BROWSING MUSIC AND SOUND USING GESTURES IN A SELF-ORGANIZED 3D SPACE

Gabrielle Odowichuk
University of Victoria
Department of Electrical and Computer Engineering

George Tzanetakis
University of Victoria
Department of Computer Science

ABSTRACT

As digital music and sound collections increase in size there has been a lot of work in developing novel interfaces for browsing them. Many of these interfaces rely on automatic content analysis techniques to create representations that reflect similarities between the music pieces or sounds in the collection. Representations in 3D have the potential to convey more information but can be difficult to navigate using the traditional ways of providing input to a computer such as a keyboard and mouse. Utilizing sensors capable of sensing motion in 3-dimensions, we propose a new system for browsing music in augmented reality. Our system places audio files in a virtual cube. The placement of the files into the cube is realized through the use of audio feature extraction and self-organizing maps (SOMs). The system is controlled using gestures, and sound spatialization is utilized to provide auditory cues about the topography of the music or sound collection.

1. INTRODUCTION

Advances in technology have drastically changed how we interact with music. The increasing capabilities of personal computaers have allowed listeners access to digital music collections of significant size. As the number of available songs increases, searching and browsing through this music becomes difficult. The conventional hierarchy of "Artist-Album-Track" and the spreadsheet interface of music software such as iTunes ared still the dominant ways of organizing and navigating digital music collections. While this method is effective for finding a specific song when one knows exactly what they are looking for, it does not allow for effective browsing through music collections when there is no specific target song. To address this issue, browsing interfaces that are based on organizing music tracks spatially based on their automatically computed similarity have been proposed. Content-based browsing has some advantages over traditional systems, many of which stem from the fact that users can browse music aurally, and no longer require a pictorial or textual representation. By removing the need for a keyword representation, we can possibly access music that has no associated text, or text available only in a different language. This type of audio browsing can also be useful for music creators or videogame audio designers who need to sort through large collections of sound clips or sound effects. Accessing music information aurally makes sense intuitively, and even allows people with vision or motion disabilities improved access to the world of music [18]. We describe a novel interface for browsing music and sound collections based on automatically computed similarity, spatially arranging the audio files in 3D using self-organizing maps (SOMs), and browsing the sonified space using 3D gestural controllers.

2. RELATED WORK

Novel interfaces for browsing music began to appear about ten years ago with SOMs being one of the first algorithms to be used for music clustering and visualization [4]. The early development of applications demonstrating these concepts such as the Sonic Browser [9], Marsyas 3D [17] and Musescape [15] was fueled by advances in the field of Music Information Retrieval (MIR). Each system uses direct sonification rather than button triggered playback as a means of music browsing to create a continuous stream of sound while navigating the music space. In Pampalk, Dixon and Widmer [12] and Knees et al [5], a visualization of the organized music collection is proposed in which the clustered songs are represented as islands, where the height of each island is relative to the number of songs in each cluster, and the terrain itself is based on a 2D SOM. In each of these applications, navigation is achieved using a mouse or joystick. In Ness et al. [11], the authors explored the use of various controllers for interfacing with self-organized music collections. These interfaces include multi-touch smartphones, motion trackers like the wiimote, and web-based applications. While advances in self-organized browsing progressed, the use of augmented reality in musical applications was being developed [13]. Often, augmented reality (AR) is understood be related to display technologies. However, AR can be applied to any senses, including hearing. In Azuma et al. [1], a mixed-reality continuum is presented, with Augmented reality defined as virtual objects added to a real space. Another good example of early combinations of self-organized music collections and augmented virtual spaces is the "Search Inside the Music" program [7]. This application allows users to browse through a virtual 3D space of songs and also showed the songs on each album visualized with the cover art. The key contribution of this work is the utilization of gestural 3D control for interacting with a 3D self-organized map of music.

3. ORGANIZING MUSIC IN A 3D SPACE

One of the main goals of Music Information Retrival is to to approximately model the concept of "similarity" in music. Similarity can be determined by using manually assigned metadata, however MIR often also focuses on extracting features directly from the audio signal. A variety of methods have been proposed in self-organized music browsers to project high-dimensional feature data, such as Principle Component Analysis [3]. Although this system could be implemented with other reduction methods, the most common approach to organizing music collections is that of the Self Organizing Map [6]. In this case, a set of features is extracted from an audio file, producing a single high-dimensional feature vector representing each song. The feature vector corresponding to a piece of music or a sound is then then mapped to a corresponding set of coordinates in a discrete grid. Feature vectors from similar audio files will be mapped either to the same grid location or neighbouring ones. The resulting map reflects both an organization of the data into clusters as well as a mapping that preserves the topology of the original feature space.

The goal of feature extraction is to produce a vector of numbers known as as features that represent a piece of audio. By choosing how the vectors are computed, we are able to come up with numbers that are similar when they correspond to perceptually similar sounds or music tracks. As described in [16], we extract features such as Flux, Rolloff, MFCCs (Mel-Frequency Cepstral Coefficients), pitch histograms and rhythm-based features. These audio features are extracted for very short periods of audio (usually under 25ms). An entire song would therefore have an array of numbers for each feature, depicting how these features change over time. To model large collections of songs, this sequence of feature vectors representing each song needs to be summarized into a single feature vector characterizing the music at the song level. To shorten the length of our feature vectors and simplify the calculations each sequence of a particular feature is summarized down to two single values: the mean and standard deviation. That way both the central tendency of the feature and the deviation from it are modelled. Finally the features are normalized to have values between 0 and 1 across the dataset.

$$V_k = [v_0, v_1, ..., v_N] \tag{1}$$

The resulting feature vector V is calculated for each audio file in our collection, and is given in Equation 1 where k is the song index, n is the number of features, and v_n is a normalized feature.

Most of the previous work in the area of self-organized music browsing involves SOMs that lie on a 2D grid. This has a nice corespondance with the majority of human-computer interfaces, like the mouse or touch screen tablets, which allow the user to navigate a 2D space. With the recent popularity of 3D Gestural controllers like the Kinect, exploring a 3D SOM is a natural extension of the current models. Luckily, the algorithm used to create 2D self-organizing maps is easily modified for any number of dimensions.

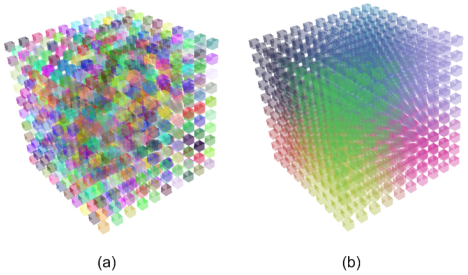

Figure 1. A 3D self organizing map before (a) and after (b) training with an 8-color dataset

The self-organizing map is a type of artificial neural network, meaning that it is inspired by interactions between biological neurons. Our neural network begins with a set of objects referred to as nodes. Each node has an associated weight vector, W, as shown in equation 2, and spatial placement $P = [x, y, z]$. Although the nodes in figure 1 have been spaced evenly within a cube, these nodes could hypothetically be placed in other, more arbitrary formations. Initially, the weights of each node are set randomly. As the organization process progresses, the weights of each node will begin to align more closely with their neighbours and also more closely with our song features. This process is depicted in Figure 1, where each node has weight vector visualized as a colour. Initially, the weights shown in this figure are random (a). As the SOM is trained with 8 distinct colours, the weights of each node become organized (b).

$$W_k = [w_0, w_1, ..., w_N] \tag{2}$$

The training process involves selecting a song to train the map with and determining which node represents that song the best. Similarity between songs and nodes is calculated as the euclidian distance between the song features and node weights, as shown in Equation 3. The smallest distance corresponds to best-matching node or best-matching unit (BMU). Now each node in the vicinity of the BMU is updated with a new set of weights, adjusted to become more like our BMU. Equation 4 described how this adjustment is made. V(t) is the feature vector, W(t) is the weights vector, and L(t) is a learning function, which decays over time and allows the organizing algorithm to settle.

$$d = \sqrt{\sum_{i=0}^{N} (V_i - W_i)^2} \tag{3}$$

$$W(t+1) = W(t) + L(t)(V(t) - W(t)) \tag{4}$$

By iteratively training our SOM, our resulting nodes reside in a space where nearby nodes have similar weight vectors. Each song is mapped to the most similar node, resulting in a set of songs residing in a space where nearby songs have similar feature vectors. In Figure 2, you can see that songs from similar genres will tend to be near one another. Note that the self-organizing map algorithm has

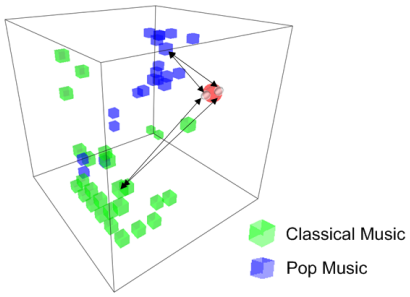

Figure 2. 3D SOM with two genres and user-controlled cursor

no knowledge of the genre labels and their spatial organization is an emergent property of the mapping and the underlying audio features.

4. NAVIGATING THROUGH THE COLLECTION

Once our songs have been organized into a virtual 3D space, user interaction becomes a significant consideration. Since the use of 3D sensors was one of the primary motivations behind this work, our focus has been on using sensors capable of reporting gesturally-produced position data for two or more points. How we go about using that captured motion is another point of discussion, and we present here only a few of the many possible ideas for expanding and refining user-interaction. Previous work has been done into user interaction with 2D visualizations for music browsing [8], and similar same concepts can be applied to the 3D scenario. We utilize two controllers: the radiodrum and the Kinect.

The radiodrum[2] is a controller with a long history known mostly in the computer music community. It is a music controller that rapidly tracks the position of the tips of two drumsticks in 3D space, and has been used to navigate and fade between two pieces of music in self-organizing maps [10]. The recent popularity of the XBox Kinect, an infrared 3D motion sensing device, has been a catalyst into further researching intuitive intelligent uses of gestural control. The Kinect provides a type of sensing in some ways similar to that of the radiodrum. It enables tracking of the hands and body of a user and does not require any hardware to be touched by the user. The radiodrum has much higher temporal precision and therefore feels more interactive. However it has the disadvantages that it is not widely available, is expensive and does not have as much spatial resolution as the Kinect. We believe that for this application scenario the Kinect is the better choice as it is mass produced, cheap, and enables very natural interaction. On the other hand the radiodrum has helped us understand better the timing requirements for such as an interface. Using our sensors for control data we want to sonify the organized sounds as we move our 3D cursors about. The simplest way to do this is to simply play back songs from whichever node is currently closest from one cursor, and only one song plays back at a time.

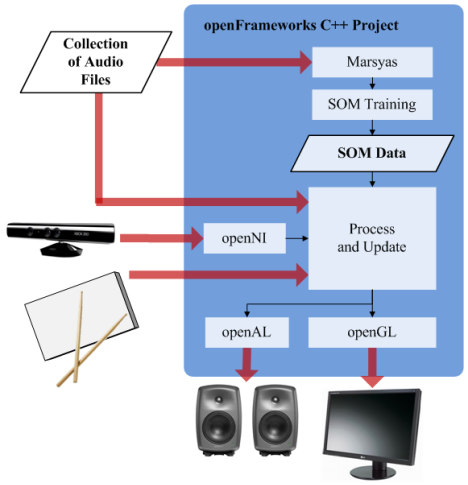

Figure 3. Implementation Diagram

The other hand could then be free to perform other types of control gestures. Another content-aware browser [14] presented a different method for playing back songs. In this case, the user can manipulate the centre point and radius of an encompassing circle, and any songs within the circle will play simultaneously. To modify this method for our purposes, the two cursors were made to act as the bounding points for a variable-size sphere. Nodes with positions within the user-controlled sphere are sonified, with a gain relative to their nearness to the center of the sphere. Once the cursor data from the sensors is mapped to playback in the auditory representation of our sound collection we need a richer gestural language to enhance the user control. For example, once music exploration is complete and the user has found a song they would like to listen to, they will want to select a song to listen to. Our simple way of implementing this functionality is to use timers, so that if we preview a song for longer than a set duration it will trigger song selection. Each node is sonified with a loudness based on its position relative to the cursor. By creating listening points that surround our cursor, we are able to perform multi-channel panning. As shown in figure 2, two smaller points are situated on either side of the user's current position, representing the the two listening points required for a stereo reproduction. This spatialization gives an aural sense of space and direction for navigation of our music collection.

5. IMPLEMENTATION

The hardware required for this music browser is simple: a controller, a computer, and a sound system. Figure 3 demonstrates the application design and interactions between the devices and software libraries. The C++ project uses a creative toolbox called openFrameworks, which allows easy access to other libraries like openGL to create visualizations, Marsyas for audio feature extraction, and openNI for sensor communications. The SOM data file is a small text file containing a list of songs with their accompanying metadata and SOM position.

6. FUTURE WORK

Future work will involve performing user evaluations that could help to answer questions about browsing music with this system. Three-dimensional SOMs have the possibility to represent richer topological spaces, reflecting more accurately the relationship between songs in our music collection. Furthermore, using 3D gesture-based controllers to navigate a 3D space would seem to offer advantages over using a joystick or other 2D controllers. However, without the proper evaluation provided by a user study any claims we can make are purely speculative. Further evaluation of this system is required, in which the time it takes to complete tasks of browsing for certain music will be measured. Quantitative comparisons between 3D and 2D SOMs can also be performed, where the distance between similar songs are compared for the same set of songss.

7. SUMMARY

The self-organized map has become a popular method for organizing songs based on similarity. This type of music browser not only reflects the way that how we interact with music is changing, it also reflects how our interaction with technology and computers is changing. By expanding previous work with self-organized music collections and adding a third dimension, it is possible to convey additional information and browse extra songs. Additionally, navigating this type of map is a good example of the advantages 3D gestural sensors like the radiodrum and the Kinect have in specific control contexts and the more natural interaction they enable.

8. REFERENCES

[1] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre, "Recent advances in augmented reality," *Computer Graphics and Applications, IEEE*, vol. 21, no. 6, pp. 34 –47, 2001.

[2] B. Boie, M. V. Mathews, and A. Schloss, "The Radio Drum as a Synthesizer Controller," in *Proceedings of the International Computer Music Conference*, 1989, pp. 42–45.

[3] P. Cano, M. Kaltenbrunner, F. Gouyon, and E. Batlle, "On the use of fastmap for audio retrieval and browsing," 2002.

[4] M. Frühwirth and A. Rauber, "Self-organizing maps for content-based music clustering," in *In Proceedings of the 12th Italien Workshop on Neural Nets (WIRN01), Vietri sul Mare*. Springer, 2001.

[5] P. Knees, M. Schedl, T. Pohle, and G. Widmer, "An innovative three-dimensional user interface for exploring music collections enriched with meta-information from the web," in *In MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*. ACM Press, 2006, pp. 17–24.

[6] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464 –1480, sep 1990.

[7] P. Lamere and D. Eck, "Using 3d visualizations to explore and discover music," in *ISMIR*, 2007.

[8] A. S. Lillie, "Musicbox: Navigating the space of your music," Master's thesis, Massachussets Institute of Technology, September 2008.

[9] D. O. Maidin and M.Fernstrom, "The best of two worlds: Retrieving and browsing," *Proceedings of the Conference on Digital Audio Effects*, 2000.

[10] J. Murdoch and G. Tzanetakis, "Interactive content-aware music browsing using the radio drum," *Multimedia and Expo, IEEE International Conference on*, vol. 0, pp. 937–940, 2006.

[11] S. Ness and G. Tzanetakis, "Audioscapes: Exploring surface iinterfaces for music exploration," 2009.

[12] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," 2003.

[13] I. Poupyrev, R. Berry, J. Kurumisawa, K. Nakao, M. Billinghurst, C. Airola, H. Kato, T. Yonezawa, and L. Baldwin, "Augmented groove: Collaborative jamming in augmented reality," in *SIGGRAPH Conference Abstracts and Applications*, vol. ACM: pp. 77, 2000.

[14] J. L. Sebastian Heise, Michael Hlatky, "Soundtorch: Quick browsing in large audio collections," in *Audio Engineering Society Convention 125*, 10 2008.

[15] G. Tzanetakis, "Musescape: An interactive content-aware music browser," in *In Proc. International Conference on Digital Audio Effects*, 2003.

[16] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 293 – 302, jul 2002.

[17] G. Tzanetakis, "Marsyas3d: a prototype audio browser-editor using a large scale immersive visual and audio display," in *In Proc. International Conference on Auditory Display*, 2001.

[18] G. Tzanetakis, M. S. Benning, S. R. Ness, D. Minifie, and N. Livingston, "Assistive music browsing using self-organizing maps," in *Proc. Int. Conference on PErvasive Technologies Related to Assistive Environments (PETRAE)*, 2009.