Content-aware visualizations of audio data in diverse contexts

by

Steven Ness
B.Sc., University of Alberta, 1994

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in the Department of Computer Science

Content-aware visualizations of audio data in diverse contexts

by

Steven Ness
B.Sc., University of Alberta, 1994

Supervisory Committee

Dr. G. Tzanetakis, Supervisor
(Department of Computer Science)

Dr. S. Ganti, Departmental Member
(Department of Computer Science)

Supervisory Committee

Dr. G. Tzanetakis, Supervisor
(Department of Computer Science)

Dr. S. Ganti, Departmental Member
(Department of Computer Science)

ABSTRACT

The visualization of the high-dimensional feature landscapes that are encountered when analyzing audio data is a challenging problem and is the focus of much research in the field of Music Information Retrieval. Typical feature sets extracted from sound have anywhere from dozens to hundreds of dimensions and have complex interrelationships between data elements. In this work, we apply various modern techniques for the visualization of audio data to a number of diverse problem domains, including the bioacoustics of Orcinus Orca (killer whale) song, partially annotated chant traditions including Torah recitation and the the analysis of music collections and live DJ sets. We also develop a number of graphical user interfaces to allow users to interact with these visualizations. These interfaces include Flash-enabled web applications, desktop applications, and novel interfaces including the use of the Radiodrum, a three-dimension position sensing musical interface.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

my parents, Fern and Randy Ness for their unfailing love and support,

Dr. George Tzanetakis, for support, cool ideas and encouragement.

UVIC, for funding me with a UVIC Fellowship.

The world is transitory. You will find stability only on the path of Karma Yoga. Only action can take a man to God and give him liberation. ... Brave ones, all of you, continue to work! Through Karma alone will you be able to change the world. It is the only way.

Babaji

DEDICATION

To my parents, Fern and Randolph Ness.

Chapter 1

Introduction

In this work, I will describe my work in applying advanced audio feature extraction, analysis and visualization tools to a variety of different problem domains. These domains include the study of Orca vocalizations, chant traditions from around the world, and the analysis and visualization of large music collections.

Although these application areas are quite different, the tools and techniques that we use to study each of them are very similar. There are two distinct types of tools that will be demonstrated, the first are tools to extract features and analyze audio. The second set of tools are web-based and allow users from around the world to collaboratively view and analyze the results obtained from the first set of tools.

An aspect characterizing this work is the need to collaborate with domain experts, and a large amount of the effort in this project is devoted to interfaces that allow domain experts with varying degrees of computer sophistication to access and make sense of the extracted data that our tools produce. Thus, the core part of this work is to bring together tools, data and scientists together into a highly effective collaborative team.

This work draws on ideas and concepts from many disciplines. Because of this it is essential to include not only definitions of these concepts, but also the fields from which them come. These are presented in the Glossary chapter.

1.1 Motivation

Web-based software has been helping connect communities of researchers since its inception. Recently, advances in software and in computer power have dramatically

widened its possible applications to include a wide variety of multimedia content. These advances have been primarily in the business community, and the tools developed are just starting to be used by academics. We have been working on applying these technologies to ongoing collaborative projects that we are involved in [NWMT08]. By leveraging several new technologies including *Flash*, *haXe*, *AJAX* and *Ruby on Rails*, we have been able to rapidly develop web-based tools. Rapid prototyping and iterative development have been key elements of our collaborative strategy. Although our number of users is limited compared to other areas of multimedia analysis and retrieval, this is to some degree compensated by their passion and willingness to work closely with us in developing these tools.

1.2 Scope of this work

The fields of Music Information Retrieval (MIR), Visualization and web-based Human Computer Interaction (HCI) are each vast topics in their own right, not to mention the application areas of Ethnomusicology and bioacoustics. So as to make the present work tractable, we will focus on three very specific problems and will apply a carefully selected subset of some of the tools in MIR, Visualization and HCI to help us in developing solutions for these areas.

In the field of the analysis of bioacoustic signals from *Orcinus Orca* vocalizations, we will use tools from MIR, including Fast Fourier Transforms (FFT) to analyze and display spectrograms, along with tools such as Mel-Frequency Cepstral Coefficients (MFCC), average zero crossing rate, and spectral centroid, along with many other such tools, to help us visualize and analyze orca vocalizations. We will be also using techniques from the fields of Visualization and HCI, including the micro-macro view, draggable panes, multi-resolution browsing, tagging and the layered presentation of data.

We will use many of these same tools in the analysis of audio from chant traditions around the world, and will in addition be using techniques such as Fundamental Frequency Estimation and Dynamic Time Warping.

In the area of the analysis of large music collections, we will use many of the previous tools, including the FFT and MFCC coefficients to extract features from the songs, and will primarily be using the technique of Self-Organizing Maps to reduce the dimensionality of the high dimension spaces created by feature extraction to two dimensions.

Chapter 2

Related Work

As this thesis is divided into the three main areas of research into partially notated chants from traditions around the world, music browsing using novel interfaces and research into orca vocalizations, I discuss the related work to each of these fields.

2.1 Partially Annotated Chants

In the section of this thesis dealing with Partially Annotated Chants we examine examples from improvised, partially improvised, partially notated and gesture-based notational chant traditions: Hungarian siratok (laments), Torah cantillation, tenth century St. Gallen plainchant, and Koran recitation. These various types of chant employ melodic formulae, which help to define syntax, pronunciation, and expression. Each of these traditions melodic framework is governed by the particular context from which they came. For example, the recitation of the Torah obeys long established rules, some more strict and some less strict.

For many generations scholars have studied music, and this discipline is known as Musicology [Ker85]. Traditionally, Musicologists would study the written score of music, and would use a variety of ethnological and statistical tools to discover correlations and differences between different scores. Computers have been applied to this task for many years [Pop92]. One book that describes the field of symbolic and empirical musicology in detail is "Empirical musicology: aims, methods, prospects" [MDC03], which gives a detailed view of a wide cross-section of this field. One recent paper describes the JRing system [Kor01] which enables the study of symbolic music with computer systems, and concludes that this system is highly flexible for studying

music.

Ethnomusicology [Gre76] has traditionally focused on the sociological [Lom62] and stylistic [Lom56] aspects of these types of music. One interesting study of the singing style of performers around the world was published in 1958 by Charles Seeger [See58]. In this article the author investigates the differences in singing style in a wide variety of cultures, and breaks down these differences to pitch, loudness, tempo, proportion, phrase-breath, timbre, and accentuation. In this study, these quantities are quantified by a human observer, which is a time-consuming and subjective method of study. Another excellent article is "Folk Song Style" by Alan Lomax [Lom59], in which folk songs from around the world are studied. In this article, one of the examples that is discussed is a comparison between American White Folk and American Negro Folk music, and the author investigates a number of different dimensions including if the music is primarily solo or choral, what the facial expressions of the singers are, and the pitch quality of the voices. One physical measuring instrument that was used in the study of ethnomusicology was the melograph [Lis63] , a device that plots a graph of sound versus time, which is typically an estimate of the pitch or fundamental frequency of a sound, or can be a description of all the partials in the sound [Hoo00]. These early studies and others [Met26] paved the way for our later work by providing an extensive qualitative language on which to base our work.

The field of Computational Ethnomusicology, where musical traditions from around the world are studied with computers, is a relatively new [TASW07] field of study, and takes advantage of the development of new algorithms to extract features directly from an audio source [Tza08]. Symbolic musicology is largely unable to deal with the complex, partially notated, musical traditions found in non-western cultures due to its reliance on an absolute symbolic representation of music. An early study of timbre in music was performed by John Grey [Gre78]. In this paper the author examines some of the many facets of timbre using the assistance of computers, one of the algorithms that is used is heterodyne analysis, which produces time-varying amplitude and frequency functions for each partial in the tone of an instrument.

Chant scholars have investigated historical and phenomenological aspects of chant formulae to discover how improvised melodies might have developed to become stable melodic entities, paving the way for the development of notation. A main aspect of such investigations has been to explore the ways in which melodic contour defines melodic identities [Kar98]. We hope that our computational tools will allow for new possibilities for paradigmatic and syntagmatic chant analysis in both culturally

defined and cross-cultural contexts. This might give us a better sense of the role of melodic gesture in melodic formulae and possibly a new understanding of the evolution from improvised to notation-based singing in and amongst these divergent chant traditions.

2.2 Novel Music Browsing Interfaces

Audioscapes is the evolution of several research efforts by our group [MT06, SHT07] to create novel content and context-aware music browsing interfaces. We have tried to combine our previous experience with knowledge from state-of-the-art systems in this domain to design a flexible framework to explore this new and fascinating interface design problem.

In the field of Music Information Retrieval, data of high dimensionality and of considerable complexity is generated. Various visualization interfaces have been proposed to make this data accessible and useful to users. Frequently these interfaces rely on automatically extracted audio features. Islands of Music [PDW03] is an example of such a visualization of audio information which uses the technique of Self-Organizing Maps (SOMs) to generate a two-dimensional representation of a collection of music. MusiCream [GG05] is an interface that allows users to interact with a music collection using a dynamic visualization interface. MusicRainbow [PG06] is a similar system that uses web-based labelling and audio similarity to visualize music collections. Another relevant system is MusicSun [PG07] which combines three different similarity measures to generate music recommendations for users. The Databionic/MusicMiner system [MUNS05] allows users to organize large collections of music and employs Emergent Self-Organizing Maps to generate visualizations of the data involved. A very large web based system for helping users find new music is part of the LastFM website `http://playground.last.fm/iom` which provides advanced functionality for music recommendation and visualization based on tag data. A 2006 review of some of the recent trends in visualization in audio based music information retrieval can be found in Cooper [CFPT06].

Self organizing maps have been used extensively in the visualization of data for audio based music information retrieval [CFPT06]. They have been used to analyze and organize music archives [RF01] [RM98b] [RM98a] [RPM02a], and to visualize the resulting music collections [RPM03] [ESG04] [RPM02b]. A particularly relevant study was that of Palmalk in his paper Islands of Music [PDW03]. SOMs have also

been used for audio retrieval, browsing and constructivist learning in several papers [CKGB02] [FR01] [HLLR00]. While the previous mentioned studies concentrated on organizing whole classes of music, SOMs have also been applied to smaller audio segments, including timbre [Toi97], energy-spectrum [Mas03], and musical time series analysis [Car98].

There has been considerable recent interest in the development of touch-based and gesture based interfaces[IU97]. This represents a movement from traditional Graphic User Interfaces (GUI) to Touch-Based User Interfaces (TUI) [Gol07]. These new forms of interfaces help to bring together the virtual world with the real world, providing a more inclusive and immersive interaction environment for users. The iPhone is a device that supports multitouch interaction, a system where multiple fingers are tracked to provide different types of functionality. For example, a touch on the surface with one finger would produce a different effect than when three fingers are used. In addition gestures such as pinching two fingers can be used for actions such as zooming.

This is a type of reality-based interaction [JGH+08], a new field that attempts to bridge the world of the virtual with users in the real world. Many such reality-based interaction models use small mobile devices. Another pertinent example is that of ThinSight [HIB+07], a technology that allows for multi-touch sensing on small, ubiquitous computing devices.

Another very popular new gesture based interface is the Wii remote controller (wiimote) [SG07], a wireless game controller that contains the traditional buttons and gamepads of other game controllers, but also contains a three dimensional accelerometer and an infrared sensor, which is capable of tracking up to four independent infrared light sources in real time. Previous research has included work to detect and track fingers [LKJK08] and also to track two handed interaction in open space [VSI08]. This type of interaction has also been explored with the WISP system [TBL07]. The wiimote has been used to control [WYC08] music generation in an interactive music performance system, and as a way to track the movement of the hands of orchestra conductors [BN08]. The wiimote has been also been used in collaborative computing scenarios as an interactive whiteboard [WL08].

Another relevant research area is surface based interaction. Surface based interaction uses an interface that resembles a tabletop, but contains some sort of projection device to render an image on the surface, and also a way of tracking one or multiple input sources on that surface. The reacTable [JKGB05] is one such collabora-

tive interface that has been used by musicians such as Björk in live musical concert settings. Other such surface based interfaces include DiamondSpin [SVFR04] and SmartSkin [Rek02]. These systems have paved the way for commercial surface computing platforms such as the Microsoft Surface [1] and the SMART Table [2]. This type of interaction has been studied in depth in a collaborative setting with multiple users with various constraints [MHM+08], with elderly users [HWI+07], and as an image classification interface [LGTS+04]. Another related project was the AudioBrowser [CTL+06] which developed a touch based interface coupled with audio feedback to help blind users access information. Another interesting study is the PHASE installation [CRL05], an interface that used haptic feedback to produce music using a game-based metaphor for interaction. There has also been research into the use of collaborative interfaces for the generation of music from large crowds of participants [FP07] in a dance club like setting. A relevant study to the present work is the MUSICTable [SGVF05], an interface that used multidimensional scaling to map music onto a two dimensional surface.

There has been considerable research on the automatic generation of music, including the generation of background music [CZJ+07] [YLC04], the creation of rhythmic patterns [Lab12] [KSH12] and more general automated music generation systems [Erb11] [UN01]. An excellent study was conducted back in 1970 by Howe [How70] about compositional considerations when creating electronic music.

A particularly relevant study was recently conducted [KSH12]. In this paper the authors describe a self-organizing map system that allows users to create rhythms co-creatively and interactively. Also closely related was the [TM07] SENEgal project, which used genetic algorithms to create rhythms from western Africa.

Often the previously described systems have their user interaction paradigm centered on the computer system. In this study we are interested in bringing the creation of music into the physical environment. The creation of new methods of interacting with the computer has seen much activity, including projects using the radiodrum [BMD07] [MT06] , SmartSkin [Rek02] and Cyber composer [ILK04]. A particularly relevant paper involved the use of large numbers of giveaway sensors in a large rave dance setting [FP07].

[1] `http://www.microsoft.com/surface`

[2] `http://www2.smarttech.com/st/en-US/Products/SMART+Table/default.htm`

2.3 Computer analysis of Orca Vocalizations

There are four distinct aspects of work related to our study of Orca Vocalizations. These are Bioacoustics, Computer Supported Collaborative Work, Crowdsourcing and Machine Learning. The following four subsections outline the work related to our research for each of these areas.

2.3.1 Bioacoustics

Bioacoustics is a scientific field of study that combines the fields of biology and acoustics. Although humans have used the sounds of animals to identify and track them for many years, one of the first researchers in this field was Ivan Regen who in 1925 systematically studied the sounds of insects [Zar29]. In the later half of the 20th century, advances in electronic means of recording and producing sound dramatically increased the breadth and scope of the field of bioacoustics. In recent years, the application of the computational tools used in Music Information Retrieval have further extended the possibilities of analyzing sound from biological sources.

For a number of years, computers have been applied to the study of orca vocalizations. One early work [VDS99] used neural networks to determine similarity between orca vocalizations. In this paper, they found that neural networks were able to predict similarity between calls with an accuracy comparable to that of expert human listeners. Another algorithm that has been applied to this problem domain is that of Dynamic Time Warping [BHDM06][BM07]. In these two papers, Brown et al. investigate the use of DTW for calculating similarity between two orca vocalizations and find that this algorithm performs very well. DTW is used frequently in the field of MIR, and this thesis discusses its use as applied to chant traditions from cultures around the world in another chapter.

The mathematical foundations of the pulsed vocalizations by orcas have also been studied [JC.08]. In this paper, formulas for their spectra are rigorously derived from the basic formulas of Fourier analysis. This paper describes in detail the complex spectra that are able to produced by orcas and the biological foundations that underlie them. Another mathematical method that has been used to describe orca vocalizations is the Hilbert-Huang transform, which has been shown [Ada06] to have advantages over using traditional Fourier based methods of calculating spectra.

One important pitfall in the automated classification of orca vocalizations [DJ06] is that there exist non-linearities in sound perception in all animals, including orcas[NRH$^+$99].

When designing tools and algorithms to study these vocalizations, it is important to recognize the differences between human and orca perception of sound.

However, most of the work in analyzing orca vocalizations has been painstakingly done by hand, sometimes using audio cassettes and expert knowledge of orca vocalizations, and sometimes by digitizing vocalizations into a computer and then analyzing spectrograms. It would be of advantage to the field to be able to apply these algorithms to a large dataset and provide collaborative visualization tools to explore it.

2.3.2 Computer Supported Collaborative Work

Although it is a positive thing that this archive of data is now available in electronic form, what would make it even more useful to scientists would be the ability to collaborate together on the process of annotating audio, running experiments and analyzing results. This type of work falls under the rubric of Computer Supported Collaborative Work [BS91] (CSCW), a field which studies multiple individuals working together with computer systems.

In this particular case, we have a number of different communities who have shown interest in this archive, these include Orcalab and its collaborators, whale biologists, developers of bioacoustic and Music Information Retrieval algorithms, and the part of the general public who enjoy listening to whales and might like to help the scientists who are working on this project.

As the scientists in this project are located all over the world, having a system that enables them to work together collaboratively even though they are separated by large distances will be a worthwhile challenge [BM02] [OO00]. There are a number of pitfalls that have been identified [Gru88] including the disparity between who performs the work and who gets the benefits, the breakdown of intuitive decision-making and the difficulty in evaluating the application. To overcome these problems, we are eliciting feedback at every stage of the design process from stakeholders, including the members of Orcalab, and we incorporate this feedback using an iterative development methodology.

In the first phase of the project, which we have just begun, our user community consists solely of the expert scientists who are part of the Orcalab and their close collaborators. Important design goals for this section of the project include the ability to view and annotate any recording in the database and to perform searches using

different criteria, such as year, time, type of call and the type of observation (acoustic/visual) along with other quantities from the daily incidence report. In addition, in this phase, we will begin to run Machine Learning algorithms on the data manually and will use the interface to present these results to the scientists involved.

In the second phase of the project we will open up access to the Orchive to the broader whale biologist and developers of bioacoustic and algorithms communities. The whale biologist community will have many of the same needs as the original Orcalab scientists, but will have less experience with this dataset, and will thus require more structured information on how to find and analyze the information they are looking for. The developers of bioacoustic algorithm community will be more interested in running different audio feature extraction and machine learning algorithms on this data, and a framework for examining and downloading these results and the original audio recordings, in order to analyze them on their local computers.

2.3.3 Crowdsourcing

In the third phase of the project we would like to invite general members of the public to participate by first providing an interface to allow them to find and listen to interesting recordings of orcas. In addition, because of the truly enormous number of recordings, it would be interesting to see if we could get the general public to help with efforts to both locate orca vocalizations on the tape and eventually even label call types.

This type of work, where non-specialists help expert scientists is called Crowdsourcing [Tra08] [Bra08] [How08] [Sur05] and has been used to great advantage in a number of research programs. One of the most successful such programs is Galaxy Zoo [SLB+08]. In this project, astronomers had collected images of many thousands of galaxies and wanted to characterize them by the chiral handedness of their spiral structure, that is, were they spinning clockwise or counterclockwise? It was assumed that there would be an even distribution of these the two chiral hands, left and right, and any deviations from this would be an important and surprising result. Even the most advanced current computer algorithms are not able to categorize galaxies based on their handedness, but humans can do this classification easily. In this paper, the authors describe their system and results, and present results that indicate that there is a hint of positive correlation for galaxies nearer than 0.5 Megaparsecs.

Another research program that benefited from crowdsourcing was the Stardust@home

project [BM06]. Stardust [ABD+97] was a NASA space mission that flew a spacecraft through the tail of comet Wild 2, collected dust from the comet and from interstellar matter using an aerogel, which is a very lightweight clear foam-like material, and then returned the satellite to earth. Comet dust was collected on one side of the aerogel, and the other side of the aerogel was exposed to interstellar matter during the entire mission. The analysis of the particles from the comet were straightforward due to the high number of particles, but the analysis of interstellar grains was much more difficult due to the small size and number of particles. The Stardust@home project allowed users from around the world to signup on a website and interactively view and annotate microscopic images of the aerogel. There was overwhelming participation by the public, and they were able to generate results that were useful to the scientists on the project [ABD+97].

There have been a number of articles that investigate the benefits of crowdsourcing. Hong [HP04] presents results that show that a group of problem solvers with a diverse background can outperform smaller groups of experts. This is of interest because there are only a very small number of orca vocalization experts in the world, but there is a large number of people who are interested in listening to orca calls, as is evidenced by the traffic on the Orca Live `http://orcalive.net` forums.

A recent article [KCS08] describes crowdsourcing with the Amazon Mechanical Turk system, a web based system where people can sign up to work on small tasks in return for micropayments. The advantage with using the Mechanical Turk system is that because people are paid for their work and have to pass a scientist defined test, the results obtained might be of higher quality. The drawback to this system is that the workers must be paid. It would be not extremely difficult to integrate the Mechanical Turk with the Orchive, and a future pilot project to test this is planned.

Because the researchers from Orcalab have devoted so much time and care to the recording and documenting the audio in the Orchive, an important design criteria is to ensure that data entered from non-specialist users will not interfere with the data from expert users and scientists. To this end, we have implemented a reputation management and filtering system, along with roles for various users.

In the current implementation of the Orchive, only expert scientists are allowed to make annotations, and only members of Orcalab and their collaborators are allowed to edit the start date of each recording. As well, users can choose exactly which user's annotations they want to view. In light of the fact that everyone in the current user community knows each other personally, reputation management is handled by

offline personal interactions. When the general public is allowed to contribute using a crowdsourcing model, this system will be expanded and enhanced.

2.3.4 Machine Learning

In addition to the presentation and collaborative annotation that the Orchive supports, we have also developed a set of Music Information Retrieval and Machine Learning tools that are available for researchers to run on the data in the Orchive. All of these tools are part of the Marsyas [Tza08] MIR framework

The first set of tools are audio feature extraction tools that have been adapted from the field of Music Information Retrieval (MIR) [FD02] to the study of orca vocalizations. In particular we have added support for the generation of spectral statistics including FFT spectrum, MFCC coefficients, Zero Crossing Rate, Spectral Centroid, Flux and Rolloff. In section 5.4 we detail some results that were obtained using some of these features and show that they perform well on this class of problems.

We also support classification of audio using Support Vector Machines [ME05], an advanced type of Machine Learning technique that finds optimal hyperplanes in high dimensional datasets, which we then use to classify audio. This classification can be as simple as either "orca", "voiceover" or "background", or can be as complex as classifying different call types or even different pods through their call repertoires.

One important design goal in this project is to make our system as flexible and extensible as possible, to this end, we have come up with a tag-based system that allows researchers to construct their own ontologies. Some of these ontologies would overlap, like "orca" or "voiceover", and some would be distinct to different areas of study. By allowing a user defined structure to emerge, we empower individual research communities to ask and answer the questions most pertinent to them. The obvious drawback to this is that within communities researchers must agree on the same language and syntax to annotate calls. We hope to provide tools within and external to the site to help encourage the collaboration necessary to converge on the same vocabulary.

Chapter 3

Computer Analysis of Partially-Annotated Chants

I will first focus on the analysis of partially annotated chants from a variety of music traditions. This section of the thesis brings together techniques from Music Information Retrieval with web-based collaborative technologies to help experts in the field of Ethnomusicology. In particular, in this chapter we make use of a new algorithm to do fundamental frequency estimation to extract the pitch of vocal performances, do a variety of analysis techniques on this data, and then build a custom designed interface for musicologists. It is by bringing together all these tools along with a flexible user interface that allows these expert users to ask new questions about these diverse chant traditions.

This research is important because the symbolic techniques used in traditional musicology research are not well suited to research into these partially notated chants from cultures around the world. One of the primary reasons for this is that although western music is often fully annotated, and the score was designed by the composer precisely for the purpose of the transmission of their music, these chants are still primarily passed on via an oral tradition that is only partially annotated. I present a tool that allows researchers to examine these chants with both the written and aural components presented in combination with each other.

3.1 Introduction to Partially-Annotated Chant research

The field of Ethnomusicology is a sub-discipline of Musicology that focuses on the study of the socio-cultural aspects of music in societies around the world. Computational Ethnomusicology [TASW07] is a new field that uses the computational techniques commonly used in Music Information Retrieval (MIR) to analyse music and audio from social and cultural traditions from around the world. In the present work I am interested in applying the tools of MIR to the analysis of partially annotated chant traditions from around the world, including Torah cantillation, Koran recitation, St. Gallen plainchant and Hungarian laments.

Our work in developing tools to assist with chant research is a collaboration with Dr. Daniel Biro, a professor in the School of Music at the University of Victoria. He has been collecting and studying recordings of chant with specific focus on how music transmission based on oral transmission and ritual was gradually changed to one based on writing and music notation. The examples studied come from improvised, partially notated, and gesture-based [Kru90] notational chant traditions: Hungarian *siratok* (laments) Torah cantillation [Zim00], tenth century St. Gallen *plainchant* [Tre82] and Koran recitation.

The application of computerized analysis tools to these partially annotated chant traditions is at an early stage, so early in fact, that many of the main research questions have not yet been formulated. It is therefore important to have, as a first step, an interface that lets a group of ethnomusicologists to explore the data and develop the research questions that will then be investigated using more automated methods.

This work presents a series of offline tools that take the audio from these chant recordings and analyze it using a variety of algorithms, experiments showing the effectiveness of these tools, and also presents a rich web-based interface to let groups of ethnomusicologists around the world investigate and explore this data and the whole collection of chants from four culturally diverse chant traditions.

3.2 Background Information for Partially-Annotated Chants

During the development of musical notation, there was an evolving relationship between the way that the music was notated, or the syntax, and the way that the music was understood by the people in the culture, or semiotics. The transmission of culture that was initially performed orally was transformed to one based on interpretation of writing or hermeneutics.

Musical notation could either function as a mental tool to help people reconstruct a melody that they remembered, or could help them construct a new melody. Chant scholars study the question of how exactly melodies, or melodic formuale, became solidified into musical material.

In our research, we examine examples from improvised, partially improvised, partially notated and gesture-based notational chant traditions including Hungarian siratok (laments), Torah cantillation, tenth century St. Gallen plainchant, and Koran recitation. We are building tools to help musicologists study examples from these various traditions.

Each of these traditions has a certain alphabet of different signs. For example, in Torah recitation, there are approximately 20 different pitch contours, or gestures. These signs, and the rules by which they can be joined together, is called by musicologists the syntagmatic structure, because these signs form the syntax of the language. Syntagmatic analysis is the analysis of the surface structure of a text, or in this case, of a song.

An additional word or two must be said about the word "gesture". In it's simplest definition, a gesture is simply the pitch contour of a sung musical phrase. The reason the word gesture was chosen by the musicologists that study these songs is because in a number of traditions, including the Torah chant, there would be a singer at the front of the room, and at the back of the room would be the old rabbis who would act sort of like conductors and would gesture with their hands as to how the words should be intoned. In time, the word "gesture" has become imbued with subtle meanings, and a precise definition of gesture is hard to pin down, but in general, it means the shape of the pitch contour.

On the other hand, the analysis of the underlying meaning of a song, where one attempts to look at the relationships between these different signs, both in one song and between songs, is known as paradigmatic analysis. In paradigmatic analysis, one

is concerned both with the meanings of the words and with the way that words or phrases are repeated and varied.

Both syntagmatic and paradigmatic analysis come to us from the field of semiotics, which is the study of signs and signifiers. Semiotics is a field that developed in the 20th century which merged the study of the syntax of language with the idea that there were deep relations between the representations of the signs that made up a language.

Although semiotics is a field quite foreign to Computer Scientists and to penetrate because of its highly specialized vocabulary it holds great promise for the study of chant. We support semiotic analysis by building tools that allow researchers to examine the syntagmatic (syntax) and paradigmatic (underlying meaning and repetition of signs) properties of these songs.

In this thesis, we explore examples from these various traditions through computational tools for paradigmatic analysis of melody (melodic formulae) and the way that various words are sung in terms of pitch contours (gesture). We can then enhance our understanding of these traditions by synthesizing the ideas of melody, melodic syntax and musical semotics into a coherent whole.

3.3 Melodic Contour Analysis Tool for Chant Research

Our tool takes in a (digitized) monophonic (one voice) or heterophonic (more than one voice) recording and produces a series of successively more refined and abstract representations of the segments it contains as well as the corresponding melodic contours. More specifically the following analysis stages are performed:

- Hand Labeling of Audio Segments
- First Order Markov Model of Sign Sequences
- F0 Estimation
- F0 Pruning
- Scale Derivation: Kernel Density Estimation
- Quantization in Pitch

- Scale-Degree Histogram
- Histogram-Based Contour Abstraction
- Dynamic Time Warping for Contour Similarity
- Plotting and Recombining the Segments

Hand Labeling of Audio Segments

The recordings are manually segmented and annotated by the expert. Even though we considered the possibility of creating an automatic segmentation tool, it was decided that the task was too subjective and critical to automate. Each segment is annotated with a word/symbol that is related to the corresponding text or performance symbols (for example cantillation marks) used during the recitation.

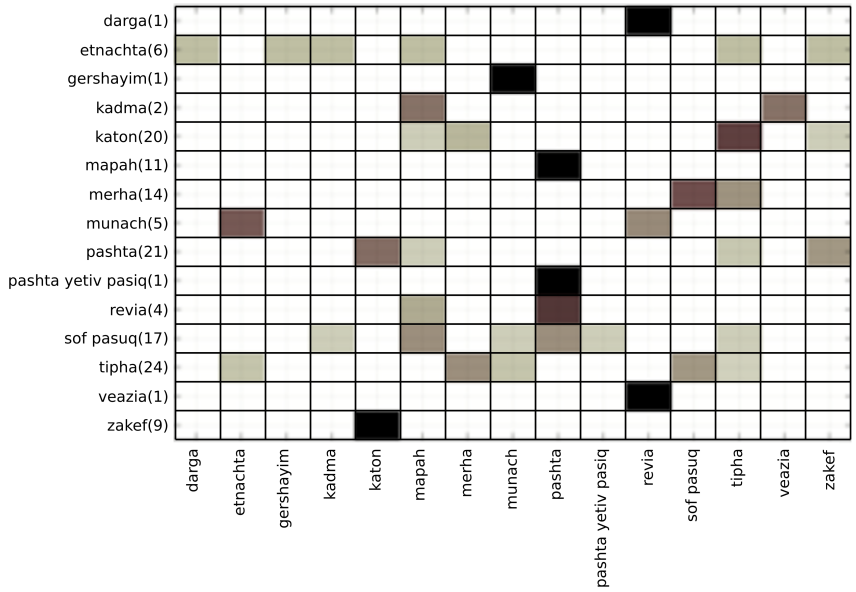

Figure 3.1: Syntagmatic analysis with a first-order Markov model of the sequence of Torah trope signs for the text Shir Ha Shirim ("Song of Songs").

First Order Markov Model of Sign Sequences

In order to study the transitions between signs/symbols we calculate a first order Markov model of the sign sequence for each recording. We were asked to perform this

type of syntagmatic analysis by Dr. Biro. Although it is completely straightforward to perform automatically using the annotation, it would be hard, if not impossible, to calculate manually. Figure 3.1 shows an example transition matrix. For a given trope sign (a row) it shows how many total times it appears in the example (numeral after row label), and in what fraction of those appearances is it followed by each of the other trope signs. The darkness of each cell corresponds to the fraction of times that the trope sign in the given row is followed by the trope sign in the given column. (NB: Cell shading is relative to the total number of occurrences of the trope sign in the row, so, e.g., the black square saying that "darga" always precedes "revia" represents 1/1, while the black square saying that "zakef" always precedes "katon" represents 9/9.) This type of analysis can help identify the syntactic role that different signs have.

F0 Estimation

After the segments have been identified, the fundamental frequency ("F0" in this case equivalent to pitch) and signal energy (related to loudness) are calculated for each segment as functions of time. We use the SWIPEP fundamental frequency estimator [Cam07] with all default parameters except for upper and lower frequency bounds that are hand-tuned for each example. For signal energy we simply take the sum of squares of signal values in each non-overlapping 10-ms rectangular window.

The SWIPEP algorithm [Cam07] uses an algorithm that is related to autocorrelation, and using a cosine as the kernel, performs an integral transform of the spectrum. An integral transform is defined as a function that takes one function and transforms it to give another function. Unlike autocorrelation, which uses the square of the magnitude of the spectrum, SWIPEP uses the square root of the magnitude of the spectrum. SWIPEP also modifies the cosine kernel in order to avoid some of the problems associated with autocorrelation. These involve first zeroing the first quarter of the first cycle of the cosine, this allows it to avoid the maximum value at zero lag that occurs when using autocorrelation. It then avoids the periodicity that autocorrelation experiences when analyzing periodic signals by multiplying the kernel by a 1/f envelope. To force the width of the main spectral lobes to match the width of the positive cosine lobes, it also normalizes the cosine kernel and applies a pitch-dependant window size.

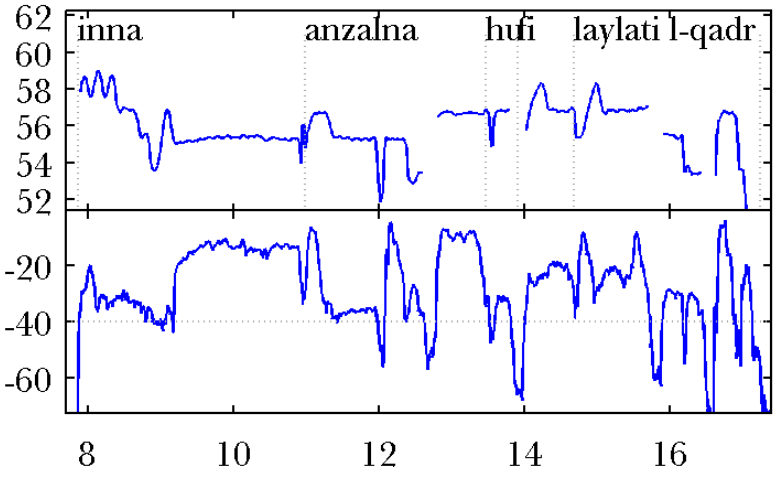

Figure 3.2: F0 contour

F0 Pruning

The next step is to identify pauses between phrases, so as to eliminate the meaningless and wildly varying F0 estimates during these noisy regions. We define an energy threshold, generally 40 decibels below each recording's maximum. If the signal energy stays below this threshold for at least 100 ms then the quiet region is treated as silence and its F0 estimates are ignored. Figure 3.2 shows an excerpt of the F0 and energy curves for an excerpt from the Koran sura ("section") Al-Qadr ("destiny") recited by the renowned Sheikh Mahmud Khalil al-Husari from Egypt.

Quantization in Pitch

Following the pitch contour extraction is pitch quantization, which is the discretization of the continuous pitch contour into discrete notes of a scale. Rather than externally imposing a particular set of pitches, such as an equal-tempered chromatic (the piano keys) or diatonic scale, we have developed a novel method for extracting a scale from an F0 envelope that is continuous (or at least very densely sampled) in both time and pitch. Our method is inspired by Krumhansl's time-on-pitch histograms adding up the total amount of time spent on each pitch [Kru90]. For this application we decided to have a pitch resolution of one cent [1], so we cannot use a simple histogram. Instead we use a statistical technique known as non-parametric

[1] One cent is 1/100 of a semitone, corresponding to a frequency difference of about 0.06%

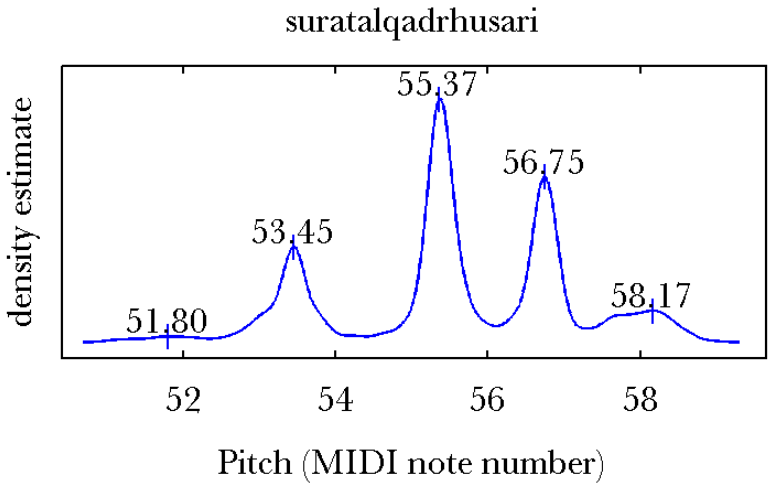

Figure 3.3: Recording-specific scale derivation

kernel density estimation, with a Gaussian kernel [2]. More specifically a Gaussian (with standard deviation of 33 cents) is centered on each sample of the frequency estimate and the Gaussians of all the samples are added to form the kernel density estimate. We chose the standard deviation of 33 cents as a result of an empirical investigation of this problem domain, of which this value gave us the most useful histogram for this application. The resulting curve is our density estimate; like a histogram, it can be interpreted as the relative probability of each pitch appearing at any given point in time. Figure 3.3 shows this method's density estimate given the F0 curve from Figure 3.2.

Scale-Degree Histogram

We interpret each peak in the density estimate as a note of the scale. We restrict the minimum interval between scale pitches which was set to 80 cents in this application using an empirical method that attempted to make sure that all peaks in the derived histogram had an associated scale degree, but that no false positives corresponding to split peaks were created. We did this by choosing only the higher peak when there are two or more very close peaks. This method's free parameter is the standard deviation of the Gaussian kernel, which provides an adjustable level of smoothness to

[2]Thinking statistically, our scale is related to a distribution given the relative probability of each possible pitch. We can think of each F0 estimate (i.e each sampled value of the F0 envelope) as a sample drawn from this unknown distribution so our problem becomes one of estimation the unknown distribution given the samples

our density estimate; we have obtained good results with a standard deviation of 30 cents.

Once we have determined the scale, pitch quantization is the trivial task of converting each F0 estimate to the nearest note of the scale. In our opinion these derived scales are more true to the actual nature of pitch-contour relationships within oral/aural and semi-notated musical traditions. Instead of viewing these pitches to be deviations of pre-existing "normalized" scales our method defines a more differentiated scale from the outset. With our approach the scale tones do not require "normalization" and thereby exist in an autonomous microtonal environment defined solely on statistical occurrence of pitch within the performance. Once the pitch contour is quantized into the recording-specific scale calculated using Kernel density estimation, we can calculate how many times a particular scale degree appears during an excerpt. The resulting data is a scale-degree histogram which is used create simplified abstract visual contour representations.

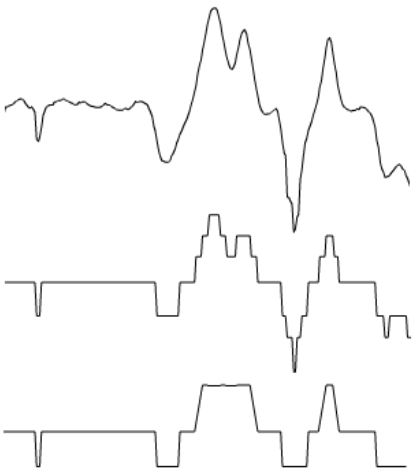

Figure 3.4: Melodic contours at different levels of abstraction (top: original, middle: quantized, bottom: simplified using 3 most prominent scale degrees

Histogram-Based Contour Abstraction

The basic idea of histogram-based contour abstraction is to only use the most salient discrete scale degrees (the histogram bins with the highest magnitude) as significant points to simplify the representation of the contour. By adjusting the number of prominent scale degrees used to represent the simplified representation the researchers

can view/listen to the melodic contour at different levels of abstraction and detail. Figure 3.4 shows an original continuous contour, the quantized representation using the recording-specific derived scale and the abstracted representation using only the 3 most prominent scale degrees. In our future work on this topic, we wish to add the ability to pitch shift the actual singers performance to our simplified pitch contour. With this we can imagine that we are interacting with the ghosts of the performers, distant from us in time and space. We anticipate this will be much more satisfying than simply playing a generated sine tone, which is rather unpleasant to listen to.

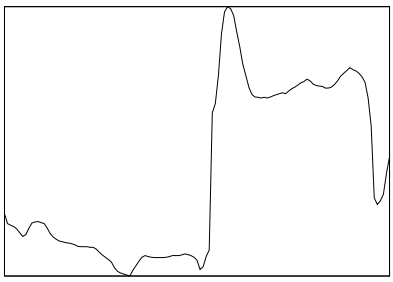

Figure 3.5: One example of the F0 contour for the pashta gesture

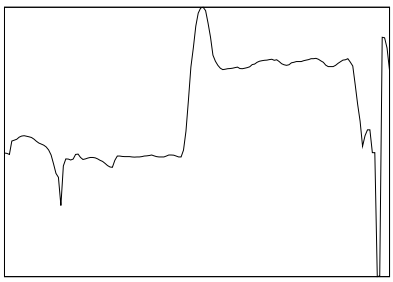

Figure 3.6: Another example of the F0 contour for the pashta gesture

In the next section we show that these simplified abstract contour representations result in better retrieval performance than the original "continuous" pitch contours.

One of the main aspects in the studying of signs in the context of chant and recitation is to what extent they convey gesture information that is invariant with respect to the underlying text. To study this question it was necessary to develop a method to compare the pitch contours of different realizations from different parts of the audio recording of the same sign.

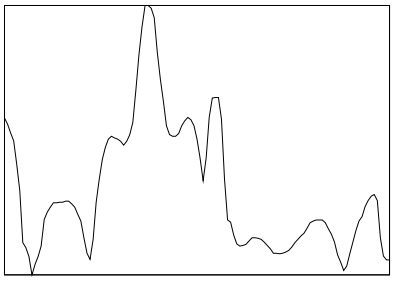

Figure 3.7: An example of the F0 contour for the sof pasuq gesture

Figure 3.8: An example of the F0 contour for the first pashta gesture but doubled in length

Dynamic Time Warping for Contour Similarity Calculation

Dynamic Time Warping (DTW) is a technique by which the similarity between two different time sequences can be measured. It allows a computer to find an optimal match between two sequences by performing a non-linear warping of one sequence to the other. The technique of dynamic programming is used for efficient implementation. An example of DTW in Music Information Retrieval is to compare the tempo variations between two different performances of a classical symphony. The DTW algorithm would identify the parts of the two symphonies that were played at the same tempo as a diagonal line, with the line varying above and below the diagonal when the tempo was different between the two pieces.

First the similarity matrix between the two pitch contours we are comparing is calculated by comparing the values of the two pitch contours at each time index. Based on the calculated similarity matrix the DTW algorithm finds the optimal alignment path of the two sequences and calculates the cost of that alignment. When the con-

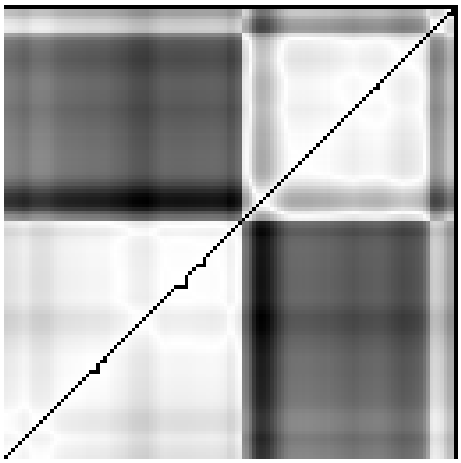

Figure 3.9: The Similarity Matrix of one pashta gesture (11pashta) compared to itself. Superimposed on the figure is the Dynamic Time Warping curve showing the optimally matching path between the two songs.

tours are similar the alignment cost will be small compared to when the contours are dissimilar. The matching process is pitch shift invariant and allows variations and tempo stretching. That way for any particular sign (pitch contour) we can sort the sign (pitch contours) by similarity.

Plotting and Recombining the Segments

To illustrate the technique we use the gestures of two separate annotated recordings of a section of the Torah. One of these was recorded in Morocco, and the other was recorded in Hungary. Figures 3.5, 3.6, 3.7 and 3.8 show the F0 contour of the sections of the audio file from a Torah recording from Hungary. Figure 3.5 shows a "pashta" sign (one of the cantillation signs from the Torah which means "Stretching out"), Figure 3.6 shows another pashta sign from further along in the audio file. Figure 3.7 shows a "sof pasuq" (a cantillation sign which means "End of verse") gesture and Figure 3.8 shows the first pashta gesture, but with the sample stretched by a factor of two.

The figures 3.9, 3.10, 3.11 and 3.12 show Similarity Matricies and the alignment paths computed using DTW for these four gestures compared to the first pashta gesture. White areas are highly similar and black areas have low similarity. In Figure 3.9 the first pashta gesture is compared to itself. The DTW curve is overlaid in black and is basically a straight diagonal line from one corner to the opposite corner, showing that the optimal path between the start and the end of the file is a direct

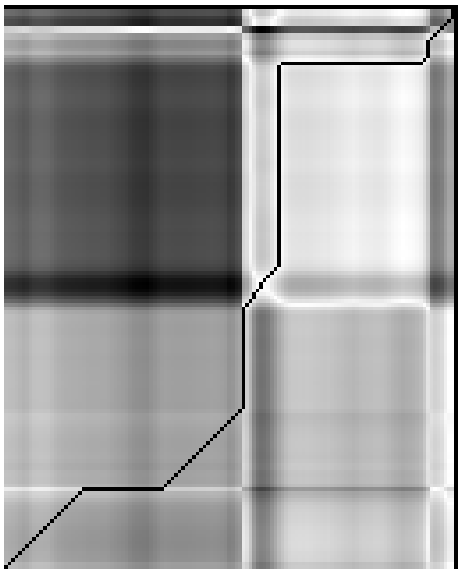

Figure 3.10: The Similarity Matrix of one pashta gesture (11pashta) compared to another pashta gesture (42pashta). Superimposed on the figure is the Dynamic Time Warping curve showing the optimally matching path between the two songs.

alignment of one file to the other. Figure 3.12 shows a similar behavior, except that the slope of the line is shallower. Figure 3.10 shows the comparison of one pashta gesture to another. This path had a DTW cost of 23.8. Figure 3.11 shows an alignment between the pashta gesture and a sof pasuq gesture. One can see that the line is not only not diagonal, but that the line is often on dark areas which denote high alignment cost.

Table 3.1 shows the average precision for particular signs for two recordings of the same excerpt from the Torah - one from Hungary and one from Morocco. The average precision here is simply the average of all the precision values for a particular sign.

Each recording contains approximately 130 realizations of each sign with a total of 12 unique signs. Two pitch contours are considered relevant to each other if they are annotated by the same sign. For each "query" contour we return a list of results which are the pitch contours sorted by the alignment cost of the DTW. Average precision emphasizes returning more relevant contours earlier. It is the average of all the precisions computed after truncating the list of returned results after each of the relevant documents in turn. Unlike traditional retrieval systems where the mean average precision can be used to characterize the overall system performance, in our case we are more interested in the individual difference in precision among different

Gesture (Hungary)	Average Precision (Hungary)	Gesture (Morocco)	Average Precision (Morocco)
tipha	0.662	katon	0.453
pashta	0.647	mapah	0.347
mapah	0.641	tipha	0.303
katon	0.604	sofpasuq	0.285
etnachta	0.601	pashta	0.242
sofpasuq	0.591	merha	0.251
merha	0.537	etnachta	0.150
revia	0.372	zakef	0.125
zakef	0.201	revia	0.091
kadma	0.200	kadma	0.043

Table 3.1: Average precision for different signs

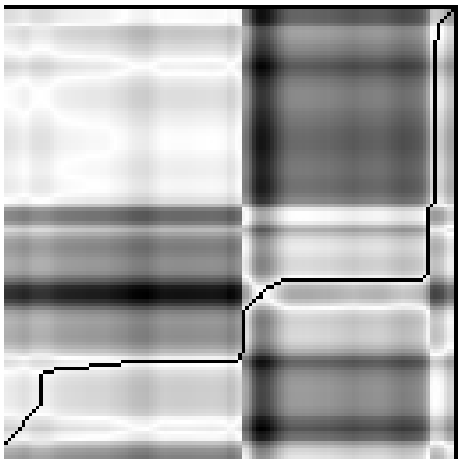

Figure 3.11: The Similarity Matrix of one pashta gesture (11pashta) compared to a sof pasuq gesture (18sofpasuq). Superimposed on the figure is the Dynamic Time Warping curve showing the optimally matching path between the two songs.

signs.

The mean average precision is calculated using the following equation:

$$AveragePrecision = \frac{\sum_{r=1}^{N} P(r)}{M}$$

Where r is the rank of each document, N is the number of documents retrieved, P is the precision of the given document, and M is the total number of relevant documents.

These differences show which signs have well-defined gestural characteristics and which signs are not interpreted consistently.

Ultimately the numbers are only meaningful after careful interpretation by an expert. For example based on Table 3.1 one can infer that the performer in the Hungarian version had more consistent interpretations of the signs than the performer in the Moroccan version.

We have also investigated the retrieval effectiveness of quantized contour representations at different levels of abstraction using the approach described above. In this case it makes sense to use Mean Average Precision across queries to explore what is the best level of abstraction for this task.

This first DTW analysis was conducted using the continuous pitch values determined by the SWIPEP algorithm. We then extended this analysis by quantizing the pitch contours, calculating the pairwise score between each contour and then cal-

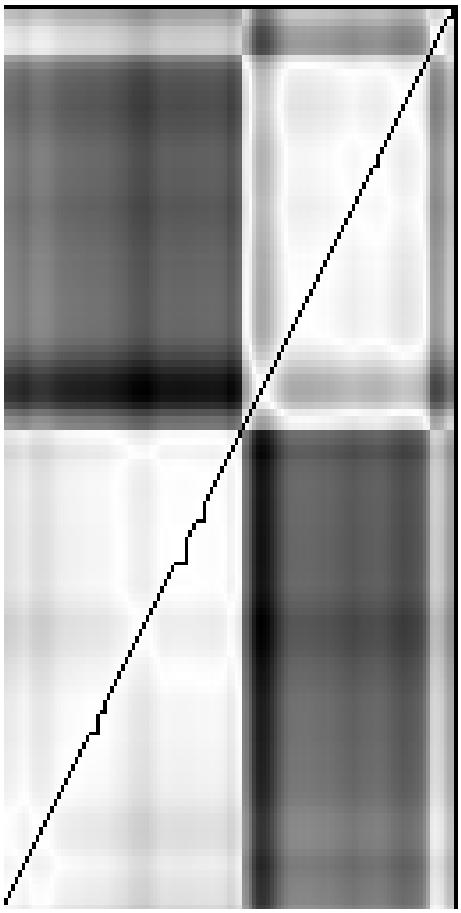

Figure 3.12: The Similarity Matrix of one pashta gesture (11pashta) compared to itself, doubled in length. Superimposed on the figure is the Dynamic Time Warping curve showing the optimally matching path between the two songs.

culating the mean average precision recall. We did this for all possible number of histogram bins, from the maximum number of scale degrees of 13, down to only the most popular histogram bin. We chose the number 13 because it provided us with the most direct comparison between our derived scale and an equal tempered scale. We then repeated this analysis with notes from a western equal-tempered scale. The total range of notes was from the A2# (the A# two octaves below middle C) to C4 (middle C). This gave a total of 16 semitones, of which we used the most common 13 scale degrees. For all of these possible histogram bin numbers, we converted all notes to these quantized values and did a pairwise DTW comparison between all of them. We then calculated the mean average precision recall for each histogram bin quantization level. These results are presented in Figure 3.13.

From this graph and Table 3.2, we can see that the optimal number of histogram

bins is 2 when notes are quantized to our derived scale. The mean average precision recall at this level is 0.493. After this, the curve quickly drops, and then remains at a steady state level of approximately 0.41. This is significantly better than using the "continuous" contour mean average precision of 0.2951. When we quantize the notes to the equal-tempered scale, the maximum value of 0.443 is also obtained with 2 histogram bins. It is important to note that the value of 0.493 that is derived when the data-driven approach of using the notes that are actual chanted is higher than the value derived from using the equal-tempered scale. Obviously, the singers do not tune themselves to a western scale. This shows the fundamental utility of our method of deriving the quantized scale from the notes that are actually sung.

These results are shown in a more intuitive way in Figure 3.14. In this figure three "sof pasuq" and three "pashta" contours were chosen, and were quantized to the derived, data-driven scale using the optimal value of 2 histogram bins. One can see that the "sof pasuq" contours have quite a different shape. This visualization shows the utility of our approach.

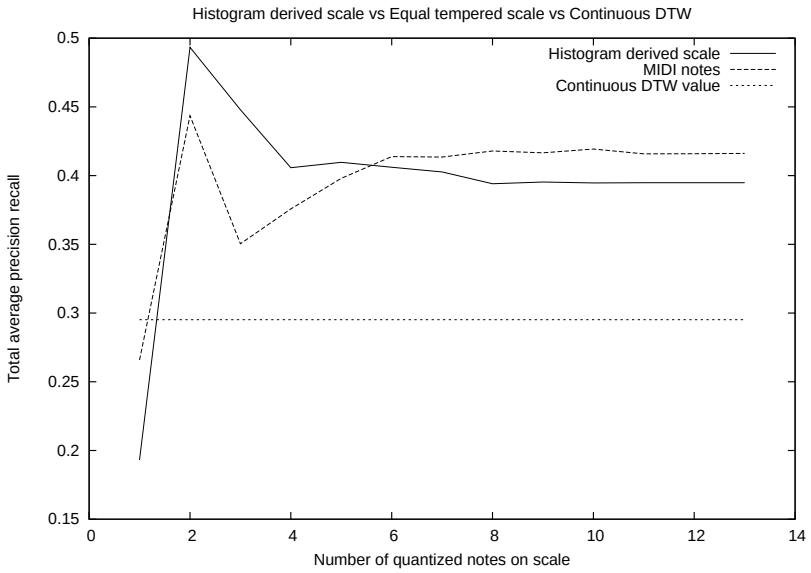

Figure 3.13: Mean average precision recall when quantizing the notes before DTW analysis. Shown are the results for quantizing to a song specific scale versus an equal tempered scale versus the mean average precision recall in the continuous case.

Number of Bins	Data Driven	Equal Temperment
1	0.1931	0.26581
2	0.4932	0.44356
3	0.4479	0.35044
4	0.4057	0.37572
5	0.4097	0.39797
6	0.4061	0.41386
7	0.4026	0.41350
8	0.3941	0.41791
9	0.3953	0.41655
10	0.3947	0.41931
11	0.3948	0.41584
12	0.3948	0.41594
13	0.3948	0.41617

Table 3.2: Table of mean average precision values when quantizing the notes before DTW analysis. Shows the calculated values for the Data-driven and Equal-temperment approaches.

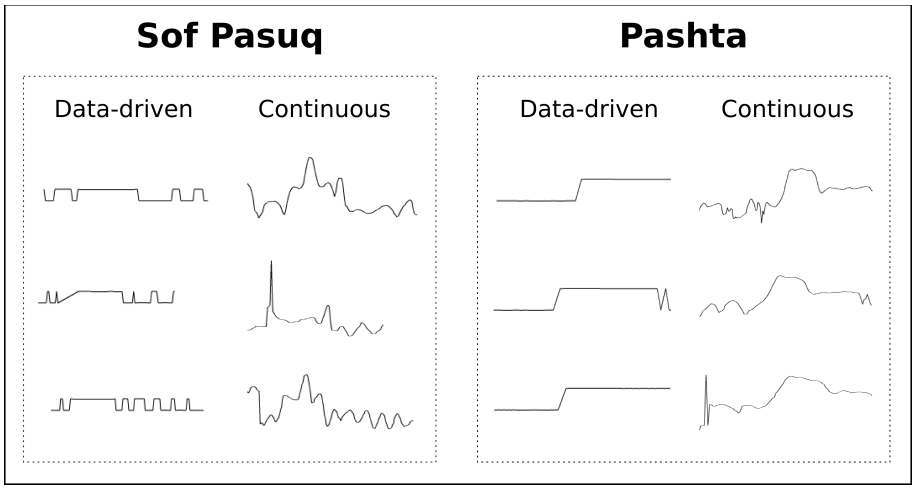

Figure 3.14: Comparison of contour quantized to the two most prevalent scale degrees in a data-driven approach to the original continuous contour. Shown are three examples of the signs "sof pasuq" and "pashta"

Cantillion interface

We have developed a browsing interface that allows researchers to organize and analyze chant segments in a variety of ways. The user manually segments each recording into the appropriate units for each chant type (such as trope sign, neumes, semantic units, or words). The pitch contours of these segments can be viewed at different levels of detail and smoothness using a histogram-based method. The segments can also be rearranged in a variety of ways both manually and automatically. That way one can compare the beginning and ending pitches of any trope sign, neume or word or compare the relationships of one neume or trope sign to its neighbors.

Our eventual goal is to explore the stability of melodic gesture and pitch content in a variety of contexts both within a given chant style and across chant styles. In addition we want to explore how the stability related to the chant texts and textual syntax. These are hard questions without clear answers. The user interface has been designed to assist and support the analysis conducted by expert musicologists without trying to impose a specific approach. Being able to categorize melodic formulae in a variety of ways allows for a larger database of their gestural identities, their functionality to parse syntax, and their regional traits and relations. A better understanding of how pitch and contour helps to create gesture in chant might allow for a more comprehensive view of the role of gesture in improvised, semi-improvised and notated

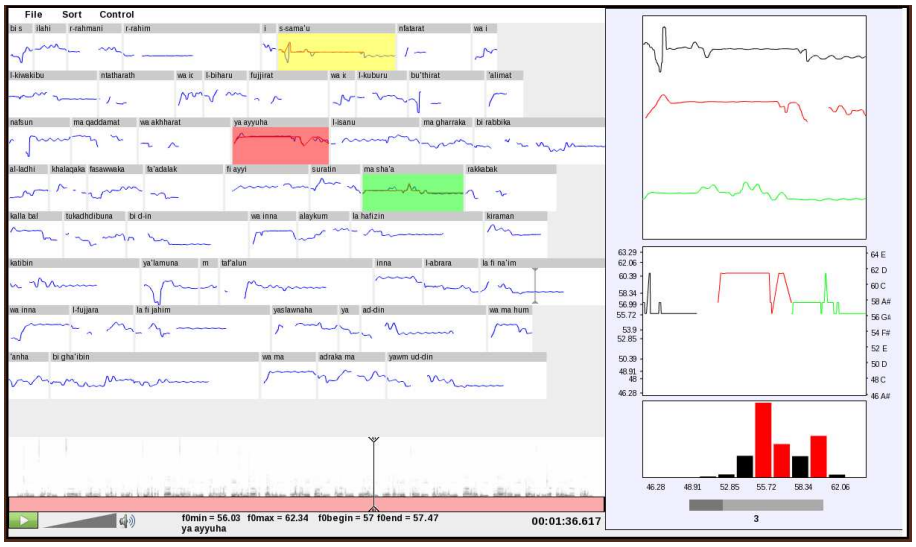

Figure 3.15: Web-based *Flash* interface to allow users to listen to audio, and to enable interactive querying of gesture contour diagrams.

chant examples.

We have chosen to implement the interface as a web-based Flash program [3]. Such web based interfaces can increase the accessibility and usability of a program, make it easier to provide updates, and can enhance collaboration between colleagues by providing functionality that lets researchers more easily communicate their results to each other. The interface (shown in Figure 3.15) has four main sections: a sound player, a main window to display the pitch contours, a control window, and a histogram window.

The sound player window displays a spectrogram representation of the sound file with shuttle controls to let the user choose the current playback position in the sound file. It also provides controls to start and pause playback of the sound, to change the volume.

The main window shows all the pitch contours for the song as icons that can be repositioned automatically based on a variety of sorting criteria, or alternatively can be manually positioned by the user. The name of each segment (from the initial segmentation step) appears above its F0 contour. The shuttle control of the main sound player is linked to the shuttle controls in each of these icons, allowing the user to set the current playback state either by clicking on the sound player window, or

[3] Cantillion Website - `http://cantillion.sness.net`

directly in the icon of interest. When the user mouses over these icons, some salient data about the sign is displayed at the bottom of the screen.

The control window has a variety of buttons that control the sorting order of the icons in the main F0 display window. A user can sort the icons in playback order, alphabetical order, length order, and also by the beginning, ending, highest and lowest F0. The user can also display the sounds in an X-Y graph, with the x-axis representing highest F0 minus lowest F0, and the y-axis showing the ending F0 pitch minus the beginning F0 pitch. Also in this section are controls to toggle a mode to hear individual sounds when they are clicking on, and controls to hide the pitch contour window leaving just the label. There are also buttons allowing the user to choose to hear the original sound file, the F0 curve applied to a sine wave, or the quantized F0 curve applied to a sine wave.[4]

When an icon in the main F0 display window is clicked, the histogram window shows a histogram of the distribution of quantized pitches in the selected sign. Below this histogram is a slider to choose how many of the largest histogram bins will be used to generate a simplified contour representation of the F0 curve. In the limiting case of selecting all histogram bins, the reduced curve is exactly the quantized F0 curve. At lower values, only the histogram bins with the most items are used to draw the reduced curve, which has the effect of reducing the impact of outlier values and providing a smoother "abstract" contour.

Shift-clicking selects multiple signs; in this case the histogram window includes the data from all the selected signs. We often select all segments with the same word, trope sign, or neume; this causes the simplified contour representation to be calculated using the sum of all the pitches found in that particular sign, enhancing the quality of the simplified contour representation.

Below the histogram window is a window that shows a zoomed-in graph of the selected F0 contours. When more than one F0 contour is selected, the lines in the graph are color coded to make it possible to easily distinguish the different selected signs.

[4]Thinking statistically, our scale is related to a distribution giving the relative probability of each possible pitch. We can think of each F0 estimate (i.e., each sampled value of the F0 envelope) as a sample drawn from this unknown distribution, so our problem becomes one of estimating the unknown distribution given the observations.

3.4 Conclusions and Future Work

The identity of chant formulae in oral and aural chant traditions is to a large extent determined by gesture and contour rather than by discrete pitches. As described in the introduction, the word gesture for musicologists has many subtle gradations of meaning, and it is the role of our tool to provide as much subtly in the visualization and description of this gesture as possible.

Computational approaches assist with the analysis of these gestures/contours and enables the juxtaposition of multiple views at different levels of detail in a variety of analytical (paradigmatic and syntagmatic) contexts. The possibilities for such complex analysis methods would be difficult if not impossible without such computer-assisted analysis. Employing these tools we hope to better understand the role of and interchange between melodic formulae in oral/aural and written chant cultures. While our present analysis investigates melodic formulae primarily in terms of their gestural content and semantic functionality, we hope that these methods might allow scholars to reach a better understanding of the historical development of melodic formulae within various chant traditions.

By combining the expert knowledge of our scientific collaborators with new multimedia web-based tools in an agile development strategy, we have been able to ask new questions that had previously been out of reach. Chant research is a challenging domain where problem seeking, that is, the search for the precise question to ask, is important. Participatory design together with content-aware visualizations and analysis tools can help researchers interact with large collections of annotated audio recordings of chant in interesting new ways. The integration of all the different components in a single web-based interface is critical for an effective system. Given the subjective interpretive nature of musicological research, each algorithm in isolation would be of little use. This necessitates the development of the system as a whole and makes evaluation harder. Ultimately there are only few experts users (one in our case) and the only feedback we can receive is through them. By including them in the design we have been able to create a system that our expert finds useful and is willing to spend significant time interacting with the tool.

There are many directions for future work. We are planning to explore the histogram-based contour simplification in conjunction with the dynamic time warping alignment process to identify what is the "optimal" simplification of the pitch contours. More careful study of the results by musicologists is also required. Making

the system available on the web can help collaborative approaches and reduce the learning curve required for usage. We also hope to make the annotation process part of the web interface and enable uploading of recordings from researchers around the world.

Chapter 4

Novel Interfaces for Music Exploration

In this chapter I present work that takes the high-dimensional feature vectors produced by music feature extraction algorithms common to Music Information Retrieval and use the algorithmic technique of Self-Organizing Maps to project those high dimensional feature vectors onto a two dimensional grid.

Based on this two-dimensional map of songs a number of interfaces were developed that allow users to browse this map. I present a system that allows a variety of different input interfaces and output views to be used to navigate this space, which include the wiimote, radiodrum, iPhone and keyboard/mouse.

In addition the generated 2-D map is combined with tag information associated with the songs to generate a novel Self-Organizing Tag Cloud. A standard tag cloud is a two-dimensional graphical representation of words related to a topic where words are typically arranged alphabetically or randomly. Our Self-Organizing Tag Cloud takes the words in a tag cloud and arranges them based on their audio similarity to each other, and does this by finding the centroid of all the tag instances over the Self-Organized Map.

4.1 Introduction to Music Exploration with Novel Interfaces

There is a growing interest in touch-based and gestural interfaces as alternatives to the dominant mouse, keyboard and monitor interaction. Content and context-aware

visualizations of audio collections have been proposed as a more effective way to interact with the increasing amounts of audio data available digitally. *Audioscapes* is a framework for prototyping and exploring how touch-based and gestural controllers can be used with state-of-the-art content and context-aware visualizations. By providing well-defined interfaces and conventions, a variety of different audio collections, controllers and visualization methods can be combined to create innovative ways of interacting with large audio collections. We describe the overall system architecture, the currently available components and specific case studies.

Personal digital music collections are continuously growing and frequently feature thousands of tracks. Browsing and navigating these large collections is challenging. The most common way of interaction is using textual meta-data such as artist name or genre. More recently tag folksonomies have also been utilized. A folksonomy is a system of classification that comes from a group of users collaboratively creating and managing tags in an effort to annotate and categorize content, and are a new phenomenon made possible by collaborative technologies on the web. A variety of visualizations based on automatically analyzed musical content have also been proposed. Tag clouds are a two-dimensional stylized visual representation of a list of words with different visual design characteristics for each word. Tag clouds are commonly ordered either alphabetically or randomly, and in this chapter I examine the utility and engageability of placing similar tags next to each other. In order to determine similarity of tags we use a self-organizing map based on acoustical features derived from songs rather than using the more common tag co-occurrence to measure similarity. To evaluate the proposed approach, a subset of the Magnatune database [1] tagged using the Tag-a-tune game-with-a-purpose. Experimental results are presented which show that using a self-organizing tag cloud is both faster and more fun.

In several cases these are content and context aware interfaces that rely on automatic analysis of the audio signal to extract high-level content information. Many of these interfaces are proof-of-concept prototypes that, although capable of demonstrating the potential of this approach, are not directly practical and usable.

Recently there has been an increasing interest in alternatives to the traditional mouse/keyboard human-computer interaction. Touch-based and gestural interfaces have changed status from research curiosities to being part of many mainstream consumer computing devices.

[1] `http://magnatune.com`

A closely related trend is the diversification of form factors beyond the traditional desktop/laptop design. Representative examples range from mobile phones to immersive displays and tabletop-surfaces.

We believe that browsing and navigation of large audio and especially music collections is a domain that would significantly benefit from the use of interfaces that go beyond [FR82] the traditional keyboard,mouse and monitor paradigm. Unfortunately, the large variety of different protocols, programming environments and operating systems make development of such interfaces more challenging. In addition, the content and context-aware visualizations required demand state-of-art signal processing and machine learning techniques which are not familiar to most researchers in human-computer interaction.

AudioScapes [NT09] is a framework developed to explore the design space of non-traditional interfaces for audio and music collection browsing based on the metaphor of a surface. In this metaphor the individual audio recordings or music tracks are mapped onto a 2-dimensional surface which can be navigated using different controller interfaces. The overall abstract architecture captures the structure of the majority of existing systems and provides significant design flexibility in the choice of individual specific components. By providing well-defined interfaces and conventions, a variety of different audio collections, controllers and visualization methods can be combined to create innovative ways of interacting with large audio collections.

Our design goal is to provide effective interaction without relying on textual metadata. There are many usage scenarios where having to know artist names and album titles or having to read text is impractical. Currently the most common approach in these cases is just playing random songs (the so-called "shuffle"). Although satisfactory for small and homogeneous collections, this approach is not particularly effective for larger audio collections. These issues become even more pronounced when the users have vision and/or motion disabilities. For example, finding a particular artist out of a long list of text using a scroll-wheel can be very difficult or even impossible for a user with motor disabilities. Similarly, reading text on a screen is not directly possible for a blind user. We have used *AudioScapes* to design and prototype interfaces for such users. Although we do not focus on textual metadata, the proposed interfaces can be used in conjunction with standard text-based interfaces.

Tagging-based systems rely on users for categorizing objects by means of tags (freely chosen words). Tags are aggregated from many users, primarily using web based interfaces, forming "folksonomies" which, although not as accurate as well-

designed ontologies, have the advantage of reflecting how users perceive the data and how their vocabulary and perception evolve over time. Tagging is simple and does not require a lot of thinking. Tags form an essential part of personalized internet radio and music community websites such as Last.fm [2]. Tag clouds are the most common way of visualizing tags. They are two-dimensional stylized visual representations of a list of words where the more prominent words are typically assigned a larger font. They are useful for quickly giving users the gist of a set of words. Tag clouds are in common usage on a number of different social networks, including Flickr [3] , del.icio.us [4] and wordle [5], but trace their origins back at least 90 years to Soviet Constructivist art [VW08]. The first true example may be that of a psychological experiment where participants were asked to create a collective mental map of landmarks in Paris [MJ76]. Later, tag clouds were featured prominently in the book *Microserfs*[Cou57], and first were introduced onto the web by Jim Flanagan in the program "Search Referral Zeitgeist". However, it was the use of tag clouds on the popular photo sharing site Flickr that made their use ubiquitous on Web 2.0 sites [Bru96]. Today, many social websites use tag clouds as a way to make large quantities of data more accessible and as a friendly interface for users.

Interacting with large music collections like most information retrieval tasks involves both querying (or direct search) in which the user has a well-defined search goal in mind as well as browsing (or indirect search) in which the goal is to explore, with some degree of serendipity, an information space. Summarization is the ability to extract the gist of a collection without going into details. Interfaces based on long sortable lists of text are effective for querying but provide little support for browsing and especially for finding music by artists that are not known to the user. In contrast, content-aware visualization-based interfaces can be quite effective for browsing, and music discovery but have weak support for direct searching. Tag clouds provide both an overview of the information space as well as direct search support. In order to satisfy all these possibly conflicting user information needs, a straightforward solution would be to provide all these three different ways of interacting with a music collection as separate views/interface components that are coupled. The disadvantage of such a design is that the user interface becomes unnecessary complicated and confusing.

In this chapter, we introduce content-aware self-organizing tag clouds, a technique

[2] `http://www.last.fm`
[3] `http://www.flickr.com`
[4] `http://dilicius.com`
[5] `http://www.wordle.net`

that attempts to support querying, browsing, and summarization using the familiar information model of a tag cloud. Tag clouds are commonly ordered either alphabetically or randomly. In some cases, tag clouds are ordered based on clustering using some kind of tag similarity metric such as tag co-occurrence. In other applications, like *wordle* ([6]), users position each word in a tag cloud by hand. In this chapter, I examine the utility and engagedness of placing similar (based on content not co-occurrence) tags next to each other using the Self-Organizing Map (SOM) [Koh95] algorithm. Specifically, we use techniques from Music Information Retrieval (MIR) to extract high-dimensional feature vectors characterizing each song, and then use the SOM algorithm to map these high dimensional feature vectors onto coordinates on a discrete 2D grid. The tags are then placed by using the centroid of the 2D grid coordinates of each set of songs associated with a particular tag. A final post-processing step using force-directed placement is utilized for better visual appearance and overlap removal.

A proof-of-concept implementation in the music collection browsing domain is described. Most existing music browsing interfaces proposed in the literature are prototype systems that have not been evaluated with user studies. This can be partly attributed to their publication in other fields in which user evaluation is not as important. However, it is also caused by the challenge of evaluating such interfaces due to the highly subjective nature of music similarity. We tried to address some of these challenges in the design of our user study and we hope that the insights gained will be of value for future research. Evaluation of the proposed interface shows that self-organizing tag clouds can result in more effective browsing especially for the case of music by artists that are unfamiliar to the user. This is supported by reporting both quantitative results as well as discussing qualitative reactions to the interface. To close, we discuss lessons learned and directions for future work.

[6] `http://www.wordle.net`

4.2 Background Information for Novel Interfaces for Music Exploration

4.2.1 Music Collection Browsing Interfaces

Currently the most common interfaces for browsing music collections such as iTunes by Apple are based on long sortable lists of text. Although effective for direct searching they provide limited support for music discovery and exploration. In the field of Music Information Retrieval, data of high dimensionality and of considerable complexity is generated. Various visualization interfaces have been proposed to make this data accessible and useful to users. Frequently these interfaces rely on automatically extracted audio features.

Islands of Music [PDW03] is an example of such a visualization of audio information which uses Self-Organizing Maps to generate a two-dimensional representation of a collection of music. *MusiCream* [GG05] is an interface that allows users to interact with a music collection using a dynamic visualization interface. *MusicRainbow* [PG06] is a similar system that uses web-based labelling and audio similarity to visualize music collections. Another relevant system is *MusicSun* [PG07] which combines three different similarity measures to generate music recommendations for users. The *Databionic/MusicMiner* system [MUNS05] allows users to organize large collections of music and employs Emergent Self-Organizing Maps to generate visualizations of the data involved. A very large web based system for helping users find new music is part of the Last.fm website [7] which provides advanced functionality for music recommendation and visualization based on a self-organized map calculated solely based on tag data. A simplified 2D grid representation with no text support based on audio content analysis has been proposed for Assistive music browsing [TBN+09].

Conceptually, the closest work to our approach is Salonen [Sal07] in a work where tag clouds are constructed using self-organizing maps. The two main differences from our work are 1) tag instead of content information is used for training the SOM 2) the lack of user evaluation. A 2006 review of visualization in audio based music information retrieval can be found in Cooper [CFPT06]. Examples of visualizations for music discovery in commercial and research systems can be found in the Visualizing Music blog [8].

[7] `http://playground.last.fm/iom`
[8] `http://visualizingmusic.com/`

Two new commerical developments in music recommendation that have recently produced interest are Apple's iTunes Genuis [9] and Pandora [10]. iTunes Genius works off of lists of song purchase histories of their customers. It has the premise that if many people who purchased song A also purchased song B, then song B is similar to song A. This simple system works well in many cases, for example, if someone is a huge Rush fan, then they will have bought a lot of Rush albums, and probably a lot of Yes and Led Zeppelin. Pandora on the other hand has a very different strategy, in which they employ a large staff of trained musicologists who, for each song in their collection, exhaustively rate this song on almost 400 different attributes. Similar songs are then found by comparing this multidimensional feature vector.

4.2.2 Tagging and Tag Clouds

Tagging systems allow users to add keywords, or tags, to resources without relying on a controlled vocabulary [JS06] and have become ubiquitous in web-based systems. It has been observed that they have the potential to enhance many types of online interaction and because of their free-form nature, they take advantage of the underlying and pre-existing social organization of web communities [MNBD06]. While controlled ontologies and taxonomies hold the promise of providing a regular and well-defined structure for organizing knowledge, in practice this taxonomic rigidity becomes too heavyweight and can stifle input and collaboration from user communities. Tag clouds are one of the most common methods of visualizing tag information.

There has been considerable research in recent years into the design, use and effectiveness of tag clouds. The *Dogear* system [MFK06], uses tags to organize social bookmarks for large enterprise organizations. A system that uses tags in an eCommerce application is described in Ganesan et al. [GSD08], and has the feature that it automatically mines tags from feedback and presents the results in a visually appealing manner. In Halvey and Keane [HK07], a variety of tag presentation techniques are evaluated, and show that the use of different techniques can affect the ease with which users can find tags. A historical look at tag clouds is presented in Viegas and Wattenburg [VW08], which looks at the development of tag clouds since their inception a decade ago, and speculates about their development in the future. A novel way of determining the size of tags in a tag cloud by examining the informational entropy

[9] `http://apple.com/genius`
[10] `http://pandora.com`

of the tag, which is then related to the emotional impact of the tag, is presented in Eda et al. [EUUY09]. Another study examined tag clouds derived from Automatic Speech Recognition (ASR) as surrogates for tag clouds generated by human listeners and found that the ASR determined tag clouds perform equally well [TLdR08]. In the paper "Seeing things in clouds" [BGN08], an extensive evaluation of different types of visual features in tag clouds, including font size, font weight, intensity, number of characters and area were investigated, and while font size and font weight had the largest impact, when multiple variables were changed at once, no one property stood out amongst the others. Tag navigation in general has been examined in detail with particular focus on "last.fm", an online social community for music [MC09]. The Qtag [LSH07] system investigates collaborative tagging in the domains of Information Filtering and Information Retrieval and presents a tag visualization model [LSH07]. A context aware browser for mobile devices that uses tag clouds is presented in Mizzaro et al. [MNV09].

Music databases are typically very large, containing thousands to millions of songs, and the number of users interested in listening to and browsing music is similarly large. The effectiveness of browsing large scale social annotations has been examined using the Effective Large Scale Annotation Browser (ELSABer) algorithm [LBY+07]. Another paper describes the Topiography system, a visualization for large scale tag cloud [FFM+08].

More artistic applications of tagging and tag clouds have also been explored, one of these is *ArsMeteo* [ABB+09], a Web 2.0 portal that allows users to collect and share digital artworks, including videos, pictures and music. Tag-based retrieval of video content has been explored using a variety of tag sources including social tags, professional metadata and automatically generated metadata [MGvSV08]. The paper "Your place or mine?" [DVWK08] explores the Many Eyes website, a place that allows for collaborative visualization of datasets, and examines the themes that reoccur across various scenarios of the use of the data in the visualizations.

There are a large number of papers that focus on the application of various mathematical techniques to tagging in general, of which we have chosen only two as a very small representative sample. The *Folksoviz* [LKJK08] project uses a statistical method for deriving subsumption relationships based on the frequency of tags in Wikipedia texts and also uses the Tag Sense Disambiguation (TSD) method for mapping each tag to a Wikipedia article.

An exploration of information seeking in the socio-semantic web shows how items

can be viewed semiotically depending on tags, topics and points of view [CZZ07].

4.2.3 Motivation and Design Goals

Music browsing and discovery in large digital collections is a particularly interesting domain with unique challenges and opportunities for interaction design. It is an activity that many computer users engage daily. As the primary goal is entertainment in many cases the user can be satisfied with little effort. For example most portable music players feature a shuffle button that just plays random songs from a collection. It is highly unlikely that the user of a text search engine would be in any way satisfied with such random retrieval. At the same time, the notion of music similarity is highly personal and subjective compared to relevance in other fields of information retrieval. Music from unfamiliar styles or cultures is typically perceived as sounding all the same by listeners and the same pair of tracks might be considered similar by one listener and dissimilar by another. As in most information retrieval tasks there is a need for both querying (direct search) and browsing (indirect search). However, in browsing, listeners frequently have only a vague idea of what they want to hear so the ability to quickly and effectively explore a large information space and discover new music by unfamiliar artists is important.

The technique of content-aware self-organizing maps, proposed in this chapter, evolved over experimentation with the ideas and techniques presented in the previous subsections and informal feedback through participatory design activities. It can be viewed as a fusion of concepts from text-based visualization interfaces and more abstract content-aware visualization interfaces. In order to motivate the design goals we briefly mention some of the issues that users raised when experimenting with various different interfaces for browsing large music collections. A rough classification of existing systems along two dimensions will be used to illustrate these issues. An additional simplification used throughout this chapter is the use of the word tags to denote any textual data associated with a particular music track. For example the genre of a song, the year of release, or the artist can be viewed as tags (albeit with some constraints such as that a track can only be associated with a single artist). Existing systems can be characterized either as tag-based (using the more generalized tag definition) or content-based. In addition they can be characterized either as simple or complex. Simple interfaces have minimum requirements in terms of screen real-estate and can be navigated using simple mouse or even just keyboard

interaction. In contrast complex interfaces require large screen real-estate and require complex user interactions and gestures to control. Existing systems are combinations of these extremes.

Traditional complex tag-based systems based on long lists of sortable text such as iTunes provide very little support for browsing, discovery and summarization, even though progress using the previously described iTunes Genius feature is being made. An alternative is visualization interfaces that are based on automatic analysis of musical content. By mapping the music collection onto a visual 2D or 3D representation they enable quick browsing and navigation especially in the case of music that is not known to the user or that has not been tagged. Simple content-based interfaces typically only provide tag data once a particular track is selected [PDW03, TBN+09, MUNS05]. User quickly learn a mental map of the representation (such as the lower corner of the display contains mostly fast, energetic rap songs) but have trouble understanding the display (for example a frequent question might be what is the meaning of the x-axis or how are the tracks placed). Some of the proposed content-based interfaces can be visually complex (for example displaying hundreds of dots representing songs in a 3D space) and require complex interactions such as 3D rotation and zooming [GG05, PG06]. Even though such interfaces make great demos they are frustrating to use regularly. Adding text/tags on the visualization further increases complexity.

Tag-clouds provide a simple, familiar interface that partly overcomes these limitations. For example they support both direct searching as well as browsing and navigation. However they come with their own problems. In order for a tag to assist search or browsing it is necessary for the user to have some notion of its meaning. For example a specialized term such as indie pop might be completely unfamiliar to a particular listener while at the same time essential to another. This problem becomes even more acute using the more generalized notion of tags that includes information such as artist or album. As one of the goals for an effective interface of music collection browsing is the discovery of new music by artists not known to the listener this is an important disadvantage. Simple tag clouds also do not provide the user with any information about the connections and similarity relations between tags. More sophisticated approaches rely on analyzing and visualizing tag similarity calculated based on co-occurrence relations. A final problem with any system based solely on tag information is that there is no way to access music tracks that have not been tagged (the so-called cold start problem). In contrast content-based visualizations

allow any track to be accessed and do not require familiarity with the music explored.

Based on these considerations, we identified five distinct design goals for our music discovery interface:

- **Simplicity:** Both the visual display and the user interaction should be simple, straightforward and familiar to users. The design should not hinder implementation on small displays, touch surfaces or general accessibility by users with special needs. Both direct and indirect search should be supported by the same user actions.
- **Discovery:** The interface should support browsing, discovery and exploration of music not familiar to the user without this support affecting direct searching and retrieval. Tracks that have not been tagged should be integrated and accessible.
- **Consistency:** Frequently multiple views of a music collection are desired. For example a listener might want to see all the artists in a particular collection or playlist as well as all the associated tags. Using existing techniques the tag clouds generated for these two views (facets) would have no relation to each other. Although clustering approaches that rely on tag similarity based on co-occurance provide a more semantically meaningful layout they can not provide layout consistency among different facets.
- **Disambiguation:** Typically there is no imposed structure or consistency in tagging especially in a highly subjective fields such as music listening. Polysemy and synonyms are well known problems of tagging. For example some music tracks might be labeled with the tag "female voice" and some with the tag "woman singing" even though they essentially refer to the same type of musical content. It is likely users will use one or the other therefore tag similarity based on co-occurance will not provide any help. Furthermore frequently tags might be completely unfamiliar to a user. Such disambiguation problems can be addressed if tags are placed based on analyzing musical content. For example a listener unfamiliar with the tag "motet" should be able to use neighboring tags such as "classical" or "vocal music" to infer the meaning.

Using these goals as requirements we propose a hybrid of text-based and content-based approaches to music collection browsing that we term content-aware self-organizing tags. In the following sections we describe the various processing steps required to

create the visualization interface and present a proof-of-concept prototype implementation. There is a lack of empirical data offering insights into the design of music discovery interfaces. We present the results of a user study evaluating our proposed design and discuss some of the methodological issues we had to deal with.

4.3 System Architecture of *Audioscapes*

The goal of *Audioscapes* is to design a framework for exploring content and context-aware user interfaces for browsing large audio collections using controllers beyond the mouse and keyboard. The abstract system architecture distills the common functional blocks required to build such interfaces. A large number of existing audio and music browsing systems fit this architecture which is shown in Figure 4.1. The underlying metaphor is that each track in an audio collection is mapped to a discrete location on a rectangular grid. More than one track can be mapped to the same location. Different algorithms for clustering and dimensionality reduction can be used to map automatically extracted audio features to the grid coordinates. Controllers take input from the user and either interact with the audio collection (for example by initiating playback or by applying digital audio effects such as pitch-shifting and time-stretching) or move around on the mapped grid representing the audio collection. Views are different ways/devices used to display the grid map. The communication between processing blocks is mainly accomplished through Open Sound Control (OSC) [WFM03] messages or alternatively custom XML or text files.

Open Sound Control is a new content format for sending musical data between computers. XML is popular tree-tree based format for encoding data.

The modular nature of the system provides flexibility and extensibility which are crucial in this exploratory domain. In the following subsections the currently available components of the framework are described.

4.3.1 Audio Processing

There are two main aspects of audio processing: digital audio effects and audio feature extraction. For digital audio effects we currently support pitch-shifting and time-stretching using a state-of-the-art phase vocoder algorithm [LD99] as well as tunable filters.

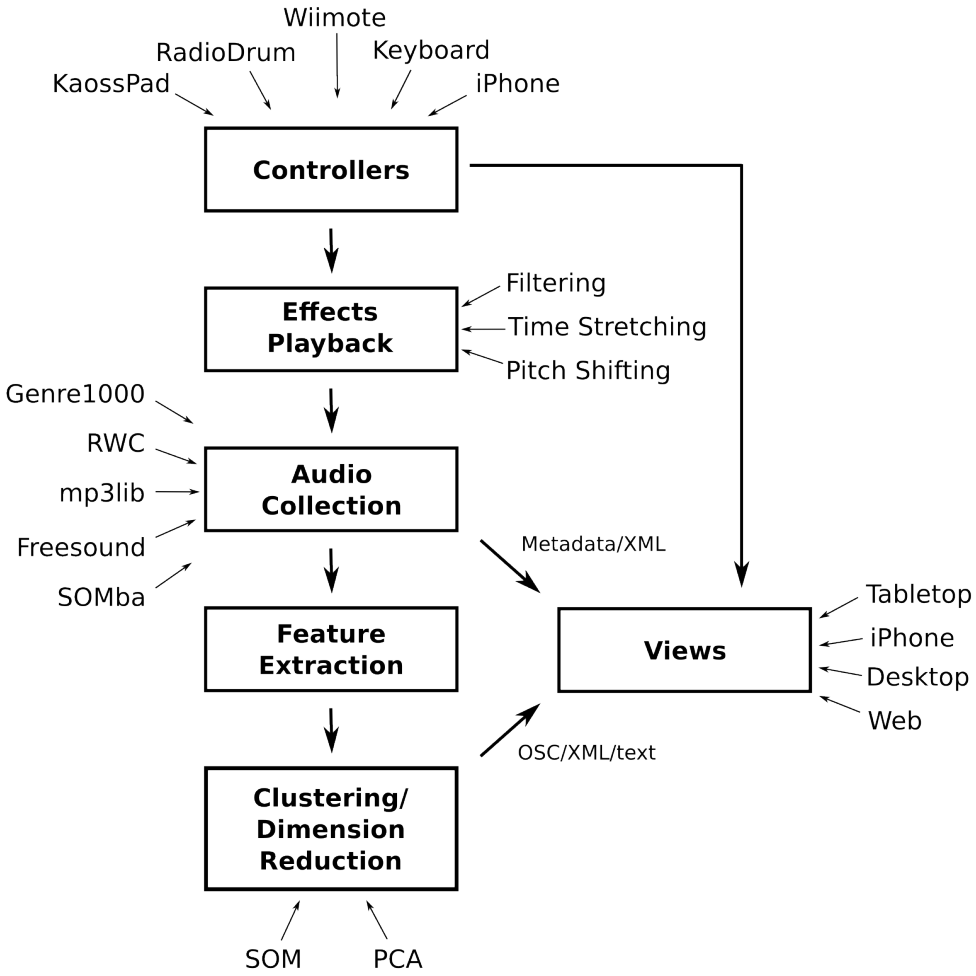

Figure 4.1: *System Architecture*

A phase vocoder is a type of digital audio filter which can scale audio in both time and frequency.

These would be useful to have available to artists to enhance their creativity with and control over the system. The goal of audio feature extraction is to represent each song in a music collection as a single vector of features that characterize musical content. Using suitable features, songs that "sound" similar should have vectors that are "close" in the high dimensional feature space. The features used are the Spectral Centroid, Rolloff, Flux and the Mel-Frequency Cepstral Coefficients (MFCC) as well as time-domain zero crossings [Tza08].

The Spectral Centroid is a measure of the "center of mass" of a spectrum, and shows where the most common frequencies are distributed in a spectrum. Rolloff is a measure of the steepness of falloff in an audio spectrum. Flux is the norm of the difference vector between two successive magnitue/power spectra. MFCC coefficients are a way to transform a standard spectrum into one that more closely approximates how the human ear perceives sound.

To capture the feature dynamics we compute a running mean and standard deviation over the past M frames:

$$m\Phi(t) = mean[\Phi(t - M + 1), .., \Phi(t)] \tag{4.1}$$

$$s\Phi(t) = std[\Phi(t - M + 1), .., \Phi(t)] \tag{4.2}$$

where $\Phi(t)$ is the original feature vector. Notice that the dynamics features are computed at the same rate as the original feature vector but depend on the past M frames (40 in our case corresponding to approximately a so called "texture window" of 1 second). This results in a feature vector of 32 dimensions at the same rate as the original 16-dimensional feature vector. The sequence of feature vectors is collapsed into a single feature vector representing the entire audio clip by taking again the mean and standard deviation across the 30 seconds (of the sequence of dynamics features) resulting in the final 64-dimensional feature vector per audio clip.

A more detailed description of the features and their motivation can be found in Tzanetakis and Cook [TC02].

For the calculation of the self-organizing map described in the next section all features are normalized so that the minimum of each feature across the music collection is 0 and the maximum value is 1. This feature set has shown state-of-the-art performance in audio retrieval and classification tasks in the Music Information Retrieval Evaluation Exchange (MIREX) 2008 [11].

The exact details of the feature extraction for the Freesound collection and the samba percussive instruments are slightly different from the process described above (which is designed for music).

[11] `http://www.music-ir.org/mirex/`

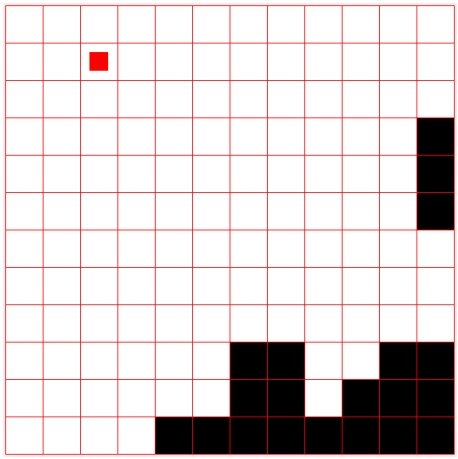

Figure 4.2: Topological mapping of musical content by the Self-Organizing Map for Classical music

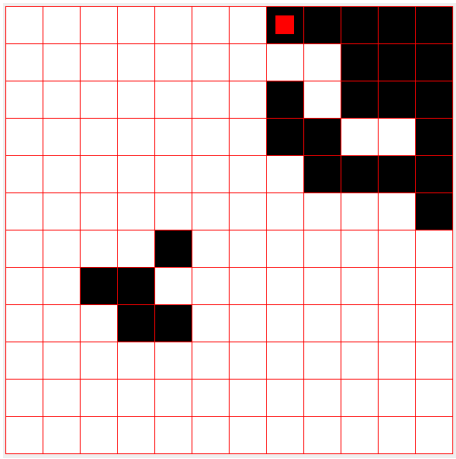

Figure 4.3: Topological mapping of musical content by the Self-Organizing Map for Metal music

4.3.2 Visualization

The primary method used for visualization is the self organizing map (SOM) which is a type of neural network used to map a high dimensional input feature space to a lower dimensional representation while preserving the topology of the high dimensional feature space. This facilitates both similarity quantization and visualization simultaneously. The SOM was first documented in 1982 by T. Kohonen, and since then, it has been applied to a wide variety of diverse clustering tasks [Koh95]. In our system the SOM is used to map the audio features (64-dimensions, which is a standard number of features often used in Marsyas to do music genre classification)

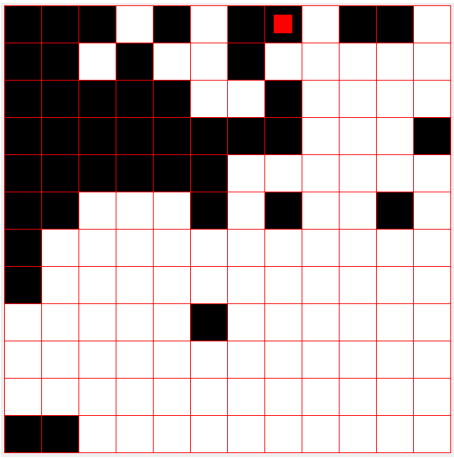

Figure 4.4: Topological mapping of musical content by the Self-Organizing Map for Hiphop music

to two discrete coordinates on a rectangular grid. The traditional SOM consists of a 2D grid of neural nodes each containing an n-dimensional vector, $\mathbf{x}(\mathbf{t})$ of data. The goal of learning in the SOM is to cause different neighbouring parts of the network to respond similarly to certain input patterns. This is partly motivated by how visual, auditory and other sensory information is handled in separate parts of the cerebral cortex in the human brain. The network must be fed a large number of example vectors that represent, as closely as possible, the kinds of vectors expected during mapping. The examples are usually applied several times. The data associated with each node is initialized to small random values before training. During training, a series of n-dimensional vectors of sample data are added to the map. The "winning" node of the map, known as the *best matching unit* (BMU), is found by computing the distance between the added training vector and each of the nodes in the SOM. This distance is calculated according to some pre-defined distance metric which in our case is the standard Euclidean distance on the normalized feature vectors.

Once the winning node has been defined, it and its surrounding nodes reorganize their vector data to more closely resemble the added training sample. The training utilizes competitive learning. The weights of the BMU and neurons close to it in the SOM lattice are adjusted towards the input vector. The magnitude of the change decreases with time and with distance from the BMU.

The update formula for a neuron with representative vector N(t) can be written as follows:

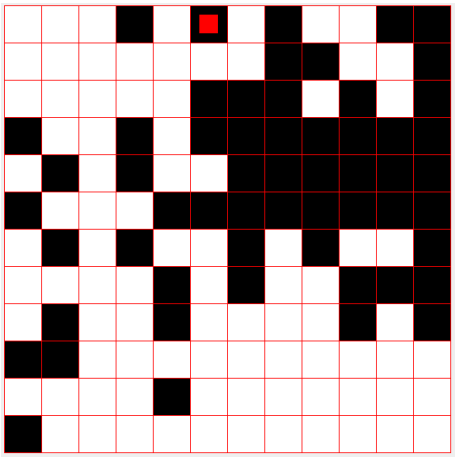

Figure 4.5: Topological mapping of musical content by the Self-Organizing Map for Rock music

$$\mathbf{N}(t+1) = \mathbf{N}(t) + \Theta(v,t)\alpha(t)(\mathbf{x}(t) - \mathbf{N}(t)) \tag{4.3}$$

where $\alpha(t)$ is a monotonically decreasing learning coefficient and $\mathbf{x}(t)$ is the input vector. The neighborhood function $\Theta(v,t)$ depends on the lattice distance between the BMU and neuron v. We utilize a Gaussian neighborhood function that shrinks over time. The time-varying learning rate and neighborhood function allow the SOM to gradually converge and form clusters at different granularities. In our implementation, $\alpha(t)$ is a linearly-decaying function with t. Once a SOM has been trained, data may be added to the map simply by locating the node whose data is most similar to that of the presented sample, ie. the winner. The reorganization phase is omitted when the SOM is not in the training mode. Another interesting property of SOMs for our application is that they can be personalized by user initialization rather than random initialization.

We also have explored other possibilities for mapping the high-dimensional space to a 2D dimensional grid. Principal Component Analysis (PCA) can be used to map the input feature space to the two "highest" (corresponding to the largest eigenvalues) principal components followed by quantization to the grid. The main advantage of the SOM is that there is no need for quantization and there is better coverage of the surface area as it mainly preserves the topology of the input space rather than the distance characteristics.

Figures 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 illustrates the ability of the extracted musical content-features and the SOM to represent musical content. Figures 4.2, 4.3,

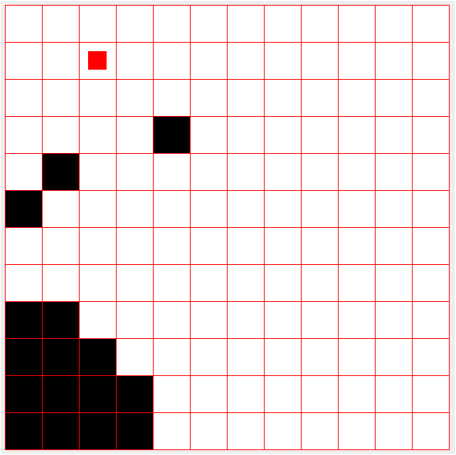

Figure 4.6: Topological mapping of musical content by the Self-Organizing Map for a selection of music by Bob Marley

4.4 and 4.5 show how different musical genres are mapped to different regions of the SOM grid (the black squares are the ones containing one or more songs from each specific genre). As can be seen Classical, Heavy Metal and HipHop are well-localized and distinct whereas Rock is more spread out reflecting its wide diversity. The SOM is trained on a collection of 1000 songs spanning 10 genres.

The figures 4.6 4.7 4.8 4.9 show how different artists are mapped to different regions of the SOM grid. The SOM in this case is trained on a diverse personal collection of 3000 songs spanning many artists and genres. It is important to note that in all these cases the only information used is the automatically analyzed actual audio signal and the locations of the genres are emergent properties of the SOM.

4.3.3 View and Control Interfaces

The common functionality among view interfaces is to display the automatically calculated grid, respond to navigation events and handle audio playback and effects. Typically the grid squares are colored darker or lighter based on the number of tracks that they contain. The most powerful view is a desktop graphical user interface written in Qt [12]. In addition to standard view functionality it provides the ability to write iTunes music library XML files, advanced coloring modes based on metadata, and continuous playback mode in which tracks change automatically when the cursor moves to a different grid square without requiring explicit clicking by the user. In ad-

[12] `http://www.qtsoftware.com/products`

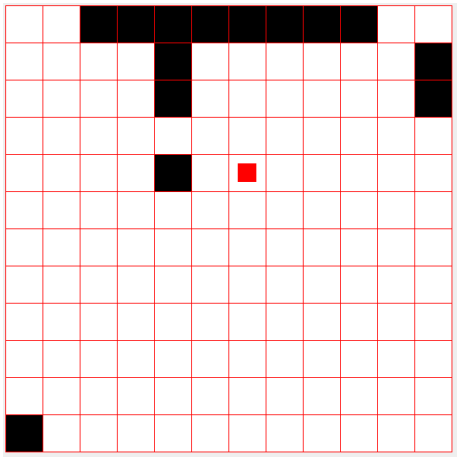

Figure 4.7: Topological mapping of musical content by the Self-Organizing Map for a selection of music by Radiohead

dition we also provide a web-interface that although more limited has the advantage that anyone on the internet can access and interact with the particular *AudioScape* deployed. As the audio is streamed, the audio collection remains on the server, this can be an important factor in commercial applications. It also provides an accessible platform to conduct demonstrations and user studies that can be updated frequently without requiring any user action.

In order to explore non-standard form factors we have developed a implementation specific to the iPhone [13]. Having a touch-based display surface facilitates spatial awareness especially for blind or limited vision users. As the user moves his/her finger across the various squares, songs from each corresponding node cross-fade with each other to help her navigate the music collection by hearing how the songs in each grid location are changing. By laying out a music collection in this spatial fashion, navigation with only the knowledge of a few reference points is needed. For example, if it is known that Rock music is in the upper left corner, and jazz music is in the lower left corner, by dragging from top to bottom along the left edge of the grid, rock music will slowly transition into jazz music. The use of multi-touch (two or more fingers) may also be used to control playback. Swiping the surface with two fingers in the right direction skips to another song in the same node while swiping left plays the previous song in that node. Figure 4.10 shows the iPhone view. Finally we have also explored tabletop views based on the desktop and web implementations on a Smart Table system and well as a Mitsubishi Diamond Touch. However we did not

[13] `http://www.apple.com/iphone`

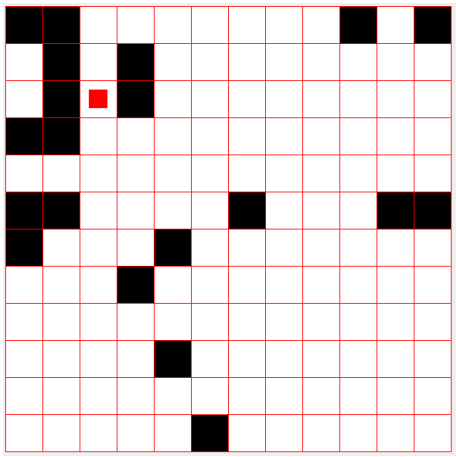

Figure 4.8: Topological mapping of musical content by the Self-Organizing Map for a selection of music by Led Zeppelin

take advantage of their multi-touch capabilities as it would require writing specialized versions of the software.

4.3.4 Control interfaces

In addition to the traditional mouse/keyboard based control, we have explored various alternative control interfaces. The radiodrum [SD01] is a novel three dimensional controller that uses capacitive sensing to detect the positions of two radio frequency oscillators, usually attached to drum sticks or other similar objects. Developed by Max Matthews and Bob Boie, it was originally designed as the first three dimensional computer mouse, but found a more pertinent application in the area of the production of computer music [NDS03]. It has also been used as a controller to facilitate the browsing of music collections [MT06]. In our prototype, we have mapped the x, y and z axes of the sticks of the radiodrum to our the user interface. Movement of the sticks in the x and y axes moves the audio track selector cursor on the GUI, and movement in z controls the volume of that track. Each stick controls a different music track. We use both of the sticks of the radiodrum, each of which can independently select a different musical track. With the addition of volume control, the interface transforms from a simple music browsing interface to a more DJ-like experience. with the performer able to control the sound mix between two different audio tracks by moving the sticks up and down on the z axis.

Some preliminary tests of this interface show that it is a potent and exciting way

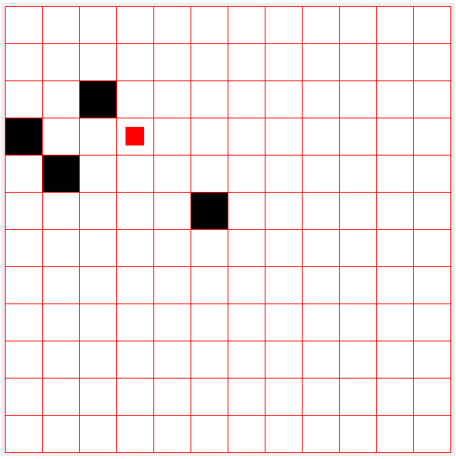

Figure 4.9: Topological mapping of musical content by the Self-Organizing Map for a selection of music by Dexter Gordon

to create mashups and mixes of a variety of different music styles, with styles like reggae, rock and hiphop.

The Wii remote, or wiimote, is a multimodal interface device developed by Nintendo for use with the Wii game system. The wiimote has traditional buttons and rumble functionality, but also contains a speaker, an 3-dimensional accelerometer, and an Infrared (IR) sensor. This IR sensor has the ability to track up to 4 independent sources of infrared light, and reports back the positions and intensities of the detected points. All data from the wiimote is sent back to the computer via Bluetooth.

In order to enhance the configurability and expandability of this project, we use the OSC protocol [WFM03] to transmit messages received from the wiimote controller. Using OSC, we have added functionality that allows for the use of multiple wiimotes at once, each one returning the positions of four different people in a space. As an example we have developed SOMba, a system for collaborative creation of Samba rhythms by dancers in a space. The rhythms and instrument sounds are arranged in a grid using a Self-Organizing Map. In the canonical SOMba system, we use two wiimotes, allowing up to 8 dancers to be tracked, but this number could easily be expanded. It is important to align the sensors of the wiimotes accurately with the performance space and with each other, so that an individual dancer is only tracked by one wiimote at once. In our current system we use the wiimote for position tracking of the dancers, but the use of OSC allows us to use a variety of other position trackers quite easily.

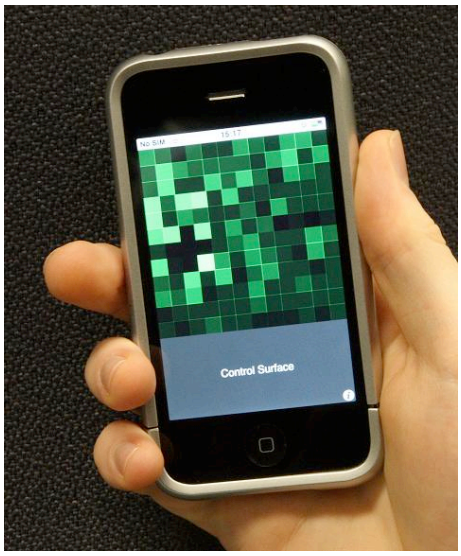

Figure 4.10: iPhone control interface

4.3.5 Data Collections and Implementation

In order to explore different configurations we have created *AudioScapes* for several large audio data collections which are known to the Music Information Retrieval (MIR) and Computer Music community. The Freesound Project [14] is a huge collection of sound effects, music and environmental songs all licenced under the Creative Commons licence.

The website makes extensive use of folksonomy based methods of audio classification primarily through a process of collaborative tagging of audio files as well as through geotagging, sample packs, and a remix tree view. The audio in this database is an wide variety of formats, including .wav, .mp3, .aiff and .au, amongst others, and is recorded at many different sample rates and bit depths. For our current research, we selected a well-behaved subset of the audio in the Freesound database, and converted all the audio to a common format, sample rate, and bit depth.

The RWC (Real World Computing) Music Database [GN03] is a database of music that has been copyright cleared and made available to the Music Information Retrieval community.

It contains large amounts of high quality audio samples and musical pieces. There are large numbers of short samples of audio from different musical instruments around the world, including both tonal and percussive instruments.

[14] `http://freesound.org`

The music in the RWC database is from a wide variety of genres, with many classical and jazz pieces, as well as a sampling of the genres of popular, rock, dance, jazz, latin, classical, marches, world music, vocals, and traditional Japanese music. A collection of 1000 music tracks from 10 genres was gathered and described in [TC02].

The music in this collection includes such diverse genres as blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock. It has been used extensively by the Music Information Retrieval community. We have also used a large database of 3000 30-second snippets from the personal collection of one of the authors. Another collection consists of samples of percussion instruments of Samba music.

The Marsyas [15] audio processing software framework has been used for the audio feature extraction, digital audio effects, calculation of the SOM, the desktop graphical user interface and handling of controller data. The calculation of the SOM in Marsyas is carried out after the feature extraction in a data-flow architecture. This architecture allows for easy integration and rapid execution of all the required processing steps, and is readily extensible to support a diverse set of different experimental parameters. The Qt interface to Marsyas is used for display of the generated data and user interaction.

For the web based interface we employ an *XHTML/CSS* and *Flash* based interface. The user is presented with a simple *XHTML/CSS* web page that has been designed to be standards compliant which will facilitate accessibility by the research community on a wide variety of different web browsers and computer platforms. The *Flash* based interface is written in the haXe [MSP08] programming language, which compiles the ECMAScript language *haXe* down to *Flash* bytecodes. The *Flash* interface presents a simple interface to the user with an interface similar to the Qt based MarGrid interface.

The Open Sound Control (OSC) protocol [WFM03] is used in this project to facilitate communication between the various components of the system. The main MarGrid Qt interface presents an OSC listener which all other controllers send messages to. The wiimote Bluetooth messages are translated to OSC to update the grid position. In a similar manner, the MIDI messages from the RadioDrum are translated into OSC messages. This system architecture lets us easily add new controllers to our system. Because OSC is able to send its messages over a network, we are able to run the controller programs on different machines than the machine producing the audio, thus facilitating experimentation and collaboration between multiple users.

[15] `http://marsyas.sourceforge.net`

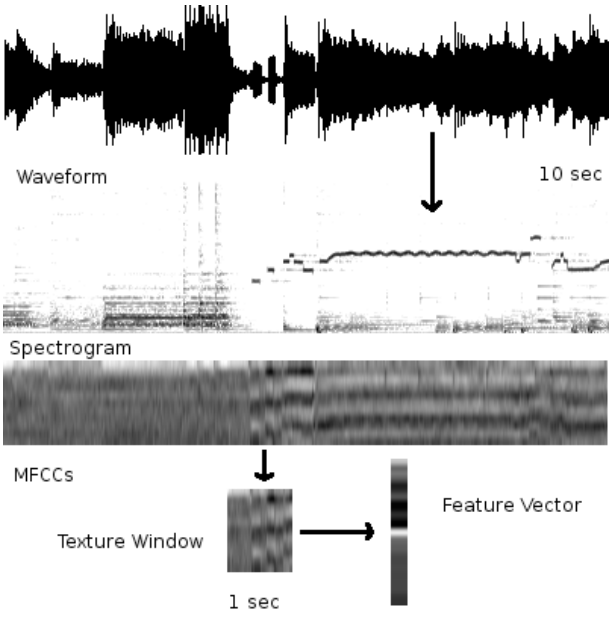

Figure 4.11: *Audio Feature Extraction*

4.4 Self-Organizing Tag Clouds as a Novel Music Exploration Tool

We describe a new method for organizing music tag clouds that makes a persistent map that takes into account the musical similarity between songs. Figure 1 shows the various stages of the process of creating a content-aware self-organizing tag cloud. The first step (Figure 4.11) uses techniques from the field of Music Information Retrieval (MIR) to calculate a high-dimensional feature vector representation for each track in the music collection. Once all the feature vectors are calculated each track is mapped onto a discrete position on a 2D grid using a Self-Organizing Map (Figure 4.12). Each generalized tag is associated with a set of tracks that have been annotated with it. As the tracks have been mapped to feature vectors and subsequently to 2D grid coordinates each tag can be associated with a set of 2D grid coordinates. The self-organized map process ensures that neighboring points (tracks) will have similar high-dimensional audio features and therefore similar musical content. In the third step the tags are placed on the centroids of their corresponding set of 2D grid coordinates. Their placement will reflect the underlying musical content but results in visual overlap between them (Figure 4.13). The final step (Figure 4.14) is applying a force-based layout drawing algorithm to reduce overlap and result in a

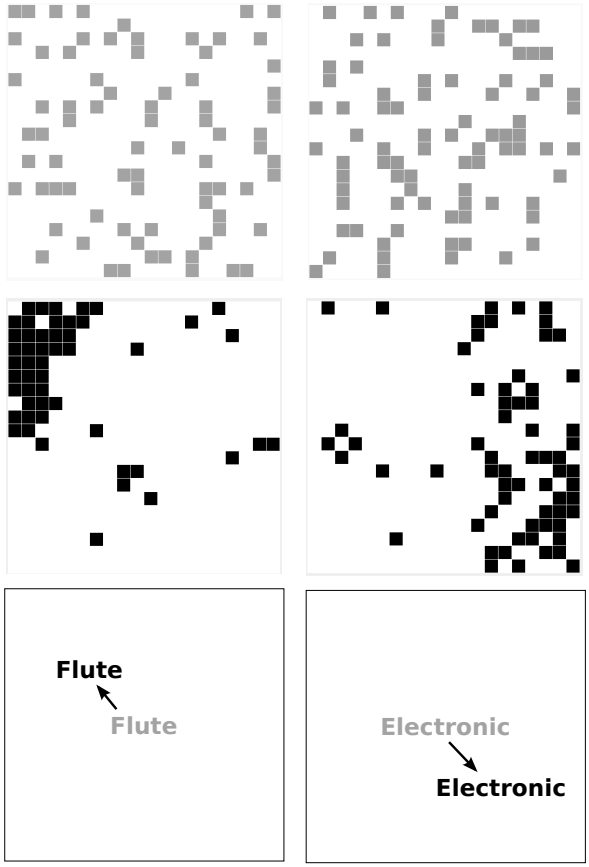

Figure 4.12: *Self Organizing Map*

more aesthetically pleasing tag cloud.

4.4.1 Self-Organizing Maps for Layout of Tag Clouds

For creating the visualization layout we utilized the self-organizing map (SOM) which is a type of neural network used to map a high dimensional input feature space to a lower dimensional representation while preserving the topology of the high dimensional feature space. This facilitates both similarity quantization and visualization simultaneously. The theory behind the SOM was described in detail in an earlier chapter.

Once the self-organized map of tracks is created, each track is mapped to a set of 2D coordinates (x, y). The centroid of this set of 2D coordinates is used as the position of the corresponding tag. To generate the content-aware self-organized tag cloud we iterate over each of the songs in the collection and place it using the centroid. Figure

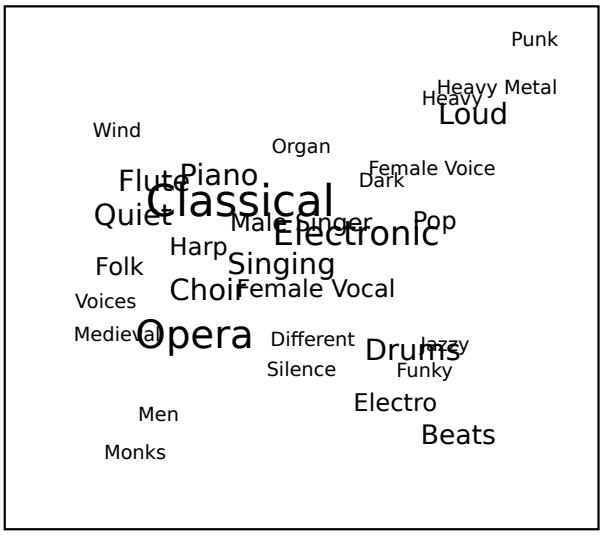

Figure 4.13: *Self-Organizing Tag Cloud before Mass-Spring-Damper*

4.12 shows how the self-organizing map can move tags using an underlying content space. The top of the figure shows a random distribution of songs for the tags "Flute" and "Electronic". As can be seen after the SOM processes the songs corresponding to each tag are clustered although there is some overlap (for example an electronic piece might contain flute) and the corresponding tag positions are moved.

This initial layout contains many overlapping words, so the position of each tag is repositioned using a mass, spring and damper force-based algorithm for drawing [EGK+01]. In our implementation each tag is anchored to its original position using a spring and an electrostatic-like force is applied between every pair of tags that is proportional to the inverse of their squared distance. Therefore tags that are close and overlapping will be pushed away while still trying to remain close to their original location. An additional wall force term was added to keep all tags within the designated window. Figures 4.13,4.14 shows the self-organizing tag cloud before and after applying the Mass-Spring damper algorithm.

In order to present this data to users, we designed an interactive web-based interface which utilizes a multiple facet view of synchronized tags clouds [HFS+08]. Three views/panes that are coupled are provided: 1) a track view 2) an artist view and 3) a tag view. The interface is shown in Figure 4.15. On the right is the tag pane, and in this figure the tags are organized according to the Self-Organizing Map. In the center is the artist pane, which shows the names of the artists that created the tracks, and the songs are shown in the left hand pane. Because there are many more tags than

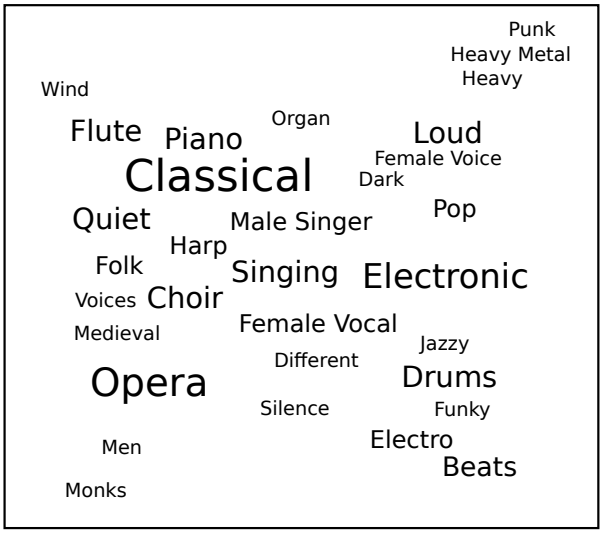

Figure 4.14: *Self-Organizing Tag Cloud After Mass-Spring-Damper*

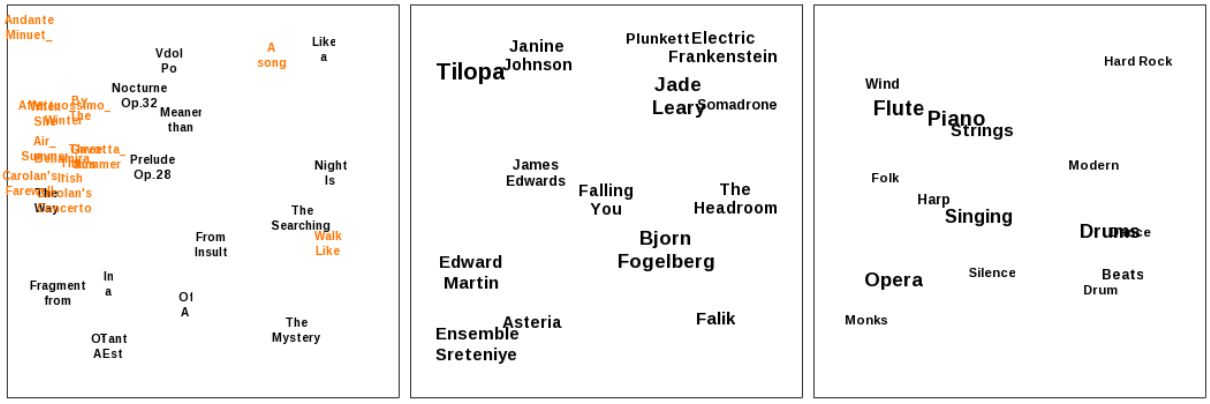

Figure 4.15: Play Tag Now! Interface

can be shown in either the tag pane or the song pane, only a small subset of the tags are displayed at any one time, and above each pane is a "Shake" button which selects a different random subset of tags to show the user. The display area is partitioned into a 5x5 grid and tags in each subgrid are rotated during each "Shake". In Figure 4.15, the user has clicked on the "Flute" tag in the tag window, which then displays all the songs that have the tag "Flute" associated with them in orange. To give the user a feeling of the overall song organization, a subset of tracks is shown in black in the track pane.

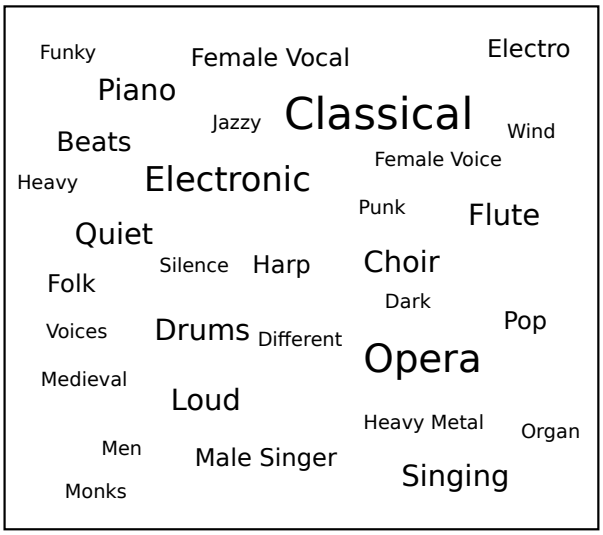

Figure 4.16: Random Tag Cloud

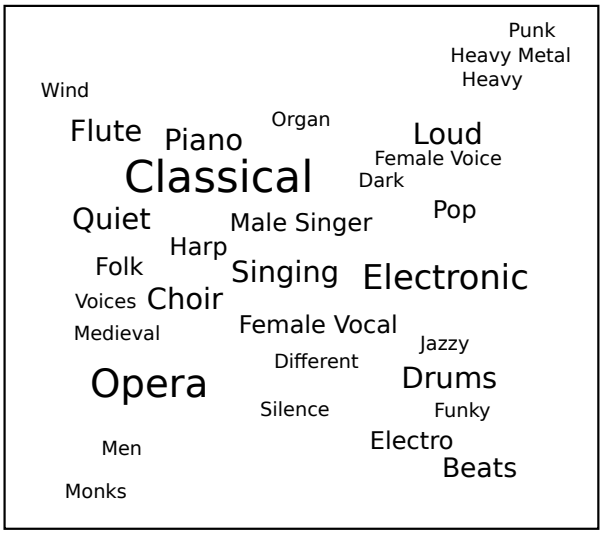

Figure 4.17: SOM Tag Cloud1

Figure 4.18: Alphabetical Tag Cloud

4.5 Evaluation of Self-Organizing Tag Clouds

The goal of the evaluation was to compare the effectiveness of three different ways of tag clouds visualization: random placement, alphabetical layout, and the self-organized tag clouds based on music content similarity proposed in this chapter. Figures 4.16, 4.18 and 4.17 show these three configurations.

The evaluation of music browsing and discovery interfaces is challenging. Frequently the effectiveness of the interface is evaluated indirectly through specific tasks which tend to be more related to directed search. Examples of such tasks include finding a song of a particular genre or artist or using the interface to create a playlist. One methodological problem with such tasks is that the stopping condition from the task is dictated by the experiment designer and does not take into account individual patterns of usage. Although the artificial nature of such tasks for evaluation is to some extent unavoidable, they can be designed to be more open and flexible. Throughout our participatory design process and pilot study we observed that users tend to have very different behaviors when interacting with large collections. Some users are easily satisfied with a quick match that approximately corresponds to what they are looking for while others spend considerable time refining and narrowing their search in order to find a much more tighter match.

In order to accommodate this variety in usage patterns we decided to let the users specify the stopping criterion rather than the experimenter. For all tasks the users were asked to indicate when they locate a music track that was similar to the provided query according to them. In contrast other user studies frequently ask the user to keep looking for tracks until the "correct" answer is found and measure the time to complete such task. In our opinion our approach provides a more valid assessment of browsing effectiveness across different usage patterns but has the unfortunate side effect of largely varying task completion times among users. For this experiment, we studied a total of 14 users selected from Graduate students and Professors in the Department of Computer Science at the University of Victoria. Therefore, in order to compare the three configurations across different users we normalized the task completion times. For each task the "slowest" configuration was used to normalize the task completion times for a particular user. As an example a user with task completion times $(A : 60, B : 14, C : 30)$ in seconds would have normalized task completion times of $(A : 1, B : 0.23, C : 05)$ which can be interpreted that the time to complete the task using configuration B was 0.23 faster than using configuration

A (or 14 seconds for configuration B compared to 60 seconds for configuration A).

When examining the results of the user study, we found that the average time to complete the tasks varied widely between different users. The time to complete all the tasks averaged across all the participants was 436 seconds with a standard deviation of 224.35 seconds. The shortest time to complete all the tasks was 202 seconds, and the longest time was 1162 seconds. The times did not follow a normal distribution curve.

Another issue we had to wrestle with was whether the configuration used was known to the users. The problem was that unless the users were aware of the underlying content-based mapping they would not expect similar tags to be located near each other. Therefore we decided that the users would see which configuration was used each time. Given that the differences between each configuration are obvious anyway we don't think that this choice affected our results. We collected both quantitative data such as task completion times as well as qualitative data. The following sections describe the experimental setup and results.

4.5.1 Experimental Setup

Fourteen participants were recruited from graduate Computer Science students. Three were female and 11 were male. All subjects had normal or corrected-to-normal vision, enjoyed listening to music and were experienced computer users. None of the participants had previous knowledge of the Magnatunes [16] dataset.

The Magnatagatune dataset is a new dataset for MIR applications that contains a large collection of songs that has associated tags generated by users. It contains songs from the Magnatunes record label that aims to treat both artists and customers fairly by releasing music under permissive licences. Magnatunes made available a large number of songs to the scientific community for use in research. The *Tagatune game* [LvA09] is a new game-with-a-purpose [vAD08] in which two users are both presented with a song. Both users are then asked to guess what tag the other user would select for this song, and if both users agree, the tag is added to the song. This tag is then not allowed to be used for this song by subsequent pairs of players. The Magnatagatune dataset contains over 25,000 songs and over 180 of the most common tags derived from the Tagatune game and is freely available for research. [17] One

[16] `http://magnatunes.com`

[17] `http://tagatune.org/Magnatagatune.html`

Table 4.1: Task 1 - Similar song, different tag

Sorting	Mean	SE
Random	0.73	0.33
SOM	0.48	0.34
Alphabetical	0.73	0.32

thousand clips were randomly chosen from the Magnatunes dataset and were placed on a self-organizing map using music feature similarity. These 1000 clips represented 278 songs by 24 artists and have 188 different tags.

Subjects were instructed in the use of the interface and were then asked to first practice with the interface for 5-10 minutes. Subjects then performed each task in sequence. During all tasks, subjects were encouraged to speak aloud and their comments were recorded along with the completion times for each task. After the experiments, users filled in a 5-point System Usability Survey (SUS) and were interviewed.

4.5.2 Task 1

In Task 1, subjects were asked the question: "Play a classical music track by clicking on the classical tag." After the corresponding tag was clicked, a piece of classical music was played. The subjects were than asked to: "Find another track that sounds similar, according to you, using a different tag." and the time required was recorded. The mean and standard deviations of response times are detailed in Table 4.1. From this table we can see that the SOM condition had the lowest mean normalized time, of 0.48, which is considerably better than the mean normalized times of either the Random or Alphabetical conditions. A one-way between subject ANOVA was conducted.

A one-way ANOVA, or Analysis of Variance between groups, is a statistical techniques that is used to test hypotheses about a certain experiment in order to reject or tentatively accept the hypothesis. In it, one examines the hypothesis the means of two experimental variables are equal, under the assumption that the two variables follow a standard distribution. It is a commonly used technique in the analysis of experimental results in a wide number of fields, including Human Computer Interactions.

There was no significant effect of tag cloud configuration for this task ($F(2,39)=2.66$, $p=0.0830>0.05$). This task is more representative of direct searching and therefore participants did not need to utilize the underlying representation. For example a user could locate the "Baroque" tag visually in any configuration and know that it

Table 4.2: Task 2 - Similar song, different artist

	Sorting	Mean	SE
Task 2a			
	Random	0.9	0.23
	SOM	0.4	0.28
	Alphabetical	0.53	0.31
Task 2b			
	Random	0.70	0.34
	SOM	0.42	0.24
	Alphabetical	0.69	0.34

probably contains similar tracks.

4.5.3 Task 2

In Task 2, subjects were asked the question: "Play Asteria by clicking on Asteria in the artists pane." After the corresponding tag was clicked, a piece of music by the artist Asteria was played. The subjects were asked to: "Find another track that sounds similar, according to you, using a different artist." and the response time was recorded. The subject was then asked to "Find another track that sounds similar, according to you, using a tag."

The mean and standard deviations of response times are detailed in Table 4.2. From this table we can see that for Task 2a the SOM condition had the lowest mean normalized time, of 0.4, which is considerably better than the mean normalized times of either the Random or Alphabetical conditions. A one-way between subjects ANOVA was conducted showing that this result was statistically significant ($F(2,39)=12.38$, $p<0.001$). For Task 2b, the SOM condition also had the lowest mean normalized time (0.42), which was considerably better than the mean normalized times of either the Random or Alphabetical conditions. We also found that this result was statistically significant ($F(2,39)=3.56$, $p<0.05$). This task was the most representative of browsing unfamiliar music as the participants had no knowledge of the artists involved. It also illustrates the importance of visual consistency in different facets. For example as can be seen in Figure 4.15 the tags "Monks" and "Opera" are probably relevant for the artists "Asteria" and "Ensemble Sreteniye".

Table 4.3: Task 3 - Similar song, user guided search

	Sorting	Mean	SE
Task 3a			
	Random	0.68	0.33
	SOM	0.52	0.33
	Alphabetical	0.68	0.35
Task 3b			
	Random	0.83	0.26
	SOM	0.60	0.29
	Alphabetical	0.65	0.31

4.5.4 Task 3

In Task 3, subjects were asked the question: "In the next exercise, I'll ask you to find a song that you enjoy using the interface in any way you like." The response time was intentionally not recorded. The subjects were then asked to "Find another song that sounds similar to your selection, according to you in any way you want using the interface." as well as "Find a song that sounds very different to it, according to you.". Both response times were recorded.

The mean and standard deviations of response times are detailed in Table 4.3. From this table we can see that for Task 3a the SOM condition had the lowest mean normalized time, of 0.52, which is considerably better than the mean normalized times of either the Random or Alphabetical conditions. A one-way between subjects ANOVA was conducted showing no statistically significant difference ($F(2,39)=1.04$, $p=0.3640>0.05$). For Task 2b, the SOM condition also had the lowest mean normalized time (0.42), which was considerably better than the mean normalized times of either the Random or Alphabetical conditions. This result was also not statistically significant ($F(2,39)=2.71$, $p=0.0794>0.05$). For similar reasons to the ones described in Task 1 this can be attributed to knowledge of tag semantics. At the same time it shows that there is no significant penalty in any of these tasks by using the self-organizing tag cloud.

4.5.5 System Usability Survey

After the conclusion of the three timed tasks, the participants were asked to fill out a short System Usability Survey [Bro96] consisting of 6 questions, each rated on a 5 point scale, where "1" was labelled "Strongly disagree" and "5" was labelled "Strongly

Table 4.4: System Usability Survey

Question	1	2	3	4	5	Mean	SE
1	0	1	3	8	2	3.79	0.8
2	5	7	1	1	0	1.86	0.86
3	5	3	3	1	2	2.43	1.45
4	0	0	2	6	6	4.29	0.73
5	0	2	1	4	7	4.14	1.1
6	0	2	0	6	6	4.14	1.03

agree". The 6 questions were as follows:

1. I thought the application was easy to use
2. I needed to learn a lot before I could accomplish tasks with the application
3. I think people would need technical support to be able to learn how to use the application
4. I think most people would learn to user the application very quickly
5. Overall, accomplishing tasks using the self-organizing map was easier than with other methods
6. Overall, accomplishing tasks using the self-organizing map was more fun than with other methods

The results from the survey are detailed in Table 4.4. On average users rated Question 4 highest, which indicated that they thought most other people would be able to learn the application quickly. This question also had the lowest variance. In Table 4.4 we detail all the responses from the participants, and we can see that 2 participants chose the middle check box, 6 chose the next one to the right, and 6 chose the checkbox labelled "Strongly agree".

In a similar vein, participants also rated questions 5 and 6 highly, although notably, two participants rated this question as one box to the right of "Strongly Disagree". This shows that certain users found our interface facile to use and fit in well with their expectations of an interface to explore music collections, but for other users it did not. Different people enjoy different ways of interacting with media, some are more spatially oriented, and others prefer to have options presented in a linear form. In subsequent versions of this application, we would like to explore the possibility of

using different visual design strategies to make this an enclusive environment for a wide community of users.

For Question 2, the average response was 1.85, which means that on average, users mostly strongly disagree that they would have to learn a lot before accomplishing tasks with this application. It is important to include negative examples on such a user study to ensure that participants are not just choosing answers to questions randomly, and this question performs this control function.

4.5.6 Interview

We also carried out an interview with all the participants after the SUS survey. The participants were first asked which of the three conditions they felt took the least amount of time to complete, and were then asked which of the three conditions they found most fun. We then asked the participants to feel free to give us feedback on the software and algorithms.

Of the 14 users, 10 users felt the Self-Organized Map condition was the fastest, 2 users felt the random condition was the fastest, 1 felt the alphabetical condition was the fastest, and 1 expressed no preference. When asked which condition was the most fun, 9 users felt the Self-Organized Map condition was the most fun, 2 users felt the random condition was the most fun, 1 felt the alphabetical condition was the most fun, and 2 expressed no preference. This type of qualitative data is very useful in testing interfaces such as this, because of the subtle and not easily measurable behaviour of people in a task as subjective as music listening and browsing.

4.6 Conclusions and Future Directions for Music Exploration

In this chapter we describe our investigations in designing an interface for content-aware music browsing and discovery based on faceted self-organizing tag clouds. The experimental results show that self-organizing tag clouds can result in more effective retrieval especially in the case of browsing unknown artists and relating different facets such as artists and tags. We also discuss evaluation issues for music browsing interfaces. The proposed interface provides a simple, consistent interface for music discovery that can easily be adapted to small screen real-estate and touch surfaces.

There are many directions for future work. We are planning to explore visualizing tag-based similarities as edges between tags with proportional thickness. Another interesting direction is the use of more complex layout algorithms that take into account the shape of words to approximate the aesthetic seen in manually created tag clouds. Several of the user study participants suggested using the same interface for tag annotation. Another interesting possibility is the use of self-organizing tag clouds for collaborative music browsing and comparison of collections between different listeners. Finally, although we focus on music browsing in this work, we hope the ideas in this chapter can be applied to any application domain where the underlying objects that are tagged can be automatically analyzed based on their content.

Chapter 5

Computer Assistance for Analysis of Orca Vocalizations

In this chapter I use techniques from Music Information Retrieval and Machine Learning on the problem domain of the vocalizations of Orcas. I present a web-based system that I built called "The Orchive" that takes a huge archive of over 20,000 hours of orca vocalizations and and presents it to researchers around the world. This system allows users to view and listen to any recording, and also lets them annotate these recordings. These annotations would typically be either the call type or the name of the matriline or pod that made these vocalizations. This web based system then links into a set of powerful Music Information Retrieval and Machine Learning algorithm that allow researchers to extract features from the audio, train machine learning classifiers and make predictions on other audio files.

5.1 Introduction to Orca Vocalizations and The Orchive

The whale species *Orcinus orca*, commonly known as Killer Whales [JFB00], are large toothed whales found around the world, in places as far afield as Antarctica and Alaska[EDSW09]. There are three distinct types of Orcas, Transients, Residents and Offshores, each of which have different feeding behaviours and different styles of communication. The vocalizations of orcas are complex and diverse, and consist of a wide variety of vocalizations, which include echolocation clicks, tonal whistles and pulsed calls [DFS00].

Around Vancouver Island there are two distinct communities of Orcas, the Northern Residents, which have a range north of Campbell River, and the Southern Residents, which spend time around Victoria. Orcalab is a research station located on Hanson Island, a small island up near the top of Vancouver Island, directed by Dr. Paul Spong and Helena Symonds and which studies the Northern Resident population of Orcas. Orcalab has been recording Orca vocalizations for over 20 years, and has amassed a huge archive of over 20,000 hours of audio on magnetic cassette tapes. In a collaboration with them, we are digitizing and analyzing this huge and rich archive in a project called "The Orchive".

Although these recordings contain large amounts of Orca vocalizations, the recordings also contain other sources of audio, including voice-overs describing the current observing conditions, boat and cruise-ship noise, and large sections of silence. Finding the Orca vocalizations on these tapes is a labor-intensive and time-consuming task.

In the current work, we present a web-based collaborative system to assist with the task of identifying and annotating the sections of these audio recordings that contain Orca vocalizations. This system consists of a dynamic and user-informed front end written in *XHTML/CSS* and *Flash* which lets a researcher identify and label sections of audio as Orca vocalization, voice-over or background noise. By using annotation boot-strapping [Tza04], an approach inspired by semi-supervised learning, we show that it is possible to obtain good classification results while annotating only a small subset of the data. This is critical as it would take several human years to fully annotate the entire archive, a daunting task even with the use of crowdsourcing to help annotate the archive.

Once the data is annotated it will be easier to focus on data of interest such as all the orca vocalizations for a particular year without having to manually search through the audio file to find the corresponding relevant sections.

5.2 Relevance of this work to Orcas and their Vocalizations

This work is of importance to a number of different scientific communities. The primary scientific community that will benefit from this work will be cetacean biologists. In order to study the rich archive orca vocalizations that have been recorded by Orcalab, researchers must travel to Hanson Island, search through the lab books

and incidence reports to find which recordings contain the data they are interested in, locate the physical cassette tape corresponding to this recording, and then either manually listen to the tape, or perhaps digitize the tape and analyze it in the computer. Each researcher typically then keeps the annotations and data generated from this procedure themselves, if future researchers want to obtain this data for further analysis, they must first be aware of the fact that this researcher has the data, and then request it from them.

With the distributed collaborative system we have designed, not only can these biologists easily listen to any recording in the entire archive from any internet connected computer in the world, and compare different recordings, they can also add their annotations to the system. These annotations can be either private or public, if they are for use in a publication, after the article has been accepted for publication, the researcher can make their private annotations public.

Another scientific community that will receive benefits from this archive are the developers of bioacoustic algorithms. These scientists are typically computer scientists with interests in Music Information Retrieval and bioacoustics. This archive represents a site where researchers can get large amounts of high quality and uniformly collected data.

Another group of scientists that have expressed interest in the Orchive are Environmental and Conservation scientists [FOH04]. Of particular interest is the effect of boat noise on cetaceans and on the marine environment in general. For these researchers, the data they will be most interested in is the frequency and nature of orca vocalizations, and the intensity and spectral characteristics of boat noise. There are large differences in the intensity and frequency content of boat noise depending on the type of boat that creates it, speed pleasure craft often create a high pitched noise that quickly moves away, tug boats have a lower pitched sound and take a long time to move through an area, and cruise ships make a loud and distinctively high pitched sound. Analyzing the effects of these various types of boat noise will help researchers to establish guidelines for boat noise as it affects this sensitive population of marine mammals.

Another group of scientists that this work will benefit are those studying the social organization of whale communities [BOE+90]. There have been studies that investigate the transmission of culture in orca societies [DFS00] and have found evidence of this through the examination of dialect change. In a similar vein, other studies have investigated social learning [JS00] in communities of orcas. With a large database

such as this, more of these type of studies will be made possible in the future.

5.3 The Orchive

The *Orchive* ([1]) is a web-based collaborative system designed to assist with the task of identifying and annotating sections of audio recordings that contain orca vocalizations. This system consists of a dynamic front end written in *XHTML/CSS* and *Flash*. The interface allows the user to annotate regions of the recording with any tag they choose, for example "orca" and "voiceover". It automatically assigns the "background" label to unlabeled regions of the audio. In voiceover sections the observer that started the tape recording talks about the details of the particular recording such as the geographic location of the Orcas, the time of the day, the weather conditions and other items of note. A sample section of audio with voiceover, orca vocalizations and background is shown in Figure 5.3. Although we also provide ways for users to enter more a detailed classification scheme, such as the type of orca calls, in practical terms this basic classification to three categories is very important to the researchers involved. We are working on classification algorithms to automate this process of segmentation and labelling.

This web server then runs audio feature extraction and performs supervised and semi-supervised learning using the *Marsyas* [TC00] ([2]) open source software framework for audio analysis.

OrcaAnnotator is a Model-View-Controller system containing well-defined and well-separated sections, each of which presents a uniform interface to the other sections of the system. Each part is made to be a simple and well-defined unit, making them easier to test and maintain.

The primary mode of communication with the user is via an *XHTML/CSS* and *Flash* based interface. The user is presented with a simple and attractive *XHTML/CSS* web page that has been designed to be standards compliant which will facilitate accessibility by the research community on a wide variety of different web browsers and computer platforms. The *Flash* based interface is written in the haXe [MSP08] programming language, which compiles the ECMAScript language *haXe* down to *Flash* bytecodes. The *Flash* interface presents a simple interface to the user with a spectrogram of the sound file, shuttle and volume controls, a time display, and an interface

[1] `http://orchive.cs.uvic.ca`
[2] `http://marsyas.sness.net`

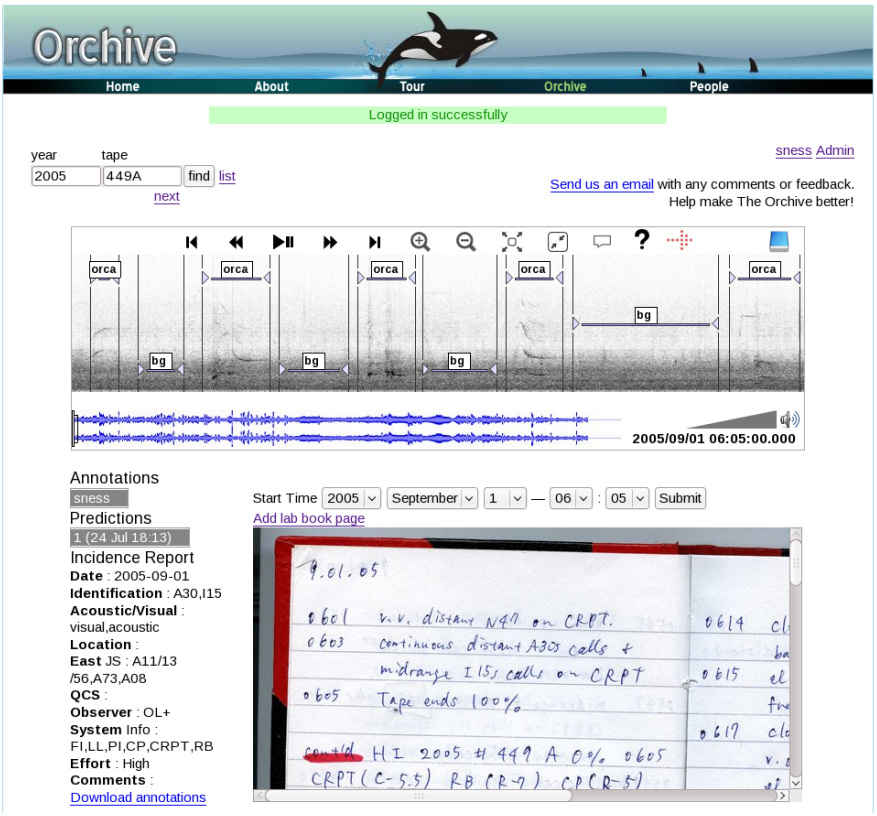

Figure 5.1: An annotated region of audio from the *Orchive*, with regions of background and orca vocalization shown. Unlabeled regions are automatically assigned a label of background noise.

for labeling the audio file. We used the labeling functionality in *Audacity* [MD02] as a model for our user-interaction paradigm. To add a label, the user simply clicks and drags the mouse on the label region. This creates a label with left and right extents, and a text region where the user can enter a text description of the audio. In addition, a pull-down menu with labels can be used for quick annotation.

Labels are saved to the database with the user that created them and the time that they were created. This user can be an actual user on the system, or can be labeled with *Marsyas* and the name and parameters of the classifier that was used for labeling. *Marsyas* contains a number of machine-learning classifiers, including Gaussian (MAP), Gaussian Mixture-Model (GMM), and Support Vector Machines (SVM). We used the "bextract" program which is part of *Marsyas*, which now includes a new Timeline module that allows the import of human-annotated sections of audio into *Marsyas* as a start for a bootstrapping approach. A variety of standard audio

feature extraction algorithms such as Mel-Frequency Cepstral Coefficients (MFCC) as well as various types of spectral features are also provided. The integration of machine learning and audio signal processing is essential in creating a semi-automatic annotation interface.

To provide communication between the *Flash* user-interface and the *Marsyas* classifier algorithms, we have employed the *Ruby on Rails* web framework[THB+06]. *Ruby on Rails* allows for quick and easy development and deployment of websites, and it provides a tight interface layer to an underlying database like *MySQL*.

Ruby on Rails also has the advantage that it makes it simple to build REST based applications[Fie]. REST is the model on which the internet is built and has the ability to minimize latency and network communication, while simultaneously maximizing the independence and scalability of network services. *Ruby on Rails* queries the database for user data, label data and locations of audio files. It then generates all the *XHTML/CSS* files displayed to the user and sends the required XML data to the *Flash* application. Once the user submits their annotated data back to the web server, it first stores this data in the database and then queues this data for *Marsyas* to run in a separate background process, perhaps on another machine, or network of machines. Once *Marsyas* completes processing the audio, the results are automatically sent back to the web server using REST web services.

Being able to segment and label the audio recordings into the three main categories (voiceover, orca vocalizations and background noise) is immensly useful to researchers working with this vast amount of data. For example background noise comprises approximately 64% of the recordings, and is much higher in some individual recordings. Fully annotating the data even using a well-designed user interface is out of the question given the size of the archive. To address this problem we have designed a semi-supervised learning system that only requires manual annotation of a small percentage of the data and utilizes machine learning techniques to annotate the other part. This recording-specific annotation bootstrapping can potentially be used with other types of time-based multimedia data.

5.4 Annotation Bootstrapping applied to Orca Vocalizations

Annotation bootstrapping is inspired by semi-supervised learning [CSZ06]. It has been shown that unlabeled data, when used in conjuction with a small amount of labeled data, can produce considerable improvements in learning accuracy. The acquisition of labeled data for a learning problem often requires manual annotation which is a time consuming process so semi-supervised learning can significantly reduce annotation time for large multimedia archives.

We extend the idea of semi-supervised learning to take advantage of the strong correlation between feature vectors from the same audio recording. In the *Orchive* each audio recording has a duration of 45 minutes and corresponds to a particular date and time. There is considerable consistency within a recording as the same person is doing the voiceover sections, the mixing settings are the same and the orcas that are vocalizing typically come from the same group. A recording-specific bootstrap classifier is trained as follows: a small percentage of the specific audio recording is manually annotated and used to train a recording-specific classifier. This classifier is then used to label the remaining parts of the recording. Due to the consistency of the recording this classifier will be to some extent overfitted to the recording and will not generalize well to other recordings. However, that is not a problem in our case as we are mainly interested in obtained labels for the entire recording. This process is repeated for each recording. Once all the recordings have been semi-automatically fully labeled then feature extraction is performed for the entire archive and a generalizing classifier is trained using the full dataset.

In order to explore whether this idea would work for our data, we created a representative database consisting of 10 excerpts from our recordings with each excerpt lasting between 5 and 10 minutes. Table 5.1 shows classification results using 10-fold cross-validation for each particular recording using a recording specific classifier as well as using a classifier trained on the entire dataset.

Cross validation is a technique used to test the accuracy of machine learning classifiers by separating a dataset into examples used to train the classifier and examples used to test the classifier. In 10 fold cross-validation, one breaks up a dataset into 10 folds, each containing 1/10 of the dataset. One then trains the machine learning classifier with 9 of the folds, and tests with the remaining fold. This process is repeated for all combinations of folds, training on 9 folds each time, and testing on the

Table 5.1: Recording-specific classification performance

	Naive bayes % correct		SMO % correct	
	self	train with remaining	self	train with remaining
446A	89.42	93.10	95.00	73.39
446B	63.45	77.66	85.85	70.23
447B	75.46	57.32	82.02	68.17
448A	52.18	61.02	81.57	62.24
448B	84.63	67.62	83.64	67.87
449B	82.24	51.85	86.41	75.72
450A	94.66	90.91	96.12	91.58
450B	83.65	96.27	99.29	94.92
451A	70.92	89.58	97.04	78.72
451B	74.18	33.73	82.34	50.88

Table 5.2: Classification performance using annotation-bootstrapping (SVM classifier)

% data used	%correct	F-measure
100	82.38	0.876
10	81.98	0.874
5	82.04	0.874
1	79.95	0.864
0.1	78.08	0.857
0.01	71.42	0.800

remaining fold.

Two classifiers are used: a simple Naive Bayes classifier (NBS), as well as a Support Vector Machine (SVM). The results shown are based on the use of the standard Mel-Frequency Cepstral Coefficients (MFCC) as audio features. The "self" column shows the classification accuracy results of using a recording-specific classifier, whereas the "remaining" columns shows the classification accuracy results using the remaining nine recordings. As can be seen, recording-specific classifier can generate significantly better results than generalized classifiers, which is not surprising as they adapt to the specific data of the recording. This justifies the use of their annotation results to labeled the unlabeled parts of the audio recording.

The goal of annotation bootstrapping is to only label a small part of each recording to train a recording-specific classifier which is then used to annotate the remainder of the recording. Table 5.2 shows the results in terms of classification accuracy and F-measure over the entire dataset for different amounts of labeled data. As one can see the classification accuracy remains quite good, even when only a small percentage of the data is labeled and annotation bootstrapping is used to label the rest. The first row shows the classification accuracy when all the data is used for training.

The F-measure is the weighted harmonic mean of precision and recall, and is described by the following formula:

$$F = \frac{2 * precision * recall}{precision + recall}$$

Figure 5.4 shows graphically how the classification accuracy increases with the amount of labeled data used for training. In both the table and the figure, the classifier used is a Support Vector Machine (SVM) and for evaluation a variation of

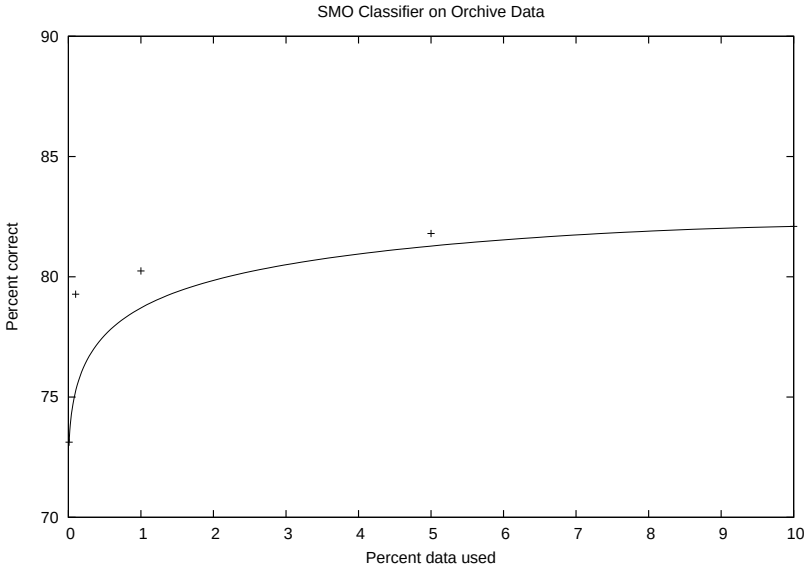

Figure 5.2: Graph of classification accuracy as percentage of labeling required. Data shown is for the performance of SMO classifier for different percentages of data used to train the classifier.

10-fold cross-validation where each of the 10 recordings is held out for testing, the remaining ones are used for training, and the process is iterated 10 times. We also experimented with different choices of window size for the feature calculation as well as different audio feature parametrization but there was no significant difference in the obtained results.

To make the importance of annotation bootstrapping concrete, fully annotating the archive would take approximately 2 and half years (assuming 24/7 manual annotation) whereas using one percent annotation bootstrapping would take 3 months (assuming 24/7 manual annotation) without significantly affecting the ability of the system to succesfully label all the recordings in the 3 classes of interest.

5.5 Conclusions for Computer Assisted Analysis of Orca Vocalizations

By combining the expert knowledge of our scientific collaborators with new multimedia web-based tools in an agile development strategy, we have been able to ask new questions that had previously been out of reach. The large and multi-dimensional datasets in both the chant community and in orca vocalization research provide challenging fields for study, and new web-based technologies provide the flexibility to allow true collaboration between scientific partners in widely disparate fields of study. We described an automatic technique for simplifying melodic contours based on kernel density estimation. Annotation of the archive of orca vocalizations is very time-consuming, we proposed annotation bootstrapping and show that it is an effective technique for automatically annotating recordings.

Chapter 6

Conclusions

In this thesis, I have explored a variety of different areas where tools and ideas from Music Information Retrieval (MIR) and web based collaborative software can be used. In addition, developed tools that support these investigations have been presented.

Although the application areas, from Orca vocalizations, to chant traditions from around the world, to browsing of large music collections at first seem quite disparate, the same tools and analysis can be applied to all of them. These tools, which include algorithms such as Audio Feature Extraction, Self-Organizing Maps, Support Vector Machines, were developed in a variety of areas, and only recently have been applied to the analysis of audio signals and music. This thesis outlines only the preliminary stages of applying these tools to these three application areas, and in the future, we anticipate that many more useful results will arise from more detailed analysis with these and other tools.

The use of online, web-based tools to present these results to scientists is a very important new development in the dissemination of advanced algorithms, developed in Computer Science, to researchers in other fields, such as Biology, Ethnomusicology and Music. In the past, these powerful algorithms were often too difficult to use and therefore basically unavailable to be used by these other scientists. However, web-based tools are often easier to use and are less intimidating for less sophisticated users, and thus open up whole new avenues for collaboration between disciplines.

In this thesis I present results that show the power of these methods, first, by demonstrating that real users find these systems useful, but also in terms of the results of machine learning algorithms on various datasets. These results are summarized below.

In the chapter on analyzing chants, I present results in Table 3.1 that show the

average precision of gestures performed by two speakers, one in Morocco and one in Hungary. From this table we see that the average precision is consistently higher with the speaker from Hungary, which can be interpreted to mean that this speaker has a more consistent vocal interpretation of the signs. Later, in 3.13 we show results that compare our method of simplifying a pitch contour is superior to using a continuous pitch contour. This figure also shows that our method of using a data derived scale is better than using an equally tempered scale, which can be easily seen by comparing the upper two curves. In addition, we show that the optimal number of scale degrees to quantize to using our method is 2 scale degrees.

AudioScapes is an extensible framework and architecture for surface-based interfaces for browsing large audio and music collections. Given the exploratory nature of the work we have not yet been able to conduct detailed quantitative user studies which are planned for the future. It is our hope that the developed interfaces have the potential to make browsing of audio collections much more effective. As it is difficult to convey how the system works in paper we have collected videos and web demonstrations on a web-page `http://audioscapes.sness.net`.

In the chapter on novel music browsing interfaces we present a variety of novel interfaces to help people explore their music collections. One of these is a new organizational method for tag clouds that uses Self-Organizing Maps. We conducted a user study that compared traditional tag clouds, with tags organized either alphabetically or randomly to this new method, and obtained promising results. In Table 4.2 results are presented from an experiment where subjects were asked the question: "Play Asteria by clicking on Asteria in the artists pane." After the corresponding tag was clicked, a piece of music by the artist Asteria was played. The subjects were asked to: "Find another track that sounds similar, according to you, using a different artist." and the response time was recorded. The subject was then asked to "Find another track that sounds similar, according to you, using a tag.". The mean and standard deviations of response times are detailed in Table 4.2. Both Task 2a and Task2b had results that were statistically significant, with an ANOVA analysis of the results from Task 2a being ($F(2,39)=12.38$, $p<0.001$) and for Task 2b being ($F(2,39)=3.56$, $p<0.05$). Of all the tasks, this task was the most representative of browsing unfamiliar music as the participants had no knowledge of the artists involved.

During this experiment, we also conducted a System Usability Survey with 6 questions rated on a 5 point scale. The results from the survey are detailed in Table 4.4, and show that participants enjoyed using the interface and found it an interesting

and fun way to browse their music collections. The results from this usability study are significant because for a task that involves people listening to music collections, it is important to consider not just quantitative measures, such as time to completion, but also qualitative measures, which can more accurately gauge people's experiences. By combining both quantitative and qualitative measures within this study, we hope to combine the relative strengths of both methods.

In this thesis, I have done preliminary investigations and applications of some advanced algorithms to three diverse problem domains, using easily accessible web-based tools. I look forward to the further extension of these results, both by myself and the scientific community at large.

Chapter 7

Glossary

ANOVA (Science) - Analysis of Variance between groups - A group of statistical techniques to test hypotheses based on experimental results. It tests the distribution of the means of variables to see if they are the same, assuming the variables are normally distributed.

Computational Ethnomusicology (Musicology) - a new field that uses the computational techniques commonly used in Music Information Retrieval (MIR) to analyse music and audio from social and cultural traditions from around the world.

Cross-validation (Computer Science) - A method used to test the accuracy of machine learning classifiers by separating a dataset into examples used to train the classifier and examples used to test the classifier.

Ethnomusicology (Musicology) - a sub-discipline of Musicology that focuses on the study of the socio-cultural aspects of music in societies around the world.

Folksonomy (Computer Science) - a system of classification that comes from a group of users collaboratively creating and managing tags in an effort to annotate and categorize content

Flux (MIR) - In this work, is the norm of the difference vector between two successive magnitue/power spectra.

Gesture (Musicology) - In it's simplest definition, a gesture is simply the pitch

contour of a sung musical phrase. However, this word has become imbuded with subtle meanings in Ethnomusicology, and can be thought of as not just the pitch contour, but is also related to the gesture of the conductor of the piece of music and the intention of the performer.

Hermeneutic (Musicology) - The study of interpretation, particularily the interpretation of the Bible and Torah.

Heterophonic (Musicology) - Two voice singing a melody (or harmonies of that melody) at one time

MFCC coefficients (MIR) - A way to transform a standard spectrum into one that more closely approximates how the human ear perceives sound.

Monophonic (Musicology) - One voice singing a melody at one time

Music Information Retrieval (Computer Science) - also known by the acronym MIR, a new field of study where one applies tools from areas such as Digital Signal Processing, Audio Feature Extraction and Machine Learning to help people understand and retrieve information from music or audio.

Ontology (Computer Science) - a formal representation of ideas or concepts and the relation between them. This Computer Science definition of ontology is a subset of the broader philosophical idea of ontology, which is a study of what exists and what can exist.

Open Sound Control (Computer Science + Music) - A new content format for sending musical data between computers

Paradigmatic (Musicology) - In a text or a song, the relationships between symbols, and the analysis of how symbols relate to each other both in one text and amongst a group of texts.

Pashta (Musicology) - One of the cantillation signs from the Torah - It means "Stretching out", because its shape leans forward.

Phase Vocoder (Computer Science + Music) - A type of digital audio filter which can scale audio in both time and frequency

Rolloff (MIR) - A measure of the steepness of falloff in an audio spectrum

Semiotics (Musicology) - The study of signs and signifiers

Self-Organizing Map (Computer Science) - A technique that maps a high dimensional feature space to a lower space. It is a similar technique to artifical neural networks.

Siratok (Musicology) - A form of lament song found in Hungary

Sof Pasuq (Musicology) - One of the cantillation signs from the Torah. It means "End of verse"

Spectral Centroid (MIR) - A measure of the "center of mass" of a spectrum.

Syntagmatic (Musicology) - In a text or song, the symbols and the way that they can be joined together. The surface structure of a text or a song.

Tag Cloud (Computer Science) - a two-dimensional graphical representation of words related to a topic where words are typically arranged alphabetically or randomly. Words that are of higher relevance are usually shown in a larger font.

Wii remote (Computer Science) - The Wii is a new computer gaming system developed by Nintendo, which has a novel input controllers containing not just buttons, but a 3D accellerometer, an infrared camera, a speaker and a vibration generator. The Wii remote is wireless, communicating with the base station via bluetooth, which can be received by a standard computer with a bluetooth receiver.

XML (Computer Science) - eXtensible Markup Language - A popular tree-tree based format for encoding data

Chapter 8

Web Links

The following are web links to the various web-based tools developed in the course of this research:

Cantillion : `http://cantillion.sness.net`.

Orchive : `http://orchive.cs.uvic.ca`

Audioscapes : `http://audioscapes.sness.net`

Chapter 9

Publications from this Research

Presented here is a list of all the publications that come from the research presented in this thesis:

Steven R. Ness, Daniel Peter Biro and George Tzanetakis Computer-assisted cantillation and chant research using content-aware web visualization tools, Multimedia Tools and Applications, Accepted for publication

S. R. Ness, A. Theocharis, G. Tzanetakis, L. G. Martins Improving Automatic Music Tag Annotation Using Stacked Generalization Of Probabilistic SVM Outputs, ACM Multimedia 2009

Steven R. Ness, George Tzanetakis, Audioscapes: exploring surface interfaces for music exploration - ICMC 2009

Steven R. Ness, George Tzanetakis, SOMba : Multiuser music creation using Self-Organizing Maps and motion tracking - ICMC 2009

Steven R. Ness, Daniel Peter Biro, George Tzanetakis : Content-aware web browsing and visualization tools for cantillation and chant research, 7th International Workshop on Content-Based Multimedia Indexing

George Tzanetakis, Manjinder Singh Benning, Steven R. Ness, Darren Minifie, Nigel Livingston : Assistive Music Browsing using Self-Organizing Maps - PETRA 2009 : 2nd International Conference on PErvasive Technologies Related to Assistive

Environments

Steven R. Ness, Matthew Wright, Luis Gustavo Martins, George Tzanetakis: Chants and Orcas: semi-automatic tools for audio annotation and analysis in niche domains. ACM Multimedia 2008: 9-16

Daniel Peter Biro, Steven Ness, Matthew Wright, W. Andrew Schloss and George Tzanetakis Decoding the Song: Histogram-Based Paradigmatic and Syntagmatic Analysis of Melodic Formulae in Hungarian Laments, Jewish Torah Trope, Tenth Century Plainchant and Koran Recitation EMUS Expressivity in MUsic and Speech : IRCAM - Institut de Recherche et de Coordination Acoustique/Musique - Paris, France

Bibliography

[ABB+09] Edoardo Acotto, Matteo Baldoni, Cristina Baroglio, Viviana Patti, Flavio Portis, and Giorgio Vaccarino. Arsmeteo: artworks and tags floating over the planet art. In *Proc. ACM conf. on Hypertext and hypermedia*, pages 331–332, 2009.

[ABD+97] K.L. Atkins, D.E. Brownlee, T. Duxbury, C.-W. Yen, P. Tsou, and J.M. Vellinga. Stardust: Discovery's interstellar dust and cometary sample return mission. volume 4, pages 229–245, 1997.

[Ada06] O. Adam. Advantages of the Hilbert Huang transform for marine mammals signals analysis. *J. Acoust. Soc. Am.*, 120:2965–2973, Nov 2006.

[BGN08] Scott Bateman, Carl Gutwin, and Miguel Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *HT '08: Proc. ACM conf. on Hypertext and hypermedia*, pages 193–202, 2008.

[BHDM06] J. C. Brown, A. Hodgins-Davis, and P. J. Miller. Classification of vocalizations of killer whales using dynamic time warping. *J. Acoust. Soc. Am.*, 119:34–40, Mar 2006.

[BM02] Erin Bradner and Gloria Mark. Why distance matters: effects on cooperation, persuasion and deception. In *CSCW '02: Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 226–235, 2002.

[BM06] A.J. Westphal B.M. Mendez, N. Craig. Stardust@home: Enlisting students and the public in the search for interstellar dust. 2006.

[BM07] J. C. Brown and P. J. Miller. Automatic classification of killer whale vocalizations using dynamic time warping. *J. Acoust. Soc. Am.*, 122:1201–1207, Aug 2007.

[BMD07] Manjinder Singh Benning, Michael McGuire, and Peter Driessen. Improved position tracking of a 3-d gesture-based musical controller using a kalman filter. In *NIME '07: Proc. of the 7th Int. Conf. on New interfaces for musical expression*, pages 334–337, New York, NY, USA, 2007. ACM.

[BN08] David Bradshaw and Kia Ng. Tracking conductors hand movements using multiple wiimotes. *Automated solutions for Cross Media Content and Multi-channel Distribution, 2008. AXMEDIS '08. Int. Conf. on*, pages 93–99, Nov. 2008.

[BOE+90] M.A. Bigg, P.F. Olesiuk, G.M. Ellis, J.K.B. Ford, and K.C. Balcomb III. Social organization and genealogy of resident killer whales (orcinus orca) in the coastal waters of british columbia and washington state. Technical report, 1990.

[Bra08] Daren C. Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75–90, 2008.

[Bro96] J. Brooke. *SUS: a "quick and dirty" usability scale.* 1996. P. W. Jordan, B. Thomas, B. A. Weerdmeester and A. L. McClelland.

[Bru96] Peter Brusilovsky. Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6(2-3):87–129, 1996.

[BS91] Liam J. Bannon and Kjeld Schmidt. Cscw: Four characters in search of a context. In *Studies in Computer Supported Cooperative Work*, volume 8 of *Human Factors in Information Technology*, pages 3–17. North-Holland, 1991.

[Cam07] A. Camacho. *A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music.* PhD thesis, University of Florida, 2007.

[Car98] O.A.S. Carpinteiro. A self-organizing map model for analysis of musical time series. In *Proc. Brazilian Symposium on Neural Networks*, pages 140 – 5, 1998.

[CFPT06] Matthew Cooper, Jonathan Foote, Elias Pampalk, and George Tzanetakis. Visualization in audio-based music information retrieval. *Computer Music Journal*, 30(2):42–62, 2006.

[CKGB02] Pedro Cano, Martin Kaltenbrunner, Fabien Gouyon, and Eloi Battle. On the use of fastmap for audio retrieval and browsing. In *Proc. of the International Symposium on Music Information Retrieval*, Paris, France, 2002.

[Cou57] D. Coupland. *Microserfs*. HarperCollins, 1957.

[CRL05] Roland Cahen, Xavier Rodet, and Jean-Philippe Lambert. Sound navigation in phase installation: Producing music as performing a game using haptic feedback, 2005.

[CSZ06] O. Chappelle, B. Scholkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[CTL+06] Xiaoyu Chen, Marilyn Tremaine, Robert Lutz, Jae woo Chung, and Patrick Lacsina. Audiobrowser: a mobile browsable information access for the visually impaired. *Universal Access in the Information Society*, 5(1):4–22, 06/01 2006.

[CZJ+07] Rui Cai, Lei Zhang, Feng Jing, Wei Lai, and Wei-Ying Ma. Automated music video generation using web image resource. In -, volume 2, pages 737–740 –, Honolulu, HI, United States, 2007.

[CZZ07] Jean-Pierre Cahier, L'Hédi Zaher, and Manuel Zacklad. Information seeking in a "socio-semantic web" application. In *ICPW '07: Proc. of the 2nd Int. Conf. on Pragmatic web*, pages 91–95, 2007.

[DFS00] V.B. Deecke, J.K.B. Ford, and P. Spong. Dialect change in resident killer whales (orcinus orca): implications for vocal learning and cultural transmission. *Animal Behaviour*, 60(5):619–638, 2000.

[DJ06] V. B. Deecke and V. M. Janik. Automated categorization of bioacoustic signals: avoiding perceptual pitfalls. *J. Acoust. Soc. Am.*, 119:645–653, 2006.

[DVWK08] Catalina M. Danis, Fernanda B. Viegas, Martin Wattenberg, and Jesse Kriss. Your place or mine?: visualization as a community component. In *Proc. CHI 2008*, pages 275–284, 2008.

[EDSW09] J. A. Estes, D. F. Doak, A. M. Springer, and T. M. Williams. Causes and consequences of marine mammal population declines in southwest Alaska: a food-web perspective. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 364:1647–1658, Jun 2009.

[EGK+01] John Ellson, Emden R. Gansner, Eleftherios Koutsofios, Stephen C. North, and Gordon Woodhull. Graphviz - open source graph drawing tools. *Graph Drawing*, pages 483–484, 2001.

[Erb11] J. Erb. Musicalc, the program for music fans. *HC Mein Home Computer*, (11):98 – 100, 1984/11/.

[ESG04] Pampalk E., Dixon S., and Widmer G. Exploring music collections by browsing different views. *Computer Music Journal. Summer 2004; 28(2): 49–62*, 2004.

[EUUY09] Takeharu Eda, Toshio Uchiyama, Tadasu Uchiyama, and Masatoshi Yoshikawa. Signaling emotion in tagclouds. In *WWW '09: Proc. of the 18th Int. Conf. on World Wide Web*, pages 1199–1200, 2009.

[FD02] J. Futrelle and S. Downie. Interdisciplinary communities and research issues in music information retrieval. In *Proc. Int. Conf. on Music Information Retrieval(ISMIR)*, 2002.

[FFM+08] Ko Fujimura, Shigeru Fujimura, Tatsushi Matsubayashi, Takeshi Yamada, and Hidenori Okuda. Topigraphy: visualization for large-scale tag clouds. In *WWW '08: Proc. of the 17th Int. Conf. on World Wide Web*, pages 1087–1088, 2008.

[Fie] Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. Phd dissertation, University Of California, Irvine.

[FOH04] A. D. Foote, R. W. Osborne, and A. R. Hoelzel. Environment: whale-call response to masking boat noise. *Nature*, 428:910, 2004.

[FP07] M. Feldmeier and J.A. Paradiso. An interactive music environment for large groups with giveaway wireless motion sensors. *Computer Music Journal*, 31(1):50 – 67, Spring 2007.

[FR82] W. Schloss Foster, S. and A. Rockmore. Towards an intelligent editor of digital audio: Signal processing methods. *Computer Music Journal*, 6(1):42–51, 1982.

[FR01] Markus Frühwirth and Andreas Rauber. Self-Organizing Maps for Content-Based Music Clustering. In *Proc. of the 12th Italien Workshop on Neural Nets (WIRN01)*, Vietri sul Mare, Italy, 2001. Springer.

[GG05] Masataka Goto and Takayuki Goto. Musicream: New music playback interface for streaming, sticking, sorting, and recalling musical pieces. In *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2005.

[GN03] Masataka Goto and Takuichi Nishimura. RWC music database: Music genre database and musical instrument sound database. In *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2003.

[Gol07] F. Golshani. TUI or GUI–it's a matter of somatics. *Multimedia, IEEE*, 14(1), 2007.

[Gre76] John Greenway. *Ethnomusicology / John Greenway.* 1976.

[Gre78] John M. Grey. Timbre discrimination in musical patterns. *Journal of the Acoustical Society of America*, 1978.

[Gru88] Jonathan Grudin. Why cscw applications fail: problems in the design and evaluationof organizational interfaces. In *CSCW '88: Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, pages 85–93, 1988.

[GSD08] Kavita A. Ganesan, Neelakantan Sundaresan, and Harshal Deo. Mining tag clouds and emoticons behind community feedback. In *WWW '08: Proc. of the 17th Int. Conf. on World Wide Web*, pages 1181–1182, 2008.

[HFS+08] Maria-Elena Hernandez, Sean M. Falconer, Margaret-Anne Storey, Simona Carini, and Ida Sim. Synchronized tag clouds for exploring semi-structured clinical trial data. In *CASCON '08: Proc. of the 2008 Conf. of the center for advanced studies on collaborative research*, pages 42–56, 2008.

[HIB+07] Steve Hodges, Shahram Izadi, Alex Butler, Alban Rrustemi, and Bill Buxton. Thinsight: versatile multi-touch sensing for thin form-factor displays. In *UIST '07: Proc. of the 20th annual ACM symposium on User interface software and technology*, pages 259–268, New York, NY, USA, 2007. ACM.

[HK07] Martin J. Halvey and Mark T. Keane. An assessment of tag presentation techniques. In *WWW '07: Proc. of the 16th Int. Conf. on World Wide Web*, pages 1313–1314, 2007.

[HLLR00] Timo Honkela, Teemu Leinonen, Kirsti Lonka, and Antti Raike. Self-organizing maps and constructive learning. In *Proc. of ICEUT'2000, Beijing, August 21-25*, pages 339–343, 2000.

[Hoo00] Ki Mantle Hood. Ethnomusicology's bronze age in y2k. pages 365–375, 2000.

[How70] H.S. Howe. Compositional considerations in electronic music. volume 18, pages 690 –, New York, NY, USA, 1970.

[How08] Jeff Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business, 2008.

[HP04] Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

[HWI+07] Juha Häikiö, Arto Wallin, Minna Isomursu, Heikki Ailisto, Tapio Matinmikko, and Tua Huomo. Touch-based user interface for elderly users. In *MobileHCI '07: Proc. of the 9th Int. Conf. on Human computer interaction with mobile devices and services*, pages 289–296, New York, NY, USA, 2007. ACM.

[ILK04] H.H.S. Ip, K.C.K. Law, and B. Kwong. Cyber composer: hand gesture-driven intelligent music composition and generation. In -, pages 46 – 52, Honolulu, HI, USA, 2004.

[IU97] Hiroshi Ishii and Brygg Ullmer. Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proc. CHI 1997*, 1997.

[JC.08] Brown JC. Mathematics of pulsed vocalizations with application to killer whale biphonation. *J Acoust Soc Am.*, 123(5):2875–83, 2008.

[JFB00] G.E. Ellis J.K.B. Ford and K.C. Balcomb. *Killer Whales 2nd ed.* University of British Columbia Press, Vancouver, 2000.

[JGH+08] Robert J.K. Jacob, Audrey Girouard, Leanne M. Hirshfield, Michael S. Horn, Orit Shaer, Erin Treacy Solovey, and Jamie Zigelbaum. Reality-based interaction: a framework for post-wimp interfaces. In *Proc. CHI 2008*, pages 201–210, New York, NY, USA, 2008. ACM.

[JKGB05] Sergi Jordà, Martin Kaltenbrunner, Günter Geiger, and Ross Bencina. The reactable*. In *Proc. of the International Computer Music Conference (ICMC 2005)*, Barcelona, Spain, 2005.

[JS00] V.M. Janik and P.J.B. Slater. The different roles of social learning in vocal communication. *Animal Behaviour*, 60:1–11, 2000.

[JS06] Ajita John and Doree Seligmann. Collaborative tagging and expertise in the enterprise. In *WWW '06: Proc. of the 15th Int. Conf. on World Wide Web*, 2006.

[Kar98] Theodore Karp. *Aspects of Orality and Formularity in Gregorian Chant.* Northwestern University Press, Evanston, 1998.

[KCS08] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456, New York, NY, USA, 2008. ACM.

[Ker85] Joseph Kerman. *Contemplating music : challenges to musicology / Joseph Kerman.* Harvard University Press, Cambridge, Mass. :, 1985.

[Koh95] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences.* Berlin, Heidelberg, 1995.

[Kor01] Andreas Kornstadt. The jring system for computer-assisted musicological analysis. In *In Second International Symposium on Music Information Retrieval (ISMIR 2001*, pages 93–98, 2001.

[Kru90] Carol L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford, 1990.

[KSH12] S. Kasahara, R. Saegusa, and S. Hashimoto. Rhythm generation with associative self organizing maps. *Transactions of the Information Processing Society of Japan*, 48(12):3649 – 57, 2007/12/.

[Lab12] N. Labordus. Design techniques for rhythm generators. ii. *Electronics Australia*, 45(12):78 – 81, 1983/12/.

[LBY+07] Rui Li, Shenghua Bao, Yong Yu, Ben Fei, and Zhong Su. Towards effective browsing of large scale social annotations. In *WWW '07: Proc. of the 16th Int. Conf. on World Wide Web*, pages 943–952, 2007.

[LD99] J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, (7(3)):323–332, 1999.

[LGTS+04] Roberto Lopez-Gulliver, Hiroko Tochigi, Tomohiro Sato, Masami Suzuki, and Norihiro Hagita. Senseweb: collaborative image classification in a multi-user interaction environment. In *MULTIMEDIA '04: Proc. of the 12th annual ACM Int. Conf. on Multimedia*, pages 456–459, New York, NY, USA, 2004. ACM.

[Lis63] George List. The musical significance of transcription. *Ethnomusicology*, 1963.

[LKJK08] Kangpyo Lee, Hyunwoo Kim, Chungsu Jang, and Hyoung-Joo Kim. Folksoviz: a subsumption-based folksonomy visualization using wikipedia texts. In *WWW '08: Proceeding of the 17th Int. Conf. on World Wide Web*, pages 1093–1094, 2008.

[Lom56] Alan Lomax. Notes on a systematic approach to the study of folk song. *Journal of the International Fold Music Council*, 1956.

[Lom59] Alan Lomax. Folk song style. *American Anthropologist*, 1959.

[Lom62] Alan Lomax. Song structure and social structure. *Ethnomusicology*, 1962.

[LSH07] Sung Eob Lee, Dong Kwan Son, and Steve SangKi Han. Qtag: tagging as a means of rating, opinion-expressing, sharing and visualizing. In *SIGDOC '07: Proc. of the 25th annual ACM Int. Conf. on Design of communication*, pages 189–195, 2007.

[LvA09] Edith Law and Luis von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proc. CHI 2009*, pages 1197–1206, 2009.

[Mas03] M. Masugi. Energy spectrum-based analysis of musical sounds using self-organizing map. *IEICE Transactions on Information and Systems*, E86-D(9):1934 – 8, Sept. 2003.

[MC09] Cédric S. Mesnage and Mark J. Carman. Tag navigation. In *SoSEA '09: Proc. of the 2nd international workshop on Social software engineering and applications*, pages 29–32, 2009.

[MD02] Dominic Mazzoni and Roger B. Dannenberg. A fast data structure for disk-based audio editing. *Comput. Music J.*, 26(2):62–76, 2002.

[MDC03] Stephen McAdams, Philippe Depalle, and E. Clarke. Analysing musical sounds (? para?tre). In N. Clarke, E. / Cook, editor, *Empirical Musicology: Aims, Methods, Prospects*. Oxford University Press, Oxford, 2003.

[ME05] Michael Mandel and Daniel Ellis. Song-level features and support vector machines for music classification. In *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2005.

[Met26] M. Metfessel. Technique for objective studies of the vocal art. *PSTCHOL. MONOG*, 1926.

[MFK06] David R. Millen, Jonathan Feinberg, and Bernard Kerr. Dogear: Social bookmarking in the enterprise. In *Proc. CHI 2006*, pages 111–120, 2006.

[MGvSV08] Mark Melenhorst, Marjan Grootveld, Mark van Setten, and Mettina Veenstra. Tag-based information retrieval of video content. In *UXTV '08: Proceeding of the 1st Int. Conf. on Designing interactive user experiences for TV and video*, pages 31–40, 2008.

[MHM+08] P. Marshall, E. Hornecker, R. Morris, N. Sheep Dalton, and Y. Rogers. When the fingers do the talking: A study of group participation with varying constraints to a tabletop interface. *Horizontal Interactive Human Computer Systems, 2008. TABLETOP 2008. 3rd IEEE International Workshop on*, pages 33–40, Oct. 2008.

[MJ76] Stanley Milgram and D. Jodelet. *Psychological Maps of Paris*, pages 104–124. Holt, Rinehart and Winston, 1976. W.I.H. Proshanksy and L. Rivlin(eds).

[MNBD06] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proc. of the 17th Conf. on Hypertext and hypermedia*, pages 31–40, 2006.

[MNV09] Stefano Mizzaro, Elena Nazzi, and Luca Vassena. Collaborative annotation for context-aware retrieval. In *ESAIR '09: Proc. of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 42–45, 2009.

[MSP08] L. McColl-Sylvester and F. Ponticelli. *Professional haXe and Neko*. Wiley Publishing, Inc., Indianapolis, IN, 2008.

[MT06] Jennifer Murdoch and George Tzanetakis. Interactive content-aware music browsing using the radio drum. In *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2006.

[MUNS05] Fabian Mörchen, Alfred Ultsch, Mario Nöcker, and Christian Stamm. Databionic visualization of music collections according to perceptual distance. In *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2005.

[NDS03] B. Nevile, P. Driessen, and W.A. Schloss. A new control paradigm: software-based gesture analysis for music. *Communications, Computers and signal Processing, 2003. PACRIM. 2003 IEEE Pacific Rim Conference on*, 1:360–363 vol.1, Aug. 2003.

[NRH+99] S. Nummela, T. Reuter, S. Hemil?, P. Holmberg, and P. Paukku. The anatomy of the killer whale middle ear (Orcinus orca). *Hear. Res.*, 133:61–70, Jul 1999.

[NT09] Steven R. Ness and George Tzanetakis. Audioscapes: exploring surface interfaces for music exploration. In *Proc. of the International Computer Music Conference (ICMC 2009)*, 2009.

[NWMT08] S. Ness, M. Wright, L.G. Martins, and Tzanetakis.G. Chants and Orcas: Semi-automatic tools for Audio Annotation and Analysis in Niche Domains. In *Proc. ACM Multimedia*, Vancouver, Canada, 2008.

[OO00] G.M. Olson and J.S. Olson. Distance matters. *Human-Computer Interaction*, 15(2/3):139–178, 2000.

[PDW03] E. Pampalk, S. Dixon, and G. Widmer. Exploring music collections by browsing different views. In *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2003.

[PG06] Elias Pampalk and Masataka Goto. Musicrainbow: A new user interface to discover artists using audio-based similarity and web-based labeling. In *Proc. of the 7th Int. Conf. on Music Information Retrieval (ISMIR '06)*, pages 367–370, Victoria, Canada, October 2006.

[PG07] Elias Pampalk and Masataka Goto. Musicsun: A new approach to artist recommendation. In *Proc. of 8th Int. Conf. on Music Information Retrieval*, Vienna, Austria, 2007.

[Pop92] Anthony Pople. Computer music and computer-based musicology. *Comput. Educ.*, 19(1-2):173–182, 1992.

[Rek02] Jun Rekimoto. Smartskin: An infrastructure for freehand manipulation on interactive surfaces. In *Proc. CHI 2002*, 2002.

[RF01] A. Rauber and M. Fruhwirth. Automatically analyzing and organizing music archives, 2001.

[RM98a] A. Rauber and D. Merkl. Creating an order in distributed digital libraries by integrating independent self-organizing maps, 1998.

[RM98b] A. Rauber and D. Merkl. Organization of distributed digital libraries: a neural network-based approach, 1998.

[RPM02a] A. Rauber, E. Pampalk, and D. Merkl. Content-based music indexing and organization, 2002.

[RPM02b] A. Rauber, E. Pampalk, and D. Merkl. Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarities, 2002.

[RPM03] A. Rauber, E. Pampalk, and D. Merkl. The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models. *Journal of new Music Research*, 32(2):193–210, June 2003.

[Sal07] J. Salonen. Self-organising map based tag clouds - creating spatially meaningful representations of tagging data. In *Proc. of the 1st OPAALS conference, 26-27 November 2007, Rome, Italy*, 2007.

[SD01] W.A. Schloss and P. Driessen. New algorithms and technology for analyzing gestural data. In *Proc. IEEE Pacific Rim Conference*, 2001.

[See58] Charles Seeger. Singing style. *Western Folklore*, 1958.

[SG07] Torben Schou and Henry J. Gardner. A wii remote, a game engine, five sensor bars and a virtual reality theatre. In *OZCHI '07: Proc. of the 19th Australasian conference on Computer-Human Interaction*, pages 231–234, New York, NY, USA, 2007. ACM.

[SGVF05] Ian Stavness, Jennifer Gluck, Leah Vilhan, and Sidney Fels. The musictable: A map-based ubiquitous system for social interaction with a digital music collection. In *ICEC*, pages 291–302, 2005.

[SHT07] Jennifer Murdoch Stephen Hitchner and George Tzanetakis. Music browsing using a tabletop display. In *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2007.

[SLB+08] Anze Slosar, Kate Land, Steven Bamford, Chris Lintott, Dan Andreescu, Phil Murray, Robert Nichol, Jordan M. Raddick, Kevin Schawinski, Alex Szalay, Daniel Thomas, and Jan Vandenberg. Galaxy zoo: Chiral correlation function of galaxy spins. *Mon. Not. R. Astron. Soc*, Sep 2008.

[Sur05] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[SVFR04] Chia Shen, Frédéric D. Vernier, Clifton Forlines, and Meredith Ringel. Diamondspin: an extensible toolkit for around-the-table interaction. In *Proc. CHI 2004*, pages 167–174, New York, NY, USA, 2004. ACM.

[TASW07] G. Tzanetakis, Kapur. A, W.A Schloss, and M. Wright. Computational ethnomusicology. *Journal of Interdisciplinary Music Studies*, 1(2), 2007.

[TBL07] Bernie C. Till, Manjinder Singh Benning, and Nigel Livingston. Wireless inertial sensor package (wisp). In *NIME '07: Proceedings of the 7th international conference on New interfaces for musical expression*, pages 403–404, New York, NY, USA, 2007. ACM.

[TBN+09] George Tzanetakis, Manjinder Singh Benning, Steven R. Ness, Darren Minifie, and Nigel Livingston. Assistive music browsing using self-organizing maps. In *PETRA '09: Proc. of the 2nd Int. Conf. on PErvsive Technologies Related to Assistive Environments*, pages 1–7, 2009.

[TC00] G. Tzanetakis and P. Cook. Marsyas: A framework for audio analysis. *Organized Sound*, 4(3), 2000.

[TC02] G. Tzanetakis and P. Cook. Musical Genre Classification of Audio Signals. *IEEE Trans. on Speech and Audio Processing*, 10(5), July 2002.

[THB+06] Dave Thomas, David Hansson, Leon Breedt, Mike Clark, James Duncan Davidson, Justin Gehtland, and Andreas Schwarz. *Agile Web Development with Rails, 2nd Edition*. Pragmatic Bookshelf, Flower Mound, TX, 2006.

[TLdR08] Manos Tsagkias, Martha Larson, and Maarten de Rijke. Term clouds as surrogates for user generated speech. In *SIGIR '08: Proc. of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 773–774, 2008.

[TM07] D. Tzimeas and E. Mangina. Senegal: a ga system for generating rhythms of western africa. In -, volume vol.1, pages 25 – 30, Kowloon, China, 2007.

[Toi97] P. Toirvainen. Optimizing self-organizing timbre maps: two approaches. *Music, Gestalt, and Computing. Studies in Cognitive and Systematic Musicology*, pages 337 – 50, 1997.

[Tra08] John Travis. Science and commerce: Science by the masses. *Science*, 319(5871):1750–1752, 2008.

[Tre82] Leo Treitler. The early history of music writing in the west. *Journal of the American Musicological Society*, 35, 1982.

[Tza04] George Tzanetakis. Song specific bootstrapping of singing voice structure. In *Proc. Int. Conf. on Multimedia and Exposition ICME*, TaiPei, Taiwan, 2004. IEEE.

[Tza08] G. Tzanetakis. *Marsyas-0.2: A case study in implementing music information retrieval systems*, chapter 2, pages 31–49. Intelligent Music Information Systems: Tools and Methodologies. Information Science Reference, 2008. Shen, Shepherd, Cui, Liu (eds).

[UN01] T. Unemi and E. Nakada. A tool for composing short music pieces by means of breeding. In *2001 IEEE Int. Conf. on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace*, volume vol.5, pages 3458 – 63, Tucson, AZ, USA, 2001.

[vAD08] Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.

[VDS99] J.K.B. Ford V.B. Deecke and P. Spong. Quantifying complex patterns of bioacoustic variation: use of a neural network to compare killer whale (orcinus orca) dialects. *Journal of the Acoustical Society of America*, 105:2499–2507, 1999.

[VSI08] L. Vlaming, J. Smit, and T. Isenberg. Presenting using two-handed interaction in open space. *Horizontal Interactive Human Computer Systems, 2008. TABLETOP 2008. 3rd IEEE International Workshop on*, pages 29–32, Oct. 2008.

[VW08] Fernanda B. Viégas and Martin Wattenberg. Timelines : Tag clouds and the case for vernacular visualization. *interactions*, 15(4):49–52, 2008.

[WFM03] M. Wright, A. Freed, and A. Momeni. Opensound control: State of the art 2003. In *Int. Conf. on New Interfaces for Musical Expression (NIME'03)*, Montreal, Canada, 2003.

[WL08] Zhixun Wang and James Louey. Economical solution for an easy to use interactive whiteboard. *Frontier of Computer Science and Technology, 2008. FCST '08. Japan-China Joint Workshop on*, pages 197–203, Dec. 2008.

[WYC08] Elaine L. Wong, Wilson Y. F. Yuen, and Clifford S. T. Choy. Designing wii controller: a powerful musical instrument in an interactive music performance system. In *MoMM '08: Proc. of the 6th Int. Conf. on Advances in Mobile Computing and Multimedia*, pages 82–87, New York, NY, USA, 2008. ACM.

[YLC04] Min-Joon Yoo, In-Kwon Lee, and Jung-Ju Choi. Background music generation using music texture synthesis. In *Entertainment Computing - ICEC 2004. Third Int. Conf.. Proc. (Lecture Notes in Comput. Sci. Vol.3166)*, pages 565 – 70, Eindhoven, Netherlands, 2004.

[Zar29] Boris Zarnik. Zivot i rad ivana regena. *Priroda*, pages 1–7, 1929.

[Zim00] Heidi Zimmermann. *Untersuchungen zur Musikauffassung des rabbinischen Judentums*. Peter Lang, Bern, 2000.