Multimed Tools Appl (2010) 48:207–224
DOI 10.1007/s11042-009-0357-x

Computer-assisted cantillation and chant research using content-aware web visualization tools

Steven R. Ness · Dániel Péter Biró · George Tzanetakis

Published online: 2 October 2009

Abstract Chant and cantillation research is particularly interesting as it explores the transition from oral to written transmission of music. The goal of this work to create web-based computational tools that can assist the study of how diverse recitation traditions, having their origin in primarily non-notated melodies, later became codified. One of the authors is a musicologist and music theorist who has guided the system design and development by providing manual annotations and participating in the design process. We describe novel content-based visualization and analysis algorithms that can be used for problem-seeking exploration of audio recordings of chant and recitations.

Keywords Multimedia annotation · Multimedia analysis · Audio feature extraction · Semi-automatic annotation · Machine learning

1 Introduction

In recent years there has been increasing research activity in the areas of multimedia learning and information retrieval. Most of it has been in traditional specific domains, such as sports video [7], news video [8] and natural images. There is broad interest in these domains and in most cases there are clearly defined objectives such as identifying highlights in sports videos, explosions in news video or sunsets in natural

S. R. Ness (✉) · G. Tzanetakis
Department of Computer Science, University of Victoria, Victoria, British Columbia, Canada
e-mail: sness@sness.net

G. Tzanetakis
e-mail: gtzan@cs.uvic.ca

D. P. Biró
School of Music, University of Victoria, Victoria, British Columbia, Canada
e-mail: dpbiro@uvic.ca

images. Our focus in this paper is a niche domain that shares the challenge of effectively accessing large amounts of data but has specific characteristics that preclude the use of existing multimedia tools.

Although there is much related work little of it is directly relevant to our particular application domain. There is a lot of work on melodic similarity but most of it is based on symbolic representations such as music notation or MIDI files [6] and therefore not applicable in our case. Even in the cases where audio recordings are used [4] there are no interactive visualizations of the results making their use by expert musicologists limited. An earlier version of the web-based system described in this paper that did not have support for content-based similarity retrieval of pitch contours was been described in Ness et. al [14].

The goal of this project is to develop tools to study chants from various traditions around the world including Hungarian *siratok* (laments) [10], Torah cantillation [18], tenth century St. Gallen plainchant [9, 12], and Koran recitation [13]. These diverse traditions share the common theme of having an origin in primarily non-notated melodies which then later became codified. The evolution and spread of differences in the oral traditions of these different chants are a current topic of research in Ethnomusicology [17].

It has proved difficult to study these changes using traditional methods and it was decided that a combined approach, using field recordings marked up by experts, mathematical models for analyzing the fundamental frequency content of the audio, automatic alignment for pitch contour similarity and a flexible graphic user interface, would help figure out what questions needed to be asked. Unlike traditional multimedia data where most users, including the developers of tools, can be used as annotators, in our case any type of annotation requires highly trained experts. In addition it is a problem seeking domain where there are no clearly defined objectives and formulating problems is as important as solving them. We believe that despite these challenges it is possible to develop semi-automatic tools that can assist researchers in formulating questions regarding how symbols are used in chant and recitation.

The number and variety of chants that are available to be studied is vast, in our current work, we concentrate on four different chant traditions. Our collaborators have collected a large database of recordings and are actively engaged in field research where they are collecting more recordings. In addition, the analytical approach we describe in this paper could also be applied to many other chant traditions from around the world, including chants of rapidly disappearing indigenous cultures. The recordings are currently mostly tape based, however some newer recordings are directly recorded in digital form. For the experiments described in Section 3 we used 19 different recordings with a total duration of 26.8 min.

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 [14]. 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.

Classical tools to analyze chant recordings primarily consist of listening to recordings on tape and on simple digital audio players and annotating them by hand. The primary method of investigation of chant traditions involves listening to chants, and qualitatively grouping similar chants together. There are existing computational tools that could be used for this purposes. For example, conceivably ethnomusicologists could use a combination of the Praat software [1] for speech analysis and custom written Matlab scripts to do this analysis. This would require programming expertise which most of them do not have.

In the current work, we developed a computer based tool to help with the study of chant traditions using a process of participatory design, where the ethnomusicology domain expert interacts regularly and frequently with the developers of the software tools. This is a niche application, designed to be used by experts in ethnomusicology, which makes traditional human computer interaction user studies and evaluations challenging because of the rarity of users with the required expertise in ethnomusicology. To compensate for this, our primary mode of evaluation of the interface was through feedback with the domain expert through an iterative process of design and development. In addition we provide experimental results that show the effectiveness of the analysis tools used. These include a comparison with traditional methods of pitch contour representation (interval-based contour abstraction) used in the different but related field of query-by-humming (QBH) [3, 5]. More details of this process can be found in Section 3.9.

In this paper, we examine of the effects of using different quantization levels on the mean average precision recall of different trope signs. We then compare the results when we use a data-driven approach of examining the most common notes that the singers use versus notes derived from the western equal tempered scale. This data-driven approach is the one used in our interactive chant research software, Cantillion.

2 Chant research

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 [11] notational chant traditions: Hungarian siratok (laments),[1] Torah cantillation [19],[2] tenth century St. Gallen plainchant [15],[3]

[1] Archived Examples from Hungarian Academy of Science (1968–1973).

[2] Archived Examples from Hungary and Morocco from the Feher Music Center at the Bet Hatfatsut, Tel Aviv, Israel.

[3] Godehard Joppich and Singphoniker: Gregorian Chant from St. Gallen (Gorgmarienhtte: CPO 999267-2, 1994).

and Koran recitation.[4] This work falls under the more general area of Computational Ethnomusicology [17].

Although Dr. Biro has been studying these recordings for some time and has considerable computer expertise for a professor in music, the design and development of our tools has been challenging. This is partly due to difficulties in communication and terminology as well as the fact that the work is exploratory in nature and there are no easily defined objectives. The tool has been developed through extensive interactions with Dr. Biro with frequent frustration on both sides. At the same time, a wonderful thing about expert users like Dr. Biro is that they are willing to spend considerable time preparing and annotating data as well as testing the system and user interface which is not the case in more traditional broad application domains.

We used a number of different datasets in the current work, all of which were recorded onto tape and then subsequently digitized at a sample rate of 44100 Hz and with 16 bit precision. The recordings used in this paper included five recordings of Torah chant from two different reciters, four recordings of Koran recitation from three singers, two recordings of Gregorian chant from a choir, and one recording of a Hungarian lament.

3 Melodic contour analysis

Our tool takes in a (digitized) monophonic (audio in which only one voice is present) or heterophonic recording (audio in which two or more voices elaborate the same melody simultaneously) 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

3.1 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.

[4]Examples from Indonesia and Egypt: in Approaching the Koran (Ashland: White Cloud, 1999).

3.2 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 difficult and time consuming to calculate manually. Figure 1 shows an example transition matrix. For a given trope sign (a row) it shows how many total times does it appear 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.

The symbols above are the cantillation symbols from Torah chant, these symbols are well established and form a stable and describable grammar. Different traditions, for example Koran recitation, would have different sets of symbols, which may or may not be of such a proscribed vocabulary, but generally have definite regularities in their usage.

The above Markov Model pre-processing analysis shows that there are certain well defined rules about which signs follow other signs, which is of general interest

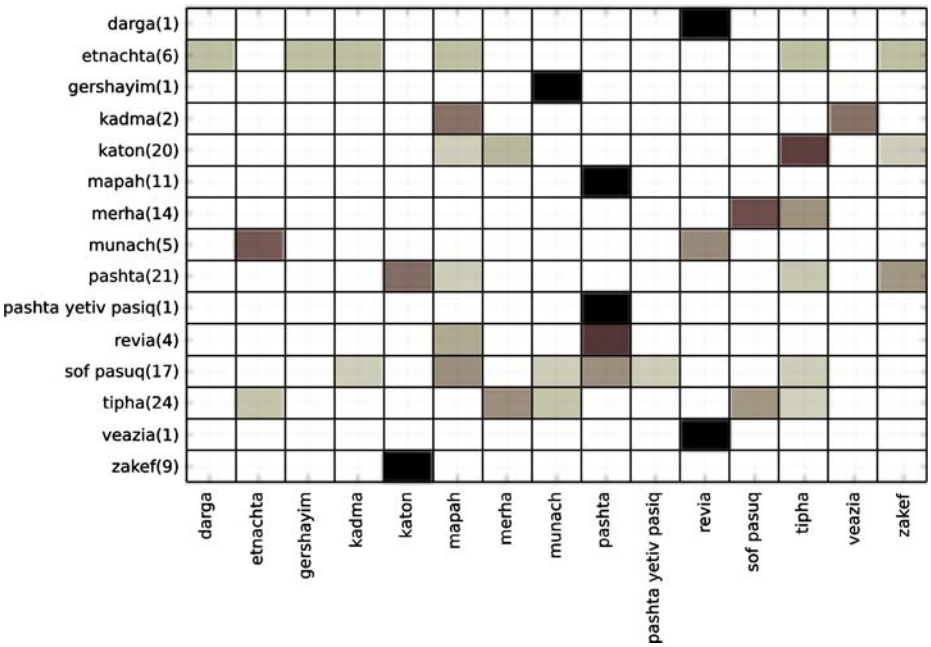

Fig. 1 Syntagmatic analysis with a first-order Markov model of the sequence of Torah trope signs for the text Shir Ha Shirim ("Song of Songs")

to ethnomusicologists. This analysis is meant to be a qualitative measure of the regularity of the grammar of the chant, and can give clues to ethnomusicologists about which sign relationships will be most profitable to investigate.

3.3 F0 estimation

After the segments have been hand-labeled and 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 [2] 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 [2] uses an algorithm that is related to autocorrelation, and using a cosine as the kernel, performs an integral transform of the spectrum. 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.

The SWIPEP algorithm is a powerful algorithm for estimating the fundamental frequency of audio signals, and in a comparison against twelve other leading F0 estimation algorithms it outperformed all of them [2]. It performs especially well when compared against the traditional autocorrelation approach in that it makes less errors, including less octave errors, a common problem encountered when using traditional autocorrelation.

3.4 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 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.

3.5 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)

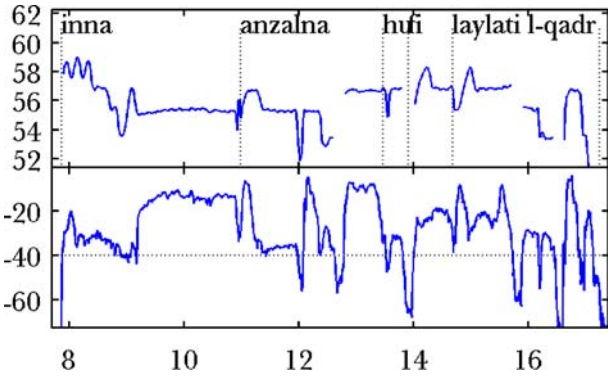

Fig. 2 F0 contour: the *top half* of the graph shows the F0 contour as estimated by the SWIPEP algorithm. The *x-axis* shows time in seconds from the start of the audio file, and the *y-axis* shows the pitch of the contour in MIDI note numbers. The *bottom half* of the graph shows the signal energy of the audio, with the *x-axis* describes time in seconds, and the *y-axis* shows the energy of the audio in decibels

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 [11]. We demand a pitch resolution of one cent,[5] so we cannot use a simple histogram. Instead we use a statistical technique known as non-parametric kernel density estimation, with a Gaussian kernel.[6] 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. 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 shows this method's density estimate given the F0 curve from Fig. 2.

In quantizing the pitch, the size of the excerpt chosen can influence both the number of peaks and the location of these peaks. In the current work, we have chosen to use the entire file as the dataset for doing pitch quantization. In the user interface presented below, we allow the user to choose subsets of signs which can then be viewed at different levels of quantization granularity. This process is necessary because of the two step nature of our analysis process, where first the audio file is analyzed, and then this analysis is presented to the user, who can then perform further analysis on the audio. We are in the process of developing a new interface that will overcome this drawback, and will allow the user to directly interact with the first analysis procedure.

3.6 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 (currently 80 cents by default) 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 our density estimate; we have obtained good results

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

[6]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.

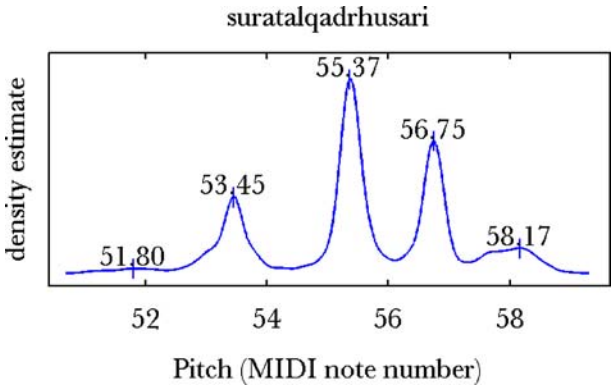

Fig. 3 Recording-specific scale derivation

with a standard deviation of 33 cents. Note that this method has no knowledge of octaves.

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 a temporal unfolding of the given melodic context. 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.

3.7 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 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 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.

To our knowledge, our use of the Histogram-Based Contour abstraction is novel, the classical approach is that of an interval-based contour abstraction of the pitch

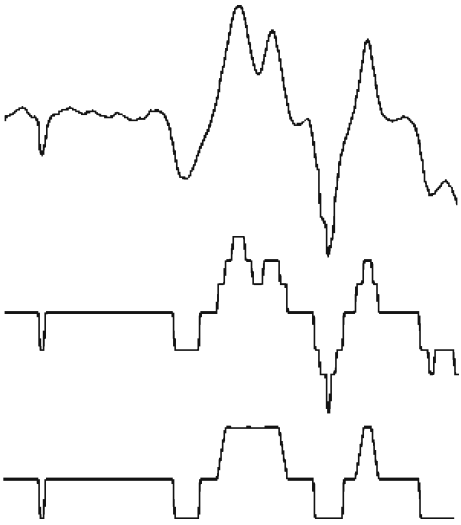

Fig. 4 Melodic contours at different levels of abstraction (*top*: original, *middle*: quantized, *bottom*: simplified using 3 most prominent scale degrees

contour. This interval-based method is a commonly used method in Query-by-Humming experiments (QBH) [3]. The interval-based contour abstraction representation simplifies a contour by describing the relation of each tone to the next in terms of how many scale degrees exist between one note and the next. For example, using the note sequence AED, there would be a step of +4 between the A and E, and a step of −1 between the E and D. This representation can then be simplified by quantizing it to a smaller number of step. The simplest interval abstraction has three levels, "goes up" (+1), "goes down"(−1), and "remains the same" (0). One can then successively subdivide the upper and lower ranges, giving 5 levels, 9 levels, 11 levels, and so on. The result of interval-based quantization is a string of the quantized interval differences between one pitch value and then next. The resulting strings of values for each trope are then compared to one another using the technique of Dynamic Time Warping, as described in the next section.

3.8 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. 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 contours 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.

3.9 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. Figure 5a, b, c and d show the F0 contour of the sections of the audio file from a Torah recording from Hungary. Figure 5a shows a pashta sign, Fig. 5b shows another pashta sign from further along in the audio file. Figure 5c shows a sof pasuq gesture and Fig. 5d shows the first pashta gesture, but with the sample stretched by a factor of two.

The Fig. 6a, b, c and d 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 Fig. 6a 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 alignment of one file to the other. Figure 6d shows a similar behavior, except that the slope of the line is shallower. Figure 6b shows the comparison of one pashta gesture to another. This path had a DTW cost of 23.8442. Figure 6c 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 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. Each recordings contains approximately 130 realizations of each sign with a total of 12

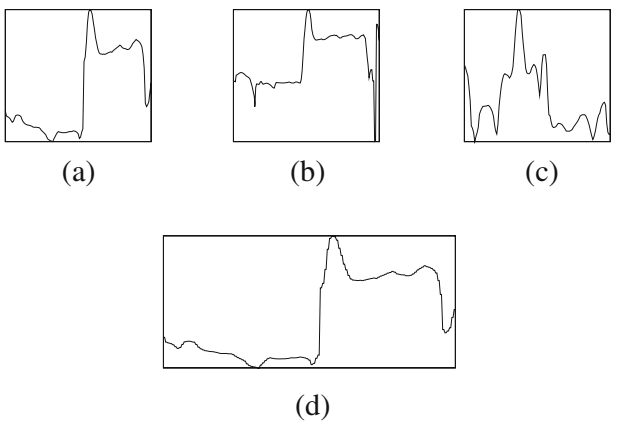

Fig. 5 F0 contours of 4 different gestures from a Torah recitation recorded in Hungary. The first two show different versions of the pashta gesture (11 pashta and 42 pashta) and the third shows the gesture for sof pasuq (18 sof pasuq). The last is a version of the first pashta gesture (11 pashta) with each audio sample doubled, which effectively stretches the contour by a factor of two. **a** F0 Countour of 11 Pashta. **b** F0 Contour of 42 Pashta. **c** F0 Contour of 18 Sof Pasuq. **d** F0 Contour of 11 Pashta Doubled

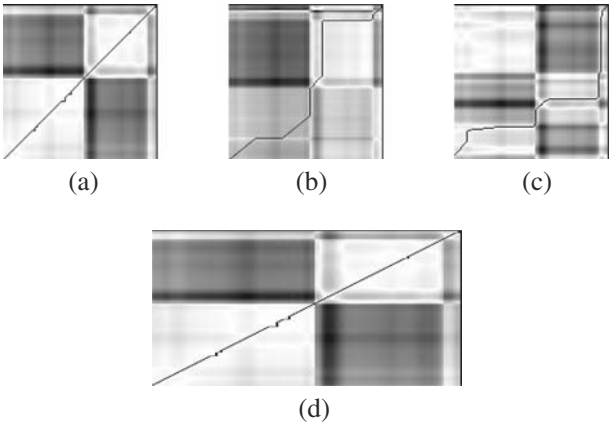

Fig. 6 Shown above are Similarity Matrices of the above four gestures compared with the first pashta gesture. Superimposed on the figures is the Dynamic Time Warping curve showing the optimally matching path between the two songs. **a** DTW of 11 Pashta vs 11 Pashta. **b** DTW of 11 Pashta vs 42 Pashta. **c** DTW of 11 Pashta vs 18 Sof Pasuq. **d** DTW of 11 Pashta vs 11 Pashta Doubled

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 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 cases we are more interested in the individual difference in precision among different signs. 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 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

Table 1 Average precision for different signs

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

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 calculating 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 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 Fig. 7.

From this graph and Table 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. The term continuous contour refers to when the original contour of the song is quantized to the closest equivalent equal-tempered, or MIDI, note. 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, and this can be easily understood by realizing that 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 Fig. 8. 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

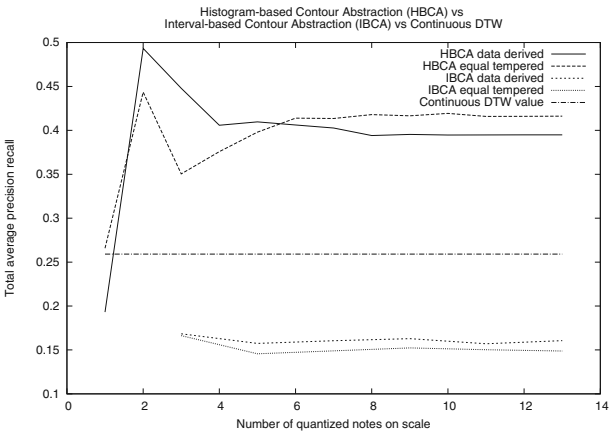

Fig. 7 Mean average precision recall when quantizing the notes before DTW analysis. Shown are the results for quantizing to a song specific scale (Histogram derived scale) versus an equal tempered scale (MIDI notes) for both Histogram-based Contour Abstraction (HBCA) and Interval-based Contour Abstraction (IBCA) approaches. In addition, the mean average precision recall in the continuous case is also shown

Table 2 Table of mean average precision values when quantizing the notes before DTW analysis

Number of bins	HBCA data driven	HBCA equal temperment	IBCA data driven	IBCA equal temperment
1	0.1931	0.2658		
2	0.4932	0.4435		
3	0.4479	0.3504	0.1684	0.1663
4	0.4057	0.3757		
5	0.4097	0.3979	0.1575	0.1456
6	0.4061	0.4138		
7	0.4026	0.4135	0.1605	0.1490
8	0.3941	0.4179		
9	0.3953	0.4165	0.1629	0.1522
10	0.3947	0.4193		
11	0.3948	0.4158	0.1571	0.1503
12	0.3948	0.4159		
13	0.3948	0.4161	0.1606	0.1488

Shown are the calculated values for the Data-driven and Equal-temperment approaches using both the Histogram-based Contour Abstraction (HBCA) and Interval-based Contour Abstraction (IBCA) approaches

"sof pasuq" contours have quite a different shape. This visualization shows the utility of our approach.

In order to compare these results with those of the classical method of Interval-based Contour Abstraction (IBCA), we first quantized the continuous contour using either the data-derived scale mentioned previously or to an equal tempered scale. We then generated the interval-based contour abstraction for each trope using first 3, then 5, 7, 9, 11 and 13 different interval quantization levels. These strings of interval differences for each trope were then compared against each other using the same Dynamic Time Warping (DTW) technique used above. Average precision-recall values were then generated for each of the interval quantization levels. The results of this are shown in Table 2 and are presented graphically in Fig. 7.

From these results, one can immediately see that our proposed method of Histogram-based Contour Abstraction (HBCA) outperforms the traditional method of Interval-based Contour Abstraction (IBCA) by a large margin. In addition, one can see a small improvement in the IBCA approach when using a scale derived from the data, as opposed to using an equal-tempered scale. However, it must be stated that this is a very small difference, and may not be statistically significant. Further investigation in this area with larger sample sizes is required.

3.10 Cantillion interface

We have developed a browsing interface that allows researchers to organize and analyze chant segments in a variety of ways (http://cantillation.sness.net). Each recording is manually segmented 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

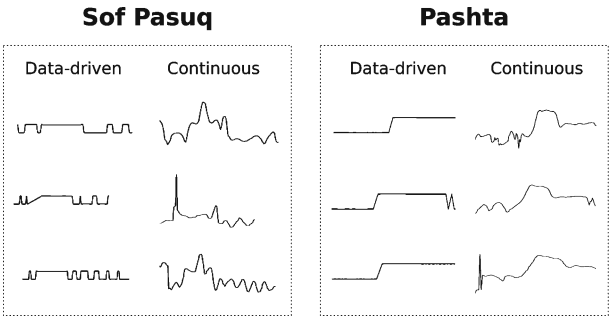

Fig. 8 Comparison of contour quantized to the two most prevalent scale degrees in a data-driven approach to the original continuous contour. Shown are three examples of the signs "sof pasuq" and "pashta"

both manually and automatically. The audio analysis (pitch extraction and dynamic time warping) are performed using the Marsyas audio processing framework[7] [16].

The interface (Fig. 7) 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. 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 way.

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 outliers 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. Figure 7 shows a screenshot of the browsing interface.

In the current work we implemented a mode that allows the researcher to sort the samples based on the Dynamic Time Warping cost from one sample to the other. The interface allows the user to select an arbitrary gesture from the interface, and then perform a sorting of all other gestures to it. In the example shown in Fig. 9 the user has chosen a "revia", and has sorted all the other gestures based on their DTW-based alignment distance from this first revia. One can see that the gesture closest to this revia is another revia gesture from a different section of the audio file.

[7] http://marsyas.sourceforge.net

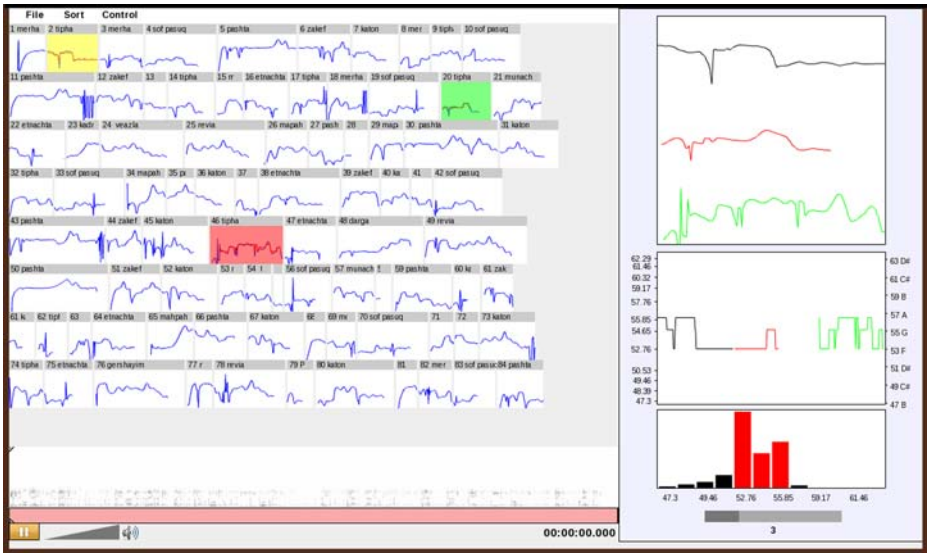

Fig. 9 Web-based *Flash* interface to allow users to listen to audio, and to enable interactive querying of gesture contour diagrams

We are currently developing an addition to the Cantillion interface to allow us to visualize subsets of signs at different quantization levels, and to compare these to the original continuous contour. This interface uses a checkbox list to allow the user to select different types of signs, and then displays these contours in the main interface pane. The user can select multiple quantization levels and can compare them for many signs at once, which allows the user to quickly perform an analysis similar to the full pair-wise comparison described above, but interactively, and therefore using the knowledge and skills of ethnomusicologists.

4 Summary and discussion

The identity of chant formulae in oral/aural chant traditions is to a large extent determined by gesture/contour rather than by discrete pitches. 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 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 we only have few experts users (one in our case) and the only feedback we can receive is through them. By including them in the design we having been able to create a system that our expert finds useful and is willing to spend significant time interacting with it.

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.

Acknowledgements We would like to thank Matt Wright for initial work on this project and Emiru Tsunoo for the Marsyas implementation of dynamic time warping and similarity matrix computation used in the paper. We would also like to thank the National Sciences and Engineering Research Council (NSERC) and Social Sciences and Humanities Research Council (SSHRC) of Canada for their financial support.

References

1. Boersma P (2001) Praat, a system for doing phonetics by computer. Glot Int 5(9/10):341–345
2. Camacho A (2007) A sawtooth waveform inspired pitch estimator for speech and music. PhD thesis, University of Florida
3. Dannenberg RB, Birmingham WP, Pardo B, Hu N, Meek C, Tzanetakis G (2007) A comparative evaluation of search techniques for query-by-humming using the musart testbed. J Am Soc Inf Sci Technol 58(5):687–701
4. Duggan B, O' Shea B, Cunningham P (2008) A system for automatically annotating traditional irish music field recordings. In: Int workshop on content-based multimedia indexing (CBMI). IEEE, Piscataway
5. Ghias A, Logan J, Chamberlin D, Smith BC (1995) Query by humming: musical information retrieval in an audio database. In: MULTIMEDIA '95: proceedings of the third ACM international conference on multimedia. ACM, New York, pp 231–236
6. Hanna P, Ferraro P (2007) Polyphonic music retrieval by local edition of quotiented sequences. In: Int workshop on content-based multimedia indexing (CBMI). IEEE, Piscataway
7. Hauptman A, Witbrock M (1997) Informedia: news-on-demand multimedia information acquisition and retrieval. MIT, Cambridge
8. Hauptman A et al (2003) Informedia at trec 2003: analyzing and searching broadcast news video. In: Proc of (VIDEO) TREC 2003, Gaithersburg, MD
9. Karp T (1998) Aspects of orality and formularity in Gregorian chant. Northwestern University Press, Evanston
10. Kodaly Z (1960) Folk music of Hungary. Corvina, Budapest
11. Krumhansl CL (1990) Cognitive foundations of musical pitch. Oxford University Press, Oxford
12. Levy K (1998) Gregorian chant and the Carolingians. Princeton University Press, Princeton
13. Nelson K (1985) The art of reciting the Koran. University of Texas Press, Austin

14. Ness S, Wright M, Martins L, Tzanetakis G (2008) Chants and orcas: semi-automatic tools for audio annotation and analysis in niche domains. In: Proc ACM multimedia, Vancouver, Canada
15. Treitler L (1982) The early history of music writing in the west. J Am Musicol Soc 35
16. Tzanetakis G (2008) Marsyas-0.2: a case study in implementing music information retrieval systems, chapter 2. In: Shen S, Shepherd J, Cui B, Liu L (eds) Intelligent music information systems: tools and methodologies. Information science reference, pp 31–49
17. Tzanetakis G, Schloss KAW, Wright M (2007) Computational ethnomusicology. J Interdiscip Music Studies, 1(2), 2007
18. Wigoder G et al (1989) Masora, the encyclopedia of Judaism. MacMillan, New York
19. Zimmermann H (2000) Untersuchungen zur Musikauffassung des rabbinischen Judentums. Peter Lang, Bern

Steven R. Ness is a M.Sc. student in Computer Science at the University of Victoria. He is working with Dr. George Tzanetakis in the field of Music Information Retrieval, and for his M.Sc. is focussing on the areas of Self-Organizing Maps, Music Feature Extraction and User Interface Design to analyze and visualize the vast amounts of information present in audio, music and bioacoustic signals.

Dániel Péter Biró is Assistant Professor of Composition and Music Theory at the School of Music at the University of Victoria. In July 2004 Dániel Péter Biró completed his Ph.D. in Composition at Princeton University. He first started his musical studies at the Bartók Conservatory in Budapest, Hungary. From 1991-1992 he was a Fulbright scholar in Frankfurt, Germany where he studied at the Hochschule für Musik in Frankfurt. He later studied in Bern and Vienna. In 1995, he did folk music research at the Academy of Science in Budapest. His works have been performed at the Alte Oper-Frankfurt, at the Konzerthaus in Vienna, at the Bartók Festival in Szombathely, Hungary and

have been broadcast on Swiss, Austrian, German, and on Italian public radio. He received an opera commission from the Neue Horizonte Bern/Schlachthaus Theater in Bern, Switzerland in 1998. In 1999, he was awarded the Hungarian Government's Kodály Award for Hungarian Composers. In 2000, he received grants from Center for Near Eastern Studies and the Association of Princeton Graduate Alumni for purposes of Hebrew study and dissertation research at Haifa University, Israel. In 2001, his piece The Crossing (Daf), that was based on a text by Franz Kafka, was performed as a commissioned piece of the Stuttgart Opera. In 2002, he was a fellow at the Atlantic Center of the Arts. In 2003 he received a dissertation research grant from the Princeton University Program in Judaic Studies. In the summer of 2003, he was awarded a Summer Research Grant from the Princeton Council on Regional Studies, enabling him to take part in the Sommerakademie at the Schloss Solitude in Stuttgart, Germany. There he worked with the Ensemble SurPlus, which performed the first part of his composition Mishpatim (Laws). In 2004 he presented his work at the Internationale Ferienkurse für Neue Musik in Darmstadt, Germany. In 2005 he was a fellow at the Mannes Institute for Advanced Studies in Music Theory in New York. Mishpatim (Laws) was also performed by the Aventa Ensemble at the University of Victoria in February 2006. In August 2006 the second version of Mishpatim (Laws) was performed by the ensemble recherche at the Darmstadt International Summer Courses for New Music: The composition was commisioned by the Internationale Musikinstitut Darmstadt. In Darmstadt he lectured on his music. In 2006 Dániel Péter Biró was a faculty fellow at the University of Victoria Centre for Studies in Religion and Society: there he researched early Jewish and Christian chant traditions. In 2007 his composition Simanim (Signs/Traces) was performed by members of the Frankfurt Radio Symphony Orchestra: this composition was commissioned by the German Radio (HR). Dániel Péter Biró was recently commisioned by Vancouver New Music to write a piece for solo viola, three voices, seven instruments and electronics. This new composition, supported through a grant of the British Columbia Arts Council and the Canada Council for the Arts will be premiered in their 2008/2009 season.

George Tzanetakis is an assistant Professor of Computer Science (also cross-listed in Music and Electrical and Computer Engineering) at the University of Victoria, Canada. He received his PhD degree in Computer Science from Princeton University in May 2002 was a PostDoctoral Fellow at Carnegie Mellon University working on query-by-humming systems with Prof. Dannenberg and on video retrieval with the Informedia group. His research deals with all stages of audio content analysis such as feature extraction, segmentation, classification with specific focus on Music Information Retrieval (MIR). His pioneering work on musical genre classification is frequently cited and received an IEEE Signal Processing Society Young Author Award in 2004. He is the principal designer and developer of the Marsyas open source audio processing framework which has been used for MIR projects both in academia and industry. He has presented tutorials on MIR and audio feature extraction at several international conferences. He is also an active musician and has studied saxophone performance, music theory and composition. More information can be found at http://www.cs.uvic.ca/ gtzan.