Soundscape Generation for Virtual Environments using Community-Provided Audio Databases

Nathaniel Finney Jordi Janer[*]

Abstract

This research focuses on the generation of soundscapes using unstructured sound databases for the sonification of virtual environments. The design methodology incorporates the use of concatenative synthesis to construct a sound environment using online community-provided sonic material, and an application of this methodology is described in which sound environments are generated for *Google Street View* using the online sound database *Freesound.* Furthermore, the model allows for the creation of augmented soundscapes using parameterization models as an input to the resynthesis paradigm, which incorporates multiple source and textural layers. A subjective evaluation of this application was performed to compare the immersive properties of the generated soundscapes with those of real recordings, and the results suggest a general preference to the generated soundscapes incorporating sound design principles. The potential for further research and future applications in the area of augmented reality are discussed incorporating emerging web media technologies.

Introduction

The advancement of online virtual reality and augmented reality technologies in conjunction with the increase of community-provided sonic material, encourages a development of autonomous soundscape generation tools that allow for the enhancement of the immersive experience of the user.

With the use of unstructured databases as the audio input to the soundscape generator, such a model may be considered both a framework for community interaction and an ecological acoustics preservation tool, while taking advantage of the growing databases of sonic material [10]. However, such unstructured databases pose challenges to the autonomous generation of immersive sonic environments due to the ranging characteristics of the material. Differences such as sound quality and reverberation may result in an incongruent sonic ambience and thereby reduce the immersive quality of the soundscape. Synthesis methods such as wavelet analysis and resynthesis and large grain concatenative synthesis are proposed in order to alleviate the effect of these variations and to allow for parameterization-based sound design algorithms.

Applying parameterization to the synthesis paradigm furthermore allows for augmentation of the soundscape according to either predefined or user defined criteria, and thereby expands the potential for adaptive and creative soundscape design for virtual environments and augmented reality.

The design of sonic environments requires a fundamental classification strategy for both soundscapes and sound sources. The classification of soundscapes is most widely performed using R. Murray Schafer's referential taxonomy, which incorporates socio-cultural attributes and ecological acoustics [12]. In the work of Birchfield et. al., the generation of the soundscape is based on a probabilistic selection of audio tracks separated according the Schafer concept of keynote sounds and signal sounds, where the material has been retrieved prior to generation with a lexical search using WordNet [8]. Birchfield et. al. incorporates dynamically changing probabilities in response to user behavior for triggering audio tracks, which results in an evolving sound environment that they propose reflects the sonic diversity of the equivalent natural environment [7]. Probabilistic models for soundscape generation are a demonstration of the notion that a generated soundscape is greatly enhanced by variety and evolution, for which research has shown to be a contributing factor to the sense of presence in VE's [13], [15].

The sense of presence is defined as the feeling of being situated in an environment despite being physically situated in another [14]. In quantifying and/or qualifying the sense of presence, or the similar term immersion, different authors have proposed and evaluated various components that make up this highly subjective sense [14], [15], [13]. In order to experience some degree of presence in a VE, the psychological states involvement and immersion are necessary precursors. Involvement is defined as the focus on a coherent set of stimuli, and immersion is the actual perception of envelopment in an environment [14].

[*] Music Technology Group, Universitat Pompeu Fabra, Barcelona. His work was partially supported by the ITEA2 Metaverse1 Project (http://www.metaverse1.org).

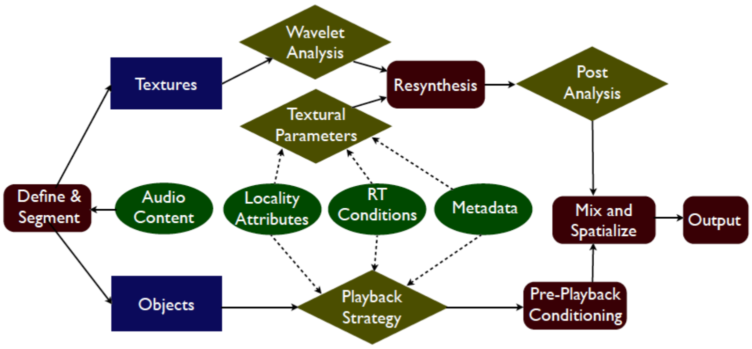

Figure 1: Overview of the system functionality with given inputs (green circles), analysis blocks (diamonds), sound groups (rectangles) and processes (rectangles with soft corners).

The model proposed in this research applies these fundamental soundscape design principles to a generation methodology and is evaluated according the resulting sense of presence of the user. The application developed in this research includes only a portion of the applicable soundscape design principles described by previous authors, and therefore suggests that as web audio generation technologies advance, this methodology carries a large potential for creating immersive online media.

Design Methodology

The principles of sound design and soundscape characterization as laid out by R. Murray Shafter [12] form the underlying hypothesis of the design methodology described in this section, that the treatment of the soundscape in terms of two layers, textures (background) and objects (foreground), establishes a framework for algorithmic design and interaction principles more suitable for information delivery and perceptual comfort than with a unified montage of sound sources.

The textural elements and object elements are segmented from sound files retrieved from a database, and categorized into one of these two layers using their semantic identifiers, which are either extracted from tags or a recognition model. Objects are those sounds that are meant or expected to draw attention from the user, and may include indicators, soundmarks or informational content such as church bells or non-diegetic sounds such as narration. Textural elements are determined to be those sources that form the ambience such as birds and wind, and tend to be more stochastic in nature while drawing minimal focus from the user.

Figure 2 shows the overview of the system functionality, where the two sound groups, *textures* and *objects* are seen to be handled separately until the final mixing and spatialization. Solid arrows signify transfer of audio content, while dashed arrows are informational streams supplied to the various function blocks. A detailed description of the functionality of the individual blocks may be found in [9].

The blocks *Textural Parameters* and *Playback Strategy* contain models for parameterizing the synthesis of the soundscape, and thereby allow for manipulation and augmentation according to a desired resultant soundscape. These functions may be considered to be the design portion of the model, which may be predefined or allow for user or contextual manipulation.

For the concatenation of the original audio files, an MFCC-based bayesian information criterion (BIC) segmentation procedure is used to scan a recording for optimal segmentation points [6]. Using an MFCC calculation takes into account the spectral properties of the signal and is based on the Mel-scale, which correlates the frequency spectrum of the signal with the perceptual attribute of timbre. A minimum segmentation length is selected depending on the signal classification, based on Schafer's taxonomy, and the grain sizes longer than that length are chosen according to the optimal segmentation points from the MFCC-BIC calculation.

Current Implementation

As an initial implementation of the autonomous soundscape generation model, a static photorealistic environment was chosen in order to evaluate the use of recorded sounds from unstructured databases in an application that must be consistent with the natural environment. The Google Street View application contains panoramic images from discrete locations within a city, in which the user is free to rotate and can translate to different locations.

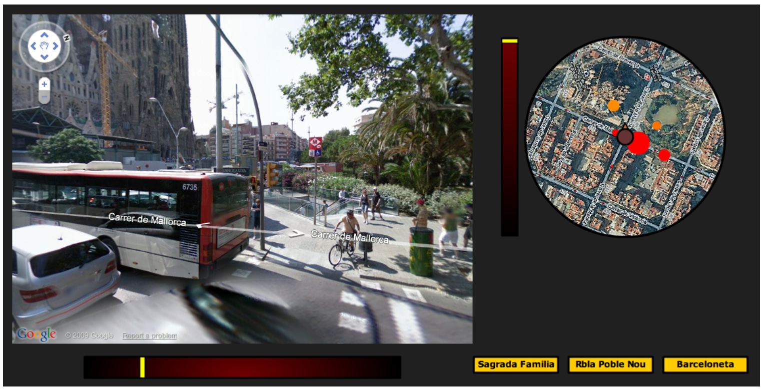

Figure 2: GUI.

Table 1: Questionnaire for subjective evaluation

1. How well could you identify individual sound sources in the soundscape?
2. How well could you actively localize individual sound sources in the soundscape?
3. How compelling was your sense of objects moving through space?
4. How much did the auditory aspects of the environment involve you?
5. How compelling was your sense of movement (turning) inside the virtual environment?
6. How much did your experiences in the virtual environment seem consistent with your real-world experiences?

The sound database used for feeding the soundscape generator was The Freesound Project [4], and as many samples as possible were selected with origins in Barcelona and Spain for purposes of authenticity. For the purposes of this study involving the Street View application, the sound objects are retrieved manually in order to focus on the autonomous generation aspect, and the automatic retrieval implementation is reserved as an avenue for future work.

The application includes dynamic changes of audio level and panning in the individual source and texture layers according to the location and user orientation. The application uses Actionscript 3.0 for synthesis and user interaction, and pre-analyzes the audio files offline using MATLAB to determine segmentation points and individual segment selection probabilities. The user interface is shown in figure 2 and the demo application is available online [1].

A subjective evaluation of the application was performed using eight participants and three generated soundscapes in comparison with three quadraphonic recordings at the relevant locations. The participants were able to control the selected location and orientation freely, and were not informed of which soundscapes were generated and which were real recordings. Questions considered relevant for soundscapes in photorealistic environments were selected from the questionnaire proposed by Singer et. al. for evaluating the sense of presence in virtual environments [14] (see figure 1), where the user was asked to rate each response on a five-point scale from *Not at all* to *Very Well.*

The results of the analysis showed that the subjects generally rated the generated soundscapes higher than the recorded soundscapes for each of the individual questions, by a margin of $15 - 20\%$. While there were not enough subjects for drawing statistically significant conclusions, the results demonstrate that the generated soundscapes are at least acceptable in comparison with actual recordings. The normalized responses to the individual questions for the three locations are shown in figure 3.

As many of the questions were related to the identification of sources, localization and dynamism of sound objects, the spatialization and mixing of the separate layers seem to have had a pronounced effect. The participants were more easily able to distinguish and localize objects in the sound scene and perceived a higher degree of movement of objects with the generated soundscapes using separated source layers and individual layer spatialization.

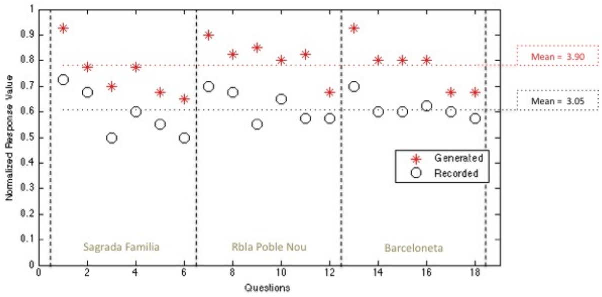

Figure 3: Normalized responses averaged over all participants in the subjective evaluation for each question.

Discussion

The results of the evaluation of the design methodology employed in this study suggest that the use of techniques for separating object and textural layers and the application of level and panning algorithms based on user orientation, heighten the user's sense of presence in the virtual environment in comparison with recordings of an entire soundscape. The use of soundscape design principles and parameterization models to augment the generated soundscape with reference to user orientation and locality attributes will be studied further in subsequent implementations of the model, with the focus on developing applications that may be integrated in online virtual environments.

In the current application, soundscape generation and real-time manipulation were performed with Actionscript 3.0, and offline analyses for segmentation were conducted using MATLAB. Retrieval of audio material from the online community database was performed manually in this study. Future implementations of the model are to be developed with the intention to incorporate sonification of virtual environments directly in the web browser with technologies such as the new extension to the HTML5 media API [2], which allows for manipulating raw audio data.

The current implementation of the soundscape generation model has relatively modest real-time audio requirements, primarily consisting of the reproduction of pre-segmented audio files with spatialization and dynamic level adjustments related to the individual sound objects and textural layers. With the incorporation of the capability to apply dynamic filtering to the individual sound objects and textural layers, the sound design potential of the model would be highly improved. Furthermore, as many augmented applications target portable devices, in which the audition experience is reproduced through headphones, a binaural spatialization algorithm using HRTF's [5] [3], would likely benefit the user's sense of presence. Such spectral filtering models could be implemented as time-domain filters within the proposed HTML5 API [2]. However if many sources are present, methods for optimizing binaural 3D rendering may require further development [11].

Such advances in web media technology are likely to provide a means for further developing the functionality and efficacy of the soundscape generation model, and allow for increased abilities to augment the soundscapes according to desired criteria.

References

[1] `http://dev.mtg.upf.edu/soundscape/media/StreetView/streetViewSoundscaper2_0.html`.

[2] `https://wiki.mozilla.org/Audio_Data_API`.

[3] `http://recherche.ircam.fr/equipes/salles/listen`.

[4] Freesound.org. `http://www.freesound.org`.

[5] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, pages 99–102, 2001.

[6] X. Anguera and J. Hernando. Xbic: Real-time cross probabilities measure for speaker segmentation. *Univ. California Berkeley, ICSIBerkeley Tech. Rep*, 2005.

[7] D. Birchfield, N. Mattar, and H. Sundaram. Design of a generative model for soundscape creation. In *International Computer Music Conference, Barcelona, Spain*, 2005.

[8] C. Fellbaum et al. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.

[9] N. Finney. Autonomous generation of soundscapes using unstructured sound databases. Master's thesis, Universitat Pompeu Fabra, 2009.

[10] J. Janer, N. Finney, G. Roma, S. Kersten, and X. Serra. Supporting soundscape design in virtual environments with content-based audio retrieval. *Journal of Virtual Worlds Research*, 2(3), 2009.

[11] M. Noisternig, T. Musil, A. Sontacchi, and R. Höldrich. 3d binaural sound reproduction using a virtual ambisonic approach. In *IEEE International Symposium on Virtual Environments*, 2003.

[12] R. Murray Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books, 1994.

[13] S. Serafin. Sound design to enhance presence in photorealistic virtual reality. In *Proceedings of the 2004 International Conference on Auditory Display*, pages 6–9, 2004.

[14] Michael J Singer and Bob G. Witmer. Measuring presence in virtual environments: A presence questionnaire. *PRESENCE*, 7:225–240, 1998.

[15] P. Turner, I. McGregor, S. Turner, and F. Carroll. Evaluating soundscapes as a means of creating a sense of place. In *International Conference on Auditory Display*, 2003.