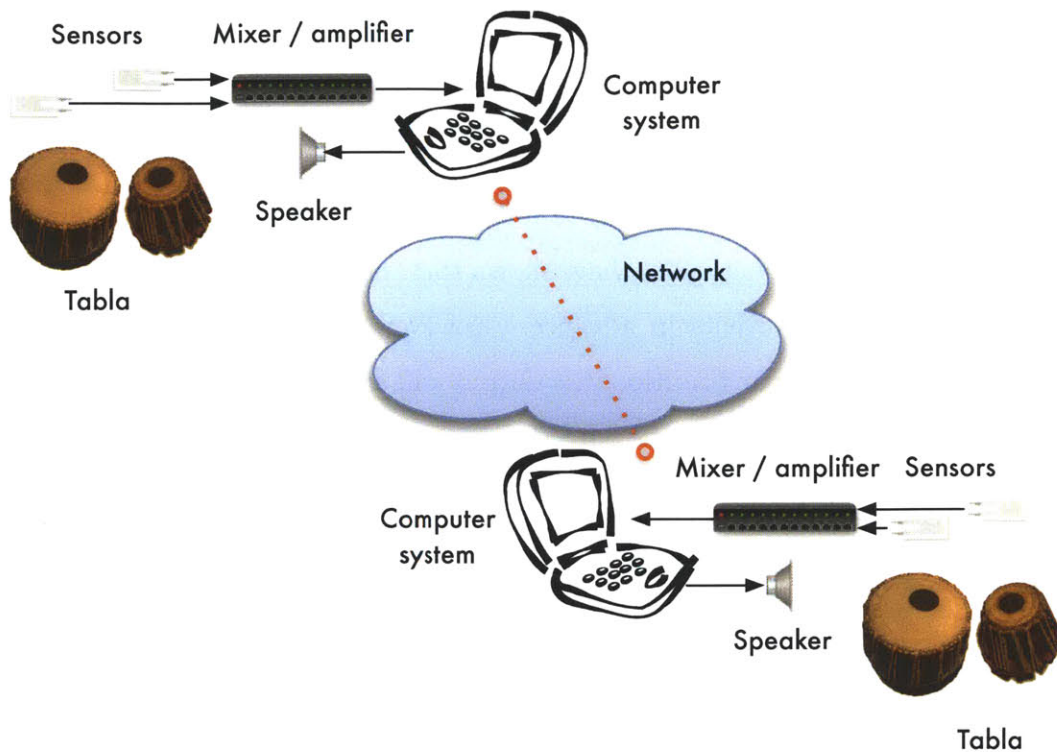


Figure 5-1: TablaNet: a real-time online musical collaboration system for Indian percussion (Sarkar and Vercoe, 2007)

5.3 Network Music Performance: a *Killer App*?

One of the most promising application seems to be a system for network music performance. Network latency has a physical limit that is bounded by networking overhead and the speed of light, which cannot be reduced by conventional means. A predictive approach is a creative way to deal with network lag by generating and synthesizing future musical events based on the previous events that are transmitted across the network. In this scenario the automatic music prediction engine is used in a dynamic realtime environment within a closed feedback loop involving machines and humans.

In my master's thesis at the MIT Media Lab, I designed TablaNet 2007, an online collaboration system for the tabla based on recognition and prediction.

In this work I was striving to overcome network latency for online musical collab-

oration so that musicians remain perceptually synchronized with each other as they play in real-time over the internet. I developed a system that is tuned to a particular musical instrument and musical style—the tabla in North Indian (Hindustani) music. The system includes hardware and software components, and relies on standard network infrastructure (no low latency required). Software modules at the near-end process the acoustic signal and recognize individual drum strokes. Symbolic events are sent over the network, and high-level rhythmic features are extracted and stored as generative grammar rules at the far-end. Rules are also applied to each incoming event in order to estimate the most likely next event so it can be synthesized at the far-end. Each performer has both the input and output parts of the system.

Because the system predicts one event in advance, the amount of lag that automatic music prediction compensates for is a function of the musical content itself, in particular its tempo. In fact the amount of lag compensation varies continuously based on the individual timing of successive musical events. For musical passages that have a series of onsets in rapid succession, we may consider encoding them as a single musical primitive in order to allow for longer lag times and improved prediction accuracy.

The system was tested live in a lecture-demonstration during a workshop at the MIT Media Lab with collaborators in India and at the Center for Computer Research in Music and Acoustics at Stanford University. The system compensated delays up to 100ms between the US East Coast and the West Coast, and up to 300ms with India. Tempo variations (faster and slower) were demonstrated with performers at each end recounting their experience at the end of the performance. It turned out that the system was able to compensate for latencies far above any audio-based system, and trade-offs (i.e. prediction errors) were “manageable” during improvised live jamming. According to the audience members, the system was able to convey the gist, or high-level musical intentions, of the performers in the appropriate musical style.

5.4 Contributions

This dissertation results in scientific contributions in the realm of music perception and cognition. It also contributes to the design, engineering, and technology of computational music systems. Finally, as an artistic component to this work, it uncovers a new perspective that emphasizes the role of culture in computer music.

In this work, I:

- Defined a method to quantify and compare cultural complexity in music using information theory.
- Introduced a measure of *musical entropy* to analyze the predictability of a piece of music based on its context.
- Conducted a study that supports the importance of *computational enculturation* for music prediction.
- Proposed design principles for a *culturally sensitive* computational representation of music.
- Designed a software framework for *automatic music prediction* that takes musical culture into account.
- Developed *cultural plug-ins* that incorporate an appropriate representation of several musical instruments in various traditions, including tabla (Hindustani), drums (Blues), bansuri (Hindustani), and flute (Blues).
- Invented the concept of *prelag* (negative lag) when music prediction anticipates musical expectation in the cognitive realm.

5.5 Future Work and Final Remarks

This work is by no means complete. In addition to filling a gap in human knowledge the purpose of a doctoral thesis seems to be to uncover more questions and unknowns. Music prediction is a new field of inquiry in music technology. At the time of “going to press” I am aware of several computer music departments that have received grants and advertised research positions for the study of computational music prediction. It is my hope that culture will be taken into consideration when attempting to design music prediction algorithms, and that claims of a *universal* system won’t be taken lightly.

Among the follow-ups that this work could benefit from, and the new fields of scholarship that this work opens, some merit attention:

- A system that switches context (i.e. selects the appropriate cultural plug-in) automatically based on the musical content (i.e. genre) of the incoming audio.
- An audio-based prediction system that works on audio samples instead of musical symbols.
- Multi-dimensional prediction that models pitch, rhythm, loudness, timbre, and other parameters of music.
- A system that would create appropriate musical representations for each genre without the intervention of a human designer.
- A computational model of music perception and cognition that would lead to a fully autonomous Artificial Intelligence that would learn through enculturation and perform prediction like a human would.

This work brings together years of research in what I call *culturally sensitive music technology*. I strongly believe that tools influence creative output (through what we call their affordances), and I think it unfair to provide the world with tools biased

towards a particular culture. I don't believe it is necessary to provide the 'containers' and the content that will cater to all the cultures found in the world—local designers will often do a better job in making the right choices for their own community and culture—but I strongly believe that it is of the utmost importance to be aware of it.

Appendix A

Glossary of Terms

Element Basic building block of music, e.g. pitch, beat, timbre.

Event Discrete time-based element of music, e.g. note, percussive stroke.

Symbol Label and associated parameters of a musical event.

Primitive Low-level perceptual representation of a musical element, e.g. pitch, tactus, loudness.

Construct High-level or cognitive, usually learnt, combination of musical primitives, e.g. scale, rhythm, dynamicity.

Structure Generic term for a musical primitive or construct, or a combination thereof.

Representation Mental or computational mapping corresponding to a musical structure.

Culture (also Tradition) Set of pieces of music that share a common representation.

Genre Set of pieces of music in a particular tradition that share common generative rules.

Form Overall structure or plan of a piece of music.

Style Characteristics that are specific to the performance of a particular piece of music.

Expectation Cognitive capacity for predicting music in a listening context.

Anticipation Cognitive capacity for predicting music in an active context, i.e. while performing.

Prediction In general, computational estimation of forthcoming musical events.

Bibliography

- M.H. Allen. Tales tunes tell: Deepening the dialogue between” classical” and” non-classical” in the music of india. *Yearbook for traditional music*, 30:22–52, 1998.
- P. Allen and R.B. Dannenberg. Tracking musical beats in real time. *Proceedings of the 1990 International Computer Music Conference*, pages 140–143, 1990.
- C. Bartlette, D. Headlam, M. Bocko, and G. Velikic. Effect of network latency on interactive musical performance. *Music Perception*, 24(1):49–62, 2006.
- B. Bel. The beginnings of computer music in india. *Computer Music Journal*, 22(4): 9–11, 1998.
- B. Bel. Rationalizing musical time: syntactic and symbolic-numeric approaches. In Edited by C. Barlow, editor, *The Ratio Book*, pages 86–101. Feedback Papers, 2000.
- B. Bel. *Acquisition et représentation de connaissances en musique*. Édilivre-Aparis, 2008.
- B. Bel and J. Kippen. Bol processor grammars. *Understanding music with AI: perspectives on music cognition table of contents*, pages 366–400, 1992.
- L. Bernstein. *The unanswered question: Six talks at Harvard*. Harvard Univ Pr, 1976.
- J. Blacking, R. Byron, and B. Nettl. *Music, culture, & experience: selected papers of John Blacking*. University of Chicago press, 1995.
- C. Chafe. Statistical Pattern Recognition for Prediction of Solo Piano Performance. In *Proc. ICMC, Thessaloniki*, 1997.
- C. Chafe, B. Mont-Reynaud, and L. Rush. Toward an intelligent editor of digital audio: Recognition of musical constructs. *Computer Music Journal*, 6(1):30–41, 1982.
- C. Chafe, M. Gurevich, G. Leslie, and S. Tyan. Effect of Time Delay on Ensemble Accuracy. In *Proceedings of the International Symposium on Musical Acoustics*, 2004.
- W. Chai. *Melody retrieval on the web*. PhD thesis, Massachusetts Institute of Technology, 2001.

- P. Chordia. Segmentation and Recognition of Tabla Strokes. In *Proc. of ISMIR (International Conference on Music Information Retrieval)*, 2005.
- M. Clayton. *Time in Indian Music: Rhythm, Metre, and Form in North Indian Râg Performance*. Oxford University Press, 2000.
- J.E. Cohen. Information theory and music. *Behavioral Science*, 7(2):137–163, 1962.
- N. Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. *118th Audio Engineering Society Convention, Barcelona*, pages 28–31, 2005.
- D. Conklin. Music generation from statistical models. In *Proceedings of the AISB 2003 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, pages 30–35. Citeseer, 2003.
- D. Conklin and IH Witten. Prediction and entropy of music. Master’s thesis, 1990.
- D. Conklin and I.H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.
- A. Cont. Modeling musical anticipation: From the time of music to the music of time. 2008.
- D. Cope. *Computers and musical style*. AR Editions, Inc., 1991.
- D. Cope. Computer modeling of musical intelligence in emi. *Computer Music Journal*, 16(2):69–83, 1992.
- D. Cope. *Experiments in musical intelligence*, volume 1. AR Editions Madison, WI, 1996.
- I. Cross. Is music the most important thing we ever did? music, development and evolution. *Music, mind and science*, pages 10–39, 1999.
- I. Cross. Music, cognition, culture, and evolution. *Annals of the New York Academy of Sciences*, 930(1):28–42, 2001.
- I. Cross. Music and social being. *Musicology Australia*, 28(1):114–126, 2005.
- R.E. Cumming. The interdependence of tonal and durational cues in the perception of rhythmic groups. *Phonetica*, 67(4):219–242, 2010.
- A. Daniélou. *Music and the power of sound: The influence of tuning and interval on consciousness*. Inner Traditions Rochester, VT, 1995.
- R. B. Dannenberg. Toward Automated Holistic Beat Tracking, Music Analysis, and Understanding. In *ISMIR 2005 6th International Conference on Music Information Retrieval Proceedings*, pages pp. 366–373. London: Queen Mary, University of London, 2005.

- R.B. Dannenberg. Music representation issues, techniques, and systems. *Computer Music Journal*, 17(3):20–30, 1993.
- R.B. Dannenberg, D. Rubine, and T. Neuendorffer. The resource-instance model of music representation. *Computer Science Department*, page 482, 1991.
- I. Deliège and J.A. Sloboda. *Musical beginnings: Origins and development of musical competence*. Oxford University Press, USA, 1996.
- I. Deliège and J.A. Sloboda. *Perception and cognition of music*. Psychology Pr, 1997.
- D. Dempster. Is there even a grammar of music? *MusicaeScienti*, 2(1):55–65, 1998.
- D. Deutsch. *The psychology of music*. Academic Pr, 1999.
- A.E. Diaz Andrade and C. Urquhart. Icts as a tool for cultural dominance: prospects for a two-way street. *The electronic journal of information systems in developing countries*, 37(0), 2009.
- P.F. Driessen, T.E. Darcie, and B. Pillay. The effects of network delay on tempo in musical performance. *Computer Music Journal*, 35(1):76–89, 2011.
- B. Duane. Information content in melodic and non-melodic lines. ICMPC, 2010.
- S. Dubnov, G. Assayag, O. Lartillot, and G. Bejerano. Using machine-learning methods for musical style modeling. *Computer*, 36(10):73–80, 2003.
- S. Dubnov, S. McAdams, and R. Reynolds. Predicting human reactions to music on the basis of similarity structure and information theoretic measures of the sound signal. In *Proc. AAAI Fall Symp. Style and Meaning in Language, Music, Art and Design*, pages 37–40, 2004.
- S. Dubnov, S. McAdams, and R. Reynolds. Structural and affective aspects of music from statistical audio signal analysis. *Journal of the American Society for Information Science and Technology*, 57(11):1526–1536, 2006.
- R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley-Interscience, 2000.
- R. Dunbar. *Grooming, gossip, and the evolution of language*. Harvard Univ Pr, 1996.
- T. Eerola. *The dynamics of musical expectancy: cross-cultural and statistical approaches to melodic expectations*. PhD thesis, 2003.
- T. Eerola, T. Himberg, P. Toiviainen, and J. Louhivuori. Perceived complexity of western and african folk melodies by western and african listeners. *Psychology of Music*, 34(3):337–371, 2006.
- D.P.W. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

- M.M. Farbood. *A quantitative, parametric model of musical tension*. PhD thesis, Massachusetts Institute of Technology, 2006.
- S. Feld and D. Brenneis. Doing anthropology in sound. *American Ethnologist*, 31(4): 461–474, 2004.
- A. Freed and A. Schmeder. Features and future of open sound control version 1.1 for nime. In *NIME'09: Proceedings of the 9th Conference on New Interfaces for Musical Expression*, 2009.
- O.K. Gillet and G. Richard. Automatic Labelling of Tabla Signals. In *Proc. of the 4th ISMIR Conf.*, 2003.
- B. Gold, N. Morgan, and D. Ellis. *Speech and audio signal processing*. Wiley Online Library, 2000.
- E. Gómez, A. Klapuri, and B. Meudic. Melody description and extraction in the context of music content processing. *Journal of New Music Research*, 32(1):23–40, 2003.
- M. Goto. An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds. *Journal of New Music Research*, 30(2):159–171, 2001.
- F. Gouyon and S. Dixon. A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–54, 2005.
- S. Gulati, V. Rao, and P. Rao. Meter detection from audio for indian music. In *International Symposium on Computer Music Modeling and Retrieval*, 2011.
- A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.
- M. Hamanaka, K. Hirata, and S. Tojo. Melody expectation method based on gttm and tps. In *ISMIR 2008: proceedings of the 9th International Conference of Music Information Retrieval*, page 107, 2008.
- S. Handel. *Listening: An introduction to the perception of auditory events*. The MIT Press, 1993.
- D.E. Hast, J.R. Cowdery, and S.A. Scott. *Exploring the world of music: an introduction to music from a world music perspective*. Kendall/Hunt Publishing Company, 1999.
- H. Helmholtz. On the sensations of tone as a physiological basis for the theory of music. *Dover Publications, New York*, 1954.
- E.O. Henry. Folk song genres and their melodies in india: Music use and genre process. *Asian music*, 31(2):71–106, 2000.

- D.R. Hofstadter. *Gödel, Escher, Bach: an eternal golden braid*. Harvester Press, 1979.
- D. Huron. Design principles in computer-based music representation. *Computer representations and models in music*, pages 5–39, 1992.
- D. Huron. The humdrum toolkit: Software for music research. *Center for Computer Assisted Research in the Humanities, Ohio State University, copyright*, 1999, 1993.
- D. Huron. Is music an evolutionary adaptation? *Annals of the New York Academy of Sciences*, 930(1):43–61, 2001.
- D.B. Huron. *Sweet anticipation: Music and the psychology of expectation*. The MIT Press, 2006.
- J.R. Iversen, A.D. Patel, and K. Ohgushi. Perception of rhythmic grouping depends on auditory experience. *The Journal of the Acoustical Society of America*, 124: 2263, 2008.
- R. Jackendoff and F. Lerdahl. *A Generative Theory of Tonal Music*. MIT Press, 1996.
- R. Jackendoff and F. Lerdahl. The capacity for music: What is it, and what’s special about it? *Cognition*, 100(1):33–72, 2006.
- M. Kahrs and K. Brandenburg. *Applications of digital signal processing to audio and acoustics*, volume 437. Kluwer Academic Pub, 1998.
- K. Keniston. Software localization: Notes on technology and culture. *Retrieved from the World Wide Web: <http://web.mit.edu/kken/public/papers.htm>*, 1997.
- K. Keniston. Cultural diversity or global monoculture. *Understanding the Impact of Global Networks on Local, Social, Political and Cultural Values*, <http://www.mpp.rdg.mpg.de/woodsh.htm>, 1999.
- Y.E. Kim, W. Chai, R. Garcia, and B. Vercoe. Analysis of a contour-based representation for melody. In *Proc. International Symposium on Music Information Retrieval*, page 20. Oct, 2000.
- J. Kippen and B. Bel. Modelling Music with Grammars: Formal Language Representation in the Bol Processor. *Computer Representations and Models in Music, Ac. Press ltd*, pages 207–232, 1992.
- J. Kippen and B. Bel. Computers, composition and the challenge of “new music” in modern india. *Leonardo Music Journal*, pages 79–84, 1994.
- A. Klapuri. *Signal processing methods for the automatic transcription of music*. PhD thesis, 2004.

- L. Knopoff and W. Hutchinson. Entropy as a measure of style: The influence of sample length. *Journal of Music Theory*, 27(1):75–97, 1983.
- N. Kogan. Reflections on aesthetics and evolution. *Critical Review*, 11(2):193–210, 1997.
- A. Krishnaswamy. Inflexions and microtonality in south indian classical music. *Proc of Frontiers of Research on Speech and Music (FRSM), Annamalainagar, India*, 2004a.
- A. Krishnaswamy. Results in music cognition and perception and their application to indian classical music. *Proceedings of FRSM-2004, Chidambaram, India*, 2004b.
- C.L. Krumhansl. Music psychology and music theory: Problems and prospects. *Music Theory Spectrum*, pages 53–80, 1995.
- C.L. Krumhansl. Rhythm and pitch in music cognition. *Psychological bulletin*, 126(1):159, 2000.
- C.L. Krumhansl. *Cognitive foundations of musical pitch*. Number 17. Oxford University Press, USA, 2001.
- E.W. Large, C. Palmer, and J.B. Pollack. Reduced memory representations for music. *Cognitive Science*, 19(1):53–93, 1995.
- S. Larson and S. McAdams. Musical forces and melodic expectations: Comparing computer models and experimental results. *Music Perception*, 21(4):457–498, 2004.
- O.E. Laske. In search of a generative grammar for music. *Perspectives of New Music*, 12(1/2):351–378, 1973.
- D.J. Levitin. *This is your brain on music: The science of a human obsession*. Dutton Adult, 2006.
- T. Mäki-Patola. Musical Effects of Latency. *Suomen Musiikintutkijoiden*, 9:82–85, 2005.
- L.C. Manzara, I.H. Witten, and M. James. On the entropy of music: An experiment with bach chorale melodies. *Leonardo Music Journal*, pages 81–88, 1992.
- E.H. Margulis. A model of melodic expectation. *Music Perception*, 22(4):663–714, 2005.
- E.H. Margulis and A.P. Beatty. Musical style, psychoaesthetics, and prospects for entropy as an analytic tool. *Computer Music Journal*, 32(4):64–78, 2008.
- J. McCormack. Grammar based music composition. *Complex systems*, 96:321–336, 1996.
- M. Merleau-Ponty. *Phénoménologie de la perception*. Gallimard, Paris, 1945.

- L.B. Meyer. Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, 15(4):412–424, 1957.
- L.B. Meyer. *Emotion and meaning in music*. University of Chicago Press, 1961.
- G. Miller. Evolution of human music through sexual selection. *The origins of music*, pages 329–360, 2000.
- M. Minsky. Music, mind, and meaning. *Computer Music Journal*, 5(3):28–44, 1981.
- M.L. Minsky. Why people think computers can't. *AI Magazine*, 3(4):3, 1982. URL <http://web.media.mit.edu/~minsky/papers/ComputersCantThink.txt>.
- S.J. Mithen. *The singing Neanderthals: The origins of music, language, mind, and body*. Harvard Univ Pr, 2005.
- A. Moles. *Information theory and esthetic perception*. U. Illinois Press, 1968.
- BCJ Moore. *An introduction to the psychology of hearing (Academic, San Diego)*. 1997.
- S.J. Morrison and S.M. Demorest. Cultural constraints on music perception and cognition. *Progress in brain research*, 178:67–77, 2009.
- S.J. Morrison, S.M. Demorest, and L.A. Stambaugh. Enculturation effects in music cognition. *Journal of Research in Music Education*, 56(2):118–129, 2008.
- E. Narmour. *The analysis and cognition of basic melodic structures: The implication-realization model*. University of Chicago Press, 1990.
- E. Narmour. The top-down and bottom-up systems of musical implication: Building on meyer's theory of emotional syntax. *Music Perception*, pages 1–26, 1991.
- E. Narmour. *The analysis and cognition of melodic complexity: The implication-realization model*. University of Chicago Press, 1992.
- E. Narmour. *Hierarchical expectation and musical style*. 1999.
- E. Narmour. Music expectation by cognitive rule-mapping. *Music Perception*, pages 329–398, 2000.
- A.V. Oppenheim and R.W. Schafer. *Digital Signal Processing*. NJ: Prentice-Hall, 1975.
- A.V. Oppenheim, R.W. Schafer, J.R. Buck, et al. *Discrete-time signal processing*, volume 2. Prentice hall Upper Saddle River, NJ, 1989.
- J.F. Paiement, S. Bengio, and D. Eck. Probabilistic models for melodic prediction. *Artificial Intelligence*, 173(14):1266–1274, 2009a.

- J.F. Paiement, Y. Grandvalet, and S. Bengio. Predictive models for music. *Connection Science*, 21(2-3):253–272, 2009b.
- C. Palmer. Sequence memory in music performance. *Current Directions in Psychological Science*, 14(5):247–250, 2005.
- C. Palmer and C.L. Krumhansl. Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance; Journal of Experimental Psychology: Human Perception and Performance*, 16(4):728, 1990.
- A.D. Patel. Music, language, and the brain. 2010.
- A.D. Patel, J.R. Iversen, M.R. Bregman, I. Schulz, and C. Schulz. Investigating the human-specificity of synchronization to music. In *Proceedings of the 10th International Conference on Music and Cognition. Sapporo, Japan*, pages 100–104, 2008.
- A.D. Patel et al. Language, music, syntax and the brain. *Nature neuroscience*, 6(7):674–681, 2003.
- M.T. Pearce. The construction and evaluation of statistical models of melodic structure in music perception and composition. *City University, London*, 2005.
- M.T. Pearce and G.A. Wiggins. Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–405, 2006.
- J.G.S. Pearl. Hypothetical universe: a functionalist critique of Ier Dahl-Jackendoff.
- P.Q. Pfordresher and C. Palmer. Effects of hearing the past, present, or future during music performance. *Attention, Perception, & Psychophysics*, 68(3):362–376, 2006.
- J. Piaget. The origins of intelligence in children. 1952.
- J.R. Pierce. *Symbols, signals and noise: The nature and process of communication*. Harper, 1961.
- S. Pinker. How the mind works, 1997.
- R.C. Pinkerton. Information theory and melody. *Scientific American*, 1956.
- G.E. Poliner, D.P.W. Ellis, A.F. Ehmann, E. Gómez, S. Streich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1247–1256, 2007.
- DJ Povel and H. Okkerman. Accents in equitone sequences. *Percept Psychophys*, 30(6):565–72, 1981.
- LR Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

- N. Ramanathan. Sruti-its understanding in the ancient, medieval and modern periods. *Journal of the Indian Musicological Society*, 12:5–6, 1981.
- C. Roads and P. Wieneke. Grammars as representations for music. *Computer Music Journal*, 3(1):48–55, 1979.
- C. Roads, A. Piccialli, G.D. Poli, and S.T. Pope. *Musical signal processing*. Swets & Zeitlinger, 1997.
- M.A. Rohrmeier and S. Koelsch. Predictive information processing in music cognition. a critical review. *International Journal of Psychophysiology*, 2012.
- S.J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1995.
- J.R. Saffran. Statistical language learning mechanisms and constraints. *Current directions in psychological science*, 12(4):110–114, 2003.
- M. Sarkar and B.L. Vercoe. Tablanet: a real-time online musical collaboration system for indian percussionndian percussion. Master’s thesis, Massachusetts Institute of Technology, 2007.
- E.D. Scheirer and B.L. Vercoe. Saol: The mpeg-4 structured audio orchestra language. *Computer Music Journal*, 23(2):31–51, 1999.
- E.G. Schellenberg. Expectancy in melody: Tests of the implication-realization model. *Cognition*, 58(1):75–125, 1996.
- N. Schuett. The Effects of Latency on Ensemble Performance, 2002.
- C.E. Shannon and W. Weaver. *The mathematical theory of information*. University of Illinois Press, 1949.
- G. Shaw, M. Bodner, and J. Patera. Innate brain language and grammar: Implications for human language and music. *Stochastic point processes*, page 287, 2003.
- S.J. Simon. *A Multi-dimensional entropy model of jazz improvisation for music information retrieval*. PhD thesis, University of North Texas, 2006.
- J.A. Sloboda. The musical mind: The cognitive psychology of music. 1985.
- J.A. Sloboda. *Exploring the musical mind: cognition, emotion, ability, function*. Oxford University Press, USA, 2005.
- B. Snyder. *Music and memory*. MIT Press Cambridge, Mass, 2000.
- J.L. Snyder. Entropy as a measure of musical style: the influence of a priori assumptions. *Music Theory Spectrum*, pages 121–160, 1990.
- C. Stevens. Cross-cultural studies of musical pitch and time. *Acoustical science and technology*, 25(6):433–438, 2004.

- J. Sundberg and B. Lindblom. Generative theories in language and music descriptions. *Cognition*, 4(1):99–122, 1976.
- G.N. Swift. South indian “gamaka” and the violin. *Asian music*, 21(2):71–89, 1990.
- M. Szczerba and A. Czyzewski. Pitch detection enhancement employing music prediction. *Journal of Intelligent Information Systems*, 24(2):223–251, 2005.
- JL Trivino-Rodriguez and R. Morales-Bueno. Using multiattribute prediction suffix graphs to predict and generate music. *Computer Music Journal*, 25(3):62–79, 2001.
- S. Ullman, M. Vidal-Naquet, E. Sali, et al. Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, 5(7):682–687, 2002.
- B.L. Vercoe. A realtime auditory model of rhythm perception and cognition. In *2nd Int. Conf. on Music Perception and Cognition*, pages 307–326, Sep. 1990.
- BL Vercoe. Computational auditory pathways to music understanding. *Perception and cognition of music*, pages 307–326, 1997.
- B.L. Vercoe, W.G. Gardner, and E.D. Scheirer. Structured audio: Creation, transmission, and rendering of parametric sound representations. *Proceedings of the IEEE*, 86(5):922–940, 1998.
- KG Vijayakrishnan. *The grammar of Carnatic music*, volume 8. Walter de Gruyter, 2007.
- T. Weyde. Grouping, smilarity and the recognition of rhythmic structure. *icmc2001*, 2001.
- I.H. Witten, L.C. Manzara, and D. Conklin. Comparing human and computational models of music prediction. *Computer Music Journal*, 18(1):70–80, 1994.
- P.C.M. Wong, V. Ciocca, A.H.D. Chan, L.Y.Y. Ha, L.H. Tan, and I. Peretz. Effects of culture on musical pitch perception. *PloS one*, 7(4):e33424, 2012.
- M. Wright and D. Wessel. An Improvisation Environment for Generating Rhythmic Structures Based on North Indian “Tal” Patterns. *International Computer Music Conference, Ann Arbor, Michigan*, 1998.
- Matthew Wright, Adrian Freed, and Ali Momeni. Open sound control: State of the art 2003. In *Proceedings of the New Interfaces for Musical Expression Conference*, pages 153–159, Montreal, 2003.
- J.E. Youngblood. Style as information. *Journal of Music Theory*, 2(1):24–35, 1958.
- I. Zukerman and D.W. Albrecht. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11(1):5–18, 2001.