3D INTERACTIVE ENVIRONMENT FOR MUSIC COLLECTION NAVIGATION

Rebecca Stewart, Mark Levy, and Mark Sandler

Centre for Digital Music, Dept. of Electronic Engineering,
Queen Mary, University of London
London, UK
{rebecca.stewart|mark.levy|mark.sandler}@elec.qmul.ac.uk

ABSTRACT

Previous interfaces for large collections of music have used spatial audio to enhance the presentation of a visual interface or to add a mode of interaction. An interface using only audio information is presented here as a means to explore a large music collection in a two or three-dimensional space. By taking advantage of Ambisonics and binaural technology, the application presented here can scale to large collections, have flexible playback requirements, and can be optimized for slower computers. User evaluation reveals issues in creating an intuitive mapping between between user movements in physical space and virtual movement through the collection, but the novel presentation of the music collection has positive feedback and warrants further development.

1. INTRODUCTION

The traditional classification scheme of song/album/artist has begun to be superseded by more intuitive representations derived from content analysis or text metadata of a large digital music collection such as discussed in [1] and applied in [2]. Researchers in music information retrieval have been creating new paradigms for exploring and retrieving audio data that try to incorporate spatial audio into the experience. However, whether because of a lack of familiarity with spatial audio techniques or because it may be the afterthought to the project, the best spatial audio techniques are not always applied. This paper proposes a black-box spatial audio interface which allows a user to explore a two or three dimensional environment populated with songs intelligently arranged according to an external algorithm.

1.1. Previous Interactive Playlists

The MIT Media Lab published several interfaces in the 1990s designed to exploit human audio scene analysis and stream segregation. AudioStreamer [3] presented a user with three simultaneously playing sound sources, primarily recordings of news radio programs, spatially panned to static locations, directly in front and 60 degrees to either side of the listener, using HRTFs. Non-contact sensors built into the chair track the primary sound source the listener is paying attention to.

Dynamic Soundscape [4] spatially arranged audio from different points in time of the same file around a listener's head to assist in quickly finding specific portions of the audio file without needing to listen to the entire file sequentially. Though it is not explicitly stated in [4], it appears that the moving audio is spatialized using HRTFs. Head motion for indicating a source position and head-tracking is detected via sensors on the headset.

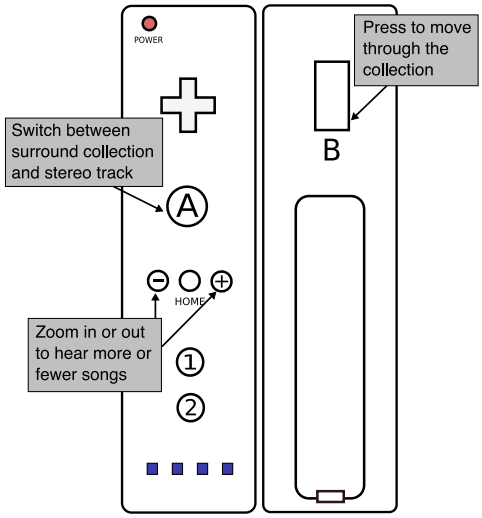

Figure 1: Functions of the buttons on the Nintendo Wii remote used in the interface.

Non-individualized HRTFs are also used in [5], a spatial audio user interface for generating music playlists from a music collection arranged in a typical hierarchical structure. As common errors with non-individualized HRTFs are front-back confusion, the authors allowed the sounds to occur in only 3 static positions across the front.

Hamanaka and Lee [6] continued with Music Scope headphones, an audio-only interface meant to assist users in choosing a single song from ten songs or listening to a multi-track recording and interacting with individual tracks. Various sensors mounted on the headphones track the user's movements. While the motions a user makes to use the interface are described in detail, the paper fails to mention how the sounds are spatialized.

In 2006, [7] added a visual and audio interface to [8] to allow a user to explore a collection of music arranged by similarity. The music was spatialized for a 5.1 loudspeaker setup, not for headphone listening, and used basic sound libraries such as OpenAL, most likely using the built-in panning algorithms to place the sound in space, but specific details are again not given.

This is just a small sampling of the different audio interfaces meant to interact with audio files that share a few traits, notably the use of non-individualized HRTFs to spatialize the sound sources and the approaches used to overcome common errors introduced

by this method. Interfaces were limited to front positions to aid localization or kept static, possibly to decrease the computational load and other problems introduced by interpolating HRTFs. Others describe an interface in detail, but fail to adequately describe how sounds are spatialized when the 3D sound is an integral contribution to the interface.

While many audio interface designers put a considerable amount of thought into how a user will interact with information presented through spatial audio, it appears that the techniques chosen to synthesize that spatial audio are seldom considered to the same extent. We are presenting here an interface primarily driven by spatial audio that takes advantage of advances in interactive 3D audio systems acknowledging that the choices made at the design stage for how a sound will be placed in the virtual environment has a direct consequence on the final design.

2. INTERACTIVE COLLECTION NAVIGATION TOOL

A prototype of our interface is created in Max/MSP and interfaces to a Nintendo Wii remote. A user hears a number of songs, ideally three to four, over headphones arranged spatially around their head. Each song is only about 30 to 90 seconds long and loops continuously. The user then navigates through the songs with the remote and chooses a song to listen to in its full stereo version.

The interface is intended to assist a user in selecting a song from a large collection without any visual feedback. It is blind to how or why the songs are arranged in the space, but only knows the coordinates for each song in the collection and the location of the user. This allows the interface to be used in conjunction with any playlist generator or similarity map that can generate a unique two or three-dimensional coordinate for each song.

The user can choose to interpret the interface from one of two essentially equivalent viewpoints. If the user perceives him or herself to be static and that the songs are moving, then they point at a song to bring it towards them. If they perceive that they are mobile and moving around the songs which are in fixed positions, then they point in the direction they would like to move. As shown in Figure 2, when they are close enough to a song, the remote vibrates, indicating that they can now listen to that song in stereo.

Non-individualized HRTFs commonly create front-back confusion for many listeners with head-tracking being the usual method to overcome the errors. However, we are limiting the equipment to headphones and a processing unit, here a desktop computer, with the intent to use this interface in mobile applications with minimum equipment. The hope is that users will be able to resolve front-back confusion and other localization problems by moving in the environment and changing the directional cues they receive.

We chose to use a Nintendo Wii remote as the controller because of the diverse range of data it can convey and the easy access to its data through the aka.wiiremote external for Max/MSP [9]. The remote has 3 accelerometers, an IR camera, 7 buttons, 4 buttons arranged in a directional cross, 4 LEDs, the ability to vibrate, and a speaker to play audio. We are not using most of the capabilities of the remote, but are using data from the accelerometers, 4 of the buttons, and the vibration mechanism. While it can be difficult to extract precise directional information from the accelerometers, using the more absolute measurements from the IR camera requires two IR emitters. By using only the accelerometers, which use Bluetooth to communicate with the computer, there is no absolute direction that the remote needs to be pointed towards. The user can be facing towards or away from the computer and it has no effect on the direction of movement within the interface. The remote works best when it is positioned relative to the headphones.

The data from the accelerometers is processed to extract a general direction that the remote in pointing towards in three dimensions. The user then moves with a constant velocity in the direction dictated by the remote. The remote can easily move without the user intending it to, so the [B] button is used to indicate when a movement is intentional. The accelerometer data is only read when the button is pressed.

2.1. User Interface

The user experiences multiple songs spatially arranged outside and around their head playing simultaneously and continuously. The Wii remote moves the listener through the collection or the collection around the listener, depending on the point of view. When a song is sufficiently close to the user, the remote vibrates indicating that the song can be listened to in stereo as shown in Figure 2. When the user is finished with the stereo song, they can return to the two/three-dimensional world and select another song.

Since the song data is not directly produced by the interface, there is no guarantee that the songs are always perfectly arranged. The most common issues are that data can be clustered so that a large number of songs are playing at the same time or that the data is too sparse and the user feels lost and cannot find a song to listen to. This is eased by the zoom function illustrated in Figure 3 which increases or decreases the listening area. The user can hear

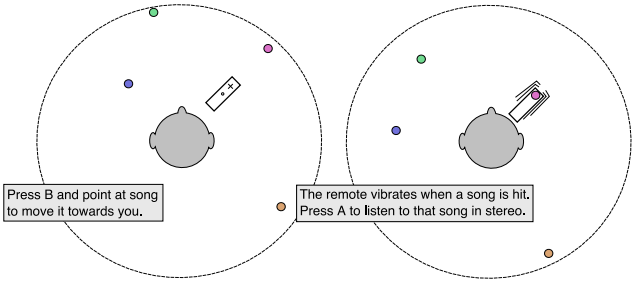

Figure 2: When a song is close enough to the user, the remote vibrates indicating that the song can now be heard in stereo.

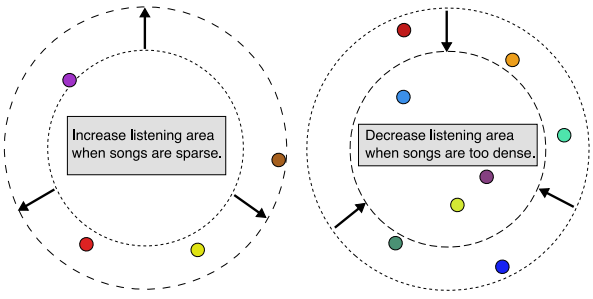

Figure 3: Illustration of the how the zoom function can be used to navigate through dense or sparse data.

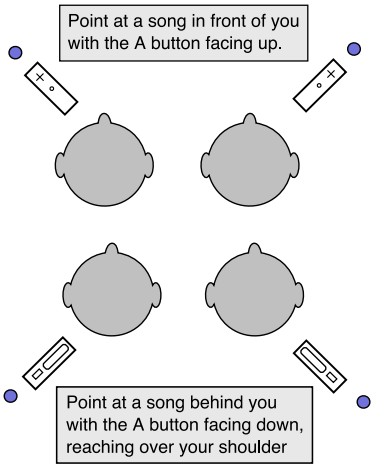

Figure 4: Instructions for how to access songs in front of and behind the user.

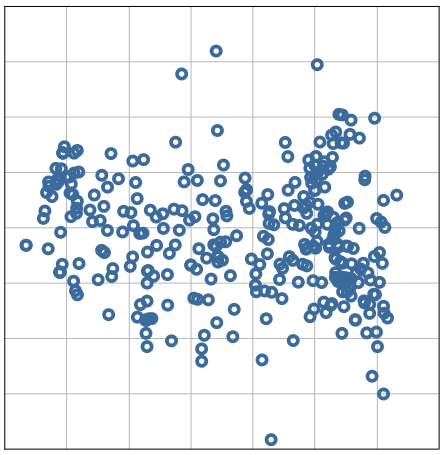

Figure 5: A overview of the spatial distribution of the songs for the data set used during evaluation. Each song is a circle; the scale is merely relational amongst the songs, there are no absolute units.

the songs within a circle surrounding them, when the [+] button is pressed the radius of the circle shrinks, allowing only the closest songs to be heard. When the [-] button is pressed, the radius grows allowing more songs to be heard.

2.2. Spatial Audio

The spatial audio engine uses a specific method of rendering an Ambisonics B-format sound field to a binaural signal called virtual Ambisonics as described in [10]. We used a set of externals for Max/MSP produced by the ICST for the Ambisonics encoder and decoder, encoding and decoding to the third order [11]. The horizontal information is decoded to eight loudspeaker feeds which are each convolved with the non-individualized HRTF for the loudspeaker's position. The HRTFs are from the set of compact HRTFs produced by MIT [12].

We encode and then decode with Ambisonics before convolving with HRTFs to lend flexibility in a number of ways. The audio playback is not limited to headphone listening but can use the highly variable number of loudspeaker configurations for Ambisonics. While third order encoding and decoding is used here, a lower order could also be used to decrease the computational complexity. An advantage particularly pertinent to the navigation of large music collections is that a constant number of HRTFs independent of the number of sound sources are convolved without any need for interpolation. While there are psychoacoustical limits on the number of sound sources that should be playing around a listener as explored in [13], in this configuration the use of HRTFs does not inherently limit the number of sources. Though it was not implemented here, rotations of the entire sound field would be very simple requiring a matrix operation on the B-format signal before decoding [14].

2.3. Data Set

In order to test and evaluate the interface a data set is needed. Our example data set is a collection of tracks positioned in a two-dimensional space structured by mood. Since the seminal work in [15], psychologists have regarded emotions as being well-expressed as points in very low-dimensional spaces, with as few as two or three dimensions being regarded as sufficient to capture the basic meta-feelings from which regular emotions are in some sense comprised. The first two axes, generally regarded as capturing most of the variance observed in the configuration of emotions in various contexts, are usually known as *activity* (from mild to intense) and *valence* (from unpleasant to pleasant). Recent work has gone as far as postulating explicit mechanisms in the brain for the production of these two meta-feelings [16]. When traditional emotion words are visualized in the space defined by these axes, a *circumplex* arrangment frequently emerges, with emotions distributed loosely around a circle [17]. The circumplex of emotions has been widely explored in traditional laboratory studies of music listening [18, 19, 20].

We applied dimension reduction techniques to a large corpus of tens of thousands of mood words mined from social tags for a collection of several thousand tracks to create an updated low-dimensional emotion space for music. Details of our tag data set, dimension reduction methods, and the resulting configuration of emotions, which differs in some respects from those found in previous studies of music, are described in detail in [21].

For the present study we took a simple approach to positioning tracks in this space. To create a two-dimensional arrangement of tracks, we first chose the plane defined by our first and third most significant axes, which correspond roughly to activity and valence. We then mapped each track to the centroid of the three emotion words most frequently applied to it in our data set of social tags, where the weight associated with each word was the number of times the track had been tagged with it. We use here a set of 320 songs distributed across two dimensions as seen in Figure 5.

3. EVALUATION

The interface is evaluated by 12 users, 4 female and 8 male, with varying experience with the Wii or other gaming consoles. Users were given time to become familiar with the interface, usually taking about 10 to 15 minutes, then answered 4 questions rating various aspects of the interface on a scale of 1 to 7 and 3 questions

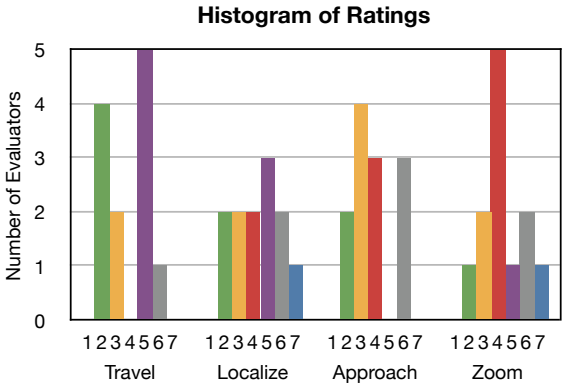

Figure 6: Histogram showing the distribution of user ratings for each question, "how easy it is to travel through the collection," "how easy is it to localize a song," "how easy is it to approach a song," and "how useful is the zoom function." For all questions, 1 is the most negative response with 7 being the most positive.

with open responses. Though the users did not have a visual interface, their movements and actions were monitored on a visual interface behind the user.

Histograms of the responses to the four questions with rating scales can be seen in Figure 6. The responses were evenly split with six people giving a two or three and six rating a five or six for the question "how easy is it to travel through the collection" with one being the most difficult and seven being the easiest. All the users found the interface confusing when there were no songs within the listening range, so they couldn't figure out where they were moving. Those that rated the task higher seemed to learn the interface faster, but several suggested that a feedback mechanism to give a sense of the boundaries of the collection would be helpful.

Some users found the zoom functions beneficial especially when the data was sparse and no songs could be heard. A few found the zoom function confusing, but this seemed related to how individual users approached the interface. Some relied on the zoom function more than changing their position to locate new songs while others tended to move through the collection and use the zoom functions sparingly. The users that used the zoom often exposed some bugs in the application that caused difficulties with the interface, so those that didn't use the zoom as much tended to have a more positive experience with the interface.

Unsurprisingly, the most common complaint about localizing songs in the song environment was front-back confusion. Once users were aware of this error they found they could move more easily through the environment by moving forwards or backwards to resolve localization issues. Overall, the users didn't find the environment difficult to understand and could separate the songs in space easily, but they had difficulties then expressing their intention to move to a specific place with the remote.

Users often found it difficult to approach a single song; this may stem from a number of places. The data set is not normalized so some songs may be significantly louder or softer than others, causing them to be perceived closer or farther than their actual coordinates in relation to other songs. Some users were more adept at moving within the world in a certain direction such as in front or to the sides, so when a song moved out of the preferred region, it was difficult to approach.

When asked about the perceived viewpoint, three users felt that the songs are static and that they are moving around the collection and nine felt that the songs were moving while they stayed in the same place. Though, some users commented that they felt the difference to be ambiguous and perceived both viewpoints at different times.

The users had a wide range of responses when asked whether they would prefer a visual interface. Four felt strongly that there should be while three others responded strongly that interface should remain as audio only, noting that a visual component would inherently change the application. The other five users felt that with some improvements in the current interface, a visual component would not be needed. Most felt that the only visual information they needed was global location information, i.e. where they were located in relation to the entire data set.

Though the numbers seem to reflect a rather average view by the users in the evaluation, there was a great deal of enthusiasm for the interface. Users felt that it has great potential and was a unique way of interacting with a collection of music, but did not think the mappings between the remote and movements in the virtual space were yet intuitive enough.

4. FUTURE WORK

We used a third order Ambisonics encoder in this implementation of the interface merely because it was the highest order the encoder could handle and the computer had no problems with the load. However, this application may not need such a high order to still be effective, and certainly would not need as high of an order to encode height information. A lower order would save computation time and also decrease the number of virtual speaker feeds and HRTFs needed, allowing the interface to function with slower processors.

The mappings between the remote and movement through the virtual space could be arranged in a number of configurations that were not tried here. In the current configuration, users never rotate their point of view. In a manner of speaking they are always looking north when they move about the world, whether forwards, backwards, or sideways. If a user could rotate the entire sound field, then they ideally could move songs to an easily accessible region, such as directly in front without changing their location within the environment. This also might ease front-back confusion. Since the audio is encoded into B-format, transformations such as rotations are easily applied [14].

If it is found that self-movement is not enough to overcome the errors introduced by non-individualized HRTFs, then head-tracking could be added without greatly increasing the computational load because it would only involve rotations. As discussed above, this is a simple operation in the Ambisonics domain.

Since this interface is intended as a black-box for any data set that assigns a unique coordinate to each song, this environment needs to be tested with different data sets.

5. CONCLUSIONS

We have described an interface to navigate through a large collection of music using primarily audio information. A user can navigate through a collection of music that is arranged according to any algorithm that can assign a unique coordinate to each song in the collection. The interface allows a user to listen to a varying

number of mono songs spatially arranged the listener's head. Controls such as a zoom function let the user control how many songs are heard. When the user receives haptic feedback through the remote, they can listen to the closest song in its full stereo version. The selected song could then be used for further processing, such as automatically generating a playlist.

By using only audio information, tag-based or textual information is never used in the interface. Incorrect song title, artist, or other tags is inconsequential to the user, though if the user wishes to learn this information about the songs, then either a visual display or text-to-speech function will be needed. However, the quality of the audio files accessed by the interface does have a strong impact on the interaction, especially with localization. If the files are not normalized and vary greatly in volume, the songs will not be perceived in the correct locations relative to each other. This may be further exasperated in diverse data sets where, for example, highly processed pop music may be heard near more dynamic, less compressed classical string music.

By using a virtual Ambisonics approach, much of the computational load associated with moving sound sources using HRTFs is eliminated. While the common problem of front-back confusion when using non-individualized HRTFs still exists, the ability of a user to move and change the spatial cues presented to them helps alleviate problems. The user cannot yet rotate the sound field which may further help reduce localization errors, but such operations are easily implemented with B-format signals.

Twelve users evaluated the interface and gave positive feedback about the information presented, but reported difficulties with interacting with the data set. The evaluations confirmed that the mapping of user movement to virtual movement through the music collection is not yet ideal. This is not surprising when using the Wii remote. The remote has a large number of sensors and buttons that can be used to convey user information, in particular the three accelerometers, but this freedom in expression can be difficult to interpret and map in an intuitive manner. We have experienced similar problems with third party games developed for the Wii gaming console. Nintendo certainly knows how to use the remote that it has developed far better than many third party developers. It may be advantageous to move away from the Wii remote and towards a more traditional gaming interface as it is a more familiar controller for most users. This interface may also be applied to mobile devices such as phones which increasingly include sensors such as accelerometers.

6. ACKNOWLEDGMENTS

This research was supported by EPSRC grant EP/E017614/1 (Online Music Recognition And Searching), Queen Mary, University of London, and the Overseas Research Awards Scheme. The assistance of Tony Stockman in the evaluation of the interface is greatly appreciated.

7. REFERENCES

[1] Adam Berenzweig, Beth Logan, Daniel P.W. Ellis, and Brian Whitman, "A large-schale evaluation of acoustic and subjective music similarity measures," *Computer Music Journal*, vol. 28, no. 2, pp. 63–76, Summer 2004.

[2] Elias Pampalk, Tim Pohle, and Gerhard Widmer, "Dynamic playlist generation based on skipping behavior," in *ISMIR'05*, London, UK, September 2005, pp. 634–637.

[3] Chris Schmandt and Mullins. Atty, "AudioStreamer: exploiting simultaneity for listening," in *CHI'95*, May 1995.

[4] Minoru Kobayashi and Chris Schmandt, "Dynamic Soundscape: mapping time to spae for audio browsing," in *CHI'97*, March 1997.

[5] Jarmo Hiipakka and Gaëtan Lorho, "A spatial audio user interface for generating music playlists," in *ICAD '03*, Boston, MA, USA, July 2003.

[6] Masatoshi Hamanaka and Seunghee Lee, "Music Scope Headphones: natural user interface for selection of music," in *ISMIR'06*, 1006.

[7] Peter Knees, Markus Schedi, Tim Pohle, and Gerhard Widmer, "An innovative three-dimensional user interfacefor exploring music collections enriched with meta-information from the web," in *ACM Multimedia*, 2006.

[8] Elias Pampalk, "Islands of music: analysis, organization, and visualization of music archives," M.S. thesis, Vienna University of Technology, 2001.

[9] Masayuki Akamatsu, "aka.wiiremote Max/MSP external for Wii remote," available at http://www.iamas.ac.jp/ aka/max, March 2008.

[10] Markus Noisternig, Alois Sontacchi, Thomas0 Musil, and Robert Hóldrich, "A 3D ambisonice based binaural sound reproduction system," in *AES 24th International Conference on Multichannel Audio*, 2003.

[11] ICST, "Max/MSP Ambisonics externals," available at http://www.icst.net.

[12] William Grant Gardner and Keith Martain, "HRTF measurements of a KEMAR dummy-head microphone," Tech. Rep., MIT Media Lab, May 1994.

[13] Gaëtan Lorho, Juha Marila, and Jarmo Hiipakka, "Feasibility of multiple non-speech sounds presentation using headphones," in *ICAD '01*, Espoo, Finland, August 2001.

[14] Dave Malham, "Approaches to spatialisation," *Organised Sound*, vol. 3, no. 2, pp. 167–177, 1998.

[15] C. E. Osgood, G. J. Succi, and P. H. Tannenbaum, *The measurement of meaning*, University of Illinois Press, 1957.

[16] J. Posner, J. Russell, and B. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, coginitive development, and psychopathology," *Development and Psychopathology*, vol. 17, pp. 715–734, 2005.

[17] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–78, 1980.

[18] K. Hevner, "Experimental studies of the elements of expression in music," *American Journal of Psychology*, vol. 48, pp. 246–68, 1936.

[19] L. Wedin, "Dimension analysis of emotional expression in music," *Swedish Journal of Musicology*, vol. 51, pp. 119–140, 1969.

[20] G. L. Collier, "Beyond valence and activity in the emotional connotations of music," *Psychology of Music*, vol. 35, no. 1, pp. 110–131, 2007.

[21] G. Kreutz, M. Levy, and M. Sandler, "Emotion words in social tags for popular music," Submitted to *Music Perception*.